Institute of Mathematical Statistics

LECTURE NOTES–MONOGRAPH SERIES

Volume 0

Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh

Bertrand Clarke and Subhashis Ghosal, Editors

Institute of Mathematical Statistics
Beachwood, Ohio, USA

Institute of Mathematical Statistics
Lecture Notes–Monograph Series

Series Editor:
Anthony Davison

The production of the *Institute of Mathematical Statistics Lecture Notes–Monograph Series* is managed by the IMS Office: Jiayang Sun, Treasurer and Elyse Gustafson, Executive Director.

Library of Congress Control Number: ???

International Standard Book Number ???

International Standard Serial Number 0749-2170

Printed in the United States of America

Contents

Preface

This volume is an effort to draw together contributions to the main areas in which Jayanta has worked and continues to work. The papers naturally fall into 5 categories.

First, sequential estimation was Jayanta's starting point. Thus, beginning with that topic, there are two papers, one classical by Hall and Ding leading to a variant on p-values, and one Bayesian by Berger and Sun extending reference priors to stopping time problems.

Second, there are 5 papers in the general area of prior specification. Much of Jayanta's earlier work involved group families as does Sweeting's paper here for instance. There are also two papers dwelling on the link between fuzzy sets and priors, by Meeden and by Delampady and Angers. Equally daring is the work by Mukerjee with data dependent priors and the pleasing confluence of several prior selection criteria found by Ghosh, Santra and Kim. Jayanta himself studied a variety of prior selection criteria including probability matching priors and reference priors.

Third, between his work on parametric Bayes and nonparametrics, Jayanta took an interest in model selection. Accordingly, three papers on model selection come next. Bunea's work on consistency echoes Jayanta's work on consistency of the BIC. Chatterjee and Mukhopadhyay's work on data adaptive model averaging continues the direction he started under Jayanta's guidance. Chakrabarti and Samanta's work on the asymptotic optimality of predictive cross validation contrasts nicely with standard Bayes model selection, via the BIC for instance.

Fourth, there are 5 papers generally on Bayesian nonparametrics. Some are applied as in Malec and Mueller's work on semi-parametrics in small area estimation or Guo, Dey and Holsinger's work carefully using prior selection for modeling purposes. And some are more theoretical: Choi and Ramamoorthi provide a review, with some new results, on posterior consistency while James focuses on a class of priors and van der Vaart and van Zanten focus on the role of Reproducing Kernel Hilbert spaces in Bayesian nonparametrics with Gaussian process priors.

Finally, Jayanta has most recently turned his attention to high dimensional problems. On this topic, there are 5 papers from a variety of standpoints. For instance, it is possible to make unexpected use of the information in the large dimensions themselves as in Sen's work with Kendall's tau. Others focus on the parametric parts of a nonparametric model as in Ishwaran and Papana, or in Bhattacharya and Bhattachcarya. A third tack in Clarke and Seo is the focus on selecting the dimensions for use in emerging model classes. Finally, the work of Bickel, Li and Bengtsson establishes a general convergence result for computing conditional distributions.

As can be seen, some papers fit comfortably into more than one section and some only fit into a section if it is interpreted broadly. Even so, we would like to think that the papers have achieved a nice tradeoff between clustering rather nicely around the topic of each category and maintaining a reasonable diversity in line with Jayanta's work.

Despite his diverse research interests, asymptotics and their applications have been the main recurring theme of Jayanta's research since he published his first paper in sequential statistics in 1960 (at the age of 23). So, as a generality, asymptotics undergirds most of the material in this volume honoring him.

Fortunately, asymptotic thinking pervades statistical inference, even in the most applied contexts. So, this is hardly a limitation. On the other hand, asymptotics has a way of being impenetrably abstract. However, all the papers here, are, in Woodroofe's memorable phrase, written at a level that would be 'accessible to a determined graduate student'. We encourage readers to have a look at least at the introductions of papers outside their research area, just for pure love of the field and the joy of intellectual stimulation. We suspect that once some one has read the introduction, he or she will be ineluctably led to finish reading the whole paper.

As Editors, we have been delighted at the depth and quality of work our contributors submitted. They all make foundational points in the spirit of Jayanta. We believe each paper will be of interest to researchers, theoretical and applied, who confront problems that are difficult enough that conventional solutions are inadequate and closed form solutions are intractable in the several areas covered here. We are deeply grateful to all contributors for offering their finest work to this volume.

Of course, no volume such as this could have been possible without the free and anonymous labor of referees: You folks know who you are, but for the sake of confidentiality we cannot name you. We especially thank those who provided extremely prompt reports when we badly needed them. If any of you meet one of us at a conference, we owe you a drink. Probably two – you helped us immeasurably.

In terms of actually producing this volume, Jennifer Clarke provided invaluable support. She helped us repeatedly with compiling complete versions of the volume. In particular, the final, detailed copy-editing was her work; the balance of her account in the Bank of Karma is astronomical. We can't thank her enough.

Along the way there were many other people who gave us their time and expertise. Dipak Dey helped guide us as we prepared initial proposal. Rick Vitale, the former Editor for the LNMS series, also did a first rate job in explaining the details of how we had go about this kind of project, if we wanted it to be successful. Rick then gave the initial approval – thanks, Rick! Anthony Davison, the current Editor for the LNMS series, worked closely with us to get the volume in final form and then gave it final approval. We appreciate the burden that you carried for us, Anthony.

Finally, we are grateful that the IMS put its scarce resources into supporting this volume. The IMS has a tradition of honoring it's most illustrious members and, in our view, Jayanta is most assuredly among them.

It was a pleasure to put this tribute together and we hope that in some small way we have served the interests of the research world.

Bertrand Clarke
Subhashis Ghosal

November 2007

Contributors to this volume

Angers, J.-F., *Université de Montréal, Montréal, QC, Canada*,

Bayarri, M.J., *University of Valencia, Valencia, Spain*,
Bengtsson, T., *Bell Labs, Murray Hill, NJ, USA*,
Berger, J. O., *Duke University, Durham, NC, USA*,
Bhattacharya, A., *University of Arizona, Tucson, AZ, USA*,
Bhattacharya, R., *University of Arizona, Tucson, AZ, USA*,
Bickel, P., *University of California-Berkeley, Berkeley, CA, USA*,
Bunea, F., *Florida State University, Tallahassee, FL, USA*,

Chakrabarti, A., *Indian Statistical Institute, Kolkata, India*,
Chatterjee, S., *University of Minnesota, Minneapolis, MN, USA*,
Choi, T., *University of Maryland - Baltimore County, Baltimore, MD, USA*,
Clarke, B., *University of British Columbia, Vancouver, BC, Canada*,
Clarke, J., *University of Miami School of Medicine, Miami, FL, USA*,

Datta, G. S., *University of Georgia, Athens, GA, USA*,
Delampady, M., *Indian Statistical Institute, Bangalore, India*,
Dey, D., *University of Connecticut, Storrs, CT, USA*,
Ding, K., *Queen's University, Kingston, ON, Canada*,

Ghosal, S., *North Carolina State University, Raleigh, NC, USA*,
Ghosh, M., *University of Florida, Gainesville, FL, USA*,
Guo, F., *Virginia Polytechnic Institute and State University, Blacksburg, VA, USA*,

Hall, W. J., *University of Rochester, Rochester, NY, USA*,
Holsinger, K., *University of Connecticut, Storrs, CT, USA*,

Ishwaran, H., *Cleveland Clinic, Cleveland, OH, USA*,

James, L. F., *Hong Kong University of Science and Technology, Kowloon, Hong Kong*,

Kim, D., *Kyungpook National University, Taegu, Korea*,

Li, B., *Tsinghua University, Beijing, China*,

Müller, P., *M.D. Anderson Cancer Center, Houston, TX, USA*,
Malec, D., *U.S. Census Bureau, Washington, DC, USA*,
Meeden, G., *University of Minnesota, Minneapolis, MN, USA*,
Mukerjee, R., *Indian Institute of Management Calcutta, Kolkata, India*,
Mukhopadhyay, N., *Virginia Commonwealth University, Richmond, VA, USA*,

Papana, A., *Case Western Reserve University, Cleveland, OH, USA*,

Ramamoorthi, R. V., *Michigan State University, East Lansing, MI, USA*,

Samanta, T., *Indian Statistical Institute, Kolkata, India*,
Santra, U., *University of Florida, Gainesville, FL, USA*,
Sen, P. K., *University of North Carolina, Chapel Hill, NC, USA*,
Seo, D., *University of Miami School of Medicine, Miami, FL, USA*,
Sun, D., *University of Missouri-Columbia, Columbia, MO, USA*,
Sweeting, T. J., *University College London, London, UK*,

van der Vaart, A. W., *Vrije Universiteit, Amsterdam, The Netherlands*,
van Zanten, J. H., *Vrije Universiteit, Amsterdam, The Netherlands*,

IMS Lecture Notes–Monograph Series
Volume Title
Vol. 0 (0000) 1

J. K. Ghosh's Contribution to Statistics: A Brief Outline

Bertrand Clarke[1] **and Subhashis Ghosal**[2]

University of British Columbia and North Carolina State University

Abstract: Professor Jayanta Kumar Ghosh has contributed massively to various areas of Statistics over the last five decades. Here, we survey some of his most important contributions. In roughly chronological order, we discuss his major results in the areas of sequential analysis, foundations, asymptotics, and Bayesian inference. It is seen that he progressed from thinking about data points, to thinking about data summarization, to the limiting cases of data summarization in as they relate to parameter estimation, and then to more general aspects of modeling including prior and model selection.

Contents

1. Introduction

Professor Jayanta Kumar Ghosh, or J. K. Ghosh, as he is commonly known, has been a prominent contributor to the discipline of statistics for five decades.

[1] Department of Statistics, University of British Columbia, 6356 Agricultural Road, Vancouver, BC, V6T1Z2, Canada; e-mail: riffraff@stat.ubc.edu

[2] Department of Statistics, North Carolina State University, 12 Patterson Hall, 2501 Founders Drive Raleigh, NC 27695, USA; e-mail: ghoshal@stat.ncsu.edu

AMS 2000 subject classifications: Primary 62; secondary 62

Keywords and phrases: Sequential Analysis, Foundations of Statistics, Edgeworth Expansions, Second Order Efficiency, Bartlett Corrections, Noninformative prior, Semiparametric Inference, Posterior Convergence, Bayesian Nonparametrics, Model Selection

The spectrum of his contributions encompasses sequential analysis, the foundations of statistics, finite populations, Edgeworth expansions, second order efficiency, Bartlett corrections, noninformative, and especially matching, priors, semiparametric inference, posterior limit theorems, Bayesian nonparametrics, model selection and Bayesian hypothesis testing, high dimensional data analysis as well as some applied work in reliability theory, statistical quality control, modeling hydrocarbon discoveries, geological mapping and DNA fingerprinting. By itself covering such diverse topics in depth is a major career achievement. He has authored over 130 publications including three monographs and several edited volumes. His books, one entitled *Higher Order Asymptotics* and published as an IMS monograph and another entitled *Bayesian Nonparametrics*, co-authored by R. V. Ramamoorthi and published by Springer-Verlag, continue to hold respected positions for researchers in these areas. His recently published third book [34] is a fine graduate text on Bayesian inference. In addition, his service to the profession, especially as the editor of Sankhyā has been invaluable.

In spite of the variety of his work, asymptotics has been central to his thinking across a wide range of problems. Accordingly, in what follows, we outline some of his work, in roughly chronological order, focussing on those contributions which are intimately connected to asymptotics. In the course of reviewing his work, we try to characterize the progression of thinking that naturally connects the topics that J.K. Ghosh has done so much to develop.

2. Sequential Analysis

J. K. Ghosh started his research career in Sequential Analysis in the early sixties as a Graduate Student in the Department of Statistics at Calcutta University. Wald had recently introduced his *sequential probability ratio test* (SPRT), but its properties were not well understood in the composite case. This was the first topic to which Ghosh turned his attention. Through his work, many of the properties of SPRT and related procedures were established and better understood. For instance, in the testing context, double minimaxity essentially means simultaneous minimization of average type I and II error probabilities. In his first published work [26], Ghosh clarified a result of Wald on the double minimaxity of the SPRT for normal two-sided alternative hypothesis (with unknown scale) separated from the null by δ.

It is well-known that the power function is monotonic in many common families for fixed sample sizes. Ghosh established an analog of this result in [27], namely that the operating characteristic function of the (generalized) SPRT continues to be monotonic. Also in the sequential context, [28] considered the admissibility of sequential tests based on a simple identity which later became known as the Ghosh–Pratt identity. Ghosh compared the SPRT not just with the class of all tests with finite expected sample size but also within other classes, for instance, the class which requires at least one observation or which requires no more than a predetermined number of observations to reach a conclusion.

Following this, Ghosh continued to elucidate more properties of the SPRT, and its variants, which could be seen as analogs of the corresponding properties Neyman–Pearson or Bayes tests for fixed sample size. In [29], he proved that for exponential families, truncated or untruncated Bayesian sequential decision rules' terminal decisions describe regions in terms of sufficient statistics, and also showed that for testing problems, truncated generalized SPRT's form a complete class.

About two decades later, Ghosh returned to sequential problems, along with various co-authors. In [33], he studied an invariant SPRT to identify two normal populations with equal variance and obtained bounds for error probabilities. Most recently, similar bounds for an invariant SPRT with respect to an improper prior have also been obtained in [50].

Two-stage procedures are closely related to sequential procedures. Recall Stein's famous problem of finding a bounded length confidence interval for the normal mean with unknown variance. Stein proposed a two-stage procedure for doing this: In the first stage, the sample variance determines how many samples are to be taken in the second stage. An obvious shortcoming of the procedure is that the second stage sample variance is not used in the construction of the interval. So, it is natural to ask whether one can improve Stein's procedure by using the second stage sample variance. Surprisingly, it is impossible to better Stein's procedure as shown in [38].

However, the procedure can be improved in a different, and perhaps more appropriate sense. The confidence coefficient does not in general properly reflect the true sense of confidence about a parameter after observing data. For instance, if two observations are obtained from a $U(\theta, \theta+1)$ family, then the assessment of θ is very precise when the two observations differ in magnitude by nearly 1, while the assessment is much less precise if the two observations are close to each other. This means that classical confidence intervals fail to indicate the true difference in the level of confidence after observing the sample.

Motivated by this, Kiefer suggested letting the confidence coefficient depend on the data. After all, in reality, for a given random interval I, we often want to predict the indicator function $1\{\theta \in I\}$. Since this object is unknown, it is traditionally estimated by a constant, the best constant being the expectation $P_\theta(\theta \in I)$, which becomes fixed (or asymptotically fixed) for many classical intervals. However, from a prediction theory point of view, it makes more sense to let the predictor of $1\{\theta \in I\}$ depend on the observed data. The predictor considered in this way is called the random confidence coefficient associated with the confidence interval I. It is shown in [39] that the second stage sample variance can be used to boost the random confidence coefficient of a bounded length confidence interval.

3. Foundations of Statistics

From the examination of individual data points as they relate to the testing problem, Ghosh shifted his attention to data summarization, focussing on the relationship between sufficiency and invariance. Sufficiency isolates features of the collection of observations from those of the individual ones which are independent of the features of the collection. Invariance, on the other hand, summarizes data by imposing symmetry constraints. In practice, both sufficiency and invariance restrictions are applied, but their order of application is an issue of interest.

Consider a statistical model $(\mathfrak{X}, \mathscr{A}, \mathcal{P})$ where a group of transformations G is acting on the sample space and attention is limited to invariant procedures. To find a sufficiency reduction, one needs to find a sufficient sub-σ-field $\mathscr{C}$ of the invariant σ-field $\mathscr{I}$. However, in practice, it is typically easier to invoke invariance on the data after it has been reduced by sufficiency. Let $\mathscr{S}$ be a sufficient σ-field. To justify the application of invariance restriction after a sufficiency reduction, it is enough to establish $\mathscr{S} \cap \mathscr{I}$ is sufficient for $\mathscr{I}$. This problem was addressed by W. J. Hall, R. A. Wijsman and J. K. Ghosh, independently and roughly simultaneously. Once they realized they had compatible results, they published a combined paper [65]. Their main result can be described briefly as follows. A statistic T is called almost invariant if, for every $g \in G$, $T(x) = T(gx)$ a.s. Under conditions that imply that every almost invariant set is equivalent, up to null sets, to an invariant set, it follows that $\mathscr{S}$ and $\mathscr{I}$ are conditionally independent given $\mathscr{S} \cap \mathscr{I}$, and hence $\mathscr{S} \cap \mathscr{I}$ is sufficient for $\mathscr{I}$.

Another notion which relates two sequences of σ-fields in sequential experiments is that of transitivity, introduced by Bahadur. Two sequences of σ-fields $\mathscr{B}_n \subset \mathscr{A}_n$ are said to be transitive if for every $\mathscr{B}_{n+1}$-measurable function f, $\mathrm{E}(f|\mathscr{A}_n)$ is $\mathscr{B}_n$ measurable. In the usual sequential set up, $\mathscr{S}_n \cap \mathscr{I}_n$ is transitive for $\mathscr{I}_n$, where the extra index n indicates the sample size. Several implications of this result were discussed in [65].

In many application areas, sample surveys for instance, discrete models arise, where the probability is concentrated on a countable set but the models do not have common support, i.e., the support set is different for different parameter values. Clearly, such a family is not dominated and the Halmos–Savage theorem on sufficiency does not hold. Nevertheless, as shown in [2], minimal sufficient σ-fields exist and the Neyman factorization theorem holds good. These results were extended for pairwise sufficient σ-fields and condition for existence of minimal pairwise sufficient sigma-field was found in [37].

Another basic question is whether a fixed-dimensional sufficient statistic independent of sample size actually exists. In exponential families, it is well known that fixed-dimensional sufficient statistics exist. Outside of exponential families, however, sufficient statistics are hard to find. Some distinguished nonregular cases like $U(-\theta, \theta)$ provide additional examples. In [54], it is shown that if the support $(a(\theta), b(\theta))$ is shrinking or expanding, as in the support of $U(0, \theta)$ for example, then the density must be of the form $g(\theta)h(x)$ to have a real-valued sufficient statistic. If $a(\theta)$ and $b(\theta)$ are both increasing, or both decreasing, as in $U(\theta, \theta + 1)$, then an $\mathbb{R}^2$-valued sufficient statistic can exist only in special cases.

4. Asymptotics

The asymptotic point of view undergirded Ghosh's thinking, even in problems that were not primarily focussed on asymptotic properties. In a sense, much of his work on sequential analysis, Bayesian analysis and Bayesian nonparametrics are also, at least implicitly, work on asymptotics. In fact, many of the most important asymptotic ideas, such as higher order asymptotics and Edgeworth expansions, were pioneered by him. Moreover, in terms of how his thinking progressed, asymptotics can be regarded as the next natural conceptual step after thinking about data points in sequential analysis, and sufficiency or invariance as a data summarization technique. That is, once we have gathered and summarized our data, we want to see where it seems to be leading us.

Ghosh's asymptotic work can be broadly grouped into seven categories. He worked on the Bahadur–Ghosh–Kiefer representation for a quantile. He made foundational contributions to establishing the existence of Edgeworth expansions. In higher order asymptotics, Ghosh examined second order efficiency, Bartlett correction and contributed to our understanding of how Wald, Rao and likelihood ratio tests compare. Then he turned his attention to Bahadur efficiency and the vexing Neyman–Scott problem.

4.1. Bahadur–Ghosh–Kiefer Representation

Bahadur represented a sample quantile as an average of i.i.d. random variables. To get this representation, Bahadur assumed the existence of two derivatives of the c.d.f.; the second derivative is bounded and the first derivative is positive on a neighborhood of the p-th population quantile ξ_p. Then Bahadur showed that the error in the representation is $O(n^{-3/4}(\log n)^{3/4})$. The order of error in Bahadur's representation is nearly sharp, cf. the exact order $n^{-3/4}(\log n)^{1/2}(\log\log n)^{1/4}$ obtained by Kiefer.

One of the reasons this is important is that the order of error is small enough to obtain asymptotic normality for the sample quantiles. However, assuming the existence of two derivatives is somewhat strong. For instance, it rules out the location family from the double exponential density. On the other hand, for most statistical purposes, where only the asymptotic distribution is important, having an error term of order $o_p(n^{-1/2})$ is enough. Therefore it is of interest to weaken Bahadur's assumptions at the expense of weakening the conclusion to $o_p(n^{-1/2})$. This is possible, even for a variable point p_n depending on n, as shown in [30], using only the assumption of positive first derivative at ξ_p.

Actually, the idea of representing a quantile approximately as an average of i.i.d. observations occurred to Ghosh independently in the mid sixties at the same time Bahadur was working on the problem. Ghosh was looking at the problem in the more general multivariate multisample framework in connection with asymptotic normality of multivariate rank tests. He did not record his proof then since it did not extend to the multivariate set up at that time.

4.2. Edgeworth Expansions

Edgeworth expansions are natural refinements of asymptotic normality results in that they give error terms of asymptotically smaller order by including more terms in addition to the leading normal term. However, for a long time, Edgeworth expansions were only heuristically justified. In the pioneering paper [7], it is shown that under conditions of finiteness of certain moments and a condition on characteristic function known as Cramér's condition is the literature, the r-th order Edgeworth expansion of a smooth function of sample averages admits error $O(n^{-(r+1)/2})$. In particular, it follows that for the sample average, finiteness of the $2r$-th moments is required to justify an Edgeworth expansion of order r. The follow-up paper [8] relaxes some moment conditions. A thorough and lucid treatment of Edgeworth expansions and related higher order asymptotics is given in Ghosh's IMS monograph [32].

Another angle on Edgeworth expansions comes from the idea of Fisher consistency. Consider an exponential family with density proportional to $\exp[\sum_{j=1}^{k} w_j(\theta)t_j(x)]$. In this context, an estimator T_n, which is a function of the k-dimensional sufficient statistic $(\sum_{i=1}^{n} t_1(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i))$, is Fisher consistent if $T_n(w(\theta)) = \theta$. Assuming sufficient smoothness conditions and linear independence of the component functions $w_1(\theta), \ldots, w_k(\theta)$, a Fisher consistent estimator can be written as a smooth function of sample averages, and hence has an Edgeworth expansion. In [62], this Edgeworth expansion is compared with that of the MLE, which is another Fisher consistent estimator. Interestingly, for any bowl shaped loss function, the MLE has better second order risk properties than any other Fisher consistent estimator. Consequently, this gives a way to discriminate among estimators which are first order asymptotically equivalent. This property is called second order efficiency and will be discussed in the next subsection.

Edgeworth-type expansions need not be restricted to asymptotically normal estimators. Other limiting distributions can appear naturally. Recall that log-likelihood ratio type statistics are among the most common statistics converging to non-normal limits such as a chi-square distribution. For locally quadratic functions of sample averages, such as the log-likelihood ratio, asymptotic expansions have been obtained in [13]. They have a leading chi-square term. Subsequent terms appear as coefficients of powers of $n^{-1/2}$ and are finite linear combinations of chi-square distributions of degrees of freedoms p, $p+2$, $p+4$, etc., where p is the degree of freedom of the leading term. Similar expansions hold even under contiguous alternatives with non-central chi-squares replacing the chi-square leading term as shown in [14]. The subsequent terms are finite linear combinations of non-central chi-square distributions with degrees of freedoms p, $p+2$, $p+4$, and so forth.

4.3. Second Order Efficiency

Second order efficiency (also called third order efficiency by some authors) is the natural way to compare two asymptotically efficient estimators since they are first order equivalent. In particular, it was widely believed that the MLE, or some suitable variant of it, had, asymptotically, the smallest possible risk up to the second order. Ghosh, among others like Efron, Chibisov, Pfanzagl, Akahira and Takeuchi, made pioneering contributions towards rigorous justification of this assertion in [64]. His main result may be roughly described as follows: Let T_n be an efficient estimator and consider a modification $T_n' = T_n + m(T_n)/n$. Then T_n' can be beaten by $\hat{\theta}_n' = \hat{\theta}_n + g(\hat{\theta}_n)/n$, a modification of the MLE $\hat{\theta}_n$, where the function g depends on T_n and m. Here, by a better estimator, we mean that

$$\lim_{n\to\infty} n^2[\mathrm{E}_\theta\{W(T_n',\theta)\} - \mathrm{E}_\theta\{W(\hat{\theta}_n',\theta)\}] \geq 0,$$

for all $\theta \in \Theta$, for a truncated squared error loss W. This paper also contains other impressive results such as Bhattacharya-type bounds, a Bayesian connection with second order efficiency and a notion of second order asymptotic sufficiency. Similar results about second order efficiency of the MLE for Pitman closeness and any bounded bowl shaped loss function are in [63].

In addition to second order efficiency, there is a notion of second order admissibility. An estimator is second order admissible if there is no estimator which has uniformly smaller second order risk with strict inequality for at least one point. In [59], for estimators of the form $\hat{\theta}_n + g(\hat{\theta}_n)/n$, a necessary and sufficient condition for second order admissibility under squared error loss is obtained.

These second order optimality properties of modified versions of the MLE raise the issue whether the MLE has optimality properties beyond the second order. A nice counterexample in [60], however, concludes negatively. On the other hand, questions on second order admissibility go beyond the MLE to any BAN estimator $\hat{\theta}_n$ modified to $\hat{\theta}_n + g(\hat{\theta}_n) + o_p(n^{-1})$. The condition for second order Pitman admissibility is obtained in [58], and its multiparameter version in [49].

Another natural question in the context of second order admissibility is the following. If two or more statistics are separately second order admissible, for two different components of a parameter with bias $o(n^{-1})$, then, is it true that they are jointly second order admissible? The question has a curious answer given in [16]. For two dimensions, they are jointly second order admissible, but for three or more dimensions, they are not jointly second order admissible. This result is reminiscent of Stein's phenomenon on ordinary admissibility with respect to the squared error loss for estimating the normal mean. Intuitively, asymptotically, all regular experiments are normal experiments and thus a phenomenon under normality continues to hold asymptotically under any regular model. The interesting part of the result is that the phenomenon shows up in the second order.

4.4. Bartlett Correction

Bartlett introduced a remarkable technique, which bears his name, to improve the chi-square approximation to the distribution of a log-likelihood ratio statistic. The idea is embarrassingly simple: rescale the chi-square distribution with the second order expansion of the mean of the statistic. It is surprising that such a simple strategy improves the approximation so much.

In the seminal paper [9], a variant on Wilks theorem tuned to the goal of understanding the Bartlett correction was presented. Recall that Wilks theorem is the statement that the log-likelihood ratio is asymptotically chi-square. However, the chi-square is the result of squaring normals. To see how this might apply to the log-likelihood ratio statistic, let $X_1, X_2, \ldots$ be i.i.d. observations from a parametric family governed by $\theta = (\theta^1, \ldots, \theta^p)$ and let $L(\theta)$ be the log likelihood. For $j = 1, \ldots, p$, let $\hat{\theta}_j$ be the MLE of θ under the null hypothesis $\theta^1 = \theta_0^1, \ldots, \theta^j = \theta_0^j$, and let $T_j = 2n(L(\hat{\theta}_{j-1}) - L(\hat{\theta}_j))^{1/2}\text{sign}(\hat{\theta}_{j-1} - \hat{\theta}_j)$, where $\hat{\theta}_0$ stands for the unrestricted MLE. Note that squaring T_j gives the usual object in Wilks theorem with limiting chi-square behavior. However, now, without the square, $(T_1, \ldots, T_p)$ is asymptotically normal with error $O_p(n^{-3/2})$ under the grand null hypothesis. This property of T_j gives rise to the Bartlett correction in the multidimensional setting.

Another result developed in that paper is a Bayesian version of the Bartlett correction. This is a Bartlett correction to the posterior distribution, conditional on the data, obtained by letting the prior tend to the degenerate distribution at the true parameter value. The relation between the Bartlett correction and the Bayesian correction gives a deeper understanding of the Bartlett correction phenomenon and leads to a variety of generalizations.

Following this path, [41] studied the asymptotic equivalence of the frequentist and Bayesian Bartlett corrections for the likelihood ratio and the conditional likelihood ratio statistic (CLR) introduced by Cox and Reid. In particular, the conditions for equivalence are instrumental for giving a simple proof of the existence of the frequentist Bartlett correction for the CLR statistic. This was extended to the multivariate case in [40]. A variant on the likelihood ratio called the adjusted likelihood ratio (ALR) was introduced by McCullagh and Tibshirani. In [45], it was shown that the ALR statistic has behavior similar to that of the CLR statistic, in that it admits a Bartlett correction and its power under contiguous alternatives, is equivalent to that of the CLR up to the order $o(n^{-1/2})$. In terms of average power, the agreement continues up to $o(n^{-1})$.

4.5. Comparison of the Likelihood Ratio, Wald's and Rao's Statistics

The problem of comparing the likelihood ratio (LR), Wald's and Rao's tests, with regard to power has received significant attention in the statistics and econometrics literature. It is well-known that, up to the first order of approximation and under contiguous alternatives, these three tests have the same local power as dictated by the noncentral chi-square distribution. Discrimination among them, therefore, calls for comparison via higher order power. However, while the LR test is locally unbiased up to a higher order of approximation, the same does not hold in general for the other two tests. From this perspective, to make them really comparable, Ghosh suggested considering locally unbiased versions of Wald's and Rao's tests. This work, done under his supervision, eventually led to an optimum property of Rao's test in terms of third order local power. A review of these developments is available in [31].

In addition, the power properties of the three tests as well as their Bartlett adjustability, when they are developed on the basis of a quasi-likelihood rather than a true density-based likelihood was discussed in [48].

4.6. Bahadur–Cochran Deficiency

To compare the asymptotic performance of two tests, one may look at their Bahadur–Cochran relative efficiency, which is the limit, as $\delta \to 0$, of the ratio of the smallest integers which make the sizes less than δ. For many pairs of reasonable tests, the ratio turns out to be 1. To compare them at a finer level, it is sensible to look at their difference, which may be called the Bahadur–Cochran deficiency. The limit inferior (or superior) of the difference, reflecting the relative advantage of one test over other, was calculated in [12] for some common test statistics.

4.7. Neyman-Scott Problem and Semiparametric Inference

Ghosh made notable contributions in the Neyman–Scott problem also. In the Neyman–Scott problem, a new observation X_i is governed by a common parameter θ and an additional parameter ξ_i, depending on i, but only the parameter θ is of interest. The problem is notoriously difficult in that common estimators, such as the MLE, are usually inconsistent. For instance, if X_{ij} are independently normally distributed with mean μ_i and variance σ^2, $i = 1, \ldots, n$, $j = 1, \ldots, k$, then the MLE of σ^2, $(nk)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{k} (X_{ij} - \bar{X}_i)^2$, where $\bar{X}_i = k^{-1} \sum_{j=1}^{k} X_{ij}$, converges to the wrong value $(1 - k^{-1})\sigma^2$. Although it is easy to correct the MLE in this particular situation, in general identifying the correction is a hard problem.

Ghosh proposed constructing an asymptotically efficient estimator for θ by viewing the ξ_i as random variables arising from an unknown distribution G. The semiparametric model resulting from this can then be explored to find efficient estimators for θ. In addition, efficiency in the Neyman–Scott model can be defined in terms of the semiparametric model so the two models have many interesting links between them. These links may be exploited and are studied in detail in [4], [5] and [6].

5. Bayesian Inference

Ghosh was always very fond of Bayesian ideas and, later in his career, he became more convinced that the Bayesian approach to statistics is more natural and fruitful. Over the course of his investigations, Ghosh examined each aspect of the Bayesian formulation, from construction of a prior to model selection, to asymptotic properties. And again, this can be seen as a continuation from his asymptotic work. After all, once the asymptotics are developed, we want to see how they can be used in a complete inference problem and the Bayesian setting provides a unified context. Indeed, Ghosh's contributions helped speed the development of several branches of Bayesian analysis because of his asymptotic orientation.

Ghosh was always pragmatic and thought that a good statistical method should have good frequentist properties as well as sensible conditional properties. Moreover, as in the frequentist case, asymptotics often play a vital role in Bayesian inference and one of the recurring themes in Ghosh's work was the quest for frequentist properties of posterior distributions. As one of the leaders in developing objective Bayesian methods, he regularly worked to reconcile the two schools of thought. The paper [57] elaborately reviews issues and developments in objective Bayesian methodology.

Ghosh's Bayesian work can be broadly grouped into four categories. He worked on frequentist matching and other objective priors. He worked on determining the limiting behavior of posterior distributions in the parametric context. Then, he turned his attention to richer model classes, examining Bayesian nonparametrics and model selection.

5.1. Matching and Other Objective Priors

Ghosh was never very keen on the term noninformative to describe priors that are constructed through some automatic mechanism rather than through a subjective assessment of odds. His preference was to use these priors as objective or default priors in the absence of genuine subjective information. To him, such priors can be obtained by any one of various techniques including matching what a frequentist might obtain, invariance, entropy-type maximizations (reference priors), approximation, or anything that seem reasonable.

The idea of a Bayesian choosing a prior so as to match frequentist inferences was originally introduced by Peers and Welch, but the term "probability matching prior" was first used by Ghosh and Mukerjee [42] and the approach became popular after Ghosh's presentation in the Valencia meeting. The basic idea is quite simple: choose a prior so that Bayesian notions like credibility approximately agree with the corresponding frequentist notions like confidence level. However, when asymptotic normality of the posterior distribution holds (discussed in the next subsection), it means that the variability according to the posterior distribution of a parameter is asymptotically equivalent to the sampling fluctuations of the MLE in the frequentist sense. This implies that Bayes-frequentist matching occurs for any prior under minimal restrictions. Consequently, to identify a prior uniquely, first order matching of limits is not enough. Satisfyingly, agreement continues to the next order, but only if the prior is of a certain form. Thus matching can be used to characterize a prior, which may then be thought of as objective at least in the sense that it was not chosen according to the personal views of the experimenter.

Of course, neither Bayesian credibility sets nor frequentist confidence sets are unique, so when θ is a scalar, it is natural to look at one sided intervals. If W is a properly centered and normalized version of the parameter, then equating the posterior probability $P_\pi(W \leq t|X_1, \ldots, X_n)$ with the frequentist probability $P_\theta(W \leq t)$ for t and ensuring both sides are $1-\alpha$ up to $o(n^{-1/2})$ for each α leads to a differential equation. The Jeffreys' prior is the solution to this equation.

A multiparameter version of this frequentist-Bayesian matching was used in [43]. In higher dimensions, the components of W may be defined by successively computing the regression residual of the current component over the earlier components. Naturally, this depends on the ordering of the parameters, but the dependence is not present when the parameters are orthogonal. In these cases, the matching criterion leads to partial differential equations. Curiously, Jeffreys' prior is a solution in some cases, but not in all cases: the location-scale problem is an important exception. In fact, it is well known that Jeffreys' prior, which is also the left Haar measure, may be an inappropriate choice in this case, so the matching criterion genuinely leads to sensible solutions even in high dimensional cases.

The matching prior is closely related to other important objective priors such as the reference prior. Reference priors often depend on the role of the parameter; nuisance parameters are treated differently from parameters of interest. Interestingly, in the two parameter case, a cute observation of Ghosh is that the reverse reference prior, rather than the reference prior itself, is probability matching. Here, by reverse reference prior, we mean the reference prior computed by reversing the roles of the parameter of interest and the nuisance parameter. More details and discussion of other properties, such as weak minimaxity, may be found in [42].

Although matching posterior probabilities does yield useful insight, highest posterior density (HPD) regions are more efficient credible sets from a Bayesian standpoint. Accordingly, matching the coverage probability of HPD regions with the credibility is an alternative that might be more appealing to some Bayesians. When this matching is done to $o(n^{-1})$, it leads to differential equations characterizing prior distributions. These were derived in [44]. In some cases, Jeffreys' prior solves these equations and so is a matching prior in the sense of coverage probability as well. A related paper is [46]. Matching the coverage of one-sided posterior credibility intervals for parametric functions up to $O(n^{-1})$ was studied in [17].

Alternatively, instead of characterizing a prior through matching, one might ask if there is some adjustment to make matching work for any prior satisfying mild general conditions. Indeed, in [47], it is shown, with examples, that if the center of the $(1-\alpha)$-HPD ellipsoid is appropriately shifted by a $o(n^{-1/2})$ amount, where the correction is obtained by solving an equation depending on the prior, then the resulting perturbed HPD ellipsoid's coverage is $1-\alpha+o(n^{-1})$.

Of course, there are many sensible notions of objectivity for a prior other than matching. Invariance is often the driving force in group models, where a group of transformation is acting on the parameter space and the parameter of interest is the maximal invariant parametric function. In [18], a detailed study of various priors such as the Chang–Eaves prior for group models is given in the light of matching and the marginalization paradox.

5.2. Limits of Posterior Distributions

One of the most intriguing results in statistics is the Bernstein–von Mises theorem, which states that the posterior distribution of the parameter centered at the MLE and scaled by $\sqrt{n}$ times the square root of the Fisher information converges to the standard normal distribution almost surely, as the sample size increases to infinity. This parallels the frequentist result that $\sqrt{n}(\hat{\theta}-\theta_{\text{true}})$ is asymptotically normal with variance given by the inverse Fisher information. In essence, posterior normality implies that in an asymptotic sense, at least to first order, any sensible Bayesian must agree eventually with frequentist notions of variability.

Ghosh worked to extend posterior normality in a variety of directions. One natural idea is to look at higher order properties so that the usual normal limit is viewed as merely the first term in an asymptotic expansion. This parallels the sense in which an Edgeworth expansion is an improvement over the standard central limit theorem. Johnson pioneered such expansions, but the probability statements in his expansions are in terms of the true distribution of the sample. Often, a Bayesian is more interested in bounds that are uniform on sets with high probability in the marginal distribution of the sample. In [61], precise conditions were given so that Johnson's expansion of posterior distribution holds on a set with marginal probability $1 - O(n^{-r})$, where r is the extra number of terms in the expansion, i.e., not counting the leading normal. It was also shown, by counterexamples, that some of the earlier published results in the field are incorrect.

Sometimes it is meaningful to condition on a statistic rather than the full data to obtain the posterior distribution. In particular, since the sample mean is a widely used summary measure, it is natural to ask if a version of the Bernstein–von Mises theorem holds when the posterior is computed given only the mean. Provided that expectation and variance are smooth functions, and the eigenvalues of the covariance matrix are uniformly bounded and bounded away from zero, it is shown in [15] that a normal limit for the posterior distribution is obtained. The variance of the limiting distribution can equal the variance of an observation, but in general, the normal limit can differ from that in the usual Bernstein–von Mises theorem, unless the sample mean is asymptotically sufficient. The proof is based on an Edgeworth expansion for the sample mean and a local limit theorem. The idea extends to independent but non-identically distributed observations.

More broadly, the Bernstein–von Mises phenomenon in a parametric family may be seen as the convergence of the posterior density of the standardized parameter to a non-degenerate distribution, where in general, the centering need not be at the MLE, the scaling need not be $\sqrt{n}$ and the limit distribution need not be normal. Indeed, in some nonregular families such as the uniform distribution on $[0, \theta]$ or the location family of the exponential distribution, centering by the Bayes estimator and scaling by n yields an exponential limit. This leads to the following question: For which families will a limit of the posterior distribution exist? When it does exist, what is the correct centering, scaling and limiting distribution? This problem is germane to approximating posterior distributions numerically when n is large.

Under the general set up of a parametric family considered by Ibragimov and Has'minskii in their book, a very elegant characterization was given in [35] and [23] in terms of the behavior of the limiting (local) likelihood ratio process of the model, $Z_n(u) = p(X^n; \theta + r_n u)/p(X^n; \theta)$, where X^n is the observation at stage n and r_n is the appropriate normalizer for the problem. Usually $X^n = (X_1, ..., X_n)$ and n is the sample size. Let $Z(u)$ stand for the weak limit of $Z_n(u)$ and $\xi(u) = Z(u)/\int Z(v)dv$, a random probability density. Under the natural scaling in the family, the posterior distribution converges to a limit, after appropriate centering, if and only if $\xi(u) = g(u + W)$ for some fixed probability density g and a random variable W, i.e., as a random element in L_1, $\xi(\cdot)$ is a random location shift of a fixed probability density g. When this holds, g is the limit of the posterior density. Clearly, this is a stringent representation, so in many nonregular cases the posterior distribution will not have a limit.

Interestingly, in the regular cases, local asymptotic normality implies that $\xi(u) = g(u+W)$, in which g is normal and W is a random normal shift. Thus this yields a Bernstein–von Mises theorem under an extremely general condition. A similar limit theorem holds with an exponential limit whenever densities are positively supported on an interval $[a(\theta), b(\theta)]$, where the support is either expanding or contracting.

While it is disappointing to find that posterior limits exist only in relatively rare cases, it does not rule out the possibility of finding useful approximation to the posterior distributions depending on the sample size n. For change-point problems, where the the density jumps from a positive value to another positive value at an unknown location but is otherwise smooth, a useful approximation was obtained in

[24] by normalizing an approximation to the likelihood ratio process. It turns out that a certain mixture of n many truncated and shifted exponential densities is a good approximation.

5.3. Bayesian Nonparametrics

Ghosh's involvement with Bayesian nonparametrics started in the mid 90's with the paper [51] attempting to determine whether the priors used for survival analysis lead to consistency under censoring. This paper showed that for the Dirichlet process, the posterior under censoring can be represented as a Pòlya tree process whose partition depends on the data, and then consistency can be obtained from the tail-free property of Pòlya tree processes. The question is followed up in subsequent papers [53] and [36]. Since then, Ghosh has continued to be one of the most important contributors to understanding the asymptotics of Bayesian nonparametrics.

For instance, the search for a noninformative prior for infinite dimensional models, as an extension to the finite dimensional case, is ongoing. One approach is to generalize the notion of a uniform distribution. This was proposed in [19] using uniform distributions on discrete approximations to a space found by maximal ϵ-dispersed sets. Even in the parametric setting this approach is fundamental and leads to Jeffreys' prior. The approach gives consistency in infinite-dimensional cases also.

More typically, Ghosh was strongly motivated by the examples of inconsistency of posterior distributions in infinite dimensional models. While he appreciated those illuminating examples, he was always hopeful that Bayes' methods would work if priors were constructed properly. He was particularly fond of the Kullback–Leibler property which requires that the true distribution be in the support of the prior when distances are measured in terms of Kullback–Leibler divergence. That is, the prior should assign strictly positive probability to every Kullback–Leibler neighborhood around the true distribution.

Because of this, Ghosh thought the Dirichlet process was inappropriate in many contexts, despite its evident utility, since its discreteness means it fails to have anything in its Kullback–Leibler support. In [20], it was shown that a prior with the Kullback–Leibler property, such as a suitable Pòlya tree or a Dirichlet mixture process, can overcome the inconsistency property of Dirichlet processes for estimating a location parameter. Essentially the same phenomenon appears in linear regression models as shown in [1]. In that paper, the first extension of a general posterior consistency theorem to independent non-identically distributed variables is also developed.

In Bayesian nonparametrics, consistency often combines testing concepts with sieves. A celebrated result of Schwartz emphasizes the role of tests for the true density f_0 versus the complement of a neighborhood, say V, around it. The basic idea is to construct tests, by covering V^c with many small balls, say B_i's, and testing f_0 versus B_i for each i using powerful tests. One can then simply look at the maximum of all tests against each small ball, whose type II error probability is clearly under control and the type I error probability bounded by the common exponential bound for error probability multiplied by the number of small balls required to cover V^c. Thus the concept of metric metropy, which is the logarithm of the number of balls required to cover a set, comes into the picture. Generally V^c is not compact, and it is not possible to cover it by finitely many small balls. The difficulty can be overcome by using a sieve, which is a sequence of increasing subsets of a parameter space that gradually fill out the whole parameter space. One may ignore the portion of the parameter space outside the sieve as long as that part has exponentially small prior probability. Now one requires to control the metric entropy of the sieve to ensure that it does not grow faster than a small multiple of n. This style of proof gives consistency for density estimation with Dirichlet mixtures of normal kernels as shown in [21], providing a large sample justification for the most widely used Bayesian density estimator.

The approach works for density estimation with other priors in place of the Dirichlet mixtures. In [68], consistency is obtained for the logistic Gaussian prior for a density, that is, a prior on densities obtained by exponentiating and then normalizing a Gaussian process.

The importance of entropy for posterior consistency appeared in [21]. There it is seen that in the nonparametric setting prior positivity at the true density must be satisfied, but in terms of special neighborhoods given by the Kullback–Leibler number. Moreover, it must be possible to choose a sieve whose entropy grows no faster than the rate $O(n)$, while ensuring that the prior probability of the complement of sieve is exponentially small as n increases. This perception led to the derivation of the results on posterior convergence rates in [25] in the sense that the conditions for rates can be viewed as quantitative analogs of the conditions for consistency.

For instance, instead of just requiring that the prior for a fixed ϵ neighborhood in the Kullback–Leibler sense has positive probability, one now needs to show that the prior probability of the Kullback-Leibler neighborhood of radius ϵ_n is at least $e^{-n\epsilon_n^2}$, where ϵ_n is the intended rate of posterior convergence. In a similar manner, requiring that the ϵ_n- entropy of the sieve be bounded by a multiple of $n\epsilon_n^2$ is also reminiscent of the condition that the ϵ-entropy of the sieve should be bounded by a small multiple of n. Thus, for fixed ϵ_n, this reduces to the condition for consistency.

The paper also constructs a prior achieving optimal rates of convergence by bracketing densities above and below by two functions — choosing a finite collection of functions to provide upper and lower bounds for any probability density in the given class, ensuring they approximate any function within the bracket together with a control over likelihood ratios. This can be viewed as a refinement of the construction proposed in [19]. Other approaches to optimal rates are also discussed, most notably, through exponential families generated using a B-spline basis.

Many aspects of Bayesian asymptotics for infinite dimensional models are neatly summarized in the review [22], and thoroughly discussed in [52], which to date is the only book dealing with asymptotic results in Bayesian nonparametrics.

5.4. Model Selection and Bayesian Hypothesis Testing

Testing hypotheses is a major area where frequentist and Bayesian procedures often differ substantially. There is a tendency for frequentist methods to over-reject just as there is a tendency for Bayes' methods to under-reject, as in the Lindley paradox. Results such as the consistency of the Bayesian information criterion (BIC) bridge the gap somewhat because the BIC approximates Bayes' factors and is frequentist consistent for model selection under appropriate conditions in the sense that the BIC selects the correct model with probability tending to one.

These properties of the BIC are valid only if the dimension of the model p remains fixed. For many applications, especially for complex data containing numerous variables commonly arising nowadays, the BIC may fail to adequately approximate the

Bayes' factor as well as failing to be consistent, as pointed out by Stone. The main reason for the failure is ignoring certains terms in an expansion of the Bayes' factor which are not negligible when $p \to \infty$. The difficulty could be avoided by paying proper attention to these terms. In [3], a correction is proposed by introducing two more terms, one is proportional to p and the other to $\log p$, as well as changing the meaning of the sample size to the number of replications. The resulting 'generalized BIC' then selects the correct model with increasing high probability. Another generalization of BIC is developed in [11] which works in a general exponential family. These generalizations of BIC are powerful tools to overcome the challenges posed by high-dimensional data problems of contemporary statistics.

In model selection problems, the definition of optimality is often tricky. An appealing approach is comparison with the oracle, that is, with the best procedure (for a given loss function) which uses the knowledge of the correct model in making decisions. A parametric empirical Bayes' (PEB) approach approximates the Bayes' factor by deriving the rule in a parametric model but estimating the parameters in the penalty function by a penalized likelihood criterion with data dependent penalty. In the paper [66], the relative performance of a PEB, the AIC and the BIC were thoroughly studied through asymptotics and simulations under both 0-1 and prediction loss. The conclusion is that the BIC performs badly, but PEB rules perform quite satisfactorily, and so does the AIC. If Bayes' estimates are used in making prediction, instead of least squares estimators, a PEB performs better than the AIC.

One particular difficulty with the Bayes' factor is that it is undefined when improper priors are used in individual models. Various remedies are proposed in the literature, based on the idea of using a part of the information contained in the data (training portion) to make priors proper and use the remaining portion in Bayesian analysis with the obtained "proper prior". Since this typically depends on the ordering of the data, some kind of averaging, through bootstrap or cross validation, over different choices of the training portion is desired. A particularly popular candidate among these Bayes' factors is obtained by taking a geometric average. In the paper [67], such Bayes' factors are studied through asymptotics as the proportion of the training sample varies, and conditions for consistency are obtained as the total sample size goes to infinity. It is also concluded that predictive optimality of the 'geometric Bayes' factor' as it is generally claimed is not entirely correct.

There are many other significant papers on model selection authored by Ghosh. In [10], optimality of the AIC in inference about Brownian motion is shown. The reviews [56] and [55] contain wealth of information about Bayesian model selection.

6. Concluding Remarks

Overall, Ghosh's work in statistics reveals a progression. He began with individual data points, proceeded to data summarization, and then to the asymptotics of inference. Ghosh's results there were a successful attempt to map out where the accumulation of data points. In a sense, asymptotic limits are the ultimate data summarization. Then, putting it all together, Ghosh turned to the Bayesian formulation, examining each of its components, prior, model, posterior, in turn, to permit a comprehensive and unified study of the statistical problem. Indeed, his recent work on Bayesian nonparametrics is a further generalization, again a logical step because it builds on his earlier work by using ever richer model classes.

In fact, Ghosh worked in many more areas of statistics, apart from what is outlined above, as well as working on a variety of applications. These topics include distribution theory, decision theory, robustness, finite population sampling, reliability, quality control, modeling hydrocarbon discovery, geological mapping and DNA fingerprinting.

Finally, every great researcher has a strategy, a method or a drive, often summarized in a maxim, that guides or motivates their intellectual endeavors. One of Ghosh's maxims was the injunction: "Settle the question!" By this he meant formulate a question so that answering it gives you something definite for the formulation of another question. As can be inferred from the progression of his work, his questioning led him to an ever broader view of the statistical problem, culminating in a Bayesian treatment of high-dimensional models, nonparametric or not. Ghosh's injunction to settle questions has helped, and will continue to help, researchers all over the world to think deeply about the most important issues.

References

[1] Amewou-Atisso, M., Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (2003). Posterior consistency for semi-parametric regression problems. *Bernoulli* **9**, 291–312.

[2] Basu, D. and Ghosh, J. K. (1969). Sufficient statistics in sampling from a finite universe. *Bull. Inst. Internat. Statist.* **42**, 850–858.

[3] Berger, J. O., Ghosh, J. K. and Mukhopadhyay, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *J. Statist. Plann. Inference* **112**, 241–258.

[4] Bhanja, J. and Ghosh, J. K. (1992). Efficient estimation with many nuisance parameters. I. *Sankhyā Ser. A* **54**, 1–39.

[5] Bhanja, J. and Ghosh, J. K. (1992). Efficient estimation with many nuisance parameters. II. *Sankhyā Ser. A* **54**, 135–156.

[6] Bhanja, J. and Ghosh, J. K. (1992). Efficient estimation with many nuisance parameters. III. *Sankhyā Ser. A* **54**, 297–308.

[7] Bhattacharya, R. N. and Ghosh, J. K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.* **6**, 434–451.

[8] Bhattacharya, R. N. and Ghosh, J. K. (1988) On moment conditions for valid formal Edgeworth expansions. *J. Multivariate Anal.* **27**, 68–79.

[9] Bickel, P. J. and Ghosh, J. K. (1990). A decomposition for the likelihood ratio statistic and the Bartlett correction—a Bayesian argument. *Ann. Statist.* **18**, 1070–1090.

[10] Chakrabarti, A. and Ghosh, J. K. (2006). Optimality of AIC in inference about Brownian motion. *Ann. Inst. Statist. Math.* **58**, 1–20.

[11] Chakrabarti, A. and Ghosh, J. K. (2006). A generalization of BIC for the general exponential family. *J. Statist. Plann. Inference* **136**, 2847–2872.

[12] Chandra, T. K. and Ghosh, J. K. (1978). Comparison of tests with same Bahadur efficiency. *Sankhyā Ser. A* **40**, 253–277.

[13] Chandra, T. K. and Ghosh, J. K. (1979). Valid asymptotic expansions for the likelihood ratio statistic and other perturbed chi-square variables. *Sankhyā Ser. A* **41**, 22–47.

[14] Chandra, T. K. and Ghosh, J. K. (1980). Valid asymptotic expansions for the likelihood ratio and other statistics under contiguous alternatives. *Sankhyā Ser. A* **42**, 170–184.

[15] Clarke, B. and Ghosh, J. K. (1995). Posterior convergence given the mean. *Ann. Statist.* **23**, 2116–2144.

[16] DasGupta, A. and Ghosh, J. K. (1983). Some remarks on second-order admissibility in the multiparameter case. *Sankhyā Ser. A* **45**, 181–190.

[17] Datta, G. S. and Ghosh, J. K. (1995). On priors providing frequentist validity for Bayesian inference. *Biometrika* **82**, 37–45.

[18] Datta, G. S. and Ghosh, J. K. (1995). Noninformative priors for maximal invariant parameter in group models. *Test* **4**, 95–114.

[19] Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1997). Noninformative priors via sieves and packing numbers. In *Advances in Statistical Decision Theory and Applications*, 119–132, Stat. Ind. Technol., Birkhuser Boston, Boston, MA.

[20] Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999). Consistent semiparametric Bayesian inference about a location parameter. *J. Statist. Plann. Inference* **77**, 181–193.

[21] Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27**, 143–158.

[22] Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999). Consistency issues in Bayesian nonparametrics. In *Asymptotics, Nonparametrics, and Time Series* 639–667, Statist. Textbooks Monogr., **158**, Dekker, New York.

[23] Ghosal, S., Ghosh, J. K. and Samanta, T. (1995). On convergence of posterior distributions. *Ann. Statist.* **23**, 2145–2152.

[24] Ghosal, S., Ghosh, J. K. and Samanta, T. (1999). Approximation of the posterior distribution in a change-point problem. *Ann. Inst. Statist. Math.* **51**, 479–497.

[25] Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28**, 500–531.

[26] Ghosh, J. K. (1960). On some properties of sequential t-test. *Calcutta Statist. Assoc. Bull.* **9**, 77–86.

[27] Ghosh, J. K. (1960). On the monotonicity of the *OC* of a class of sequential probability ratio tests. *Calcutta Statist. Assoc. Bull.* **9**, 139–144.

[28] Ghosh, J. K. (1961). On the optimality of probability ratio tests in sequential and multiple sampling. *Calcutta Statist. Assoc. Bull.* **10**, 73–92.

[29] Ghosh, J. K. (1964). Bayes solutions in sequential problems for two or more terminal decisions and related results. *Calcutta Statist. Assoc. Bull.* **13**, 101–122.

[30] Ghosh, J. K. (1971). A new proof of the Bahadur representation of quantiles and an application. *Ann. Math. Statist.* **42**, 1957–1961.

[31] Ghosh, J. K. (1991). Higher order asymptotics for the likelihood ratio, Rao's and Wald's tests. *Statist. Probab. Lett.* **12**, 505–509.

[32] Ghosh, J. K. (1994). *Higher Order Asymptotics.* NSF-CBMS Regional Conference in Probability and Statistics **4**, IMS, Hayward.

[33] Ghosh, J. K. and Chaudhuri, A. R. (1984). An invariant SPRT for identification. *Sequential Anal.* **3**, 99–120.

[34] Ghosh, J. K., Delampady, M. and Samanta, T. (2007). *An Introduction to Bayesian Analysis, Theory and Methods.* Springer Texts in Statistics. Springer, New York.

[35] Ghosh, J. K., Ghosal, S. and Samanta, T. (1994). Stability and convergence of the posterior in non-regular problems. In *Statistical Decision Theory and Related Topics V* 183–199 Springer, New York.

[36] Ghosh, J. K., Hjort, N. L., Messan, C. and Ramamoorthi, R. V. (2006). Bayesian bivariate survival estimation. *J. Statist. Plann. Inference* **136**, 2297–2308.

[37] Ghosh, J. K., Morimoto, H. and Yamada, S. (1981). Neyman factorization and minimality of pairwise sufficient subfields. *Ann. Statist.* **9**, 514–530.

[38] Ghosh, J. K. and Mukerjee, R. (1989). Some optimality results on Stein's two-stage sampling. In *Statistical Data Analysis and Inference* 251–256, North-Holland, Amsterdam.

[39] Ghosh, J. K. and Mukerjee, R. (1990). Improvement in Stein's procedure using a random confidence coefficient. *Calcutta Statist. Assoc. Bull.* **40**, 145–152.

[40] Ghosh, J. K. and Mukerjee, R. (1991). Characterization of priors under which Bayesian and frequentist Bartlett corrections are equivalent in the multiparameter case. *J. Multivariate Anal.* **38**, 385–393.

[41] Ghosh, J. K. and Mukerjee, R. (1992). Bayesian and frequentist Bartlett corrections for likelihood ratio and conditional likelihood ratio tests. *J. Roy. Statist. Soc. Ser. B* **54**, 867–875.

[42] Ghosh, J. K. and Mukerjee, R. (1992). Non-informative priors. In *Bayesian Statistics* **4**, 195–210, Oxford Univ. Press, New York.

[43] Ghosh, J. K. and Mukerjee, R. (1993). On priors that match posterior and frequentist distribution functions. *Canad. J. Statist.* **21**, 89–96.

[44] Ghosh, J. K. and Mukerjee, R. (1993). Frequentist validity of highest posterior density regions in the multiparameter case. *Ann. Inst. Statist. Math.* **45**, 293–302.

[45] Ghosh, J. K. and Mukerjee, R. (1994). Adjusted versus conditional likelihood: power properties and Bartlett-type adjustment. *J. Roy. Statist. Soc. Ser. B* **56**, 185–188.

[46] Ghosh, J. K. and Mukerjee, R. (1995). Frequentist validity of highest posterior density regions in the presence of nuisance parameters. *Statist. Decisions* **13**, 131–139.

[47] Ghosh, J. K. and Mukerjee, R. (1995). On perturbed ellipsoidal and highest posterior density regions with approximate frequentist validity. *J. Roy. Statist. Soc. Ser. B* **57**, 761–769.

[48] Ghosh, J. K. and Mukerjee, R. (2001). Test statistics arising from quasi likelihood: Bartlett adjustment and higher-order power. *J. Statist. Plann. Inference* **97**, 45–55.

[49] Ghosh, J. K., Mukerjee, R. and Sen, P. K. (1996). Second-order Pitman admissibility and Pitman closeness: the multiparameter case and Stein-rule estimators. *J. Multivariate Anal.* **57**, 52–68.

[50] Ghosh, J. K., Purkayastha, S. and Samanta, T. (2004). Sequential probability ratio tests based on improper priors. *Sequential Anal.* **23**, 585–602.

[51] Ghosh, J. K. and Ramamoorthi, R. V. (1995). Consistency of Bayesian inference for survival analysis with or without censoring. In *Analysis of Censored Data*, IMS Lecture Notes Monogr. Ser., **27**, Inst. Math. Statist., Hayward, CA.

[52] Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics.* Springer-Verlag, New York.

[53] Ghosh, J. K., Ramamoorthi, R. V. and Srikanth, K. R. (1999). Bayesian analysis of censored data. *Statist. Probab. Lett.* **41**, 255–265.

[54] Ghosh, J. K. and Roy, K. K. (1972). Families of densities with non-constant carriers which have finite dimensional sufficient statistics. *Sankhyā Ser. A* **34**, 205–226.

[55] Ghosh, J. K. and Samanta, T. (2001). Model selection – an overview. *Current Science* **80**, (9), 1135–1144.

[56] Ghosh, J. K. and Samanta, T. (2002). Nonsubjective Bayes testing—an overview. *J. Statist. Plann. Inference* **103**, 205–223.

[57] Ghosh, J. K. and Samanta, T. (2002). Towards a nonsubjective Bayesian paradigm. Uncertainty and optimality, 1–69, World Sci. Publ., River Edge, NJ.

[58] Ghosh, J. K., Sen, P. K. and Mukerjee, R. (1994). Second-order Pitman closeness and Pitman admissibility. *Ann. Statist.* **22**, 1133–1141.

[59] Ghosh, J. K. and Sinha, B. K. (1981). A necessary and sufficient condition for second order admissibility with applications to Berkson's bioassay problem. *Ann. Statist.* **9**, 1334–1338.

[60] Ghosh, J. K. and Sinha, B. K. (1982). Third order efficiency of the MLE—a counterexample. *Calcutta Statist. Assoc. Bull.* **31**, 151–158.

[61] Ghosh, J. K., Sinha, B. K. and Joshi, S. N. (1982). Expansions for posterior probability and integrated Bayes risk. In *Statistical Decision Theory and Related Topics III* **1**, 403–456, Academic Press, New York-London.

[62] Ghosh, J. K., Sinha, B. K. and Subramanyam, K. (1979). Edgeworth expansions for Fisher-consistent estimators and second order efficiency. *Calcutta Statist. Assoc. Bull.* **28**, 1–18.

[63] Ghosh, J. K., Sinha, B. K. and Wieand, H. S. (1980). Second order efficiency of the MLE with respect to any bounded bowl-shaped loss function. *Ann. Statist.* **8**, 506–521.

[64] Ghosh, J. K. and Subramanyam, K. (1974). Second order efficiency of maximum likelihood estimators. *Sankhyā Ser. A* **36**, 325–358.

[65] Hall, W. J., Wijsman, R. A. and Ghosh, J. K. (1965). The relationship between sufficiency and invariance with applications in sequential analysis. *Ann. Math. Statist.* **36**, 575–614.

[66] Mukhopadhyay, N. and Ghosh, J. K. (2003). Parametric empirical Bayes model selection—some theory, methods and simulation. In *Probability, Statistics and Their Applications: Papers in Honor of Rabi Bhattacharya*, 229–245, IMS Lecture Notes Monogr. Ser., **41**, Inst. Math. Statist., Beachwood, OH.

[67] Mukhopadhyay, N., Ghosh, J. K. and Berger, J. O. (2005). Some Bayesian predictive approaches to model selection. *Statist. Probab. Lett.* **73**, 369–379.

[68] Tokdar, S. T. and Ghosh, J. K. (2007). Posterior consistency of logistic Gaussian process priors in density estimation. *J. Statist. Plann. Inference* **137**, 34–42.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 19

arXiv: math.PR/0000000

Objective Bayesian Analysis under Sequential Experimentation

Dongchu Sun[1] **and James O. Berger**[2]

University of Missouri-Columbia and Duke University

Abstract: Objective priors for sequential experiments are considered. Common priors, such as the Jeffreys' prior and the reference prior, will typically depend on the stopping rule used for the sequential experiment. New expressions for reference priors are obtained in various contexts, and computational issues involving such priors are considered.

Contents

1. Introduction

Bayesian analysis using objective or default priors has received considerable attention; cf. Datta and Mukerjee [17], Bernardo [6, 7] Berger and Bernardo [4], Berger [3], Ghosh, Delampady and Samanta [21], and references therein. The latter book, in particular, contains an excellent discussion of the issues and controversies involving objective priors, reflecting the many years of leadership of J.K. Ghosh in the field (along with his many coauthors). See, for example, [13, 20, 22, 23].

A common objective prior is the Jeffreys' prior [27], which is proportional to the square root of the determinant of the Fisher information matrix. The Jeffreys' prior is quite useful for a single parameter model, but can be seriously deficient for multiparameter models; this has led to preference for reference priors in multiparameter situations (cf. Berger and Bernardo [5], and Bernardo [7]).

[1] Department of Statistics, University of Missouri-Columbia, 146 Middlebush Hall, Columbia, MO 65211-6100, USA; e-mail: sund@missouri.edu; url: www.stat.missouri.edu/∼dsun

[2] Department of Statistical Science, Duke University, Box 90251, Durham, NC 27708-0251, USA; e-mail: berger@stat.duke.edu; url: www.stat.duke.edu/∼berger

AMS 2000 subject classifications: Primary 62L12, 62C10; secondary 62F15, 62L10.

Keywords and phrases: posterior distributions, sequential experimentation, expected stopping time, Jeffreys' prior, reference prior, frequentist coverage.

Almost all results on objective priors have been for fixed sample size experiments. In practice, however, statistical experiments are often conducted sequentially, with a known stopping rule (cf. Siegmund [30], and Ghosh, Sen and Mukhopadhyay [24]). Bartholomew [2] and Geisser [19] introduced the notion that objective priors for a sequential experiment should depend on the expected stopping time. Ye [38] derived the reference prior for sequential experiments when the expected stopping time depends on the parameter of interest only. In this paper we generalize Ye's result in various directions, and provide some new computational tools for use with priors that depend on expected stopping times.

The paper is arranged as follows. Section 2 reviews the Fisher information matrix for sequential experiments with a known stopping rule, derives the Jeffreys'/reference prior for illustrative one-parameter examples, and then provides an expression for multiparameter reference priors when the stopping rule satisfies a certain property. In Section 3, reference priors and matching priors (cf. Datta and Mukerjee [17]) are derived for Bar-Lev and Reiser's [1] two-parameter exponential family. Illustrations are given for normal distributions with several commonly used stopping times.

Computation of expected stopping times is often difficult, so that utilization of reference priors for sequential experiments is typically challenging. In Section 4, an approximation to the reference prior for sequential experiments is introduced which is exact under some circumstances, seems to be a reasonable approximation in general, and allows for much simpler computation.

2. Noninformative Priors with a Known Stopping Rule

2.1. Notation and the Jeffreys'-rule Prior

We assume that $X_1, X_2 \cdots$, is an i.i.d. sequence of random variables with density $f(x \mid \boldsymbol{\theta})$ that is *regular* (Walker [35].) Here $\boldsymbol{\theta}$ is a $q \times 1$ vector of unknown parameters. Let N denote a proper stopping time for the sequential experiment – see Govindarajulu [25] for a definition, which also is a source for the following well-known lemma:

Lemma 2.1. *Let $\mathbf{I}(\boldsymbol{\theta})$ be the Fisher information matrix based on X_1. Under the proper stopping time N, the Fisher information based on $(X_1, \cdots, X_N)$ is*

$$\mathbf{I}^* = E_{\boldsymbol{\theta}}(N)\mathbf{I}(\boldsymbol{\theta}). \tag{2.1}$$

The Jeffreys'-rule prior [27] for $\boldsymbol{\theta}$ is defined as the square root of the determinant of the Fisher information matrix. In the fixed sample size case, this is $\pi_J(\boldsymbol{\theta}) \propto |\mathbf{I}(\boldsymbol{\theta})|^{1/2}$. For the sequential experiment, it follows from the above lemma that Jeffreys' prior is

$$\pi_J^*(\boldsymbol{\theta}) \propto \{E_{\boldsymbol{\theta}}(N)\}^{q/2}|\mathbf{I}(\boldsymbol{\theta})|^{1/2} \propto \{E_{\boldsymbol{\theta}}(N)\}^{q/2}\pi_J(\boldsymbol{\theta})\,. \tag{2.2}$$

Example 2.1. *Let N_r be a random variable with a negative binomial distribution $NB(r,p)$, where r is a positive integer and $p \in (0,1)$. Let $X_1, X_2, \cdots$ be a sequence of Bernoulli random variables with success probability p. N_r can be viewed as a stopping time for this Bernoulli sequence as follows:*

$$N_r = \inf\{n \geq 1 : X_1 + \cdots + X_n = r\}.$$

The probability of N_r is

$$P(N_r = k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \quad \textit{for } k = r, r+1, \cdots$$

An easy computation yields $E_p(N_r) = r/p$. Since the Jeffreys' rule prior for a Bernoulli random variable is $\pi_J(p) \propto 1/\sqrt{p(1-p)}$, it follows from (2.2) that the Jeffreys' rule prior for the negative binomial distribution is

$$\pi_J^*(p) \propto \frac{r}{p}\,\pi_J(p) \propto \frac{1}{p\sqrt{1-p}}.$$

This, of course, is well known from a direct computation with the negative binomial distribution, as discussed in Geisser [8] and Bernardo and Smith ([19], Example 5.14, p. 315).

We next consider an example with a continuous stopping time.

Example 2.2. *Let $\{Z(t) : t > 0\}$ be a Brownian motion with constant drift θ and variance 1 per unit time, so $Z(t) \sim N(\theta t, t)$. Let $-\infty < a < 0 < b < \infty$, and let T_{ab} denote the random stopping time*

$$T_{ab} = \inf\{t > 0 : Z(t) \le a \textit{ or } Z(t) \ge b\}. \tag{2.3}$$

It follows from Hall [26] that

$$E_\theta(T_{ab}) \quad = \quad \begin{cases} \dfrac{1}{\theta}\left[b - (b-a)\dfrac{e^{2b\theta}-1}{e^{2(b-a)\theta}-1}\right], & \textit{if } \theta \neq 0, \\ -ab, & \textit{if } \theta = 0. \end{cases}$$

Note that the constant prior is the Jeffreys' prior based on stopping at a fixed time (Polson and Roberts [29]; Sivaganesan and Lingam [31]), from which it follows that the Jeffreys' or reference prior for this situation is

$$\pi(\theta) = \sqrt{E_\theta(T_{ab})}.$$

This is of additional interest because of the study in Brown [10], which showed that the commonly used estimate $Z(T)/T$, which is the posterior mean under a constant prior for θ, is inadmissible under estimation with squared error loss. Brown [10] further suggested that prior distributions which behaved like $|\theta|^{-1}$ as $|\theta| \to \infty$ were optimal for this situation. The Jeffreys'/reference prior has behavior $|\theta|^{-1/2}$ as $|\theta| \to \infty$, and so is not of this form, but admissibility is very dependent on the loss function used. Indeed, it can be argued that a weighted-squared error loss is appropriate for this situation, and the reference prior is likely admissible for an appropriate weight.

2.2. Reference Priors

Reference priors depend on a grouping and ordering of the parameters; see Berger and Bernardo [4, 5]. Suppose that $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(1)}, \cdots, \boldsymbol{\theta}_{(m)})$ is an m-ordered grouping, where the dimension of component $\boldsymbol{\theta}_{(i)}$ is q_i for $i = 1, \cdots, m$. Datta and Ghosh [14] considered the special case in which the (fixed sample size) Fisher information matrix is diagonal, with the diagonal elements being products of functions of the $\boldsymbol{\theta}_{(i)}$. Our first result is a generalization of their result.

Theorem 2.1. *Suppose that the Fisher information matrix corresponding to a single observation X_1 is of the form*

$$\mathbf{I}(\boldsymbol{\theta}) = diag\Big(\prod_{i=1}^{m} \boldsymbol{G}_{1i}(\boldsymbol{\theta}_{(i)}), \cdots, \prod_{i=1}^{m} \boldsymbol{G}_{mi}(\boldsymbol{\theta}_{(i)})\Big), \tag{2.4}$$

where $\boldsymbol{G}_{li}$ is a $q_i \times q_i$ matrix. Assume further that the expected stopping time is of the form

$$E_{\boldsymbol{\theta}}(N) = \prod_{i=1}^{m} g_i(\boldsymbol{\theta}_{(i)}). \tag{2.5}$$

Then the reference prior for $\boldsymbol{\theta}$ in the sequential experiment is

$$\pi_R^*(\boldsymbol{\theta}_{(1)}, \cdots, \boldsymbol{\theta}_{(m)}) \quad \propto \quad \prod_{i=1}^{m} [g_i(\boldsymbol{\theta}_{(i)})]^{q_i/2} \pi_R(\boldsymbol{\theta}_{(1)}, \cdots, \boldsymbol{\theta}_{(m)}), \tag{2.6}$$

where $\pi_R(\boldsymbol{\theta}_{(1)}, \cdots, \boldsymbol{\theta}_{(m)})$ is the reference prior based on the single observation X_1, given by

$$\pi_R(\boldsymbol{\theta}_{(1)}, \cdots, \boldsymbol{\theta}_{(m)}) = \prod_{i=1}^{m} |\boldsymbol{G}_{ii}(\boldsymbol{\theta}_{(i)})|^{1/2}. \tag{2.7}$$

Proof. The proof is essentially identical to that in Datta [12], noting that, under (2.5), the sequential Fisher information matrix has the product structure of Datta and Ghosh [13–15]. □

This theorem can also be considered to be a generalization of Ye [38], who considered the case where $E_{\boldsymbol{\theta}}(N)$ depends only on $\boldsymbol{\theta}_{(1)}$, the parameter of interest.

Berger and Bernardo [5] suggested that one should always try to use a one-at-a-time reference prior, where each component of the grouping of parameters contains only one parameter, and much of the subsequent literature has validated this suggestion. We thus take it as given here that a one-at-a-time reference prior is the desired target. The following result is an immediate corollary of Theorem 2.1.

Corollary 2.1. *Suppose that the conditions of Theorem 2.1 hold. If $q_i = 1$, for $i = 1, \cdots, m = k$, then the resulting one-at-a-time reference prior for $\boldsymbol{\theta}$ in the sequential experiment is*

$$\pi_R^*(\boldsymbol{\theta}) \quad \propto \quad \sqrt{E_{\boldsymbol{\theta}}(N)}\pi_R(\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_k).$$

For later purposes, we also note another corollary of Theorem 2.1, which applies if the dimension of each component of the grouping of parameters has dimension 2.

Corollary 2.2. *Suppose that the conditions of Theorem 2.1 hold. If all $q_j = 2$, then the reference prior for $\boldsymbol{\theta}$ in the sequential experiment is*

$$\pi_R^*(\boldsymbol{\theta}) \quad \propto \quad E_{\boldsymbol{\theta}}(N)\pi_R(\boldsymbol{\theta}_{(1)}, \cdots, \boldsymbol{\theta}_{(m)}).$$

3. A Two-Parameter Exponential Family

3.1. The Model and Reference Priors

Bar-Lev and Reiser [1] considered the following density function of the generic two-parameter exponential family:

$$f(x \mid \theta_1, \theta_2) \quad = \quad a(x)\exp\{\theta_1 U_1(x) - \theta_1 G_2'(\theta_2) U_2(x) - \psi(\theta_1, \theta_2)\}, \tag{3.1}$$

where $\theta_1 < 0$, $\theta_2 = E\{U_2(X) \mid (\theta_1, \theta_2)\}$, $G_i(\cdot), (i = 1, 2)$ are infinitely differentiable functions satisfying $G_i'' > 0$, and

$$\psi(\theta_1, \theta_2) = -\theta_1\{\theta_2 G_2'(\theta_2) - G_2(\theta_2)\} + G_1(\theta_2).$$

This is a large class of distributions, which includes, for suitable choices of G_1, G_2, U_1 and U_2, many popular statistical models such as the normal, inverse normal, gamma, and inverse gamma. Table 1, reproduced from Sun [32], indicates how each distribution arises.

Table 1

Special cases of Bar-Lev and Reiser's [1] two parameter exponential family, where $h(\theta_1) = -\theta_1 + \theta_1 \log(-\theta_1) + \log(\Gamma(-\theta_1))$.

	$G_1(\theta_1)$	$G_2(\theta_2)$	$U_1(x)$	$U_2(x)$	θ_1	θ_2
$N(\mu, \sigma^2)$	$-\frac{1}{2}\log(-2\theta_1)$	θ_2^2	x^2	x	$-1/(2\sigma^2)$	μ
Inverse Gaussian	$-\frac{1}{2}\log(-2\theta_1)$	$1/\theta_2$	$1/x$	x	$-\alpha/2$	$\sqrt{\alpha/\mu}$
Gamma	$h(\theta_1)$	$-\log\theta_2$	$-\log x$	x	$-\alpha$	μ
Inverse Gamma	$h(\theta_1)$	$-\log\theta_2$	$\log x$	$1/x$	$-\alpha$	μ

Let $X_1, X_2, \cdots$ be a sequence of random variables from (3.1). The Fisher information per observation is

$$\mathbf{I}(\theta_1, \theta_2) = \begin{pmatrix} G_1''(\theta_1) & 0 \\ 0 & -\theta_1 G_2''(\theta_2) \end{pmatrix}.$$

The two parameters θ_1 and θ_2 are orthogonal in the sense of Cox and Reid [11]. Thus the Jeffreys' prior for a single observation is

$$\pi_{J}(\theta_1, \theta_2) \propto \sqrt{|\theta_1|}\sqrt{G_1''(\theta_1)G_2''(\theta_2)}. \tag{3.2}$$

When either θ_1 or θ_2 is the parameter of interest, it is shown in Sun and Ye [33] that the one-at-a-time reference priors are

$$\pi_{R}(\theta_1, \theta_2) = \sqrt{G_1''(\theta_1)G_2''(\theta_2)}. \tag{3.3}$$

The parameter θ_2 is the expectation of $U_2(X_1)$. Bose and Boukai [9] considered inference about θ_2 in sequential experimentation with the following stopping time:

$$N_a \quad = \quad \inf\Big\{n \geq m_0 : Y_n < nG_1'\Big(-\frac{a^2}{n^2}\Big)\Big\}, \quad a \geq 0, \tag{3.4}$$

where $Y_n = n^{-1}\sum_{i=1}^n U_1(X_i) - G_2\{n^{-1}\sum_{i=1}^n U_2(X_i)\}$ and $m_0 \geq 2$ is an initial sample size. From Theorem 2 of Bose and Boukai [9], we have

$$\lim_{a\to\infty} \frac{N_a}{a} \quad = \quad \frac{1}{\sqrt{|\theta_1|}} \quad a.s. \tag{3.5}$$

$$\lim_{a\to\infty} E_{\boldsymbol{\theta}}\Big(\frac{N_a}{a}\Big) \quad = \quad \frac{1}{\sqrt{|\theta_1|}}\,. \tag{3.6}$$

Bar-Lev and Reiser [1] showed that the distribution of Y_n does not depend on the parameter θ_2. So condition (2.5) satisfies when either θ_1 or θ_2 is the parameter of interest. The following result is immediate from Theorem 2.1 or Corollary 2.1.

Fact 3.1. *(a) The Jeffreys' prior for (θ_1, θ_2) in model (3.1) with the stopping time (3.4) and when a is large is approximately*

$$\pi_J^*(\theta_1, \theta_2) \propto \sqrt{G_1''(\theta_1) G_2''(\theta_2)}. \tag{3.7}$$

(b) The one-at-a-time reference prior for (θ_1, θ_2) in model (3.1), when either θ_1 or θ_2 is the parameter of interest, the stopping time (3.4) is used, and a is large enough, is approximately

$$\pi_R^*(\theta_1, \theta_2) \propto \frac{1}{|\theta_1|^{1/4}} \sqrt{G_1''(\theta_1) G_2''(\theta_2)}. \tag{3.8}$$

Example 3.1. *Suppose $X_1, X_2, \cdots,$ are a sequence of $N(\mu, \sigma^2)$ random variables. Then $\theta_1 = -1/2\sigma^2, \theta_2 = \mu$, $G_1'(\theta_1) = -1/2\theta_1$, and $Y_n = \sum_{i=1}^n (X_i - \overline{X}_n)^2$. The stopping rule (3.4) becomes*

$$N_a = \inf\Big\{n \geq m_0 : n^{-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2 < n^2/(2a^2)\Big\}.$$

So the priors (3.2), (3.3), (3.7), and (3.8) are, respectively,

$$\pi_J(\mu, \sigma^2) \propto \frac{1}{(\sigma^2)^{3/2}},\ \pi_R(\mu, \sigma^2) \propto \frac{1}{\sigma^2},\ \pi_J^*(\mu, \sigma^2) \propto \frac{1}{\sigma^2},\ \pi_R^*(\mu, \sigma^2) \propto \frac{1}{(\sigma^2)^{3/4}}$$

or equivalently,

$$\pi_J(\mu, \sigma) \propto \frac{1}{\sigma^2},\ \pi_R(\mu, \sigma) \propto \frac{1}{\sigma},\ \pi_J^*(\mu, \sigma) \propto \frac{1}{\sigma},\ \pi_R^*(\mu, \sigma) \propto \frac{1}{\sqrt{\sigma}}.$$

3.2. Probability Matching Priors for a Sequential Experiment

Asymptotic frequentist coverage is an often-used criterion to compare objective priors; see Welch and Peers [36], Peers [28], Tibshirani [34], Datta and Ghosh [13], Datta, Ghosh, and Mukerjee [16], and Datta and Mukerjee [17] for discussion and references. The most common approach is to find a 'matching prior,' i.e., a prior which results in posterior one-sided credible intervals that are also accurate as frequentist confidence intervals. Another type of matching prior, considered by Sun and Ye [33], is a prior such that the confidence interval based on the signed squared root transformation of the log-likelihood ratio is also a Bayesian credible interval. Almost all of the literature considers the fixed sample case for i.i.d. observations; exceptions are Ye [38] and Sun [32].

For sequential experiments involving the Bar-Lev and Reiser [1] two-parameter exponential family, let $l_n(\theta_1, \theta_2)$ be the log-likelihood function of (θ_1, θ_2), given $\boldsymbol{X}_n = (X_1, \cdots, X_n)$, and let $(\hat{\theta}_{n1}, \hat{\theta}_{n2})$ be the maximum likelihood estimator of (θ_1, θ_2). Write

$$Y_n = n^{-1} \sum_{i=1}^n U_1(X_i) - G_2\{n^{-1} \sum_{i=1}^n U_2(X_i)\}.$$

Then, on $\{Y_n \in G_1'(\Theta_1)\} \cap \{n^{-1} \sum_{i=1}^n U_2(X_i) \in \Theta_2\}$, $\hat{\theta}_{n1}$ is the solution of $Y_n = G_1'(\hat{\theta}_{n1})$, and $\hat{\theta}_{n2} = n^{-1} \sum_{i=1}^n U_2(X_i)$. Define

$$\begin{aligned} I_1(\omega_1, \theta_1) &= G_1(\theta_1) - G_1(\omega_1) - G_1'(\omega_1)(\theta_1 - \omega_1), \quad \omega_1, \theta_1 \in \Theta_1, \\ I_2(\omega_2, \theta_2) &= G_2(\omega_2) - G_2(\theta_2) - G_2'(\theta_2)(\omega_2 - \theta_2), \quad \omega_2, \theta_2 \in \Theta_2. \end{aligned}$$

From the convexity of G_1 and G_2, these two functions are nonnegative. From Sun [32], the log-likelihood ratio is $l_n(\hat\theta_{n1}, \hat\theta_{n2}) - l_n(\theta_1, \theta_2) = (Z_{n1}^2 + Z_{n2}^2)/2$, where

$$\begin{pmatrix} Z_{n1} \\ Z_{n2} \end{pmatrix} = \begin{pmatrix} \{2nI_1(\hat\theta_{n1}, \theta_1)\}^{1/2}\, sgn(\theta_1 - \hat\theta_{n1}) \\ \{-2n\theta_1 I_2(\hat\theta_{n2}, \theta_2)\}^{1/2}\, sgn(\theta_2 - \hat\theta_{n2}) \end{pmatrix}$$

is a generalized signed square root of the log-likelihood ratio.

Let $P_{(\theta_1,\theta_2)}$ denote probability over $X_1, X_2, \ldots$, given (θ_1, θ_2), and, for a fixed prior $\pi(\theta_1, \theta_2)$, let $P^\pi(\cdot \mid \boldsymbol{X}_n)$ denote posterior probability given $\boldsymbol{X}_n$. Suppose we are considering a stopping time, N_a, such that $N_a \to \infty$ almost surely as $a \to \infty$. An asymptotic frequentist matching prior in this sequential setting is a prior π such that

$$(3.9) \qquad \begin{aligned} &P^\pi(Z_{N_a,1} \le c_1, Z_{N_a,2} \le c_2 \mid \boldsymbol{X}_{N_a}) \\ &\quad = \quad P_{(\theta_1,\theta_2)}(Z_{N_a,1} \le c_1, Z_{N_a,2} \le c_2) + O(a^{-1}), \end{aligned}$$

for all c_1 and c_2 in $P_{(\theta_1,\theta_2)}-$probability.

Suppose now that the stopping rule satisfies

$$(3.10) \qquad \frac{N_a}{a} \to \tau(\boldsymbol{\theta}), \quad \text{in } L_1.$$

From Sun [32], the unique prior satisfying (3.9), and hence the unique asymptotic matching prior, is

$$(3.11) \qquad \pi_m^*(\theta_1, \theta_2) \propto \sqrt{\tau(\boldsymbol{\theta}) G_1''(\theta_1) G_2''(\theta_2)}\,.$$

As an immediate example, for the stopping time defined in (3.4), property (3.6) establishes that (3.10) holds; hence the reference prior given in (3.8) is also the asymptotic matching prior, a very desirable situation.

Example 3.1 (continued). *In deriving the sequential likelihood ratio test to see if $(\mu, \sigma^2) = (\mu_0, \sigma_0^2)$, Woodroofe [37] considered the following stopping rule,*

$$(3.12) \qquad N_a = \min\Big(b_2 a, \;\inf\Big\{n \ge b_1 a : \sum_{i=1}^n X_i^2 - n - n\log(\hat\sigma_n^2) \; > 2a\Big\}\Big),$$

where $0 < b_1 < b_2 < \infty$ are two prespecified numbers, $\hat\sigma_n^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$, and $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$. Theorem 8.3 of Woodroofe [37] implies that

$$\frac{a}{N_a} \to \begin{cases} b_2, & \textit{if } \rho^2(\boldsymbol{\theta}) < 1/b_2, \\ \rho^2(\boldsymbol{\theta}), & \textit{if } 1/b_2 < \rho^2(\boldsymbol{\theta}) < 1/b_1, \\ b_1, & \textit{if } \rho^2(\boldsymbol{\theta}) > 1/b_1, \end{cases}$$

in $P_{(\theta_1,\theta_2)}-$probability, as $a \to \infty$, where

$$(3.13) \qquad \begin{aligned} \rho^2(\boldsymbol{\theta}) &= G_1(\theta_1) - G_1(-0.5) - G_1'(-0.5)(\theta_1 + 0.5) - \theta_1\theta_2^2 \\ &= \{(\mu^2 + 1)/\sigma^2 + \log(\sigma^2) - 1\}/2. \end{aligned}$$

Thus (3.11) gives an asymptotic matching prior for this situation. Note, however, that the expected stopping time is not of the form (2.5), so that we cannot assert that this prior is also a one-at-a-time reference prior.

4. Computation

If $E_{\boldsymbol{\theta}}[N]$ is available in closed form, as in the examples in this paper, computation with any of the sequential priors can be done using common MCMC techniques. Hence we only consider here the case in which $E_{\boldsymbol{\theta}}[N]$ can only be computed numerically.

4.1. Brute Force Computation

All the Jeffreys', reference, and matching priors that have been discussed for a sequential experiment are of the form $\Psi(E_{\boldsymbol{\theta}}[N])\pi_F(\boldsymbol{\theta})$, where Ψ is some operator and π_F is the corresponding prior for the fixed sample size experiment. The posterior distribution corresponding to this prior is

$$\pi^*(\boldsymbol{\theta} \mid \boldsymbol{X}_N) \quad \propto \quad \Psi(E_{\boldsymbol{\theta}}[N])\pi_F(\boldsymbol{\theta}) \prod_{i=1}^{N} f(X_i \mid \boldsymbol{\theta}), \tag{4.1}$$

where $\boldsymbol{X}_N = (X_1, \cdots, X_N)$ is the data.

The brute force method for simulating from this posterior distribution is the following Metropolis algorithm:

Step 1. Sample a proposed $\boldsymbol{\theta}'$, from the fixed sample size posterior density of $\boldsymbol{\theta}$, which is proportional to $\pi_F(\boldsymbol{\theta}) \prod_{i=1}^{N} f(X_i \mid \boldsymbol{\theta})$.

Step 2. Numerically estimate $E_{\boldsymbol{\theta}'}[N]$. For instance, one could repeatedly sample N from its distribution given $\boldsymbol{\theta}'$, by simply repeatedly simulating the sequential experiment for the given $\boldsymbol{\theta}'$, observing the N that results from each simulation, and averaging to obtain the estimate $\widehat{E_{\boldsymbol{\theta}'}[N]}$.

Step 3. Perform a Metropolis step: sample $u \sim$ uniform $(0,1)$ and, with $\boldsymbol{\theta}$ denoting the previous value the parameter, accept $\boldsymbol{\theta}'$ if

$$u \le \min\left\{1, \frac{\Psi(\widehat{E_{\boldsymbol{\theta}}[N]})}{\Psi(\widehat{E_{\boldsymbol{\theta}'}[N]})}\right\},$$

and set $\boldsymbol{\theta}'$ equal to the previous $\boldsymbol{\theta}$ otherwise.

If one cannot directly draw from the posterior in Step 1, one could instead using any MCMC scheme, e.g. Gibbs sampling or Metropolis-Hastings. If doing so, however, be sure to iterate Step 1 many times before moving on to Step 2. This is because Step 2 is typically extremely expensive, as it may involve thousands of simulations of the entire experiment simply to compute one Metropolis acceptance probability. In situations where one dependent step is much more expensive than others, it pays to iterate first on the others.

4.2. The Two-Dimensional Case

If using the Jeffreys' prior in a two-dimensional problem or the reference prior in the situation of Corollary 2.2, the posterior distribution is of the form

$$\pi^*(\boldsymbol{\theta} \mid \boldsymbol{X}_N) \quad \propto \quad E_{\boldsymbol{\theta}}[N]\, \pi_F(\boldsymbol{\theta}) \prod_{i=1}^{N} f(X_i \mid \boldsymbol{\theta})\,. \tag{4.2}$$

This allows a remarkable simplification in the computation, by introducing N as a latent variable.

To avoid confusion, we will label the latent variable as $\tilde{N}$; it is a variable with the same distribution as N, but is independent of N. Write the density of $\tilde{N}$ given $\boldsymbol{\theta}$ as $p(\tilde{N} \mid \boldsymbol{\theta})$. Then the joint density of $(\tilde{N}, \boldsymbol{\theta})$, given the data $\boldsymbol{X}_N = (X_1, \cdots, X_N)$, is proportional to

$$p(\tilde{N} \mid \boldsymbol{\theta})\tilde{N}\, \pi_F(\boldsymbol{\theta}) \prod_{i=1}^{N} f(X_i \mid \boldsymbol{\theta})\,. \tag{4.3}$$

Sampling $(\tilde{N}, \boldsymbol{\theta})$ from this distribution will result in $\boldsymbol{\theta}$ from (4.2), as can easily be seen by marginalizing over $\tilde{N}$ in (4.3).

Here is a Metropolis algorithm for sampling from (4.3).

Step 1. Sample a proposed $\boldsymbol{\theta}'$, from the fixed sample size posterior density of $\boldsymbol{\theta}$, which is proportional to $\pi_F(\boldsymbol{\theta}) \prod_{i=1}^{N} f(X_i \mid \boldsymbol{\theta})$.

Step 2. Sample a proposed $\tilde{N}'$ from $p(\tilde{N} \mid \boldsymbol{\theta}')$. This can always be done by simply simulating the sequential experiment once, given $\boldsymbol{\theta}'$.

Step 3. Perform a Metropolis step: sample $u \sim$ uniform $(0,1)$ and, letting $(\tilde{N}, \boldsymbol{\theta})$ denote the previous value the parameter, accept $(\tilde{N}', \boldsymbol{\theta}')$ if

$$u \le \min\left\{1, \frac{\tilde{N}}{\tilde{N}'}\right\},$$

and set $(\tilde{N}', \boldsymbol{\theta}')$ equal to the previous $(\tilde{N}, \boldsymbol{\theta})$ otherwise. (Note that, if $\tilde{N}' < \tilde{N}$, one would always accept the candidate.)

The reason that this is a vastly more efficient algorithm than the brute force algorithm is that one need only simulate a single draw of $\tilde{N}'$ in Step 2, whereas thousands of draws would be needed in Step 2 of the brute force algorithm to compute $\widehat{E_{\boldsymbol{\theta}'}}[N]$. Again, of course, Step 1 could be replaced by any convenient dependent MCMC scheme. Whether one then needs to iterate Step 1 before moving on to Step 2 will be context dependent.

4.3. Modified Reference Priors

The most desirable prior is the one-at-a-time reference prior given in Corollary 2.1, resulting in the posterior distribution

$$\pi^*(\boldsymbol{\theta} \mid \boldsymbol{X}_N) \quad \propto \quad \sqrt{E_{\boldsymbol{\theta}}[N]}\, \pi_R(\boldsymbol{\theta}) \prod_{i=1}^{N} f(X_i \mid \boldsymbol{\theta})\,. \tag{4.4}$$

Unfortunately, the latent variable trick is not available for sampling from this distribution.

Interestingly, however, it is frequently the case that

$$\sqrt{E_{\boldsymbol{\theta}}[N]} \approx E_{\boldsymbol{\theta}}[\sqrt{N}]\,. \tag{4.5}$$

When this is the case, the latent variable trick can be applied, and the algorithm from Section 4.2 can be utilized by simply replacing $\tilde{N}/\tilde{N}'$ in the Metropolis step with $\sqrt{\tilde{N}/\tilde{N}'}$.

In the remainder of the section, we discuss the reason that the approximation (4.5) often holds. The first is that the sampling distribution of N may be rather concentrated in a region of large N, in which case the approximation is clearly good.

Example (Bar-Lev and Reiser [1]) (continued). *For the stopping time N_a defined in (3.4), it follows from (3.5) and (3.6) that*

$$\lim_{a\to\infty} E_{\boldsymbol{\theta}}(\sqrt{N_a/a}) = 1/|\theta_1|^{1/4}.$$

We then have

$$\lim_{a\to\infty} E_{\boldsymbol{\theta}}\sqrt{\frac{N_a}{a}} \quad\propto\quad \lim_{a\to\infty}\sqrt{E_{\boldsymbol{\theta}}\Big(\frac{N_a}{a}\Big)}.$$

Example 2.1 (continued). *Let N_r have the negative binomial distribution $NB(r,p)$. Note that $E_p(N_r) = r/p$ and $Var_p(N_r) = rp/(1-p)^2$. As $r\to\infty$, we have*

$$\sqrt{N_r/r} \to 1/\sqrt{p} \text{ in probability}$$

and

$$E_p(\sqrt{N_r/r}) \to \sqrt{E_p(N_r/r)} \equiv 1/\sqrt{p}.$$

To see the difference between $E_p(\sqrt{N_r/r})$ and $\sqrt{E_p(N_r/r)}$ for moderate r, they are plotted, as a function of p, in Figure 1 for $r=1$ and $r=9$. For $r=9$, the curves are essentially indistinguishable; even for the minimal $r=1$ they are quite close.

It is also interesting to look at the posterior distributions for this example. In Figure 2, we plot the posterior densities of p for three priors

$$\begin{aligned}\pi_J(p) &\propto 1/\sqrt{p(1-p)},\\ \pi_R^*(p) &\propto 1/\sqrt{p},\\ \pi_M(p) &\propto E_p(\sqrt{N_r^*/r}).\end{aligned}$$

Here $\pi_M(p)$ is an approximate prior. For even the very small $r=2$, the posterior densities under the two priors π_R^* and π_M are quite close, yet substantially different from that under π_J. For a moderate $r=10$, the posterior densities under π_R^* and π_M are almost identical. Note that the posterior densities of p under π_J and π_R^* are Beta $(r, N_r - r + 0.5)$ and Beta $(r = 0.5, N_r - r + 0.5)$, respectively. The posterior densities of p under π_M were computed using 5000 Metropolis samples.

As a final indication of the similarity of the true and approximate reference priors in this example, and of the value of using the sequential reference priors, we compare the frequentist coverage probabilities that result from their use in obtaining confidence intervals for p. Table 2 considers the frequentist coverage of one-sided 5% and 95% Bayesian credible regions, based on the fixed sample size Jeffreys' prior π_J, the sequential Jeffreys'/reference prior π_R^* and the approximate prior π_M for various combination of r and p. The fixed sample size Jeffreys' prior performs worse then the other two, indicating the value of using the sequential versions, while the reference prior and the approximate prior are almost equally good.

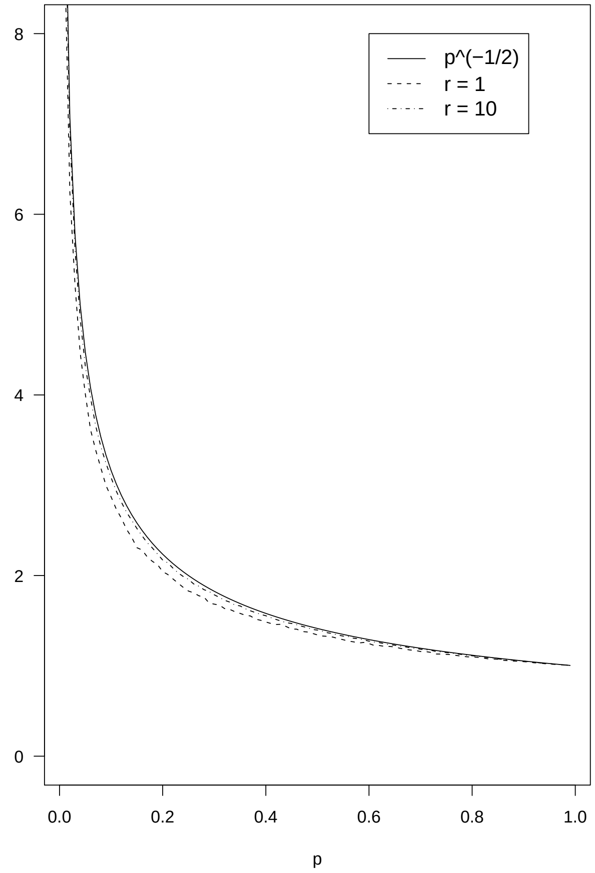


Fig 1. *Negative binomial example: comparison of* $\sqrt{E_p(N_r/r)}$ *and* $E_p(\sqrt{N_r/r})$ *for* $r = 1$ *and* $r = 9$.

Table 2

Coverage Probability of one-sided 5% (95%) Bayesian credible sets for the negative binomial Example 2.1, under the three priors $\pi_J(p) = 1/\sqrt{p(1-p)}$, $\pi_R^*(p) = 1/(p\sqrt{1-p})$, *and* $\pi_M(p) = E_p(\sqrt{N_r^*/r})$.

r	p	π_J	π_R^*	π_M
2	0.1	.1142(.9738)	.0516(.9511)	.0487(.9509)
2	0.5	.0002(.9652)	.0010(.9381)	.0008(.9455)
2	0.9	.0001(.9724)	.0003(.9700)	.0000(.9729)
8	0.1	.0751(.9642)	.0474(.9498)	.0465(.9534)
8	0.5	.0552(.9688)	.0522(.9536)	.0568(.9517)
8	0.9	.0000(.9307)	.0001(.9310)	.0002(.9339)
30	0.1	.0617(.9571)	.0508(.9497)	.0516(.9523)
30	0.5	.0556(.9594)	.0512(.9495)	.0525(.9503)
30	0.9	.0426(.9369)	.0438(.9410)	.0442(.9368)

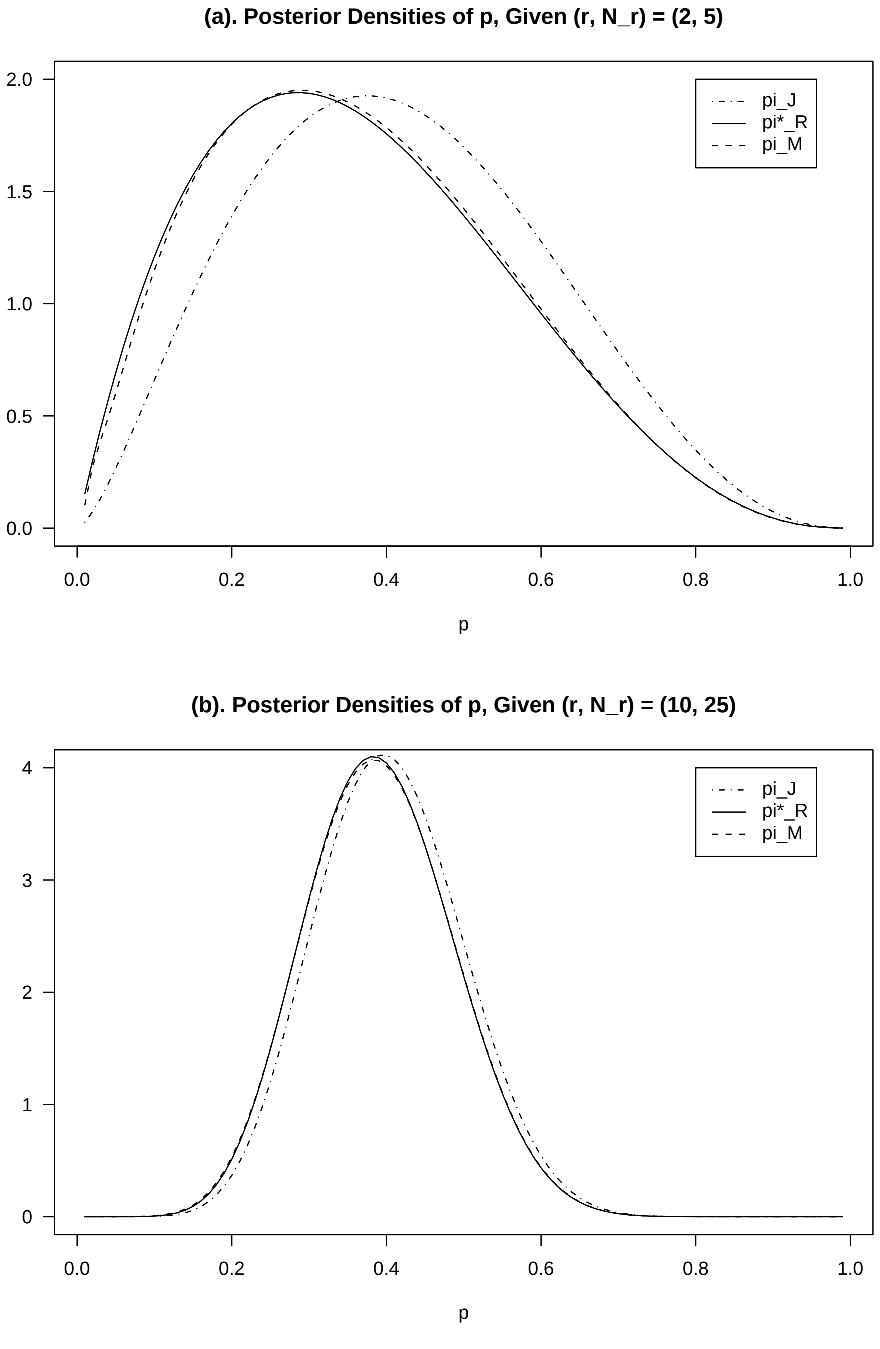


FIG 2. *Posterior densities of* p *based on the priors* $\pi_J(p) = 1/\sqrt{p(1-p)}$, $\pi^*_R(p) = 1/(p\sqrt{1-p})$, *and* $\pi_M(p) = E_p(\sqrt{N^*_r/r})$ *for* $r = 1, 10$; *(a)* $(r, N_r) = (2, 5)$; *(b)* $(r, N_r) = (10, 25)$.

Acknowledgements

This research was supported by the National Science Foundation, under grants DMS-0103265, SES-0351523, and SES-0720229, and the National Institute of Health, under grant R01-MH071418. The authors gratefully acknowledge the comments and suggestions of a referee.

References

[1] Bar-Lev, S.K. and Reiser, B. (1982). An exponential subfamily which admits UMPU test based on a single test statistic. *Ann. Statist.* **10** 979-989.

[2] Bartholomew, D. (1965). A comparison of some Bayesian and frequentist inference. *Biometrika*, **52**, 19-35.

[3] Berger, J.O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, **1**, 385–402 and 457–464.

[4] Berger, J.O. and Bernardo, J.M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc.* **84** 200-207.

[5] Berger, J.O. and Bernardo, J.M. (1992). On the development of the reference prior method (with discussion). *Bayesian Statistics,* **4**, Eds. J.M. Bernardo, J.O. Berger, A.P Dawid and A.F.M. Smith. Oxford Univ. Press, 35-60.

[6] Bernardo, J.M. (1979). Reference posterior distributions for Bayesian inference (with discussion). *J. Roy. Statist. Soc. B* **41** 113-147.

[7] Bernardo, J.M. (2005). Reference analysis. In *Handbook of Statistics* **25** (D.K. Dey and C.R. Rao, eds.), 17–90. Amsterdam: Elsevier.

[8] Bernardo, J.M. and Smith, A.F.M. (1984). *Bayesian Theory*, Wiley, New York.

[9] Bose, A. and Boukai, B. (1993). Sequential estimation results for a two-parameter exponential family of distributions. *Ann. Statist.* **21** 484-502.

[10] Brown, L.D. (1988), The differential inequality of a statistical estimation problem, *Statistical Decision Theory and Related Topics IV,* Volume **1**, 299-324.

[11] Cox, D. R. and Reid, N. (1987). Orthogonal parameters and approximate conditional inference (with discussion). *J. Roy. Statist. Soc., B,* **49** 1-39.

[12] Datta, G.S. (1996). On priors providing frequentist validity for Bayesian inference for multiple parametric functions. *Biometrika*, **83**, 287.

[13] Datta, G.S. and Ghosh, J.K. (1995). On priors providing frequentist validity for Bayesian inference. *Biometrika*, **82**, 37-45.

[14] Datta, G.S. and Ghosh, M. (1995). Some remarks on noninformative priors *J. Amer. Statist. Assoc.* **90**, 1357-1363.

[15] Datta, G.S. and Ghosh, M. (1996). On the invariance of noninformative priors. *Ann. Statist.* **24** 141- 159.

[16] Datta, G.S., Ghosh, M. and Mukerjee, R. (2000). Some new results on probability matching priors. *Calcutta Statist. Assoc. Bull.*, **50**, 179–192.

[17] Datta, G.S. and Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics.* New York: Springer.

[18] Geisser, S. (1979). Comments on "Reference posterior distributions for Bayesian inference", by J. Bernardo. *J. Roy. Statist. Soc., B,* **41**, 136-137.

[19] Geisser, S. (1984). On prior distributions for binary trials. *American Statisticians,* **38**, 244-251.

[20] Ghosh, J.K. (1994). *Higher Order Asymptotics.* Institute of Mathematical Statistics and American Statistical Association, Hayward, California, USA.

[21] Ghosh, J.K., Delampady, M., and Samanta, T. (2006). *An Introduction to Bayesian Analysis: Theory and Methods.* New York: Springer.

[22] Ghosh, J.K. and Mukerjee, R. (1992). Noninformative priors (with discussion). *Bayesian Statistics,* **4**, Eds. J.M. Bernardo, J.O. Berger, A.P Dawid and A.F.M. Smith. Oxford Univ. Press, 195-210.

[23] Ghosh, J.K. and Mukerjee, R. (1995). Frequentist validity of highest posterior density regions in the presence of nuisance parameters. *Statist. Dec.* **13**, 131−139.

[24] Ghosh, M., Sen, P.K., and Mukhopadhyay, N. (1997). *Sequential Estimation.* New York: Wiley.

[25] Govindarajulu, Z. (1981). *The sequential statistical analysis of hypothesis testing, point and interval estimation, and decision theory,* Columbus, OH: American Science Press.

[26] Hall, W.J. (1992). *A Course in Sequential Analysis.* Unpublished Lecture Notes, University of Rochester, Rochester, NY.

[27] Jeffreys, H. (1961). *Theory of Probability.* Oxford University Press.

[28] Peers, H.W. (1965). On confidence sets and Bayesian probability points in the case of several parameters. *J. Royal Statist. Soc., Ser. B* **27** 9-16.

[29] Polson, N. and Roberts, G. (1993). A utility based approach to information for stochastic differential equations. *Stochastic Processes and their Applications*, **48**, 341–356.

[30] Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals.* New York: Springer-Verlag.

[31] Sivaganesan, S. and Lingam, R. (2002). Bayes Factors for model selection with diffusion processes under improper priors. *The Annals of Institute of Statistical Mathematics,* **54**, 500–516.

[32] Sun, D. (1994). Integrable expansions for posterior distributions for a two-parameter exponential family. *The Annals of Statistics,* **22**, 1808-1830.

[33] Sun, D. and Ye, K. (1996). Frequentist validity of posterior quantiles for a two-parameter exponential family. *Biometrika*, **83**, 55-65.

[34] Tibshirani, R. (1989). Noninformative priors for one parameter of many. *Biometrika,* **76** 604-608.

[35] Walker, A.M. (1969). On the asymptotic behaviour of posterior distributions. *J. Royal Statist. Soc., Ser. B* **31** 80-88.

[36] Welch, B.N. and Peers, B. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. Ser. B* **35** 318-329.

[37] Woodroofe, M. (1982). *Nonlinear Renewal Theory in Sequential Analysis.* SIAM, Philadelphia.

[38] Ye, K. (1993). Reference priors when the stopping rule depends on the parameter of interest. *J. Amer. Statist. Assoc.* **88** 360-363.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 33
© Institute of Mathematical Statistics, 0000
arXiv: math.PR/0000000

Sequential Tests and Estimates after Overrunning Based on P-Value Combination

W. J. Hall[1] **and Keyue Ding**[2]

University of Rochester and Queen's University

Abstract: Often in sequential trials additional data become available after a stopping boundary has been reached. A method of incorporating such information from overrunning is developed, based on the 'adding weighted Zs' method of combining p-values. This yields a combined p-value for the primary test and a median-unbiased estimate and confidence bounds for the parameter under test. When the amount of overrunning information is proportional to the amount available upon terminating the sequential test, exact inference methods are provided; otherwise, approximate methods are given and evaluated. The context is that of observing a Brownian motion with drift, with either linear stopping boundaries in continuous time or discrete-time group-sequential boundaries. The method is compared with other available methods and is exemplified with data from two sequential clinical trials.

Contents

1. Introduction

Suppose a sequential trial is carried out to test a null hypothesis about a real parameter δ. Once the trial is concluded, a non-sequential trial is conducted, with a test of the same hypothesis. The trials are connected in that the amount of information in the non-sequential trial may depend on data accumulated in the sequential trial. How can the results of the two trials be combined, and a single overall test constructed? The context is that the data, or incremental information,

[1] Department of Biostatistics, University of Rochester Medical Center, Rochester, NY 14642-8630, U.S.A. e-mail: hall@bst.rochester.edu

[2] NCIC Clinical Trials Group, Queen's University, Kingston, ON K7L 3N6, CANADA

AMS 2000 subject classifications: primary 62L10; secondary 62P10.

Keywords and phrases: sequential clinical trial, delayed observations, lagged data, deletion method, ML ordering, meta analysis, double sampling.

in the non-sequential trial represent ‘overrunning’, from ‘lagged’ data from the sequential trial.

T. W. Anderson [1] considered the problem of incorporating lagged data in an accept-reject rule following a *sequential probability ratio test* and proposed an (approximate) *likelihood ratio test.* In the context of modern-day clinical trials, the problem of how to incorporate data from overrunning was raised and discussed by Whitehead [16, 17], and he gives an admittedly *ad hoc* solution, later named the *deletion method* [14]. This latter paper includes a comparison of the *deletion method* with methods described herein, under certain limited conditions. Another solution is presented in Hall and Liu [4]—actually, an extension of Anderson’s likelihood ratio method—along with a discussion of the possible structure of overrunning information in a sequential clinical trial. However, their method utilizes the *maximum-likelihood ordering* of the sample space, requiring specification of the details of the stopping rule beyond the time a stopping boundary was first reached, in contrast to *stagewise ordering.* In this paper, we focus on procedures that do not require such specification. See these references for further introductory material.

In the context of monitoring a Brownian motion with drift by periodic observations—the context considered herein—Whitehead [16] proposes treating the final analysis that incorporates the overrunning data as if it were a scheduled analysis, but ignoring the analysis that led to stopping, and hence involving a *deletion.* He uses a *stepwise ordering* (as defined in [6], for example) for computing p-values and carrying out further inference.

Here we provide another solution, based on the ‘adding weighted Zs’ method of combining p-values (Stouffer et al. [15], Mosteller and Bush [10], Liptak [7]); one p-value is derived from the sequential experiment (without the overrunning) and the other is based solely on the incremental overrunning data. We recommend weighting the two p-values using observed information. This is fully legitimate *only* if *(i)* the amount of information in the non-sequential trial (overrunning) is proportional to that available at termination of the sequential trial or *(ii)* the sequential trial was actually nonsequential (and test statistics are normally distributed). For discussion of *(i)*, see ([4], Section 2).

Another issue that arises in popular group-sequential trials is that if stopping does not occur until the last scheduled analysis, such an analysis will ordinarily not be done until lagged data are available, in which case a p-value will be computed by standard group-sequential methods with a re-scheduled final analysis. (This is consistent with the *deletion method.*) A modification of our method, which combines p-values for such trials only when stopping early, is evaluated numerically.

Another application of the combination method could be to a *double sampling* study in which the second sample size depends on the outcome—e.g., on the observed variability—of the first sample. These methods are also appropriate for a *meta-analysis* of two (or more) experiments, whether sequential or not.

Brannath, Posch and Bauer [2] proposed p-value combination rules in a different context, namely that of *adaptive group-sequential sampling.* In their setting, allowance is made for the possibility of not carrying out the second stage. (In our context, this would constitute ‘preventing overrunning’.) If the second stage is carried out, the two p-values are combined in a way that *(i)* preserves an overall significance level and *(ii)* recognizes the stopping rule. As here, the second stage p-values may be conditional on results from the first stage. They extend to multiple stages recursively. Numerical integration may be required.

The ‘adding weighted Zs’ combination method is described in Section 2 and extended to an ordered sequence of possibly dependent experiments. In Section

3, this method is applied to sequential clinical trials with overrunning. Special attention is given to the case of a constant amount of overrunning information, the case considered in [14], or to an amount proportional to the amount available at the end of the sequential trial. It is shown that the latter assumption justifies the use of weights related to the observed (and hence random) amounts of information. Otherwise, the use of such random weights leads to a null distribution of the p-value which is only approximately uniform. Still, we recommend this usage so long as the approximation is adequate.

In Section 4 we show how to use the *combination p-value method* to compute estimates and confidence intervals, and in Section 5 provide formulas for evaluating the true confidence coefficients associated with these methods, thus enabling an evaluation of approximations noted above. Some evaluations are summarized in Section 6.

In Sooriyarachchi et al. [14], the issue of reversals in the conclusions after incorporating overrunning, from rejection to acceptance of a null hypothesis or vice versa, was raised. They found, in the cases treated numerically there, that both the *deletion method* and the *combination method* might lead to an uncomfortable level of reversals, with the *deletion method* doing so less frequently. They also noted that both methods (in cases treated) sometimes lead to reduced power. We consider these issues in Section 7 and indicate a modification of the *combination method* that reduces these effects.

The methods are applied to data from the MADIT trial [8] in Section 8—for both the actual linear-boundary design and for an imagined group-sequential version. Results are compared with those from the *deletion* and *ML-ordering methods.* Results from a second example [9] are briefly summarized.

Some final comments appear in Section 9, including a summary comparison of the alternative methods for incorporating overrunning.

2. Combining P-Values by Adding Weighted Zs and an Extension

We suppose some potential data X are to be available for testing a null hypothesis about a real parameter δ belonging to an interval Δ. For each $\delta_o \in \Delta$, we consider a test of $\delta = \delta_o$ versus $\delta > \delta_o$, with p-value $p(x;\delta_o)$ when $X = x$ is observed. Suppose, for each δ_o, $P \equiv p(X;\delta_o)$ is uniformly distributed on (0,1) when $\delta = \delta_o$ and that, for each x, $p(x;\delta)$ is increasing in δ. Then $\mathcal{P} \equiv \{p(\cdot;\delta_o)|\delta_o \in \Delta\}$ defines a *proper family of p-values* for this testing problem.

This is an overly strict definition. We have restricted attention to test functions with continuous distributions, and *stochastic ordering* (increasing or nondecreasing) of p-values would allow for differing sample spaces, but the conditions given meet our application. We usually omit the word 'proper'.

The ordering is needed to avoid possible inconsistencies. Data considered to be 'more extreme' than that observed should have higher probability under an alternative hypothesis than under the null. Moreover, it facilitates construction of consistently defined confidence bounds. Simply equate the p-value for testing δ_o to γ $(1-\gamma$, resp.) and solve for δ_o to obtain a lower (upper, resp.) confidence bound with confidence coefficient $1-\gamma$. Choosing $\gamma = 0.5$ yields a median-unbiased estimate.

Suppose p_1 and p_2 are independent p-values for the same null hypothesis, and let $z(u) \equiv \bar{\Phi}^{-1}(u)$ with $\bar{\Phi}(z) = 1 - \Phi(z)$ and Φ being the standard normal distribution function. Let w_1 and w_2 be positive numbers for which $w_1^2 + w_2^2 = 1$. Then

$$p \;\equiv\; \bar{\Phi}\left(w_1\, z(p_1) \;+\; w_2\, z(p_2)\right) \tag{2.1}$$

is the *adding weighted Zs combined p-value* [7, 10, 12, 15]; also see [13]. It is readily seen that the argument of $\bar{\Phi}$ in (2.1), with p_i replaced by the random variable P_i, is distributed as standard normal under the null hypothesis, and hence P in (2.1) is distributed as $U(0,1)$. Moreover, p tends to be small whenever both p_1 and p_2 are small. More precisely, it is seen to be proper whenever p_1 and p_2 are proper—the p_i's are increasing in δ_o, so the $z(p_i)$'s are decreasing, as is any positively-weighted linear combination, and hence p is increasing in δ_o.

The interpretation should be clear: $z(p_i)$ is a standardized normal deviate that corresponds to the test statistic on which p_i is based (whether or not p_i was based on a normally distributed statistic), and the argument of $\bar{\Phi}$ in (2.1) represents a (weighted) pooling of normal deviates for the two independent tests, with p the resulting p-value.

This *combination method* may be extended to settings where p_1 and p_2 are derived from overlapping data sets but p_2 is a conditional p-value for each subset of data on which p_1 is based. A possible context is that a second experiment was designed based on the outcome of the first experiment, and a conditional test was used in the second experiment. Formally,

Proposition 2.1. *Suppose* $p_1 \equiv p_1(x;\delta)$ *and* $p_2 \equiv p_2(x,y;\delta)$*, and* $\mathcal{P}_1 \equiv \{p_1(\cdot;\delta_o)|\delta_o \in \Delta\}$ *is a family of p-values and* $\mathcal{P}_2 \equiv \{p_2(x,\cdot;\delta_o)|\delta_o \in \Delta\}$ *is, for each* $X = x$*, a family of conditional p-values. Then* $\mathcal{P}_2$ *is a family of unconditional p-values,* $p_1(X,\delta_o) \perp p_2(X,Y;\delta_o)$ *(independent) for each* δ_o*, and (2.1) defines a family of p-values.*

Proof. Since $p_2(X,Y;\delta_o)$ is conditionally $U(0,1)$ for every X, it is unconditionally $U(0,1)$, and the needed monotonicity also follows. For each δ_o, the joint distribution function of (P_1, P_2) is

$$\begin{aligned} \Pr\{P_1 \le u_1,\, P_2 \le u_2\} &= E\big\{1(p_1(X) \le u_1)\cdot E\big[1(p_2(Y,X) \le u_2 \,\big|\, X\big]\big\} \\ &= E\big\{1(p_1(X) \le u_1)\cdot u_2\big\} \;=\; u_1\, u_2\,, \end{aligned}$$

from which independence follows, and this is sufficient for the claim about p. □

Now what about the weights? Ordinarily, they might be related to sample size or information. Specifically, if the p_i's are derived from tests based on means of n_i normally distributed observations (with common variance), then a combined p with $w_i \propto \sqrt{n_i}$ would yield the same p as that from a pooling of the two samples. So far, we have only assumed the weights to be positive constants—depending neither on δ_o nor on the data. Here are some partial extensions; examples of each appear in the next section.

The weights may depend on δ_o without affecting the null distribution of P in (2.1), but the monotonicity in δ_o may be destroyed except for special choices.

The weights may be random (depending on X) without affecting the monotonicity in δ_o, but would typically disturb the uniformity of the null distribution of P.

It should be emphasized that all p-values considered above are for one-sided alternatives. After including overrunning, the usual convention of doubling them for 2-sided alternatives may be appropriate.

Finally, we note that all of this can be directly extended to an ordered set of several p-values, each involving new data and conditional on all past data, and combined in the "adding weighted Z's" fashion.

Specifically, let p_k be a p-value for the incremental stage-k data, conditional on data from all prior stages. Then define a stage-k combination p-value by replacing the argument of $\bar{\Phi}$ in (2.1) by $\sum_{i=1}^{k} w_{k:i} z(p_i)$ with stage-k weights all positive, satisfying $\sum_{i=1}^{k} w_{k:i}^2 = 1$, and $w_{k:i}^2 = w_{k-1:i}^2 \cdot (1 - w_{k:k}^2)$ for $i < k$. Equivalently, (2.1) may be applied recursively, replacing p_1 by a combined p from earlier stages with weight w_1 for this new p_1 and w_2 for the incremental data, with $w_1^2 + w_2^2 = 1$.

3. Incorporating Overrunning by Combining *P*-Values

We now assume a sequential experiment takes place, resulting in an observation of (T, X), say. We focus on the context of observing a Brownian motion $X(t)$ with drift δ, with a stopping time T and $X \equiv X(T)$ upon stopping, but other contexts may be treated similarly. After stopping, some additional data become available, represented by further observation of the process for $t_o = t_o(T, X)$ units of time. Conditional on t_o, a sufficient statistic for the overrunning data is the increment Y observed during the overrunning time increment t_o. In other words, a sequential experiment is followed by a non-sequential one, with sample size (observation time) depending on the outcome of the sequential trial. There may be additional randomness in t_o; it is sufficient to let $t_o(t, x)$ be the conditional expectation of overrunning information, given $(T, X) = (t, x)$. See ([4], Section 2) for discussion supporting t_o being a constant, $\propto \sqrt{t}$, or $\propto t$ as possible approximations to reality.

Upon reaching a stopping boundary, a p-value p_1 for a null hypothesis about the drift parameter is defined: $\delta = \delta_o$ versus $\delta > \delta_o$. And at the end of overrunning, a conditional p-value p_2 is simply $\bar{\Phi}\big((y - \delta_o t_o)/\sqrt{t_o}\big)$, given $t_o = t_o(t, x)$. A combination p-value is therefore given by (2.1). (Here, (T, X) plays the role of X in Section 2.) Hence,

Corollary 3.1. *Suppose w_1 and w_2 are positive constants for which $w_1^2 + w_2^2 = 1$. Then*

(3.1)
$$p(t, x, y; \delta_o) \;\equiv\; \bar{\Phi}\left(w_1\, z(p_1(t, x; \delta_o)) \;+\; w_2(y - \delta_o t_o)/\sqrt{t_o}\,\right)\Big|_{t_o = t_o(t,x)}, \quad \delta_o \in \Delta\,,$$

defines a family of p-values.

But how should the weights be chosen? It is tempting to choose them to be proportional to the square-root of *information* in the respective parts of the experiment. Then each summand in (3.1) would have variance or conditional variance equal to the information in that part of the experiment. Using expected information, $w_1^2 = E_{\delta_o}(T)/[E_{\delta_o}(T) + E_{\delta_o} t_o(T, X)]$ and $w_2^2 = 1 - w_1^2$. But, as noted in Section 2, this would not typically preserve the needed monotonicity of $p(\delta_o)$ in (3.1). Moreover, knowledge of the functional form of the dependence of t_o on (t, x) would be needed. If t_o were constant, this would yield

$$p(t, x, y; \delta_o) \;=\; \bar{\Phi}\left(\frac{[E_{\delta_o}(T)]^{1/2}\, z\big(p_1(\delta_o)\big) \;+\; y - \delta_o t_o}{[E_{\delta_o}(T) \;+\; t_o]^{1/2}}\right).$$

This could be used as a p-value for a single null hypothesis, but it would not be suitable for construction of confidence bounds, unless $E_{\delta_o}(T)$ was replaced by $E_{\delta_o'}(T)$ for a fixed δ_o'. Because of these limitations, we abandon this approach.

Suppose instead we use the square-roots of *observed information*, namely $\sqrt{t}$ and $\sqrt{t_o}$, yielding

$$(3.2) \qquad p(t,x,y;\delta_o) \; = \; \bar{\Phi}\left(\frac{t^{1/2}\, z\big(p_1(t,x;\delta_o)\big) \; + \; y - \delta_o t_o(t,x)}{[t \; + \; t_o(t,x)]^{1/2}} \right).$$

Monotonicity in δ_o (for each (t,x,y)) is maintained, but the uniformity of the null distribution would appear to be in doubt. However, to compute p, no knowledge of the dependency structure of t_o is required, only its observed value.

We now consider the special case of (3.1) and (3.2) with $t_o \propto t$, say $t_o = c\,t$. Since $w_1^2 = t/(t+ct) = 1/(1+c)$ and $w_2^2 = c/(1+c)$, this yields constant weights; and c is known once $T = t$ and t_o are observed. Hence, the use of observed information in this case is justified.

Corollary 3.2. *If, for some constant c, $t_o(x,t) = ct$ for all (x,t), then (3.2) defines a family of p-values.*

For the group-sequential case with up to K analyses and stagewise ordering, we modify the combination p-value (3.2): For testing δ_o, with $t_{ok} \equiv t_o(k)$, the modified p-value is defined as

(3.3)

$$p^*(t_k,x,y;\delta_o) = \begin{cases} \bar{\Phi}\left([t_k^{1/2}\, z(p_1(t_k,x;\delta_o)) + y - \delta_o t_{ok}]/(t_k + t_{ok})^{1/2} \right) & \text{if } k < K \\ p_1(t_K + t_{oK}, x + y; \delta_o) & \text{if } k = K \end{cases}$$

where $p_1(t,x;\delta_o)$ is the group-sequential stagewise p-value for testing $\delta = \delta_o$ versus larger values when the analyses are scheduled at $t_1, \ldots, t_{K-1}, t_K^o \equiv t_K + t_{oK}$ with early-stopping sets $\mathcal{S}_k$ $(k < K)$ (each the complement of an interval). This matches the *deletion method* when stopping has not occurred early.

For a group-sequential ML-ordering, the *ML-ordering method* [4] may be more suitable.

We show in Sections 5 and 6 that use of p in (3.2) or (3.3), for several choices of the dependency of t_o on (t,x) and two popular sequential designs, for constructing confidence bounds and intervals may yield adequately accurate confidence coefficients. This leads us to recommend the use of (3.2) or (3.3) as if it were a bona fide combination p-value, if the design chosen and the likely form of dependency are similar to those considered in Section 6.

One last variation permits further adjustment of the weighting: Use weights with squares proportional to T and $\rho\, t_o(T)$ for a specified weighting factor $\rho > 0$. For motivation, see Section 7.

4. Computing *P*-Values and Confidence Bounds

Here we act as if $t_o \propto t$, and discuss the use of (3.2) and (3.3) for obtaining p-values and, by inversion, confidence bounds and intervals.

For any particular null value δ_o, the *combined p-value* $p(\delta_o)$ may be computed from (3.2) or (3.3) with t (or t_k), x, y and t_o the observed values, and using software that enables computation of $p_1(\delta_o)$. For general linear boundaries, such software is available from the authors (based on formulas in [3]), and the PEST software [11] provides such output for a limited selection of linear boundaries and group-sequential modifications of them. For group-sequential boundaries with stagewsie ordering, a program—built around software for $p_1(\delta_o)$ from Jennison [5]—is available from the authors.

To obtain an upper confidence bound with confidence coefficient γ, we need to solve $p(\delta) = \gamma$ for $\delta = \hat{\delta}_U$, or equivalently, solve $z(p(\delta)) = z(\gamma)$. A little algebra leads to the equivalent problem—except in the group-sequential case with $t = t_K$—of solving $\delta - h(\delta) - [y - \sqrt{t^o}\, z(\gamma)]/t_o = 0$ where $t^o \equiv t + t_o$ and $h(\delta) \equiv \sqrt{t}\, z(p_1(\delta))/t_o$. Starting from a trial solution δ^o, and computing $h(\delta^o)$ and $h(\delta^o + \epsilon)$ for some small ϵ, an improved solution is

$$\delta \;\equiv\; \delta^o \;-\; \frac{\delta^o - h(\delta^o) - [y - \sqrt{t^o}\, z(\gamma)]/t_o}{1 \;+\; [h(\delta^o) - h(\delta^o + \epsilon)]/\epsilon}\,.$$

We find that two or three iterations provide good accuracy. (When $t = t_K$, we only need solve $p_1(t^o_K, x + y; \delta) = \gamma$.) Alternatively, a trial-and-error approach works quite satisfactorily.

5. True Confidence Coefficients

We now evaluate the true confidence coefficient for a confidence bound or interval determined by using (3.2), whether or not t_o is proportional to T. Let $\hat{\delta}_\gamma$ be an upper confidence bound determined by the method of the previous section for a nominal confidence coefficient γ. The question is: what is the true confidence coefficient? We need to evaluate, for given γ and δ, $q_\gamma(\delta) \;\equiv\; P_\delta(\,\delta < \hat{\delta}_\gamma\,)$ and determine $q^o_\gamma \equiv \inf_\delta\, q_\gamma(\delta)$.

As noted in Section 4, $\hat{\delta}_\gamma$ is the solution to $\delta - h(\delta) = [y - \sqrt{t^o} z(\gamma)]/t_o \equiv g(y, t^o, t_o, \gamma)$. Since h is decreasing in δ, the left side is increasing in δ, and hence $\delta < \hat{\delta}_\gamma$ iff $\delta - h(\delta) < g(y, t^o, t_o, \gamma)$. This latter event is equal to the event $(y - \delta t_o)/\sqrt{t_o} > [-\sqrt{t}\, z(p_1(t, x; \delta)) + \sqrt{t^o}\, z(\gamma)]/\sqrt{t_o}$. Therefore, conditioning on (T, X) and hence on T_o, we have

$$q_\gamma(\delta) = E_\delta P_\delta(\delta < \hat{\delta} | T, X) = E_\delta \Phi\left(\frac{T^{1/2}\, z(p_1(T, X; \delta)) - T^{o1/2}\, z(\gamma))}{T_o^{1/2}}\right). \tag{5.1}$$

If this combination p-value were bona fide—that is, if $T_o \propto T$—the result would be γ identically in δ.

The true confidence coefficient for an (equal-tail) confidence interval based on (3.2) may be obtained similarly. For an interval with nominal confidence coefficient γ, the true confidence coefficient is $Q^o_\gamma \equiv \inf_\delta Q_\gamma(\delta)$ where

$$Q_\gamma(\delta) \;\equiv\; q_{(1-\gamma)/2}(\delta) - q_{(1+\gamma)/2}(\delta)\,. \tag{5.2}$$

For the group-sequential modification (3.3), (5.1) needs to be modified when $T = T_K$.

6. Computational Support for Approximations

Here we report on some numerical evaluations of the validity of using (3.2) or (3.3) when t_o is not proportional to the observed stopping time t, and the validity of using the combination method only when stopping after an interim analysis. For various special cases and many values of δ, we computed (5.1) for $\gamma = 0.5$ and (5.2) for $\gamma = 0.9$ and 0.95 to see how close they are to the respective nominal values of 0.5, 0.9 and 0.95. We summarize some of the findings here.

A linear-boundary design: Consider triangular boundaries for testing $\delta = 0$ versus $\delta = 1$ with intercepts ± 5.99, slopes 0.75 and 0.25, and apex at $t = 23.97$. This design has both error probabilities 0.025. The design may be adapted for testing $\delta = 0$ versus $\neq 0$ (as prescribed by the PEST software). The resulting one-sided rejection region is the upper boundary for which the power at $\delta_1 \equiv 0.8233$ is 0.9. The expected stopping time is 7.776 at $\delta = 0$ (or 1) and 9.382 at δ_1, and has its maximum of 11.217 at $\delta = 0.5$.

We considered $t_o \propto T$, t_o constant and $t_o \propto \sqrt{T}$. For the first case, we simply verified the accuracy of our computer program, finding that the distribution of the p-value was exactly uniform, and that the true confidence coefficients matched the nominal ones exactly.

For the constant case, we considered $t_o = c\,E_{\delta_1}(T)$ with c ranging from 0.1 to 0.5. Here are selected results:

$c = 0.1$	$c = 0.5$
$0.487 < q_{.5} < 0.513$	$0.471 < q_{.5} < 0.529$
$0.900 < Q_{.9} < 0.908$	$0.899 < Q_{.9} < 0.917$
$0.950 < Q_{.95} < 0.955$	$0.949 < Q_{.95} < 0.960$

For $c = 0.1$, the true confidence coefficients Q^o_γ for nominal 90% and 95% confidence intervals are therefore correct (to 3 decimal places), and only slightly below the nominal values for $c = 0.5$. However, the median-unbiased estimate may have a few percentage points of median-bias, depending on the true δ. We also found that $q_{.5} < 0.5$ for $\delta > 0.5$ and vice versa. Computations for c-values between 0.1 and 0.5 yielded bounds between the respective ones in the display above. Results for $t_o \propto \sqrt{T}$ were uniformly better than those for t_o constant.

An O'Brien–Fleming group-sequential design: Consider an O'Brien–Fleming two-sided design for testing $\delta = 0$ with significance level 0.05 and power 0.9 at $\delta = \pm 1$, with a maximum of 5 analyses. We assume equally spaced interim analyses, at 0.2, 0.4, 0.6, 0.8 times $t_5 \equiv 10.781$, with boundary values of ± 6.6988 (obtained from [6]). Again, we considered $t_o \propto T$ for the unmodified combination p-value to confirm the accuracy of our programs. For the modified p, we considered t_o constant, namely $= c\,t_5$, and $t_o = c\,t_k$; in each case, c ranged from 0.02 to 0.1.

Here are some of the results:

	$t_o = c\,t_5$		$t_o = c\,t_k$	
	$c = 0.02$	$c = 0.1$	$c = 0.02$	$c = 0.1$
$q^o_{.5}$	0.475	0.445	0.478	0.451
$Q^o_{.9}$	0.894	0.887	0.894	0.888
$Q^o_{.95}$	0.947	0.943	0.947	0.943

Again, although $q^o_{.5}$ may be as small as 0.44 (and by symmetry $0.44 < q_{.5}(\delta) < 0.56$), we found that $q_{.5}$ was usually within ± 0.01 of 0.5. Indeed, this occurred for all but 1%, 7%, 1% and 4%, respectively (reading from left to right in the display above) of the range of δ-values within ± 2.5.

7. Reversals and Power

Sooriyarachchi et al. [14] raised concern about the frequency of reversals of acceptance and rejection conclusions after inclusion of overrunning information, but stressed their desire not to ignore such information. In simulation studies of the *deletion* and *combined p-value methods*, with constant amounts of lagged data (independent of the results at the time of stopping), they found levels of reversals that they considered worrisome, especially for the *combination method*—perhaps 3 or 4 percent. However, in popular group-sequential designs such as O'Brien–Fleming, reversals were rare and only defined when the trial stopped early, as an analysis at a final scheduled time would ordinarily await lagged data before execution.

Of more concern to us, is their finding that both methods may lead to reductions in power. Intuitively, when a rejection occurs "early", overrunning can reverse it but the chances of compensating with reversals in the other direction may be minimal.

With constant overrunning information, our computations (not reported here) confirm theirs, but we find reversals to be somewhat less frequent when overrunning information increases with stopping times, and losses in power are then rarer.

A possible compromise method is as follows: down-weight the overrunning p-value in the combination formula. By introducing a factor ρ (see end of Section 3), it is possible to maintain power and depress the frequency of reversals but still not ignore the lagged data completely. However, computations show that some situations will require extensive down-weighting (small ρ). Choice of a suitable ρ will require computational trial-and-error, with assumptions about overrunning needed. For this purpose, we provide the following formulas.

When the true drift is δ (and stopping is in continuous time), the probability of rejection upon stopping followed by acceptance after inclusion of overrunning, when $t_o \propto t$, is

(7.1)

$$P_\delta(R \to A) = \int_0^{t_{max}} \Phi\Big\{[(1+\rho c)/(\rho c)]^{1/2} z_\alpha - [1/(\rho c)]^{1/2} z_1(U,t) - \delta(ct)^{1/2}\Big\}\, dP_\delta^U(t)$$

with $z_1(U,t)$ being the standard normal deviate for which the right-hand-side tail area beyond it is $P_0^U(t)$ (the p-value when the upper boundary U is crossed at time t), $P_\delta^U(t)$ being the probability of crossing the upper boundary before the lower one prior to t, and α being the one-sided significance level for testing $\delta = 0$. For group-sequential tests, the integrator in (7.1) is $dP_\delta^U(x,t)$, indicating a need to integrate over x-values where $t = t_k$ and the upper boundary has been reached, but t may be restricted to $\{t_k | k < K\}$ since reversals at a final analysis have no role.

Similarly, $P_\delta(A \to R)$ is given by (7.1) with U replaced by L (for lower boundary) and Φ replaced by $\bar{\Phi}$. Finally, the power after inclusion of overrunning, when the power of the original design is $pow(\delta)$, is

$$ovpow(\delta) \;=\; pow(\delta) \;-\; P_\delta(R \to A) \;+\; P_\delta(A \to R)\,.$$

(Software is available from the authors.)

8. An Example: the MADIT Study

MADIT (Multicenter Automatic Defibrillator Implantation Trial [8]) was a randomized clinical trial conducted to evaluate the effectiveness of an implanted defibrillator compared with conventional drug therapy to reduce mortality associated with ventricular arrhythmias. Monitoring was based on the logrank statistic plotted against its estimated variance [17]. This behaves like a Brownian motion with drift $\delta = -\log(HR)$ where HR is the hazard ratio of the treatment-to-control arms (assuming proportional hazards). The essential features were reviewed in [4] and are summarized here.

A triangular design was used that assures a two-sided significance level of 5% and a power of 90% at a hazard ratio of 0.537 (drift = 0.6218). Monitoring was carried out weekly over the five years of the trial, thereby yielding nearly-continuous observation of the logrank process. The stopping boundaries were $u_t = 7.935+0.189t$ and $l_t = -7.935 + 0.566t$, with the early part of the lower boundary (l_t) a rejection region for superiority of the control arm.

Interpolating, the upper boundary was reached at $t = 12.145$ with $x = 10.230$, later corrected to $t = 12.037$ and $x = 10.210$. The incremental coordinates for overrunning were $t_o = 1.240$ and $y = 2.957$, showing an upturn in the sample path after reaching the boundary.

Respective p-values and estimates of the drift and of the HR are presented below, contrasting results of analyses without and with the use of the overrunning data. Values in square brackets are those reported in [4] for the *ML-ordering method*, assuming $t_o \propto \sqrt{t}$; with linear boundaries and no overrunning, stepwise ordering and ML-ordering are identical.

overrunning	*2-sided p*	*med-unb-est*	*95% confidence interval*
Inference about the drift δ			
without	0.0084	0.786	(0.204, 1.361)
with	0.0009 [0.0029]	0.938 [0.939]	(0.388, 1.484) [(0.329, 1.543)]
Inference about the hazard ratio $HR = \exp(-\delta)$			
without	0.0084	0.456	$(0.256, 0.815^{+})$
with	0.0009 [0.0029]	0.391 [0.391]	(0.227, 0.678) [(0.214, 0.720)]

Both methods reflect the upturn in the sample path during overrunning as the 'with' p-values are smaller and the estimates farther from the null values. But the combination method gives the smaller p-value and narrower confidence intervals; this may reflect the different orderings being used by the two methods.

Values reported in [8] were based on Whitehead's *deletion method*; they are identical to the 'without overrunning' values in the display above, as the *deletion method* essentially ignores overrunning when the path continues in a similar direction and there is near-continuous monitoring. (It was this observation that inspired the development of alternative methods for incorporating overrunning.) In such settings, the deletion-method p-value cannot be smaller than when computed upon first hitting an upper boundary, irrespective of the nature of the overrunning data. (For, when reaching the upper boundary at time t, with the prior analysis a short time earlier, at time t^- say, and then overrunning to x^o at a later time t^o, the deletion one-sided p is the null probability of $\{T \leq t^- \text{ and } X(T) \geq u_T\} \cup \{T \geq t \text{ and } X(t^o) \geq x^o\}$. These two events are disjoint, and the former is virtually the extremal set without overrunning.)

We now consider a group-sequential variation on MADIT as described in [4]. We pretended that an O'Brien–Fleming 5-analysis design was used for testing $\delta = 0$ versus $\neq 0$ with power 80% at a HR of 0.537. It would have stopped at the third interim analysis with the results obtained upon hitting the boundary in MADIT. Results of analyses are reported below; for comparison, values in square brackets are those reported in Table 2 of [4] using the *ML-ordering method* and assuming $t_o \propto \sqrt{t}$. Values for the *deletion method*—which treats the analysis after overrunning as a replacement for the third scheduled analysis—are also given.

In each case—i.e., without or with overrunning—results from the group-sequential combination method indicate a more significant departure from the null value of

$HR = 1$ than do those by the group-sequential ML-ordering method. At least for the 'without' results, this is attributable to the different orderings used. This time the *deletion method* gives results similar to those from *ML-ordering.* (These results are not directly comparable to those in the previous table since the pretended group-sequential design has reduced power.)

overrunning	*2-sided p*	*med-unb-est of HR*	*95% confidence interval*
Group-sequential inference about the hazard ratio			
without	0.0039 [0.0041]	0.431 [0.468]	$(0.244, 0.762)$ [$(0.29\bar{5},\ 0.77\bar{5})$]
with	0.0004 [0.0011]	0.373 [0.384]	$(0.217, 0.641)$ [$(0.221,\ 0.672)$]
Deletion method	0.0014	0.384	$(0.221, 0.680)$

Here is a brief summary of results from a second defibrillator trial, MADIT-II [9], in which the *combination method* was pre-specified. The design was again triangular, with a 5% 2-sided significance level and power 95% at a HR of 0.627: $u_t = 11.77 + 0.1273\,t$ and $l_t = -11.77 + 0.3819\,t$. This time $(t, x, t_o, y) = (45.415,\ 17.551,\ 0.483,\ 1.441)$. The results were:

overrunning	*2-sided p*	*med-unb-est of HR*	*95% confidence interval*
Inference about the hazard ratio in MADIT-II			
without	0.028	0.708	$(0.525, 0.962)$
with	0.016 [0.023]	0.688 [0.689]	$(0.511, 0.932)$ [$(0.504,\ 0.948)$]

Again, the *deletion method* of incorporating overrunning would have agreed with the 'without' analysis, and results from the *ML-ordering method* (in square brackets) are mainly intermediate.

9. Final Remarks

Proposition 1 applies to other methods of combining p-values, such as Fisher's summing of $-\log(1 - p_i)$. (For a description of such methods, see [12] or [13].) We chose the 'adding Zs' method for two reasons: *(i)* It lends itself naturally to weights—it would be unreasonable to give equal weights to a long trial and a small amount of overrunning—and *(ii)* it reduces to standard normal-theory methods when the sequential component is replaced by a non-sequential one—equivalently, if a naive analysis is done after stopping rather than one recognizing the stopping rule.

Here are some of the pros and cons of various methods for incorporating overrunning:

(a) *Deletion method:* Not suitable for near-continuous monitoring. Ignores the fact that, at the boundary-hitting stage, the monitoring statistic was in a stopping region but is the natural approach in a group-sequential trial when early stopping has not occurred. Simple to use. Results in approximate p-values and final inference. Limited computations show that a loss in power may occur.

(b) *Combination p-value method:* Makes direct use of the analysis that led to stopping. Approximate except when $t_o \propto T$, and even then for common group-sequential designs. Uses stage-wise ordering, and hence free of any direct dependence on future stopping boundaries. Needs no formal assumption about the form of t_o.

Computations show a loss in power may occur. May be modified to reduce the chance of reversal after overrunning and loss in power.

(c) *ML-ordering method:* Based on a minimal sufficient statistic, and hence ignores which boundary was first reached and when. Exact, up to needed assumptions about overrunning information (and Brownian motion approximation). Requires an assumption about the form of $t_o(t)$, but not very sensitive to it in the practical cases examined. Uses ML-ordering and hence depends on stopping boundaries beyond those when boundaries were first reached.

Sooriyarachchi et al. [14] conclude that (a) is preferable although they only considered constant amounts of overrunning whereas our focus has been on settings where the amount of overrunning information is likely to increase with increased stopping times. They highly stress the possibilities of reversals, but such possibilities cannot be avoided once one agrees to utilize lagged data. The chances can be reduced within the *combination method* by reducing the weight given to lagged data, but this would need to be considered in advance of the trial.

We recommend (c) in settings where the design is likely to be followed closely. A numerical study of reversals and power with the *ML-ordering method* will be presented elsewhere. Otherwise, we think the *combination p-value method*, possibly with a down-weighting of overrunning information, is a competitor worthy of consideration, especially when overrunning increases with increasing stopping times.

We encourage investigation of the *combination method* in other settings, including meta-analyses and double sampling.

Acknowledgements

The first MADIT trial stimulated the research reported here. We thank Arthur Moss, MD, and Boston Scientific Corporation (formerly CPI/Guidant), respectively, for leadership and sponsorship of this trial. Research was supported in part by grant R01-HL58751 from the National Heart, Lung and Blood Institute (USA) (first author) and by a grant from the Natural Sciences and Engineering Research Council of Canada (second author). We are also grateful to John Whitehead for helpful discussion and to Michael McDermott for some references to the p-value literature.

References

[1] Anderson, T. W. (1964). Sequential analysis with delayed observations. *J. Amer. Statist. Assoc.* **59** 1006-1015.

[2] Brannath, W., Posch, M. and Bauer, P. (2002). Recursive combination tests. *J. Amer. Statist. Assoc.* **97**:236-244.

[3] Hall, W. J. (1997). The distribution of Brownian motion on linear stopping boundaries. *Sequential Analysis* **16** 345-352. Addendum in: *Sequential Analysis* **17** 123-124.

[4] Hall, W. J. and Liu, A. (2002). Sequential tests and estimators after overrunning based on maximum-likelihood ordering. *Biometrika* **89** 699-707.

[5] Jennison, C. (1999). Group sequential software at website: http://www.bath.ac.uk/∼ mascj/book/programs/general.

[6] Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials.* Chapman & Hall/CRC, Boca Raton, FL.

[7] Liptak, T. (1958). On the combination of independent tests. *Magyar Tud. Akad. Mat. Kutato Int. Közl.* **3** 171-197.

[8] Moss, A. J., Hall, W. J., Cannom, D. S., Daubert, J. P., Higgins, M. D., Klein, H., Levine, J. H., Saksena, S., Waldo, A. L., Wilber, D., Brown, M. W., Heo, M.; for the Multicenter Automatic Defibrillator Implantation Trial Investigators (1996). Improved survival with an implanted defibrillator in patients with coronary disease at high risk for ventricular arrhythmia. *New England Journal of Medicine* **335** 1933-1940.

[9] Moss, A. J., Zareba, W., Hall, W. J., Klein, H., Wilber, D. J., Cannom, D. S., Daubert, J. P., Higgins, S. L., Brown, M. W., Andrews, M. L.; for the Multicenter Automatic Defibrillator Implantation Trial-II Investigators (2002). Prophylactic implantation of a defibrillator in patients with myocardial infarction and reduced ejection fraction. *New England Journal of Medicine* **346** 877-883.

[10] Mosteller, F. M. and Bush, R. R. (1954). Selected quantitative techniques. In: G. Lindzey, ed., *Handbook of Social Psychology: Vol. I. Theory and Methods.* Addison-Wesley, Cambridge, MA.

[11] MPS Research Unit (2000). *PEST: Planning and Evaluation of Sequential Trials, Version 4: Operating Manual.* University of Reading, Reading, UK.

[12] Oosterhoff, J. (1969). *Combination of One-Sided Statistical Tests*, Mathematical Centre Tract 28. The Mathematical Centre, Amsterdam.

[13] Rosenthal, R. (1978). Combining results of independent studies. *Psych. Bull.* **85** 185-193.

[14] Sooriyarachchi, M. R., Whitehead, J., Matsushita, T., Bolland, K., and Whitehead, A. (2003). Incorporating data received after a sequential trial has stopped into the final analysis: Implementation and comparison of methods. *Biometrics* **59**:701-709.

[15] Stouffer, S. A., Suchman, E. A., DeVinner, L. C., Star, R. M., Williams, R. M. (1949). *The American Soldier: Adjustment During Army Life, Vol. I.* Princeton University Press, Princeton, NJ.

[16] Whitehead, J. (1992). Overrunning and underrunning in sequential clinical trials. *Controlled Clinical Trials* **13** 106-121.

[17] Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials*, 2nd ed. revised. Wiley, New York.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 46

On predictive probability matching priors

Trevor J. Sweeting[1]

University College London

Abstract: We revisit the question of priors that achieve approximate matching of Bayesian and frequentist predictive probabilities. Such priors may be thought of as providing frequentist calibration of Bayesian prediction or simply as devices for producing frequentist prediction regions. Here we analyse the $O(n^{-1})$ term in the expansion of the coverage probability of a Bayesian prediction region, as derived in [4]. Unlike the situation for parametric matching, asymptotic predictive matching priors may depend on the level α. We investigate *uniformly predictive matching priors* (UPMPs); that is, priors for which this $O(n^{-1})$ term is zero for all α. It was shown in [4] that, in the case of quantile matching and a scalar parameter, if such a prior exists then it must be Jeffreys' prior. In the present article we investigate UPMPs in the multiparameter case and present some general results about the form, and uniqueness or otherwise, of UPMPs for both quantile and highest predictive density matching.

Contents

1. Introduction

Prior distributions that match posterior predictive probabilities with the corresponding frequentist probabilities are attractive when a major goal of a statistical analysis is the construction of prediction regions. Such priors provide calibration of Bayesian prediction or may be viewed as a Bayesian mechanism for producing frequentist prediction intervals.

It is known that exact predictive probability matching is possible in cases in which there exists a suitable transformation group associated with the model. The general group structure for parametric models starts with a group of transformations on the sample space under which the statistical problem is invariant. This group of transformations then gives rise to a group G of transformations on the parameter space. From an 'objective Bayes' point of view, it makes sense to choose a prior distribution that is (relatively) invariant under this group. In particular, this will ensure that initial transformation of the data will make no difference to

[1] Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK, e-mail: trevor@stats.ucl.ac.uk

AMS 2000 subject classifications: Primary 62F15; secondary 62E20

Keywords and phrases: Bayesian inference, predictive inference, probability matching prior, asymptotic theory

predictive inferences. The two fundamental invariant measures on the group G are the left and right Haar measures. The left (right) Haar measure is the unique left-(right-)translation invariant measure on G, up to a positive multiplicative constant. These measures give rise to invariant left and right Haar priors on the parameter space. In the decision-theoretic development, under suitable conditions it turns out that the right Haar prior gives rise to optimal invariant decision rules for invariant decision problems; see, for example, [1]. The left Haar prior, however, which coincides with Jeffreys' invariant prior, often gives inadmissible rules in multiparameter cases. These facts provide strong motivation for the use of the right Haar prior. In relation to predictive inference, following earlier work in [10] and [11] this intuition was further reinforced in [13], where it was shown that if such a group structure exists then the associated right Haar prior gives rise to exact predictive matching for all invariant prediction regions. Thus the predictive matching problem is solved for models that possess a suitable group structure when the prediction region is invariant.

When exact matching is not possible one can instead resort to asymptotic approximation and investigate approximate predictive matching. This question was explored in [4] for the case of n independent and identically distributed (i.i.d.) observations. For regular parametric families the difference between the frequentist and posterior predictive probabilities is $O(n^{-1})$ and a concise expression for this difference was obtained in [4] by using the auxiliary prior device introduced by P. J. Bickel and J. K. Ghosh in [3]. This technical device has proved to be extremely valuable for the theoretical comparison of Bayesian and frequentist inference statements, or simply as a Bayesian device for obtaining frequentist results. It has been particularly useful for deriving probability matching priors (see, for example, [8], [9] and the review in [5]) and for studying properties of sequential tests ([16]).

In order to find an approximate predictive probability matching prior, one sets the $O(n^{-1})$ discrepancy to zero and attempts to solve the resulting partial differential equation (PDE); a number of examples are given in [4]. We briefly review the main results in [4] in Section 2. Two main issues arise from this analysis. Firstly, the PDE for a predictive matching prior may be difficult to solve analytically. The second, and more fundamental, issue is that, except in special cases, the resulting matching prior will depend on the desired predictive probability level α. If there does exist a prior that gives rise to predictive probability matching for all α then we shall refer to it as a *uniformly predictive matching prior* (UPMP). Of course, in the case of a transformation model and an invariant prediction region we already know from [13] that the right Haar prior must be a solution of the PDE. It is instructive to demonstrate this directly and this is done in the Appendix for quantile matching. Since the definition of the right Haar prior depends on a specific group of transformations on the parameter space, we need to study the effect of parameter transformation on the quantities appearing in the PDE. For this reason, it is natural to regard Fisher's information, g, as a Riemannian metric tensor so that transformational properties of g and the other quantities that appear in the PDE can be studied.

In the case of quantile matching and a single real parameter, it has already been shown in [4] that if there exists a UPMP then it must be Jeffreys' invariant prior. This result therefore extends the exact Haar prior result for transformation models to the most general models for which approximate uniform matching is possible. However, it is clear from examples discussed in [4] and from the general theory for transformation models in [13] that this result will not hold in the multiparameter case. For example, the unique UPMP for the normal model with unknown mean and variance is the right Haar prior, or Jeffreys' independence prior, whereas Jeffreys' prior is the left Haar prior.

The main purpose of the present article is to investigate the general form of UPMPs whenever they exist. In particular, we explore the uniqueness or otherwise of the right Haar prior as a UPMP for quantile matching in the case of a transformation model. Although UPMPs can exist outside of transformation models, such situations would seem to occur rarely. The main results are given in Section 3 and 4. In Section 3 we explore the form of the UPMP for quantile matching. In addition to confirming that the right Haar prior is a UPMP in suitable transformation models, as discussed above, we obtain the general form of the UPMP whenever one exists and show that this prior is unique. In particular, it follows that for transformation models there are no priors other than the right Haar prior that give approximate uniform predictive quantile matching. In Section 4 we consider probability matching based on highest predictive density regions, which are particularly relevant for multivariate data. The scalar parameter case is clear-cut and was essentially treated in [4], where it was shown that if there exists a UPMP then it is unique. However, unlike quantile matching, this UPMP is not necessarily Jeffreys' prior. The situation is less straightforward in the multiparameter case. We show that, under a certain condition, if there exists a UPMP then it is unique. If this condition is not satisfied then either there will be no UPMP or there will exist an infinite number of UPMPs. This section provides predictive versions of results for highest posterior density regions obtained by J. K. Ghosh and R. Mukerjee in [8] and [9]. We end with some discussion in Section 5.

2. Review of predictive probability matching priors

We begin by introducing the notation and reviewing the main results in [4] on predictive probability matching priors. We consider only the case of i.i.d. observations in this article, but the results would be expected to hold more generally under suitable conditions. Suppose then that $X_1, X_2, \dots$ is a sequence of independent observations having the same distribution as the (possibly vector-valued) continuous random variable X with density $f(\cdot;\theta)$, where $\theta = (\theta_1, \dots, \theta_p) \in \Omega$ is an unknown parameter and Ω is an open subset of $\Re^p$. Consider the problem of predicting the next observation, X_{n+1}, based on the first n observations, $d = (X_1, X_2, \dots X_n)$. We assume regularity conditions on f and π, as detailed in [4]. In particular, the support of X is assumed to be independent of θ.

Consider first the case of univariate X. Let $q(\pi, \alpha, d)$ denote the $1-\alpha$ quantile of the posterior predictive distribution of X_{n+1} under the prior π. That is, $q(\pi, \alpha, d)$ satisfies the equation

$$P^{\pi}(X_{n+1} > q(\pi, \alpha, d)|d) = \alpha\,. \tag{2.1}$$

Let $q(\theta, \alpha)$ be the $1-\alpha$ quantile of $f(\cdot;\theta)$; that is

$$\int_{q(\theta,\alpha)}^{\infty} f(u;\theta)du = \alpha\,. \tag{2.2}$$

Write $\partial_t = \partial/\partial\theta_t$ and let $f_t(u;\theta) = \partial_t f(u;\theta)$. Define

$$\mu_t(\theta, \alpha) = \int_{q(\theta,\alpha)}^{\infty} f_t(u;\theta)du\,. \tag{2.3}$$

Finally, let $g(\theta)$ be the per observation Fisher information matrix, which we assume to be non-singular for all $\theta \in \Omega$, and let g_{st} and g^{st} be the (s,t)th elements of g and g^{-1} respectively.

Using the approach of [3] and [7] in which an auxiliary prior is introduced and finally allowed to converge weakly to the degenerate measure at θ, it follows from equations (3.3) and (3.4) in [4] that

$$P_\theta(X_{n+1} > q(\pi,\alpha,d)) = \alpha - \frac{\partial_s\{g^{st}(\theta)\mu_t(\theta,\alpha)\pi(\theta)\}}{n\pi(\theta)} + o(n^{-1})\,. \tag{2.4}$$

Here and elsewhere we use the summation convention. We will say that π is a *level-α predictive probability matching prior* if it satisfies the equation

$$\partial_s\{g^{st}(\theta)\mu_t(\theta,\alpha)\pi(\theta)\} = 0\,. \tag{2.5}$$

From (2.4), such a prior π matches the Bayesian and frequentist predictive probabilities to $o(n^{-1})$. Clearly, in general a solution of (2.5) will depend on the particular level α chosen. This is demonstrated in [4] for the specific example in which the observations are from a $N(\theta,\theta)$ distribution. Recalling the discussion in Section 1, we refer to a prior for which (2.5) holds for all α as a *uniformly predictive matching prior* (UPMP). In the case $p = 1$, it was shown in [4] that if there exists a UPMP then this prior must be Jeffreys' prior. As noted in [4], when no UPMP exists then the formula on the left-hand side of (2.5) may still be useful for comparing alternative priors.

Moving to the multiparameter case, examples in [4] illustrate that the above result on Jeffreys' prior no longer holds. An illustration of this is Example 2 in [4], which is the location-scale model $f(x;\theta) = \theta_2^{-1}f^*(\theta_2^{-1}(x-\theta_1))$. In this case there exists a UPMP given by $\pi(\theta) \propto \theta_2^{-1}$, which is the right Haar prior for this model under the location-scale transformation group, whereas Jeffreys' prior is the left-invariant prior $\pi(\theta) \propto \theta_2^{-2}$ under this group.

In the case where X is possibly vector-valued, the coverage properties of highest predictive density regions are investigated in [4]. This investigation mirrors that in [8] and [9] for highest posterior density regions. Let $m(\theta,\alpha)$ be such that

$$\int_A f(u;\theta)du = \alpha\,,$$

where $A = A(\theta,\alpha) = \{u : f(u;\theta) \geq m(\theta,\alpha)\}$ and define

$$\xi_j(\theta,\alpha) = \int_A f_j(u;\theta)du\,.$$

Let $H(\pi,\alpha,d)$ be the level-α highest predictive density region under the prior π. Then, as for quantile matching, it follows from the results in Section 5 of [4] that

$$P_\theta(X_{n+1} \in H(\pi,\alpha,d)) = \alpha - \frac{\partial_s\{g^{st}(\theta)\xi_t(\theta,\alpha)\pi(\theta)\}}{n\pi(\theta)} + o(n^{-1})\,.$$

Thus π is a level-α predictive probability matching prior if and only if it satisfies the equation

$$\partial_s\{g^{st}(\theta)\xi_t(\theta,\alpha)\pi(\theta)\} = 0\,. \tag{2.6}$$

Once again we see that in general the solution π will depend on the level α. Examples are given in [4] in which there are no priors that satisfy (2.6) for all α. Moreover, even in the case $p = 1$, if there does exist a unique prior satisfying (2.6) for all α then it is not necessarily Jeffreys' prior.

3. UPMPs: quantile matching

As discussed in Section 2 we know that when $p = 1$ and a UPMP exists for quantile matching as in (2.1), then it must be Jeffreys' prior. However, it need not be Jeffreys' prior when $p > 1$. Under a suitable group structure on the model, the results in [13] imply that the associated right Haar prior gives exact predictive matching, since the prediction region here is invariant. Thus in these cases the right Haar prior must also be a solution of equation (2.5). It is instructive to demonstrate directly that this is indeed the case.

First note from the product rule that equation (2.5) is equivalent to

$$g^{st}(\theta)\mu_t(\theta,\alpha)\partial_s\lambda(\theta) + \partial_s\{g^{st}(\theta)\mu_t(\theta,\alpha)\} = 0\,, \tag{3.1}$$

where $\lambda(\theta) = \log\pi(\theta)$. Suppose that there exists a group G of bijective transformations on the sample space under which the statistical problem is invariant. Further assume, as in [13], that $G = \Omega$, a locally compact topological group. In this case the distribution of X under θ is the same as that of θX under e, the identity element of the group, with θ regarded as an element of the transformation group G. Then there exist unique (up to a multiplicative constant) left-invariant and right-invariant Haar measures on G, giving left and right Haar priors on the parameter space. In the following we denote the right Haar prior density on Ω by π^H. The proof of the following theorem is given in the Appendix.

Theorem 3.1. *Under the above group structure the right Haar prior satisfies equation (3.1).*

Two questions naturally arise. First, if the above group structure exists then can there be UPMPs other than the right Haar prior? The answer to this question turns out to be 'no', as follows from Theorem 3.2 below. Second, if the above group structure does not exist can there still be a UPMP? The answer to this question is 'yes'. An example in the case $p = 1$ is given in Section 3 of [4] for which there is no suitable group structure but there is still a unique UPMP, which must of course be Jeffreys' prior.

We now establish the general form of the UPMP whenever it exists and show that it is unique. This is a multiparameter version of Theorem 1 in [4]. Let $F(x;\theta)$ be the distribution function of X, $l(x;\theta) = \log f(x;\theta)$ and write $F_s(x;\theta) = \partial_s F(x;\theta)$, $l_s(x;\theta) = \partial_s \log f(x;\theta)$. Define the functions

$$h_r = g^{st}\int (F_s l_r - F_r l_s)\frac{\partial l_t}{\partial x}dx\,, \tag{3.2}$$

where the integration is over the (common) support of $F(x;\theta)$. Finally write $\lambda^J = \log\pi^J$, where $\pi^J(\theta) \propto |g(\theta)|^{1/2}$ is Jeffreys' prior.

Theorem 3.2. *Suppose that there exists a UPMP, π, for quantile matching. Then π is the unique UPMP and the partial derivatives of $\lambda = \log\pi$ are given by*

$$\partial_r\lambda(\theta) = \partial_r\lambda^J(\theta) + h_r(\theta)\,. \tag{3.3}$$

Proof. We begin by expressing $g(\theta)$ in terms of the functions $\mu_t(\theta;\alpha)$ defined at (2.3). By differentiation of equation (2.2) with respect to α we see that $-f(q;\theta)\partial q/\partial\alpha = 1$, while differentiation of equation (2.3) gives

$$\partial\mu_j(\theta,\alpha)/\partial\alpha = -f_j(q;\theta)\partial q/\partial\alpha = l_j(q;\theta)\,, \tag{3.4}$$

on substitution of $\partial q/\partial\alpha$ from the previous relation. It follows that

$$(3.5)\qquad g_{ij}(\theta) = \int l_i(q;\theta)l_j(q;\theta)f(q;\theta)dq = \int_0^1 \left(\frac{\partial\mu_i(\theta,\alpha)}{\partial\alpha}\right)\left(\frac{\partial\mu_j(\theta,\alpha)}{\partial\alpha}\right)d\alpha\,.$$

Suppose that there exists a UPMP π. Differentiation of equation (3.1) with respect to α and multiplication by $\partial\mu_r/\partial\alpha$ gives the equation

$$g^{st}\frac{\partial\mu_t}{\partial\alpha}\frac{\partial\mu_r}{\partial\alpha}\partial_s\lambda + \frac{\partial\mu_r}{\partial\alpha}\partial_s\left\{g^{st}\frac{\partial\mu_t}{\partial\alpha}\right\} = 0\,.$$

Since this relation must hold for all $0 < \alpha < 1$, integration over $0 < \alpha < 1$ gives

$$(3.6)\qquad g^{st}\left\{\int_0^1 \frac{\partial\mu_t}{\partial\alpha}\frac{\partial\mu_r}{\partial\alpha}d\alpha\right\}\partial_s\lambda + \int_0^1 \frac{\partial\mu_r}{\partial\alpha}\partial_s(g^{st}\frac{\partial\mu_t}{\partial\alpha})d\alpha = 0\,.$$

But from (3.5) the left-hand side of (3.6) is $g^{st}g_{tr}\partial_s\lambda = \delta^s_r\partial_s\lambda = \partial_r\lambda$, where δ^s_r is the Kronecker delta function. Also, since $\partial_s(g^{st}g_{tr}) = \partial_s(\delta^s_r) = 0$, the product rule gives

$$0 = g^{st}\int_0^1 \partial_s\left(\frac{\partial\mu_r}{\partial\alpha}\right)\frac{\partial\mu_t}{\partial\alpha}d\alpha + \int_0^1 \frac{\partial\mu_r}{\partial\alpha}\partial_s\left(g^{st}\frac{\partial\mu_t}{\partial\alpha}\right)d\alpha$$

so that (3.6) becomes

$$\partial_r\lambda = g^{st}\int_0^1 \partial_s\left(\frac{\partial\mu_r}{\partial\alpha}\right)\frac{\partial\mu_t}{\partial\alpha}d\alpha\,.$$

This expression gives the partial derivatives of $\lambda = \log\pi$ and, furthermore, establishes that π is the unique UPMP. We now show that this expression is equivalent to (3.3).

We first obtain the partial derivatives of λ^J. From a standard result for the derivative of a matrix determinant, we have

$$\begin{aligned}\partial_r\lambda^J = \frac{1}{2}\partial_r\log|g| &= \frac{1}{2}g^{st}\partial_r g_{st}\\ &= \frac{1}{2}g^{st}\partial_r\int_0^1\left(\frac{\partial\mu_s}{\partial\alpha}\right)\left(\frac{\partial\mu_t}{\partial\alpha}\right)d\alpha = g^{st}\int_0^1 \partial_r\left(\frac{\partial\mu_s}{\partial\alpha}\right)\left(\frac{\partial\mu_t}{\partial\alpha}\right)d\alpha\,,\end{aligned}$$

again using (3.5). The difference between the rth partial derivatives of λ and λ^J is therefore

$$(3.7)\qquad \partial_r\lambda - \partial_r\lambda^J = g^{st}\int_0^1 \frac{\partial}{\partial\alpha}(\partial_r\mu_s - \partial_s\mu_r)\frac{\partial\mu_t}{\partial\alpha}d\alpha\,.$$

Differentiation of (2.2) with respect to θ_r gives, writing $q = q(\theta,\alpha)$, $q_r f(q;\theta) + \mu_r(\theta,\alpha) = 0$, from which we obtain

$$\partial_r\mu_s(\theta,\alpha) = \int_q^\infty f_{rs}(u;\theta)du - f_s(q;\theta)q_r = \int_q^\infty f_{rs}(u;\theta)du - l_s(q;\theta)\mu_r(\theta,\alpha)\,.$$

Furthermore, we have

$$\frac{\partial}{\partial\alpha}\{l_s(q;\theta)\mu_r(\theta,\alpha)\} = \frac{\partial q}{\partial\alpha}\frac{\partial l_s(q;\theta)}{\partial q}\mu_r(\theta,\alpha) + l_s(q;\theta)l_r(q;\theta)\,.$$

It now follows from these two relations that

$$\frac{\partial}{\partial\alpha}(\partial_r\mu_s - \partial_s\mu_r) = \frac{\partial q}{\partial\alpha}\left(\frac{\partial l_r(q;\theta)}{\partial q}\mu_s - \frac{\partial l_s(q;\theta)}{\partial q}\mu_r\right) .$$

Substituting into equation (3.7) gives

$$h_r = g^{st}\int\left(F_s\frac{\partial l_r}{\partial q} - F_r\frac{\partial l_s}{\partial q}\right)l_t dq \tag{3.8}$$

on the change of variables from α to q, using equation (3.4) and on noting that $\mu_s(\theta,\alpha(q,\theta)) = -F_s(q;\theta)$. Next note that the indefinite integral

$$\int F_s(q;\theta)\frac{\partial l_r(q;\theta)}{\partial q}dq = F_s(q;\theta)l_r(q;\theta) - \int l_s(q;\theta)l_r(q;\theta)dq$$

from which it follows by an integration by parts that (3.8) is equivalent to (3.2), as required. □

In the case $p = 1$ we have $h_r = 0$ so the unique UPMP is Jeffreys' prior, as given in Theorem 1 of [4]. For the location-scale model $f(x;\theta) = \theta_2^{-1}f^*(\theta_2^{-1}(x-\theta_1))$ discussed in Section 1, it can be verified that the solution to (3.3) is $\pi(\theta) \propto \theta_2^{-1}$, which is the right Haar prior for this model under the location-scale transformation group. In general a necessary condition for there to be a UPMP is that h_r be a derivative field. The condition is not sufficient, however, as Jeffreys' prior always satisfies equation (3.3) in the case $p = 1$ but we know from [4] that Jeffreys' prior is not necessarily a UPMP. When $p > 1$ the condition that h_r be a derivative field is a very strong one when the model is not transformational. We have been unable to construct a two-dimensional example that is not transformational and that satisfies this condition. Even given a model satisfying this condition, the resulting prior may still not satisfy (2.5) for all α. Thus it would seem that UPMPs rarely exist outside of transformation models. The major point of Theorem 3.2, however, is to show that if a UPMP does exist then it is unique.

Note that, whether or not a UPMP exists, when h_r is a derivative field then (3.2) defines a unique prior π which, from the proof of the Theorem 3.2, satisfies the equation $\int_0^1\left(\frac{\partial\mu_r}{\partial\alpha}\right)\left(\frac{\partial\epsilon}{\partial\alpha}\right)d\alpha = 0$, where $\epsilon(\theta,\alpha)$ is the $O(n^{-1})$ error term (2.4). Assuming that $\partial\mu_r/\partial\alpha$ is well behaved at $\alpha = 0$ and $\alpha = 1$, integration by parts shows that this is equivalent to

$$\int_0^1\frac{\partial^2\mu_r}{\partial\alpha^2}\epsilon\, d\alpha = 0$$

for all θ and r. These relations give some sort of average prediction error, but it is unclear what precise interpretation can be given to them.

Finally, when there exists a suitable group structure as discussed earlier then we know that $\partial_i\lambda$ must be $\partial_i\lambda^H$. Furthermore, since Jeffreys' prior is the left Haar prior, it follows that $h_i(\theta) = \partial_i\log\Delta(\theta^{-1})$, where Δ is the modulus of Ω and θ^{-1} is the group inverse of θ.

4. UPMPs: highest predictive density region matching

We consider now the case where X is possibly vector-valued. The question of the existence of UPMPs for highest predictive density regions in this case is not so straightforward as the quantile matching case discussed in Section 3. In particular, if there exists a suitable group structure, as in Section 3, since the prediction region $H(\pi, \alpha, d)$ defined in Section 1 is not invariant under transformation of X (unless the group is affine), the associated right Haar prior is not necessarily a UPMP. We also know that when a UPMP does exist it may not be unique. This was illustrated in Example 4 in [4] of the bivariate normal model with unknown covariance matrix; we will return to this example in Example 1 below.

The scalar parameter case is straightforward, however. For each α the prior

$$\pi(\theta) \propto g(\theta)\{\xi_1(\theta, \alpha)\}^{-1}$$

is the unique solution to (2.6). It follows that there exists a UPMP prior if and only if $\xi_1(\theta, \alpha) = Q(\theta)R(\alpha)$, as was noted in [4] where examples are given in which this condition does and does not hold. Unlike the case of quantile matching, however, the unique solution when it exists is not necessarily Jeffreys' prior. For example, in [4] it is shown that a unique UPMP exists for the $N(\theta, \theta)$ model but this is not Jeffreys' prior.

The multiparameter case is more difficult. The simplest situation is when $\xi_t(\theta, \alpha)$ is of the form

$$\xi_t(\theta, \alpha) = Q_t(\theta)R(\alpha)\,. \tag{4.1}$$

Then every UPMP will be a solution of the Lagrange PDE

$$\partial_s\{g^{st}(\theta)Q_t(\theta)\pi(\theta)\} = 0\,. \tag{4.2}$$

This equation may have no solutions or an infinite number of solutions.

Example 1. Consider the bivariate normal model with zero means and unknown standard deviations σ_1, σ_2 and correlation coefficient ρ. Let Σ be the covariance matrix of X. We work with the orthogonal parameterisation

$$T^{-1} = \begin{pmatrix} \theta_1 & 0 \\ \theta_2\theta_3 & \theta_2 \end{pmatrix},$$

where $\Sigma = TT'$ and T is the left Cholesky square root of Σ. It can then be shown that the information matrix is

$$g(\theta) = \mathrm{diag}(2\theta_1^{-2}, 2\theta_2^{-2}, \theta_1^{-2}\theta_2^2)\,.$$

Furthermore, by transforming to $Z = T^{-1}X$, it can be shown that

$$m(\theta, \alpha) = \theta_1\theta_2(1-\alpha)/(2\pi),\ \xi_1(\theta, \alpha) = \theta_1^{-1}R(\alpha),\ \xi_2(\theta, \alpha) = \theta_2^{-1}R(\alpha),\ \xi_3(\theta, \alpha) = 0\,,$$

where $R(\alpha) = -(1-\alpha)\log(1-\alpha)$. Thus $\xi_t(\theta, \alpha)$ is of the form (4.1). Therefore the UPMP priors are all the solutions of the PDE (4.2) with $Q_1(\theta) = \theta_1^{-1}$, $Q_2(\theta) = \theta_2^{-1}$ and $Q_3(\theta) = 0$. The general solution is found to be

$$\pi(\theta) \propto \theta_1^{-2}h(\theta_2^{-1}\theta_1, \theta_3)\,,$$

where h is an arbitrary positive function. Notice that the leading term θ_1^{-2} is $|g(\theta)|^{1/2}$, so Jeffreys' prior is a UPMP. In terms of $(\sigma_1, \sigma_2, \rho)$ we have

$$\theta_1 = \sigma_1^{-1},\ \theta_2 = \sigma_2^{-1}(1-\rho^2)^{-1/2},\ \theta_3 = -\rho\sigma_1^{-1}\sigma_2$$

with Jacobian of transformation $\sigma_1^{-3}\sigma_2^{-1}(1-\rho^2)^{-3/2}$. With suitable re-expression of h we find that

$$\pi(\sigma_1,\sigma_2,\rho) \propto \pi^J(\sigma_1,\sigma_2,\rho)H(\sigma_1^{-1}\sigma_2,(1-\rho^2)^{1/2})\,, \tag{4.3}$$

where $\pi^J(\sigma_1,\sigma_2,\rho) \propto \sigma_1^{-1}\sigma_2^{-1}(1-\rho^2)^{-3/2}$ is Jeffreys' prior and H is an arbitrary positive function. This is a very wide class of priors. In particular, taking $h(x,y) = x^a y^b$, we see that all priors of the form $\pi^J(\sigma_1,\sigma_2,\rho)(\sigma_1^{-1}\sigma_2)^a(1-\rho^2)^b$ are UPMPs. Taking $a=1, b=1/2$ we obtain $\sigma_1^{-2}(1-\rho^2)^{-1}$, which can be shown to be the right Haar prior arising from the group of transformations $T^{-1}X$ on the sample space, where T is a lower triangular matrix with positive diagonal elements. This group is isomorphic to Ω and since in this case the region A is invariant it follows from [13] that this prior must be a UPMP. Similarly, all right Haar priors arising from transformations of the form $T^{-1}MX$, with M a fixed non-singular matrix, are included in (4.3).

We now return to the general analysis of equation (2.6). In Theorem 4.1 below, when we say that the functions $\xi_t(\theta,\alpha)$ are *linearly independent* we shall mean that they are linear independent as functions of α for fixed θ.

Theorem 4.1. *Suppose that the functions $\xi_t(\theta,\alpha)$ are linearly independent and that there exists a UPMP, π, for highest predictive density region matching. Then π is the unique UPMP and the partial derivatives of $\lambda = \log \pi$ are given by*

$$\partial_j\lambda = -b^{ri}g_{ij}\int_0^1 \frac{\partial\xi_r}{\partial\alpha}\partial_s\left(g^{st}\frac{\partial\xi_t}{\partial\alpha}\right)\,d\alpha\,, \tag{4.4}$$

where $(b^{ri}(\theta))$ is the inverse of the non-singular matrix function $(b_{ij}(\theta))$ with (i,j)th element

$$b_{ij}(\theta) = \int_0^1 \left(\frac{\partial\xi_i(\theta,\alpha)}{\partial\alpha}\right)\left(\frac{\partial\xi_j(\theta,\alpha)}{\partial\alpha}\right)d\alpha\,. \tag{4.5}$$

Proof. We begin by showing that the matrix $(b_{ij}(\theta))$ is non-singular for all $\theta\in\Omega$ if and only if the functions $\xi_t(\theta,\alpha)$ are linearly independent. From the definition (4.5), we see that in general $(b_{ij}(\theta))$ is positive semidefinite and is therefore singular for all $\theta\in\Omega$ if and only if, for each θ, there exist functions $x^t(\theta)$, not all zero, for which $b_{ij}(\theta)x^i(\theta)x^j(\theta)=0$. This is equivalent to the condition

$$\int_0^1 \left(\frac{\partial x^t(\theta)\xi_t(\theta,\alpha)}{\partial\alpha}\right)^2 d\alpha = 0\,,$$

which in turn holds if and only if $\partial(x^t(\theta)\xi_t(\theta,\alpha))/\partial\alpha = 0$ for all θ and α. Since $\xi_t(\theta,1)=0$ it follows that a necessary and sufficient condition for the singularity of $(b_{ij}(\theta))$ is the existence of $x^t(\theta)$, not all zero, such that $x^t(\theta)\xi_t(\theta,\alpha)=0$ for all θ and α. That is, the functions $\xi_t(\theta,\alpha)$ are linearly dependent.

We now apply the product rule to (2.6) to give equation (3.1) with μ_t replaced by ξ_t. Exactly as in the proof of Theorem 3.2, we differentiate this equation with respect to α, multiply by $\partial\xi_r/\partial\alpha$ and integrate over $0<\alpha<1$ to give

$$g^{st}b_{tr}\partial_s\lambda + \int_0^1 \frac{\partial\xi_r}{\partial\alpha}\partial_s\left(g^{st}\frac{\partial\xi_t}{\partial\alpha}\right)d\alpha = 0\,.$$

Finally, under the condition of the theorem the matrix $(b_{ij}(\theta))$ is non-singular and equation (4.4) follows on multiplying both sides of the above expression by $b^{ri}g_{ij}$. □

In the case $p = 1$ we know that a UPMP exists if and only if $\xi_1(\theta, \alpha) = Q(\theta)R(\alpha)$, in which case

$$b_{11}(\theta) = \{Q(\theta)\}^2 \int_0^1 \{R(\alpha)\}^2 d\alpha \,.$$

Equation (4.4) then becomes $d\lambda/d\theta = d\log(g^{-1}\theta)/d\theta$, giving $\pi(\theta) \propto \{Q(\theta)\}^{-1}g(\theta)$ in agreement with the earlier discussion. In the multiparameter case, unlike Theorem 3.2, there does not appear to be any simple further development of (4.4). Returning to the univariate location-scale model $f(x; \theta) = \theta_2^{-1} f^*(\theta_2^{-1}(x - \theta_1))$, it can be verified that the functions $\xi_t(\theta, \alpha)$ are linearly independent and, as in Section 3, that the right Haar prior $\pi(\theta) \propto \theta_2^{-1}$ under the location-scale transformation group is the solution to (4.4). When $p > 1$ the condition that the right-hand side of (4.4) be a derivative field is very strong when the model is not transformational and we have been unable to find a two-dimensional example that is not a transformation model satisfying this condition. Again, as in Section 3, even for such an example the resulting prior may still not satisfy (2.6) for all α. Thus it would seem that unique UPMPs rarely exist outside of transformation models. As with Theorem 3.2, the major point of Theorem 4.1 is to show that, under the conditions of the Theorem, if a UPMP does exist then it is unique.

Note that when $p > 1$ Theorem 4.1 does not apply to the case (4.1) since the functions $\xi_t(\theta, \alpha)$ are linearly dependent and hence the matrix $(b_{ij}(\theta))$ is singular. A more general sufficient condition for linear dependence of the $\xi_t(\theta, \alpha)$ is

$$\xi_t(\theta, \alpha) = U_t(\theta)S(\theta, \alpha)\,. \tag{4.6}$$

Note that this is also a necessary condition for linear dependence in the case $p = 2$.

Suppose that (4.6) holds and that there exists a UPMP π. Then from equation (2.6) we see that

$$g^{st}U_t\partial_s\lambda + g^{st}U_t\partial_s \log S + \partial_s(g^{st}U_t) = 0$$

for all α, which implies that the function $g^{st}(\theta)U_t(\theta)\partial_s \log S(\theta, \alpha)$ must be free from α. Since no boundary conditions are imposed on the solutions to the resulting Lagrangian PDE, it follows that π must be one of an infinite number of solutions. Thus, under condition (4.6), either there is no UPMP or there is an infinite number of UPMPs. Note that (4.1) is a special case of (4.6).

It might appear at first sight that it is also possible to have an infinite number of UPMPs in the case of quantile matching, which would contradict the result of Theorem 1. However, using a parallel argument to that given above, we see that the structure (4.6) for $\mu_t(\theta, \alpha)$ cannot occur, as this would imply singularity of Fisher's information matrix.

Finally, the case

$$\xi_t(\theta, \alpha) = Q_t(\theta)R_t(\alpha)\,, \tag{4.7}$$

which is a generalisation of the simple case (4.1), is of some interest. It is easily seen that in this case the linear independence of the functions $\xi_t(\theta, \alpha)$ is equivalent to the linear independence of the functions $R_t(\alpha)$. Furthermore, the matrix $(b_{ij}(\theta))$ will be non-singular for all θ if and only if the matrix with (i, j)th element

$$a_{ij} = \int_0^1 \left(\frac{\partial R_i}{\partial \alpha}\right)\left(\frac{\partial R_j}{\partial \alpha}\right) d\alpha$$

is positive definite. This turns out to be the case for the location-scale models discussed earlier.

Example 2. Consider the multivariate location model with density $f(x;\theta) = f^*(x_1 - \theta_1, \ldots, x_p - \theta_p)$. The region A here is invariant under the group of transformations $x + a$, $a \in \mathcal{R}^p$ and it follows from [13] that the right Haar prior is an exact UPMP. Here the right Haar prior is also Jeffreys' prior, both being constant. We now investigate conditions under which this is the unique UPMP. As in [4], we find that $m(\theta, \alpha) = m(\alpha)$, free from θ, and $\xi_t(\theta, \alpha) = R_t(\alpha)$, which is of the form (4.7) with $Q_t(\theta) = 1$ for all t. It follows from the above discussion that Jeffreys' prior is the unique UPMP if and only if the functions $R_t(\alpha)$ are linearly independent. In that case, since both g^{st} and ξ_t are free from θ, the right-hand side of (4.4) is zero and, again, the unique UPMP is the uniform prior.

For many standard models, however, the functions $R_t(\alpha)$ will be linearly dependent. Suppose, for example, that f^* is elliptically symmetric, so that $f^*(z) = H(z'Cz)$ for some positive definite matrix C. Then it can be checked that $R_t(\alpha) = Q_t R(\alpha)$, which is of the form (4.1) with Q_t free from θ. The functions $R_t(\alpha)$ are clearly linearly dependent and hence, since we know that there exists at least one UPMP, there will be an infinite number of UPMPs. For example, in the case where f^* is spherically symmetric, we have $Q_t = Q$ and the Lagrange PDE (4.2) becomes $\sum_s \partial_s \lambda = 0$. The solutions of this equation are of the form $\pi(\theta) \propto \exp\{h(\theta_2 - \theta_1, \ldots \theta_p - \theta_1)\}$, where h is an arbitrary function. In particular, all priors of the form $\pi(\theta) \propto \exp(\sum_i a_i \theta_i)$ with $\sum_i a_i = 0$ will be uniformly matching in this case.

A similar analysis may be carried out for the multiparameter location-scale model with different location parameters, as described in [4]. Whether or not the scale parameters are assumed to be equal, there is an appropriate group of transformations for which the corresponding right Haar prior will be a UPMP. In either case $\xi_t(\theta, \alpha)$ is again of the form (4.7) so that whether or not the right Haar prior is the unique UPMP will depend on the linear independence or otherwise of the functions $R_t(\alpha)$.

When the model has no suitable group structure, we conjecture that the functions $\xi_r(\theta, \alpha)$ will always be linearly independent. To see the plausibility of this, note that the $\xi_r(\theta, \alpha)$ are linearly dependent if and only if there exist functions $x^t(\theta)$, not all zero, such that $\int_A \{x^t(\theta) l_t(x;\theta)\} f(x;\theta) dx = 0$ for all θ and α. Since the density $f(x;\theta)$ cannot be standardised by transformation, the only way that this would seem to be possible is if $x^t(\theta) l_t(x;\theta) = 0$ for all theta. However, it is easily seen by partial differentiation w.r.t θ_s that this condition leads to g being singular. This analysis therefore suggests that if the model is not transformational then there will either be no UPMP or a unique UPMP, which is then given by (4.4).

5. Discussion

Although it is known that exact matching of invariant prediction regions is achieved by the right Haar prior under a suitable group structure on the model, we have seen in Section 3 that there can be other priors that achieve approximate uniform predictive quantile matching, and that uniformly matching priors can exist when there is no suitable group structure, although these are rare. In common with other work on probability matching priors, predictive matching priors arise as solutions to a particular PDE, which in general can be very difficult to solve. However, in the case of uniform quantile matching, if a UPMP exists then it is unique and explicit formulae for its partial derivatives are available from Theorem 3.2.

Except in special cases, derivation of the UPMP for quantile matching via equation (3.3), or even verifying that the derivatives in (3.3) are consistent, will be intractable. An attractive alternative would be to use a data-dependent approximation of the UPMP based on a local prior of the form

$$\partial_r \lambda(\theta, \theta_0) = \partial_r \lambda^J(\theta) + h_r(\theta_0)\,.$$

See [14] for a derivation of data-dependent matching priors for marginal posterior distributions. Furthermore, since a data-dependent prior of this form will always exist, there may be cases for which it will be uniformly matching even when there is no α-free solution of (2.5). Although the posterior distribution arising from such a prior would not always have a strict Bayesian interpretation, use of the corresponding predictive distribution could provide a useful mechanism for constructing frequentist prediction regions with good coverage properties. It would be of interest to conduct simulation experiments in order to assess the predictive coverage afforded by such priors.

The case of highest predictive density regions is more complex. As discussed in Section 4, this will either be the unique solution or else there will be infinitely many solutions, depending on the linear independence or otherwise of the functions $\xi_r(\theta, \alpha)$. Thus in any particular example it is necessary to examine carefully the structure of the functions $\xi_r(\theta, \alpha)$. If the statistical model has a suitable group structure then this task is usually eased. One could also investigate local priors when the matrix $(b_{ij}(\theta))$ is invertible.

In the case of univariate observations, the results provide some guidance on the choice of objective prior if the main goal is to carry out Bayesian prediction and low predictive coverage probability bias is desired. In relation to the determination of an objective prior, for multivariate data the situation is less clear. When the functions $\xi_r(\theta, \alpha)$ are linearly dependent, as often occurs in transformation models, there will usually be an infinite number of UPMPs. Thus other considerations will need to be invoked in order to narrow down the choice of prior. For example, one might consider priors that are simultaneously predictive and posterior probability matching, reference priors ([2]) or priors that are minimax under suitable decision rules; in particular, for minimax prediction loss see, for example, [12] and [15].

Appendix: Proof of Theorem 3.1

Proof. Let $a \in \Omega$ and consider the transformation $\phi = a\theta$. Let $J(\theta, a) = \partial\phi/\partial a$ be the Jacobian matrix of this transformation for fixed θ. Then the right Haar prior is $\pi^H(\theta) \propto |J(\theta, e)|^{-1}$, where $|J(\theta, a)|$ is the determinant of $J(\theta, a)$; see, for example, [1].

Write $\tilde{\phi}^r_s(\theta, a) = \partial\phi_r(\theta, a)/\partial a_s$ and define $\tilde{\phi}^r_s(\theta) = \tilde{\phi}^r_s(\theta, e)$, where e is the identity element of the group, so that $\pi^H(\theta) \propto |(\tilde{\phi}^r_s(\theta))|^{-1}$. Finally, let $(a^s_r(\theta))$ be the matrix inverse of $(\tilde{\phi}^r_s(\theta))$. A standard result for the derivative of a matrix determinant then gives

$$\partial_s \lambda^H(\theta) = -a^u_r(\theta)\partial_s \tilde{\phi}^r_u(\theta)\,, \tag{5.1}$$

where $\lambda^H(\theta) = \log \pi^H(\theta)$.

Define $\phi^r_s(\theta,a) = \partial_s\phi_r = \partial\phi_r/\partial\theta_s$, with matrix inverse $\theta^s_r(\theta,a) = \partial\theta_s/\partial\phi_r$. Since the definition of the right Haar prior depends on a specific group of transformations on the parameter space, it is natural to regard Fisher's information as a Riemannian metric tensor associated with the differentiable manifold of probability densities $f(\cdot;\theta)$, $\theta\in\Omega$. This facilitates the study of the transformational properties of the quanitities $g^{st}(\theta)$ and $\mu_t(\theta,\alpha)$ appearing in the PDE (3.1). First, from the invariance of the problem under G and the contravariant tensorial property of g^{st}, we have $\bar g^{ij}(\phi) = g^{st}(\theta)\phi^i_s\phi^j_t$, where $\bar g^{ij}$ is the inverse Fisher information in the ϕ-parameterisation. Again using the invariance properties, it is seen that $\bar\mu_j(\phi,\alpha) = \mu_k(\theta,\alpha)\theta^k_j$, where $\bar\mu_j(\phi,\alpha)$ is the function (2.3) in the ϕ-parameterisation. Now write $u^s(\theta,\alpha) = g^{st}(\theta)\mu_t(\theta,\alpha)$ and $\bar u^s(\phi,\alpha) = \bar g^{st}(\phi)\bar\mu_t(\phi,\alpha)$. Then

$$\begin{aligned} \bar u^i(\phi,\alpha) &= g^{st}(\theta)\mu_k(\theta,\alpha)\phi^i_s\phi^j_t\theta^k_j \\ &= g^{st}(\theta)\mu_k(\theta,\alpha)\phi^i_s\delta^k_t = u^s(\theta,\alpha)\phi^i_s\,, \end{aligned} \tag{5.2}$$

where δ^k_t is the Kronecker delta function.

Now differentiate both sides of (5.2) with respect to a_r to give

$$\partial_s\bar u^i(\phi,\alpha)\tilde\phi^r_s(\theta,a) = u^s(\theta,\alpha)\partial\phi^i_s(\theta,a)/\partial a_r = u^s(\theta,a)\partial_s\phi^i_r(\theta,a)\,. \tag{5.3}$$

Finally, setting $a=e$ and multiplying both sides of (5.3) by $a^r_i(\theta)$ gives

$$\partial_s u^i(\theta,\alpha)a^r_i(\theta)\tilde\phi^s_r(\theta) = u^s(\theta,\alpha)a^r_i(\theta)\partial_s\tilde\phi^i_r(\theta)\,. \tag{5.4}$$

Since $(a^s_r(\theta))$ is the matrix inverse of $(\tilde\phi^r_s(\theta))$, the left-hand side of (5.4) is $\partial_s u^i(\theta,\alpha)\delta^s_i = \partial_s u^s(\theta,\alpha)$, whereas the right-hand side is $-u^s(\theta,\alpha)\partial_s\lambda^H(\theta)$ from (5.1). It follows that the right Haar prior π^H is a solution of equation (3.1) and hence of equation (2.5). □

References

[1] Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer-Verlag, New York.

[2] Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. R. Statist. Soc. B* **41** 113–147.

[3] Bickel, P. J. and Ghosh, J. K. (1990). A decomposition for the likelihood ratio statistic and the Bartlett correction- a Bayesian argument. *Ann. Statist.* **18** 1070–1090.

[4] Datta, G. S., Mukerjee, R., Ghosh, M. and Sweeting, T. J. (2000). Bayesian prediction with approximate frequentist validity. *Ann. Statist.* **28** 1414–1426.

[5] Datta, G. S. and Sweeting, T. J (2005). Probability matching priors. *Handbook of Statistics, 25: Bayesian Thinking: Modeling and Computation* (D. K. Dey and C. R. Rao, eds.), Elsevier, 91–114.

[6] Eaton, M. L. and Sudderth, W. D. (1998). A new predictive distribution for normal multivariate linear models. *Sankhya, Ser. A* **60** 363–382.

[7] Ghosh, J. K. and Mukerjee, R. (1991). Characterization of priors under which Bayesian and frequentist Bartlett corrections are equivalent in the multiparameter case. *J. Multivariate Anal.* **38** 385–393.

[8] Ghosh, J. K. and Mukerjee, R. (1993). Frequentist validity of highest posterior density regions in multiparameter case. *Ann. Inst. Stat. Math.* **45** 293–302.

[9] Ghosh, J. K. and Mukerjee, R. (1995). Frequentist validity of highest posterior density regions in the presence of nuisance parameters. *Statist. Decis.* **13** 131–139.

[10] Hora, R. B. and Buehler, R. J. (1966). Fiducial theory and invariant estimation. *Ann. Math. Statist.* **37** 643–656.

[11] Hora, R. B. and Buehler, R. J. (1967). Fiducial theory and invariant prediction. *Ann. Math. Statist.* **38** 795-801.

[12] Liang, F. and Barron, A. R. (2004). Exact minimax strategies for predictive density estimation, data compression and model selection. *IEEE Trans. Inf. Theory* **50** 2708–2726.

[13] Severini, T. A., Mukerjee, R. and Ghosh, M. (2002). On an exact probability matching property of right-invariant priors. *Biometrika* **89** 952–957.

[14] Sweeting, T. J. (2005). On the implementation of local probability matching priors for interest parameters. *Biometrika*

[15] Sweeting, T. J., Datta, G. S. and Ghosh, M. (2006). Nonsubjective priors via predictive relative entropy loss. *Ann. Statist.* **34** 441–468.

[16] Woodroofe, M. (1986). Very weak expansions for sequential confidence levels. *Ann. Statist.* **14** 1049–1067.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 60
© Institute of Mathematical Statistics, 0000
arXiv: math.PR/0000000

Data-Dependent Probability Matching Priors For Empirical And Related Likelihoods

Rahul Mukerjee

Indian Institute of Management Calcutta
e-mail: rmuk1@hotmail.com

Abstract: We consider a general class of empirical-type likelihoods and develop higher order asymptotics with a view to characterizing members thereof that allow the existence of possibly data-dependent probability matching priors ensuring approximate frequentist validity of posterior quantiles. In particular, for the usual empirical likelihood, positive results are obtained. This is in contrast with what happens if only data-free priors are entertained.

Contents

1. Introduction

Although empirical and related likelihoods have received significant attention (see [11, 14] and the references therein), their study from a Bayesian perspective began only in recent years. Lazar [10] pioneered an investigation on the validity of the empirical likelihood for posterior inference and examined, mostly by simulation, the frequentist properties of posterior empirical likelihood intervals. In another significant development, Schennach [18] proposed a Bayesian exponentially tilted empirical likelihood arising as a nonparametric limit of a Bayesian procedure which places a kind of noninformative prior on the space of distributions. Starting from a general class of empirical-type likelihoods for the population mean, Fang and Mukerjee [7] characterized its members which admit probability matching priors in the sense of allowing posterior credible sets with approximate frequentist validity.

Joka, Diamond Harbour Road, Kolkata 700 104, India

AMS 2000 subject classifications: Primary 62F25
Keywords and phrases: Edgeworth expansion, empirical likelihood, higher order asymptotics, posterior quantile

Along the line of what is traditionally done in parametric inference based on the true likelihood, Fang and Mukerjee [7] entertained only priors that are free from the data. They observed, among other things, that none of the standard empirical-type likelihoods that have been proposed and widely studied in the literature, including the usual empirical likelihood, admits a probability matching prior even with margin of error $o(n^{-1/2})$, where n is the sample size. This is somewhat disappointing and, given the popularity of these likelihoods, prompts one to investigate the consequences of working with possibly data-dependent priors with the hope that this may yield more positive results. The present article aims at exploring this issue with reference to the same general class as in [7]. Satisfyingly, it is seen that at least for the usual empirical likelihood, positive results then emerge even with margin of error $o(n^{-1})$.

Probability matching priors have been studied extensively in parametric inference - see, for example, [6], [9], [12], and [20]. A key difference with the parametric case is that here we are not working with the true density-based likelihood and, as such, a shrinkage argument, suggested originally by J. K. Ghosh to the present author, that simplifies the frequentist calculations there ([6], Ch. 1) is no longer applicable. As a result, one has to employ a direct Edgeworth expansion based on approximate cumulants.

While the present work seems to be the first attempt towards exploring the higher-order asymptotics on data-dependent priors with empirical-type likelihoods, such priors have received considerable attention in recent years in parametric inference via the true density-based likelihood. A brief indication of this literature may interest the reader. A key reference in this regard is [21], where it was found that for certain mixture models, no data-free improper prior yields a proper posterior and no data-free proper prior entails frequentist validity of posterior quantiles with margin of error $O(n^{-1})$, while both problems are solved by a data-dependent prior. Furthermore, such data-dependent priors were shown to approximate data-free priors, in addition to enjoying desirable properties like asymptotic minimaxity. Prior to [21], data-dependent priors were considered, among others, by [15] and [17] in the context of mixture models with an unknown number of components. Sweeting [19] investigated the crucial role played by data-dependent probability matching priors when the sample size is stochastic, as happens, for instance, with censoring or a stopping rule. Reid et al. [16] reviewed and discussed a notion of strong matching which requires data-dependent priors.

In the context of parametric inference, Clarke and Yuan [4] studied partial information reference priors obtained through the maximization of conditional Shannon mutual information. These priors are often data-dependent in the sense of involving statistics that are associated with nuisance parameters and capture helpful side information. The information theoretic interpretation for these priors was also discussed at length by [4]. Clarke ([3], Subsection 7.2) discussed data-dependent priors in the light of the Freedman–Purves Theorem [8], which often forms the basis of the argument put forward by orthodox Bayesians against such priors on the grounds of incoherence. He argued that the implications of this theorem are narrower than commonly appreciated, suggested a remedy in the form of a criterion of information boundedness, and observed that the data-dependent priors in [4], [16], and [21] are, indeed, information bounded.

The interested reader may refer to the papers cited in the last two paragraphs for further references on data-dependent priors in parametric inference.

2. A General Class of Empirical-Type Likelihoods: Posterior Quantiles

Let $X_1, \ldots, X_n$ be independent scalar-valued random variables from an unknown common distribution with an unknown mean θ. The parameter space for θ is the real line or an open interval thereof. The X_i are supposed to be absolutely continuous and the first four population moments are assumed to exist [2]. These assumptions justify an Edgeworth expansion used later. Let $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$, $m_s = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^s$, $s = 2, 3, \ldots$, $g_3 = m_3/m_2^{3/2}$, $g_4 = m_4/m_2^2$ and $y = y(\theta) = (n/m_2)^{1/2}(\theta - \bar{X})$. Write $\phi(\cdot)$ for the standard univariate normal density.

As in [7], we consider a general class of empirical-type likelihoods of the form

$$\begin{aligned} L(\theta) \quad \propto \quad & \phi(y)[1 + n^{-1/2}\{a_1(g_3)y + a_3(g_3)y^3\} \\ & +n^{-1}\{b_0(g_3, g_4) + b_2(g_3, g_4)y^2 + b_4(g_3, g_4)y^4 + b_6(g_3, g_4)y^6\} \\ & +o_p(n^{-1})], \end{aligned} \tag{1}$$

where the $a_i(\cdot)$ are polynomials in g_3 and the $b_i(\cdot)$ are polynomials in g_3 and g_4, the coefficients therein being constants free from n. These polynomials depend on the particular likelihood. Note that $L(\theta)$ depends on θ through $y = y(\theta)$ but does not involve any other population parameter, and that y is the standard pivotal quantity for inference on θ when the population variance is unknown. Furthermore, the terms of order $n^{-1/2}$ and n^{-1} in (1) aim at taking care of the unknown skewness and kurtosis of the population via their sample analogs g_3 and g_4 respectively. Indeed, the class (1) is very wide and, as discussed later in Section 4, covers all major empirical-type likelihoods proposed in the literature.

With reference to any likelihood in the class (1), we consider a possibly data-dependent prior of the form

$$\pi(\theta) = \exp\{\psi(\theta, m_2, g_3)\}, \tag{2}$$

where $\psi(\cdot)$ is a smooth function with functional form free from n. We aim at characterizing the $a_i(\cdot)$ and $b_i(\cdot)$ in (1) so as to allow the existence of a prior of the form (2) that entails frequentist validity of the posterior quantiles of θ, with margin of error $o(n^{-1/2})$ or $o(n^{-1})$. The class (2) can be motivated as follows. For any empirical-type likelihood as in (1) and a data-free prior, up to the first order of approximation, the quantity $(\theta - \bar{X})m_2^{-1/2}$ represents a standardized version of θ in the posterior setup; see [7]. This prompts one to consider a data-dependent prior of the form

$$\pi(\theta) = \exp\{(\theta - \bar{X})m_2^{-1/2}\chi(g_3)\}, \tag{3}$$

where the multiplier $\chi(g_3)$ is a smooth function of g_3, with functional form free of n, that aims at taking care of the population skewness (an attempt to take care of the population kurtosis would involve a more elaborate data dependent prior and a discussion of this is deferred till Section 5). Clearly, the prior in (3) is equivalent to

$$\pi(\theta) = \exp\{\theta m_2^{-1/2}\chi(g_3)\}, \tag{4}$$

because they both lead to the same posterior. The exponent in (4) is a smooth function of θ, m_2 and g_3, and the class (2) incorporates all priors that share this feature.

Let $\psi_i(t_1,t_2,t_3)=\partial\psi(t_1,t_2,t_3)/\partial t_i$, $\psi_{ij}(t_1,t_2,t_3)=\partial^2\psi(t_1,t_2,t_3)/\partial t_i\partial t_j$, $\psi_i=\psi_i(\bar{X},m_2,g_3)$ and $\psi_{ij}=\psi_{ij}(\bar{X},m_2,g_3)$, $i,j=1,2,3$. Then, analogously to the parametric case ([6], Ch. 2), the posterior density of $y=y(\theta)$, with reference to (1) and under $\pi(\cdot)$ as in (2), can be expressed as

$$\begin{aligned}\pi^*(y|X) &= \phi(y)[1+n^{-1/2}(R_1y+R_3y^3)\\ &\qquad +n^{-1}\{R_2(y^2-1)+R_4(y^4-3)+R_6(y^6-15)\}]+o_p(n^{-1})\end{aligned} \tag{5}$$

where $X=(X_1,\ldots,X_n)$ and

$$\begin{aligned}R_1 &= a_1(g_3)+m_2^{1/2}\psi_1,\\ R_2 &= b_2(g_3,g_4)+m_2^{1/2}a_1(g_3)\psi_1+\frac{1}{2}m_2(\psi_{11}+\psi_1^2),\\ R_3 &= a_3(g_3),\\ R_4 &= b_4(g_3,g_4)+m_2^{1/2}a_3(g_3)\psi_1,\\ R_6 &= b_6(g_3,g_4).\end{aligned} \tag{6}$$

The propriety of the posterior is assumed here. Let

$$\begin{aligned}u_1 &= R_1+R_3(z^2+2),\\ u_2 &= 2u_1zR_3-\frac{1}{2}u_1^2z+R_2z+R_4(z^3+3z)+R_6(z^5+5z^3+15z),\end{aligned} \tag{7}$$

where z is the $(1-\alpha)$th quantile of a standard normal variate. As in [7], recalling that $y=(n/m_2)^{1/2}(\theta-\bar{X})$, then it follows from (5) that the $(1-\alpha)$th posterior quantile of θ can be approximated by

$$\theta_1^{(1-\alpha)}(\pi,X)=\bar{X}+(m_2/n)^{1/2}(z+n^{-1/2}u_1), \tag{8}$$

or

$$\theta_2^{(1-\alpha)}(\pi,X)=\bar{X}+(m_2/n)^{1/2}(z+n^{-1/2}u_1+n^{-1}u_2), \tag{9}$$

with posterior coverage error $o_p(n^{-1/2})$ or $o_p(n^{-1})$ respectively.

3. Frequentist Coverage

3.1. Calculation of Frequentist Coverage

We next study the frequentist coverage of the interval $(-\infty,\theta_2^{(1-\alpha)}(\pi,X)]$. The steps are similar to those in [7] but more involved because of possible data-dependence of the prior; for instance, additional terms appear in the expressions for W_1 and k_2 in (15) and (21) below.

With P representing the frequentist probability, by (9) and the definition of y, the frequentist coverage is given by

$$P\{\theta\le\theta_2^{(1-\alpha)}(\pi,X)\}=P(y\le z+n^{-1/2}u_1+n^{-1}u_2). \tag{10}$$

In order to obtain an expression for the above with margin of error $o(n^{-1})$, we need stochastic expansions for y, u_1 and u_2. To this end, let E denote expectation for fixed θ and write $\sigma^2 = E(X_i - \theta)^2$, $Z_i = (X_i - \theta)/\sigma$, $\beta_s = E(Z_i^s)$, $1 \le s \le 4$, and $A_s = n^{-1/2}\sum_{i=1}^n (Z_i^s - \beta_s)$, $s = 1, 2, 3$. Then, as in [7],

$$y = -A_1 + \frac{1}{2}n^{-1/2}A_1A_2 - n^{-1}(\frac{1}{2}A_1^3 + \frac{3}{8}A_1A_2^2) + o_p(n^{-1}). \tag{11}$$

Turning next to u_1 and u_2, we note from (7) that the randomness of these quantities is only due to the R_i. For each i, let R_{i0} be obtained from R_i in (6) replacing $\bar{X}$, m_2, g_3 and g_4 therein by the corresponding population parameters θ, σ^2, β_3 and β_4 respectively, i.e.,

$$\begin{aligned} R_{10} &= a_1(\beta_3) + \sigma\psi_1^{(0)}, \\ R_{20} &= b_2(\beta_3, \beta_4) + \sigma a_1(\beta_3)\psi_1^{(0)} + \frac{1}{2}\sigma^2[\psi_{11}^{(0)} + \{\psi_1^{(0)}\}^2], \\ R_{30} &= a_3(\beta_3), \\ R_{40} &= b_4(\beta_3, \beta_4) + \sigma a_3(\beta_3)\psi_1^{(0)}, \\ R_{60} &= b_6(\beta_3, \beta_4), \end{aligned} \tag{12}$$

where $\psi_i^{(0)} = \psi_i(\theta, \sigma^2, \beta_3)$, $\psi_{ij}^{(0)} = \psi_{ij}(\theta, \sigma^2, \beta_3)$, $i, j = 1, 2, 3$. Since

$$\bar{X} = \theta + n^{-1/2}\sigma A_1, \quad m_2 = \sigma^2(1 + n^{-1/2}A_2) + o_p(n^{-1/2}), \tag{13}$$

and

$$g_3 = \beta_3 + n^{-1/2}(A_3 - 3A_1 - \frac{3}{2}\beta_3A_2) + o_p(n^{-1/2}), \quad g_4 = \beta_4 + o_p(1), \tag{14}$$

from (6) we get $R_i = R_{i0} + n^{-1/2}W_i + o_p(n^{-1/2})$, $i = 1, 3$, and $R_i = R_{i0} + o_p(1)$, $i = 2, 4, 6$, where

$$\begin{aligned} W_1 &= \sigma^2\psi_{11}^{(0)}A_1 + \sigma\{\frac{1}{2}\psi_1^{(0)} + \sigma^2\psi_{12}^{(0)}\}A_2 \\ &\quad + \{a_1'(\beta_3) + \sigma\psi_{13}^{(0)}\}(A_3 - 3A_1 - \frac{3}{2}\beta_3A_2), \end{aligned} \tag{15}$$

$$W_3 = a_3'(\beta_3)(A_3 - 3A_1 - \frac{3}{2}\beta_3A_2), \tag{16}$$

and $a_i'(\cdot)$ is the derivative of $a_i(\cdot)$. From (7), it is now evident that

$$u_1 = u_{10} + n^{-1/2}\{W_1 + W_3(z^2 + 2)\} + o_p(n^{-1/2}), u_2 = u_{20} + o_p(1),$$

where the leading terms

$$u_{10} = R_{10} + R_{30}(z^2 + 2), \tag{17}$$

and

$$u_{20} = 2u_{10}zR_{30} - \frac{1}{2}u_{10}^2z + R_{20}z + R_{40}(z^3 + 3z) + R_{60}(z^5 + 5z^3 + 15z), \tag{18}$$

are simply counterparts of (7) with the R_i there replaced by the R_{i0}.

From (10) and the stochastic expansions for u_1 and u_2 as indicated above,

$$P\{\theta \le \theta_2^{(1-\alpha)}(\pi, X)\} = P(\tilde{y} \le z + n^{-1/2}u_{10} + n^{-1}u_{20}) + o(n^{-1}), \tag{19}$$

where $\tilde{y} = y - n^{-1}\{W_1 + W_3(z^2+2)\}$. By (11), (15) and (16), the first four approximate cumulants of $\tilde{y}$ are given by

$$\begin{aligned} K_{1n} &= n^{-1/2}k_1 + o(n^{-1}), \\ K_{2n} &= 1 + n^{-1}k_2 + o(n^{-1}), \\ K_{3n} &= n^{-1/2}k_3 + o(n^{-1}), \\ K_{4n} &= n^{-1}k_4 + o(n^{-1}), \end{aligned}$$

where

$$k_1 = \frac{1}{2}\beta_3, \qquad k_3 = 2\beta_3, \qquad k_4 = 12 + 12\beta_3^2 - 2\beta_4, \tag{20}$$

and

$$\begin{aligned} k_2 &= 3 + \frac{7}{4}\beta_3^2 + 2\{a_1'(\beta_3) + a_3'(\beta_3)(z^2+2) + \sigma\psi_{13}^{(0)}\}(\beta_4 - 3 - \frac{3}{2}\beta_3^2) \\ &\quad +2\sigma^2\psi_{11}^{(0)} + 2\beta_3\sigma\{\frac{1}{2}\psi_1^{(0)} + \sigma^2\psi_{12}^{(0)}\}. \end{aligned} \tag{21}$$

The fact that W_1 and W_3 are linear in the A_i facilitates the derivation of (20) and (21). From (19), consideration of an Edgeworth expansion for $\tilde{y}$ now yields

$$P\{\theta \le \theta_2^{(1-\alpha)}(\pi, X)\} = 1 - \alpha + (n^{-1/2}\Delta_1 + n^{-1}\Delta_2)\phi(z) + o(n^{-1}), \tag{22}$$

with

$$\Delta_1 = u_{10} - k_1 - \frac{1}{6}k_3(z^2-1), \tag{23}$$

and

$$\begin{aligned} \Delta_2 &= u_{20} - \frac{1}{2}u_{10}^2 z + zu_{10}\{k_1 + \frac{1}{6}k_3(z^2-3)\} - \frac{1}{2}(k_2 + k_1^2)z \\ &\quad -(\frac{1}{24}k_4 + \frac{1}{6}k_1k_3)(z^3 - 3z) - \frac{1}{72}k_3^2(z^5 - 10z^3 + 15z). \end{aligned} \tag{24}$$

3.2. Probability Matching Conditions

The frequentist coverage in (22) equals $1 - \alpha + o(n^{-1/2})$ if and only if $\Delta_1 = 0$ identically in z and the population parameters. Since by (12), (17), (20) and (23),

$$\Delta_1 = a_1(\beta_3) + 2a_3(\beta_3) + \sigma\psi_1^{(0)} - \frac{1}{6}\beta_3 + \{a_3(\beta_3) - \frac{1}{3}\beta_3\}z^2,$$

recalling the definition of $\psi_1^{(0)}$, it is clear that the above happens if and only if

$$a_3(\beta_3) = \frac{1}{3}\beta_3 \quad \text{and} \quad \psi(\theta, \sigma^2, \beta_3) = h(\sigma^2, \beta_3) - \theta\sigma^{-1}\{a_1(\beta_3) + \frac{1}{2}\beta_3\}, \tag{25}$$

$h(\cdot)$ being any smooth function of σ^2 and β_3. Note that the first condition in (25) is on the empirical-type likelihood whereas the second condition concerns the prior. Indeed, with $\psi(\cdot)$ as in (25), it is easily seen from (2) that the specific choice of $h(\cdot)$ has no influence on the posterior. Hence, hereafter, we take $h(\sigma^2, \beta_3) = 0$, i.e.,

$$\pi(\theta) = \exp[-\theta m_2^{-1/2}\{a_1(g_3) + \frac{1}{2}g_3\}], \tag{26}$$

by (2) and (25), and continue with (25) to obtain further conditions that arise when one wishes to work with margin of error $o(n^{-1})$. Observe that the prior in (26) is actually of the form (4) that motivated the class (2) of priors. From (12), (17), (18), (20), (21) and (24), after considerable algebra, it can be seen that under (25), $\Delta_2 = C_1 z + C_3 z^3 + C_5 z^5$, where

$$\begin{aligned} C_1 &= b_2(\beta_3,\beta_4) + 3b_4(\beta_3,\beta_4) + 15b_6(\beta_3,\beta_4) - \frac{1}{2}\{a_1(\beta_3)\}^2 - \beta_3 a_1(\beta_3) \\ &\quad + \frac{1}{24}\beta_3^2 - \frac{5}{12}\beta_4 + \frac{1}{2}, \\ C_3 &= b_4(\beta_3,\beta_4) + 5b_6(\beta_3,\beta_4) - \frac{1}{3}\beta_3 a_1(\beta_3) + \frac{2}{9}\beta_3^2 - \frac{1}{4}\beta_4 + \frac{1}{2}, \\ C_5 &= b_6(\beta_3,\beta_4) - \frac{1}{18}\beta_3^2. \end{aligned}$$

Thus, under (25), the frequentist coverage in (22) equals $1-\alpha+o(n^{-1})$ for every z and every possible θ, σ^2, β_3 and β_4 if and only if C_1, C_3 and C_5 vanish identically in the population parameters, i.e., if and only if

$$\begin{aligned} b_2(\beta_3,\beta_4) &= \frac{1}{2}\{a_1(\beta_3)\}^2 + \frac{5}{8}\beta_3^2 - \frac{1}{3}\beta_4 + 1, \\ b_4(\beta_3,\beta_4) &= \frac{1}{3}\beta_3 a_1(\beta_3) - \frac{1}{2}\beta_3^2 + \frac{1}{4}\beta_4 - \frac{1}{2}, \\ b_6(\beta_3,\beta_4) &= \frac{1}{18}\beta_3^2. \end{aligned} \tag{27}$$

The conditions in (27) are again on the likelihood.

4. Implications

We now examine some major subclasses of empirical-type likelihoods in the light of the conditions obtained in the last section. These are (i) likelihoods arising from empirical discrepancy statistics ([5], Section 1) and hence from Cressie–Read discrepancy statistics [1], (ii) generalized empirical likelihoods [13], and (iii) generalized empirical exponential family likelihoods ([5], Section 4). As noted in [7] all these belong to the general class (1). Moreover, it can be shown that the forms of the $a_i(\cdot)$ and $b_i(\cdot)$ for the subclasses (i)-(iii) are as in Table 1. In this table, τ_3, τ_4, γ_3, γ_4 and μ are constants that depend on the particular likelihood. The usual empirical likelihood belongs to each of (i)-(iii) with

$$\begin{gathered} a_1(g_3) = 0, \qquad a_3(g_3) = \frac{1}{3}g_3 \\ b_0(g_3,g_4) = b_2(g_3,g_4) = 0, \\ b_4(g_3,g_4) = \frac{1}{4}g_4 - \frac{1}{2}(g_3^2+1), \qquad b_6(g_3,g_4) = \frac{1}{18}g_3^2, \end{gathered} \tag{28}$$

while Schennach's Bayesian exponentially tilted empirical likelihood [18] belongs to (iii) with $\mu = \frac{1}{8}$. From (25) and Table 1, it is clear that each likelihood in the subclass (iii) admits, with margin of error $o(n^{-1/2})$, a data-dependent probability matching prior of the form (2) for posterior quantiles. Since each of these likelihoods has $a_1(g_3) = 0$, by (26), such a prior is given by

$$\pi(\theta) = \exp(-\frac{1}{2}\theta m_2^{-1/2} g_3). \tag{29}$$

TABLE 1

Forms of the $a_i(\cdot)$ and $b_i(\cdot)$ for the subclasses (i)-(iii)

Subclass	$a_1(g_3)$	$a_3(g_3)$	$b_0(g_3, g_4)$	$b_2(g_3, g_4)$	$b_4(g_3, g_4)$	$b_6(g_3, g_4)$
(i)	0	$\tau_3 g_3$	0	0	$\tau_4 g_4 - \frac{9}{2}\tau_3^2(g_3^2+1)$	$\frac{1}{2}\tau_3^2 g_3^2$
(ii)	0	$\gamma_3 g_3$	0	0	$\gamma_4 g_4 - \frac{9}{2}\gamma_3^2 g_3^2 - 3\gamma_3 + \frac{1}{2}$	$\frac{1}{2}\gamma_3^2 g_3^2$
(iii)	0	$\frac{1}{3}g_3$	0	0	$\mu g_4 - (\mu + \frac{1}{4})(g_3^2+1)$	$\frac{1}{18}g_3^2$

TABLE 2

Simulation results on the frequentist coverage of $(-\infty, \theta_1^{(1-\alpha)}(\pi, X)]$ for generalized empirical exponential family likelihoods, with $\pi(\theta)$ as in (29).

		Sample size					Sample size			
Distribution	$1-\alpha$	8	12	16	20	$1-\alpha$	8	12	16	20
Normal(0,1)	.95	.912	.928	.933	.938	.10	.138	.123	.119	.114
	.90	.863	.877	.884	.886	.05	.088	.074	.069	.064
Uniform(0,1)	.95	.934	.944	.946	.949	.10	.112	.106	.102	.102
	.90	.887	.896	.897	.898	.05	.067	.056	.051	.051
Beta(1,2)	.95	.910	.928	.936	.938	.10	.110	.108	.106	.104
	.90	.861	.880	.888	.890	.05	.061	.055	.055	.054
Exponential(1)	.95	.850	.878	.898	.906	.10	.111	.113	.111	.111
	.90	.798	.827	.845	.854	.05	.063	.064	.063	.061
Rayleigh(1)	.95	.900	.918	.928	.931	.10	.119	.113	.111	.108
	.90	.849	.868	.878	.880	.05	.070	.064	.060	.056

From Table 1, it is also clear that none of the likelihoods in the subclasses (i)-(iii) meets the first condition in (27) because $a_1(\cdot) = b_2(\cdot) = 0$ for any such likelihood. Thus, even with possibly data-dependent priors of the form (2), none of them allows frequentist validity of posterior quantiles with margin of error $o(n^{-1})$. In Section 5, it will be seen that at least for the usual empirical likelihood this difficulty can be resolved by considering more elaborate data-dependent priors.

Before addressing this issue, we present some simulation results to indicate the finite sample implications of the aforesaid probability matching property of the prior (29) for the subclass (iii) of generalized empirical exponential family likelihoods. Since this matching holds with margin of error $o(n^{-1/2})$, it makes sense to study the simulated coverage of the interval $(-\infty, \theta_1^{(1-\alpha)}(\pi, X)]$ in this context, where $\theta_1^{(1-\alpha)}(\pi, X)$ approximates the $(1-\alpha)$th posterior quantile of θ with coverage error $o_p(n^{-1/2})$; see (8). For any likelihood in (iii), it can be seen from (6), (7), (8) and Table 1 that $\theta_1^{(1-\alpha)}(\pi, X) = \bar{X} + (m_2/n)^{1/2}\{z + \frac{1}{6}n^{-1/2}g_3(2z^2+1)\}$, under (29). The simulation results, each based on 10000 simulations are presented in Table 2. Five distributions for the population along with four choices of $1-\alpha$, namely $1-\alpha$=0.95, 0.90, 0.10 and 0.05, are considered. In all cases, except for the exponential distribution in the right tail, the convergence to the desired frequentist coverage turns out to be reasonably fast. Thus the asymptotic results, even with margin of error $o(n^{-1/2})$, are well-reflected in finite samples.

5. More Elaborate Data-Dependent Priors

Observe that the prior in (26) is equivalent to

$$\pi(\theta) = \exp[-(\theta - \bar{X})m_2^{-1/2}\{a_1(g_3) + \frac{1}{2}g_3\}], \tag{30}$$

in the sense that they both lead to the same posterior; cf. (3) and (4). This motivates us to consider possibly data-dependent priors of the form

$$\pi(\theta) = \exp\{(\theta - \bar{X})m_2^{-1/2}\chi(g_3) + \frac{1}{2}(\theta - \bar{X})^2 m_2^{-1}\lambda(g_3, g_4)\}, \tag{31}$$

where $\chi(\cdot)$ and $\lambda(\cdot)$ are smooth functions. Since (31) can possibly involve $\bar{X}$ and g_4 in addition to m_2 and g_3, it is more elaborate than (2). The introduction of $\lambda(\cdot)$ in (31) aims at taking care of the population kurtosis.

With reference to any empirical-type likelihood in the general class (1), if one considers the posterior quantiles of θ under (31), then algebra similar to but heavier than that in Sections 2 and 3 reveals the following:

(a) Frequentist validity of the posterior quantiles holds with margin of error $o(n^{-1/2})$ if and only if

$$a_3(\beta_3) = \frac{1}{3}\beta_3 \text{ and } \chi(\beta_3) = -\{a_1(\beta_3) + \frac{1}{2}\beta_3\}. \tag{32}$$

(b) Frequentist validity of the posterior quantiles holds with margin of error $o(n^{-1})$ if and only if in addition

$$\begin{aligned} b_4(\beta_3, \beta_4) &= \frac{1}{3}\beta_3 a_1(\beta_3) - \frac{1}{2}\beta_3^2 + \frac{1}{4}\beta_4 - \frac{1}{2}, \\ b_6(\beta_3, \beta_4) &= \frac{1}{18}\beta_3^2, \\ \lambda(\beta_3, \beta_4) &= \{a_1(\beta_3)\}^2 - 2b_2(\beta_3, \beta_4) + \frac{5}{4}\beta_3^2 - \frac{2}{3}\beta_4 + 2. \end{aligned} \tag{33}$$

A comparison between (25), (27) and (32), (33) shows that the last condition in (33) is new and this helps. The first condition in (32) as well as the first two conditions in (33) are on the empirical-type likelihood. From Table 1, it can be seen that any likelihood in the subclasses (i)-(iii) meets these three conditions if and only if the associated $a_i(\cdot)$ and $b_i(\cdot)$ are given by (28), which corresponds to the usual empirical likelihood. Furthermore, if (28) holds then the last conditions in (32) and (33) yield $\chi(\beta_3) = -\frac{1}{2}\beta_3$ and $\lambda(\beta_3, \beta_4) = \frac{5}{4}\beta_3^2 - \frac{2}{3}\beta_4 + 2$, so that by (31), the data-dependent prior

$$\pi(\theta) = \exp\{-\frac{1}{2}(\theta - \bar{X})m_2^{-1/2}g_3 + \frac{1}{2}(\theta - \bar{X})^2 m_2^{-1}(\frac{5}{4}g_3^2 - \frac{2}{3}g_4 + 2)\}, \tag{34}$$

ensures frequentist validity of the posterior quantiles with margin of error $o(n^{-1})$.

A comparison between the prior just obtained and the one shown in (29) is in order. The one in (29) leads to probability matching to a lower order of accuracy but at the same time enjoys the merit of being much simpler. Moreover, as the simulation results reveal, it performs quite well in finite samples. Thus a choice between the two is essentially a matter of taste. If one wishes to work with a simple prior then the one in (29) is recommended. On the other hand, if a premium is put on higher order accuracy from asymptotic considerations, then the one obtained in this section appears to be more attractive.

The connection with a result in [7] is worth noting at this stage. They worked with data-free priors and showed that a likelihood in the general class (1) admits a probability matching prior, with margin of error $o(n^{-1})$, for posterior quantiles if and only if

$$a_1(\beta_3) = -\frac{1}{2}\beta_3, \qquad a_3(\beta_3) = \frac{1}{3}\beta_3, \qquad b_2(\beta_3, \beta_4) = \frac{3}{4}\beta_3^2 - \frac{1}{3}\beta_4 + 1,$$
$$b_4(\beta_3, \beta_4) = \frac{1}{4}\beta_4 - \frac{2}{3}\beta_3^2 - \frac{1}{2}, \qquad b_6(\beta_3, \beta_4) = \frac{1}{18}\beta_3^2.$$

With the $a_i(\cdot)$ and $b_i(\cdot)$ as above, the conditions in (32) and (33) are met if and only if $\chi(\beta_3) = \lambda(\beta_3, \beta_4) = 0$. In conjunction with (31), one thus gets the flat prior and this agrees with the findings in [7].

6. Concluding Remarks

The results in Subsection 3.2 show that if $a_3(\beta_3) = \frac{1}{3}\beta_3$ then the prior in (26) ensures frequentist validity of posterior quantiles with margin of error $o(n^{-1/2})$. It is satisfying to note that under the present assumption on the existence of the first four population moments, this margin of error is actually $O(n^{-1})$; vide (22). Similarly, if the existence of the first five population moments is assumed, then under the prior in (34), the frequentist validity of posterior quantiles arising from the usual empirical likelihood holds actually with margin of error $O(n^{-3/2})$.

By (13) and (14), the data-dependent probability matching prior in (34) satisfies $\pi(\theta) = 1 + o_p(1)$. The same holds for the prior in (30) which is equivalent to the probability matching prior in (26) in the sense of yielding the same posterior. Thus these data-dependent priors approximate the flat prior which is a natural data-free prior in the present context; cf. [7] and a well-known result for fully parametric location models ([6], Ch. 2).

There is scope for extending the present results in several directions. For example, if instead of posterior quantiles, interest lies in the highest posterior density regions, then the findings in [7] show that none of the standard empirical-type likelihoods proposed in the literature, including the usual empirical likelihood, admits a data-free probability matching prior in a higher order asymptotic sense. A natural question is whether consideration of data-dependent priors, such as those of the form (31), can yield more positive results. Another important issue concerns the role of data-dependent priors in the multivariate case with vector θ. Then the algebra will be rather complicated because the empirical-type likelihoods as well as the data-dependent priors will involve multivariate pure and mixed moments in place of m_2, g_3 and g_4, and one would need to consider multivariate Edgeworth expansions for the frequentist calculations. It is hoped that the present work will generate interest in these directions.

Acknowledgments

Thanks are due to a referee and the editors for their very constructive suggestions. This work was supported by a grant from the Center for Management and Development Studies, Indian Institute of Management Calcutta.

References

[1] Baggerly, K. A. (1998). Empirical likelihood as a goodness-of-fit measure. *Biometrika* **85** 535–547.

[2] Bravo, F. (2003). Second-order power comparisons for a class of nonparametric likelihood-based tests. *Biometrika* **90** 881–890.

[3] Clarke, B. (2007). Information optimality and Bayesian modeling. *J. Econometrics* **138** 405–429.

[4] Clarke, B. and Yuan, A. (2004). Partial information reference priors: derivation and interpretations. *J. Statist. Plann. Inf.* **123** 313–345.

[5] Corcoran, S. A. (1998). Bartlett adjustment of empirical discrepancy statistics. *Biometrika* **85** 967–972.

[6] Datta, G. S. and Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics.* Springer-Verlag, Berlin.

[7] Fang, K. T. and Mukerjee, R. (2006). Empirical-type likelihoods allowing posterior credible sets with frequentist validity: higher-order asymptotics. *Biometrika* **93** 723–733.

[8] Freedman, D. and Purves, R. A. (1969). Bayes method for bookies. *Ann. Math. Statist.* **40** 1177–1186.

[9] Ghosh, J. K. and Mukerjee, R. (1992). Non-informative priors (with discussion). In *Bayesian Statistics* **4** (eds. J.M. Bernardo *et al.*), Oxford University Press, 195–210.

[10] Lazar, N. A. (2003). Bayesian empirical likelihood. *Biometrika* **90** 319–326.

[11] Mittelhammer, R., Judge, G. and Miller, D. (2000). *Econometric Foundations.* Cambridge University Press, London.

[12] Mukerjee, R. and Reid, N. (1999). On a property of probability matching priors: matching the alternative coverage probabilities. *Biometrika* **86** 333–340.

[13] Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* **72** 219–255.

[14] Owen, A. B. (2001). *Empirical Likelihood.* Chapman and Hall, London.

[15] Raftery, A. (1996). Hypothesis testing and model selection via posterior simulation. In *Markov Chain Monte Carlo in Practice* (eds. W. R. Gilks *et al.*), Chapman and Hall, London, 163–188.

[16] Reid, N., Mukerjee, R. and Fraser, D. A. S. (2003). Some aspects of matching priors. In *Mathematical Statistics and Applications: Festschrift for Constance van Eeden* (eds. M. Moore *et al.*), IMS Lecture Notes - Monograph Series **42** 31–43.

[17] Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Roy. Statist. Soc. B* **59** 731–792.

[18] Schennach, S. M. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika* **92** 31–46.

[19] Sweeting, T. J. (2001). Coverage probability bias, objective Bayes and the likelihood principle. *Biometrika* **88** 657–675.

[20] Sweeting, T. J. (2005). On the implementation of local probability matching priors for interest parameters. *Biometrika* **92** 47–57.

[21] Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *J. Roy. Statist. Soc. B* **62** 159–180.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 71

Probability Matching Priors for Some Parameters of the Bivariate Normal Distribution

Malay Ghosh[1], Upasana Santra[1] and Dalho Kim[2]

University of Florida and Kyungpook National University

Abstract: This paper develops some objective priors for certain parameters of the bivariate normal distribution. The parameters considered are the regression coefficient, the generalized variance, and the ratio of the conditional variance of one variable given the other to the marginal variance of the other variable. The criterion used is the asymptotic matching of coverage probabilities of Bayesian credible intervals with the corresponding frequentist coverage probabilities. The paper uses various matching criteria, namely, quantile matching, matching of distribution functions, highest posterior density matching, and matching via inversion of test statistics. One particular prior is found which meets all the matching criteria individually for all the parameters of interest.

Contents

1. Introduction

Bayesian methods are becoming increasingly popular in the theory and practice of statistics. It is needless to say that other than the likelihood, the key component in any Bayesian analysis is the selection of priors. With sufficient information from past experience, expert opinion or previously collected data, subjective priors are ideal, and indeed should be used for inferential purposes. However, often even without adequate prior information, one can use Bayesian techniques efficiently with some 'default' or 'objective' priors. Not too surprisingly, the catalog of such priors, over

[1] Department of Statistics
University of Florida, Gainesville, Florida, USA
[2] Department of Statistics
Kyungpook National University, Taegu, Korea

AMS 2000 subject classifications:
Keywords and phrases: Credible intervals, distribution functions, first order, generalized variance, highest posterior density, likelihood ratio, posteriors, propriety, quantiles, regression coefficient, second order

the years, has also become prohibitively large, and one needs to set some specific 'objectivity' criterion for the solution of such priors.

One such criterion which has found some appeal to both frequentists and Bayesians is the so-called 'probability matching' criterion. Simply put, this amounts to the requirement that the coverage probability of a Bayesian credible region is asymptotically equivalent to the coverage probability of the frequentist confidence region up to a certain order. An excellent monograph on this topic is due to Datta and Mukerjee [5] which provides a thorough and comprehensive discussion of various probability matching criteria. Other review papers include [6], [7] and [9].

Again, as one might expect, there are several probability matching criteria. The matching is accomplished through either (a) posterior quantiles, (b) distribution functions, (c) highest posterior density (HPD) regions, or (d) inversion of certain test statistics. However, priors based on (a), (b), (c), or (d) need not always be identical. Specifically, it may so happen that there does not exist any prior satisfying all four criteria.

In this article, we consider the bivariate normal distribution where the parameters of interest are the (i) regression coefficient, (ii) the generalized variance, i.e. the determinant of the variance-covariance matrix, and (iii) the ratio of the conditional variance of one variable given the other divided by the marginal variance of the other variable. We have been able to find a prior which meets all four matching criteria for every one of these parameters.

As is well known, matching priors are obtained by solving certain differential equations. In Section 2 of this paper, we have introduced an orthogonal reparameterization ([3], [8]) of the bivariate normal parameters which greatly simplifies these equations, resulting thereby in easily accessible solutions. Here, we have also introduced a quantile matching prior which works for all the parameters given in (i)-(iii). Section 3 establishes the distribution matching property of the prior found in Section 2, once again for all three parameters of interest. Section 4 establishes the HPD matching property of such priors, while Section 5 confirms matching by inversion of the likelihood ratio statistic. The propriety of the posteriors and some numerical computations are given in Section 6, while some final remarks are made in Section 7.

2. The Orthogonal Reparameterization

Let $(X_{1i}, X_{2i}), i = 1, \dots, n$, be independent and identically distributed random variables having a bivariate normal distribution with means μ_1 and μ_2, variances ${\sigma_1}^2(>0)$ and ${\sigma_2}^2(>0)$, and correlation coefficient $\rho(|\rho|<1)$. We use the transformation (see, e.g., [1], [4] or [10])

$$(2.1)\qquad \beta = \rho\sigma_2/\sigma_1,\ \theta = \sigma_1\sigma_2(1-\rho^2)^{1/2} \text{ and } \eta = \sigma_2(1-\rho^2)^{1/2}/\sigma_1.$$

With this reparameterization, the bivariate normal distribution can be rewritten as

$$(2.2)\qquad f(X_1, X_2) = (2\pi\theta)^{-1}\exp\left\{-\frac{1}{2}\left\{\frac{\left(X_2-\mu_2-\beta\left(X_1-\mu_1\right)\right)^2}{\theta\eta} + \frac{\eta(X_1-\mu_1)^2}{\theta}\right\}\right\}$$

where β is the regression coefficient of X_2 on X_1, $\theta^2 = {\sigma_1}^2{\sigma_2}^2(1-\rho^2)$ is the determinant of the variance-covariance matrix, and $\eta^2 = V(X_2|X_1)/V(X_1)$.

With the above reparameterization, the Fisher Information matrix reduces to

$$\mathbf{I}(\mu_1,\mu_2,\beta,\theta,\eta) = \begin{pmatrix} \boldsymbol{A} & \mathbf{0} \\ \mathbf{0} & \mathrm{Diag}(\eta^{-2},\theta^{-2},\eta^{-2}) \end{pmatrix}, \tag{2.3}$$

where

$$\mathbf{A} = \begin{pmatrix} \frac{\beta^2}{\theta\eta}+\frac{\eta}{\theta} & -\frac{\beta}{\theta\eta} \\ -\frac{\beta}{\theta\eta} & \frac{1}{\theta\eta} \end{pmatrix}.$$

This establishes immediately the mutual orthogonality of (μ_1,μ_2), β, θ and η in the sense of [3] and [8]. Such orthogonality is often referred to as 'Fisher Orthogonality'.

Since the parameters of interest are orthogonal to (μ_1,μ_2), and it is customary to use a uniform $(\Re^2)$ prior on (μ_1,μ_2), we shall consider only priors of the form

$$\pi_0(\mu_1,\mu_2,\beta,\theta,\eta) \propto \pi(\beta,\theta,\eta), \tag{2.4}$$

and find π such that the matching criteria given in (a)-(d) are all satisfied for β, θ and η, each individually. This we are going to explore in the next three sections.

Here we also state a lemma which is used repeatedly in the sequel. The proof is based on the independence of $X_2-\mu_2-\beta(X_1-\mu_1)$ and $X_1-\mu_1$ along with the fact that $X_2-\mu_2-\beta(X_1-\mu_1)\sim N(0,\theta\eta)$ and $X_1-\mu_1\sim N(0,\theta/\eta)$.

Lemma 2.1. *For the bivariate normal density given in (2.2),*

$$E(\partial\log f/\partial\beta)^3=0,\ E[(\partial\log f/\partial\beta)(\partial^2\log f/\partial\beta^2)]=0; \tag{2.5}$$

(2.6)
$$E(\partial^3\log f/\partial\beta^3)=0,\ E(\partial^3\log f/\partial\beta^2\partial\theta)=(\theta\eta^2)^{-1},\ E(\partial^3\log f/\partial\beta^2\partial\eta)=\eta^{-3};$$

$$E(\partial^3\log f/\partial\beta\partial\theta^2)=0, E(\partial^3\log f/\partial\beta\partial\eta^2)=0; \tag{2.7}$$

$$E(\partial\log f/\partial\theta)^3=2/\theta^3,\ E[(\partial\log f/\partial\theta)(\partial^2\log f/\partial\theta^2)]=-2/\theta^3; \tag{2.8}$$

(2.9)
$$E(\partial^3\log f/\partial\theta^3)=4/\theta^3,\ E(\partial^3\log f/\partial\theta^2\partial\eta)=0,\ E(\partial^3\log f/\partial\theta\partial\eta^2)=(\theta\eta^2)^{-1};$$

(2.10)
$$E(\partial\log f/\partial\eta)^3=0,\ E[(\partial\log f/\partial\eta)(\partial^2\log f/\partial\eta^2)]=-\eta^{-3}, E(\partial^3\log f/\partial\eta^3)=3\eta^{-3}.$$

We consider in this section quantile matching priors, i.e., priors which ensure approximate frequentist validity of one-sided Bayesian credible intervals based on posterior quantiles of a one-dimensional interest parameter. The pioneering research in this area is due to Welch and Peers [14] and Peers [11], while the recent stimulus comes from Stein [12] and Tibshirani [13]. Specifically, one considers here priors $\pi(\cdot)$ for which the relation

$$P_\theta\{\theta_1 \le {\theta_1}^{1-\alpha}(\pi,X)\} = 1-\alpha+o(n^{-\frac{r}{2}}) \tag{2.11}$$

holds for $r=1$ or 2 and for each $\alpha\,(0<\alpha<1)$. In the above, $\theta=(\theta_1,\ldots,\theta_p)^T$ is the unknown parameter, θ_1 is the one-dimensional parameter of interest, ${\theta_1}^{1-\alpha}(\pi,X)$

is the $(1-\alpha)$th posterior quantile of θ_1 based on the prior π and data $X = (X_1, \ldots, X_n)^T$, while $P(\cdot|\theta)$ denotes the conditional probability given θ, the usual frequentist probability. A prior satisfying (2.11) with $r = 1$ is called a *first order* probability matching prior, while one with $r = 2$ is called a *second order* probability matching prior. Clearly, second order probability matching priors constitute a subclass of first order probability matching priors.

A second order quantile matching prior which works for β, θ and η is given by $\pi(\mu_1, \mu_2, \beta, \theta, \eta) \propto (\theta\eta)^{-1}$. Back to the original parameterization, $((\mu_1, \mu_2, \sigma_1, \sigma_2, \rho))$, this reduces to the prior

$$\pi(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = {\sigma_1}^{-2}(1-\rho^2)^{-1}. \tag{2.12}$$

This prior has been identified in [2] as the right-Haar prior as well as the one-at-a-time reference prior for any arbitrary ordering of these parameters. Indeed, this prior provides exact rather than just asymptotic matching for a variety of parameters of interest including the ones considered here. Moreover, as shown in this paper, when β is the parameter of interest, any prior of the form ${\sigma_1}^{-(3-a)}(1-\rho^2)^{-1}$, $a > 0$, for $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ achieves exact matching, while when θ is the parameter of interest, both the priors ${\sigma_1}^{-2}(1-\rho^2)^{-1}$ and ${\sigma_1}^{-1}{\sigma_2}^{-1}(1-\rho^2)^{-3/2}$ achieve exact matching.

However, Berger and Sun [1] have not explored other matching criteria. While the quantile matching property is quite desirable for construction of one-sided credible intervals, the HPD matching or matching via inversion of test statistics seems more appropriate for the construction of two-sided credible intervals. It is also true that priors which satisfy a particular matching criterion, for example, the quantile matching criterion, satisfies other matching criteria. As mentioned in the introduction, we will explore various other matching criteria in the subsequent sections. In this way, we will lend further justification to one of the matching priors of [2] - the one that we have developed here, as well.

3. Matching Via Distribution Functions

In this section, we target priors $\boldsymbol{\pi}$ which achieve matching via distribution functions of some standardized variables. More specifically, when θ_1 is the parameter of interest, $(\theta_2, \ldots, \theta_p)^T$ is the vector of nuisance parameters, $\hat{\theta_1}$ is the MLE of θ_1 with $n^{-\frac{1}{2}}I^{11}$ $\left(I = ((I_{jj'})), I^{-1} = ((I^{jj'}))\right)$ as its asymptotic variance, we consider the random variable $y = \sqrt{n}(\theta_1 - \hat{\theta_1})/(I^{11})^{1/2}$. Specifically, if P^π denotes the posterior of y given the data X, what we want to achieve is the asymptotic matching

$$E[P^\pi(y \le w|X)|\theta] = P(y \le w|\theta) + o(n^{-1}). \tag{3.1}$$

Under orthogonality of θ_1 with $(\theta_2, \ldots, \theta_p)$, it follows from (3.2.5) to (3.2.7) of [5] that such a prior π is of the form $I_{11}^{1/2} g(\theta_2, \ldots, \theta_p)$, where in addition one needs to satisfy the two differential equations

(3.2)

$$\begin{aligned} A_1 =& \frac{\partial^2}{\partial{\theta_1}^2}\left(I^{11}\pi(\theta)\right) - 2\frac{\partial}{\partial\theta_1}(I^{11}\pi(\theta)) - \sum_{s=2}^{p}\sum_{v=2}^{p}\frac{\partial}{\partial\theta_s}\left\{E\left(\frac{\partial^3 \log f}{\partial{\theta_1}^2\partial\theta_s}\right)I^{11}I^{sv}\pi(\theta)\right\} \\ &- \sum_{s=2}^{p}\sum_{v=2}^{p}\frac{\partial}{\partial\theta_1}\left\{E\left(\frac{\partial^3 \log f}{\partial\theta_1\partial\theta_s\partial\theta_v}\right)I^{11}I^{sv}\pi(\theta)\right\} = 0 \end{aligned}$$

and

$$A_2 = \sum_{s=2}^{p}\sum_{v=2}^{p} \frac{\partial}{\partial \theta_s}\Big\{E\big(\frac{\partial^3 \log f}{\partial {\theta_1}^2 \partial \theta_s}\big) I^{11} I^{sv} \pi(\theta)\Big\} = 0. \tag{3.3}$$

In the given problem, when β is the parameter of interest, (3.3) reduces to

$$\frac{\partial}{\partial \beta}\Bigg(\eta^2\Big\{\theta^2 E\big(\frac{\partial^3 \log f}{\partial \beta^2 \partial \theta}\big) + \eta^2 E\big(\frac{\partial^3 \log f}{\partial \beta^2 \partial \eta}\big)\Big\}\pi(\beta,\theta,\eta)\Bigg) = 0. \tag{3.4}$$

By (2.6) of Lemma 2.1, (3.4) simplifies to $\partial/\partial\beta\{(\theta+\eta)\pi(\beta,\theta,\eta)\} = 0$ which holds trivially for any prior $\pi(\beta,\theta,\eta)$ which does not depend on β, including the prior $\pi(\beta,\theta,\eta) \propto (\theta\eta)^{-1}$, the one found in Section 2. Again, with β as the parameter of interest, for any prior $\pi(\beta,\theta,\eta)$ which does not depend on β, (3.2) simplifies to $\partial/\partial\theta\{(\theta\eta^2)^{-1}\eta^2\theta^2\pi(\beta,\theta,\eta)\} + \partial/\partial\eta\{\eta^{-3}\eta^2\eta^2\pi(\beta,\theta,\eta)\} = 0$, i.e.,

$$\frac{\partial}{\partial \theta}[\theta\, \pi(\beta,\theta,\eta)] + \frac{\partial}{\partial \eta}[\eta\, \pi(\beta,\theta,\eta)] = 0.$$

Once again $\pi(\beta,\theta,\eta) \propto (\theta\eta)^{-1}$ will satisfy (3.3). To verify that a prior $\pi(\beta,\theta,\eta)$ which does not depend on β satisfies (3.2) with $\theta_1 = \beta$, we need to verify that

$$\frac{\partial}{\partial \beta}\Bigg(\eta^2\Big\{\mathrm{E}\big(\frac{\partial^3 \log f}{\partial \beta \partial \theta^2}\big)\theta^2 + \mathrm{E}\big(\frac{\partial^3 \log f}{\partial \beta \partial \eta^2}\big)\eta^2\Big\}\Bigg) = 0, \tag{3.5}$$

which reduces to

$$\eta^2\theta^2 \frac{\partial}{\partial \beta}\mathrm{E}\big(\frac{\partial^3 \log f}{\partial \beta \partial \theta^2}\big) + \eta^4 \frac{\partial}{\partial \beta}\mathrm{E}\big(\frac{\partial^3 \log f}{\partial \beta \partial \eta^2}\big) = 0.$$

From (2.7) of Lemma 2.1, $\mathrm{E}\big(\partial^3 \log f/\partial\beta\partial\theta^2\big) = \mathrm{E}\big(\partial^3 \log f/\partial\beta\partial\eta^2\big) = 0$, so that (3.2) holds trivially for such a prior. Hence matching via distributions is achieved with any prior of the form $\pi(\mu_1,\mu_2,\beta,\theta,\eta) \propto h(\theta,\eta)$, and in particular $h(\theta,\eta) \propto (\theta\eta)^{-1}$

Next, when θ is the parameter of interest, to find a prior satisfying (3.1) one needs to first solve

$$\begin{aligned} &\frac{\partial^2}{\partial \theta^2}\Bigg(\theta^2\pi(\cdot)\Bigg) - 2\frac{\partial}{\partial \theta}\Bigg(\theta^2 \frac{\partial \pi(\cdot)}{\partial \theta}\Bigg) - \frac{\partial}{\partial \theta}\Bigg(\theta^4 \mathrm{E}\big(\frac{\partial^3 \log f}{\partial \theta^3}\big)\pi(\cdot)\Bigg) \\ &\qquad - \frac{\partial}{\partial \theta}\Bigg(\theta^4 \mathrm{E}\big(\frac{\partial^3 \log f}{\partial \theta^3}\big)\pi(\cdot)\Bigg) = 0. \end{aligned} \tag{3.6}$$

Hence, (3.6) simplifies to

$$\frac{\partial^2}{\partial \theta^2}\Bigg(\theta^2\pi(\cdot)\Bigg) - 2\frac{\partial}{\partial \theta}\Bigg(\theta^2 \frac{\partial \pi(\cdot)}{\partial \theta}\Bigg) - 12\frac{\partial}{\partial \theta}\Bigg(\theta\pi(\cdot)\Bigg) = 0.$$

Any prior $\pi(\cdot) \propto \theta^{-1}g(\beta,\eta)$ will satisfy this equation. Such a prior also satisfies

$$\frac{\partial}{\partial \theta}\big(\theta^4 E\big(\frac{\partial^3 \log f}{\partial \theta^3}\big)\pi(\cdot)\big) = 0. \tag{3.7}$$

Finally when η is the parameter of interest, again, for finding a prior satisfying (3.1), one needs to solve

$$\begin{aligned}\frac{\partial^2}{\partial\eta^2}\Big\{\eta^2\pi(\cdot)\Big\} - 2\frac{\partial}{\partial\eta}\Big\{\eta^2\frac{\partial\pi(\cdot)}{\partial\eta}\Big\} - \frac{\partial}{\partial\theta}\Big\{E(\frac{\partial^3\log f}{\partial\theta\partial\eta^2})\eta^2\theta^2\pi\Big\} \\ - \frac{\partial}{\partial\eta}\Big\{E(\frac{\partial^3\log f}{\partial\eta\partial\beta^2})\eta^2\eta^2\pi\Big\} = 0\end{aligned} \tag{3.8}$$

and

$$\frac{\partial}{\partial\eta}\big(\eta^4E(\frac{\partial^3\log f}{\partial\eta^3})\pi(\cdot)\big) = 0. \tag{3.9}$$

Again, by Lemma 2.1, the prior $\pi(\beta,\theta,\eta) = \theta^{-1}\eta^{-1}$ satisfies the matching property.

4. Highest Posterior Density (HPD) Matching Priors

We now turn our attention to HPD matching priors for each one of the parameters β, θ and η. In general, if $\tilde{\theta}$ is the parameter (real or vector-valued) of interest, then a HPD region is of the form $\{\tilde{\theta} : \pi(\tilde{\theta}|X) \geq k\}$, where $\pi(\tilde{\theta}|X)$ is the posterior of $\tilde{\theta}$ under the prior π and data X. We will consider priors which ensure that HPD regions with credibility level $1-\alpha$ also have asymptotically the same frequentist coverage probability, the error of approximation being $o(n^{-1})$.

We first consider a HPD region for β. In view of the orthogonality result derived in Section 2, such a prior $\pi_0(\mu_1,\mu_2,\beta,\theta,\eta) \propto \pi(\beta,\theta,\eta)$ must satisfy (see [5], (4.4.1))

$$\begin{aligned}\frac{\partial}{\partial\theta}\left(\eta^2\theta^2\mathrm{E}(\frac{\partial^3\log f}{\partial\beta^2\partial\theta})\pi\right) + \frac{\partial}{\partial\eta}\left(\eta^4\mathrm{E}(\frac{\partial^3\log f}{\partial\beta^2\partial\eta})\pi\right) \\ + \frac{\partial}{\partial\beta}\left(\eta^4\mathrm{E}(\frac{\partial^3\log f}{\partial\beta^3})\pi\right) - \frac{\partial^2}{\partial\beta^2}\left(\eta^2\pi\right) = 0.\end{aligned} \tag{4.1}$$

Again, by Lemma 2.1, (4.1) reduces to

$$\frac{\partial}{\partial\theta}(\theta\pi(\cdot)) + \frac{\partial}{\partial\eta}(\eta\pi(\cdot)) - \frac{\partial^2}{\partial\beta^2}(\eta^2\pi(\cdot)) = 0. \tag{4.2}$$

Clearly the prior $\pi(\beta,\theta,\eta) \propto (\theta\eta)^{-1}$ satisfies (4.2).

Next consider θ as the parameter of interest. Now one needs to solve

$$\begin{aligned}\frac{\partial}{\partial\beta}\left(\eta^2\theta^2\mathrm{E}(\frac{\partial^3\log f}{\partial\theta^2\partial\beta})\pi\right) + \frac{\partial}{\partial\eta}\left(\theta^2\eta^2\mathrm{E}(\frac{\partial^3\log f}{\partial\theta^2\partial\eta})\pi\right) \\ + \frac{\partial}{\partial\theta}\left(\theta^4\mathrm{E}(\frac{\partial^3\log f}{\partial\theta^3})\pi\right) - \frac{\partial^2}{\partial\theta^2}\left(\theta^2\pi\right) = 0.\end{aligned} \tag{4.3}$$

By (2.8) and (2.9) of Lemma 2.1, (4.3) simplifies to

$$-2\frac{\partial}{\partial\theta}(\theta\pi) - \frac{\partial^2}{\partial\theta^2}(\theta\pi) = 0, \tag{4.4}$$

which is satisfied by the prior $\pi(\beta,\theta,\eta) \propto (\theta\eta)^{-1}$.

Finally, when η is the parameter of interest, we need to solve

$$\begin{aligned}\frac{\partial}{\partial\beta}\left(\eta^4\mathrm{E}(\frac{\partial^3\log f}{\partial\eta^2\partial\beta})\pi\right)+\frac{\partial}{\partial\theta}\left(\eta^2\theta^2\mathrm{E}(\frac{\partial^3\log f}{\partial\eta^2\partial\theta})\pi\right)\\ +\frac{\partial}{\partial\eta}\left(\eta^4\mathrm{E}(\frac{\partial^3 logf}{\partial\eta^3})\pi\right)-\frac{\partial^2}{\partial\eta^2}\left(\eta^2\pi\right)=0.\end{aligned}\tag{4.5}$$

From Lemma 2.1, (4.5) reduces to

$$\frac{\partial}{\partial\theta}(\theta\pi)+\frac{\partial}{\partial\eta}(\eta\pi)-\frac{\partial^2}{\partial\eta^2}(\eta^2\pi)=0.\tag{4.6}$$

Again $\pi(\beta,\theta,\eta)\propto(\theta\eta)^{-1}$ will do.

5. Matching Priors Via Inversion of Test Statistics

One traditional way to derive frequentist confidence intervals is inversion of certain test statistics. The most popular such test is the likelihood ratio test. But tests based on Rao's score statistic or the Wald statistic are also of importance, and are first order equivalent (i.e., up to $o(n^{-1/2})$) to the likelihood ratio tests. We consider here only the likelihood ratio test.

We begin with the general case when θ_1 is the parameter of interest, while $\theta_2,\ldots,\theta_p$ are the nuisance parameters. Let $\theta=(\theta_1,\ldots,\theta_p)$, and let $l(\theta)$ denote the usual log-likelihood. The corresponding profile log-likelihood for θ_1 is given by $l^*(\theta_1)=l(\theta_1,\hat{\theta_2}(\theta_1),\ldots,\hat{\theta_p}(\theta_1))$, where $\hat{\theta_j}(\theta_1)$ is the MLE of θ_j given $\theta_1(j=2,\ldots,p)$. Then the likelihood ratio statistic for θ_1 is given by

$$M_{LR}{}^*(\theta_1,X)=2[l(\hat{\theta})-l^*(\theta_1)].\tag{5.1}$$

Then from Yin and Ghosh [15] (also from [5], (5.2.18)), a likelihood ratio matching prior π is obtained by solving

(5.2)

$$\begin{aligned}&\sum_{s=2}^{p}\sum_{u=2}^{p}\frac{\partial}{\partial u}\left\{I_{11}^{-1}I^{su}E(\frac{\partial^3\log f}{\partial\theta_1^2\partial\theta_s})\right\}+\\ &\frac{\partial}{\partial\theta_1}\left(I_{11}^{-1}\left\{\frac{\partial\pi}{\partial\theta_1}-\pi(\theta)(I_{11}^{-1}E(\frac{\partial\log f}{\partial\theta_1}\frac{\partial^2\log f}{\partial\theta_1^2})-\sum_{s=2}^{p}\sum_{u=2}^{p}I^{su}E(\frac{\partial^3\log f}{\partial\theta_1\partial\theta_u\partial\theta_s}))\right\}\right)=0.\end{aligned}$$

In the present case when β is the parameter of interest, (5.2) reduces to

(5.3)

$$\begin{aligned}&\frac{\partial}{\partial\theta}\left(\eta^2\theta^2(\theta\eta^2)^{-1}\pi(\theta)\right)+\frac{\partial}{\partial\eta}\left(\eta^4\eta^{-3}\pi(\theta)\right)+\frac{\partial}{\partial\beta}\left(\eta^2\left\{\frac{\partial\pi}{\partial\beta}\right.\right.\\ &\left.\left.-\pi\left\{\eta^2E((\frac{\partial\log f}{\partial\beta})(\frac{\partial^2\log f}{\partial\beta^2}))-\theta^2E(\frac{\partial^3\log f}{\partial\beta\partial\theta^2})-\eta^2E(\frac{\partial^3\log f}{\partial\beta\partial\theta^2})\right\}\right\}\right)=0.\end{aligned}$$

From Lemma 2.1, (5.3) reduces to

$$\frac{\partial}{\partial\theta}(\theta\pi)+\frac{\partial}{\partial\eta}(\eta\pi)+\eta^2\frac{\partial}{\partial\beta}(\frac{\partial\pi}{\partial\beta})=0,$$

i.e.,

$$\frac{\partial}{\partial\theta}(\theta\pi) + \frac{\partial}{\partial\eta}(\eta\pi) + \eta^2\frac{\partial^2\pi}{\partial\beta^2} = 0.$$

Again $\pi \propto (\theta\eta)^{-1}$ provides a solution.

Next, if θ is the parameter of interest, the LR matching prior π for θ is obtained by solving the differential equation

(5.4)

$$\frac{\partial}{\partial\beta}\left\{\eta^2\theta^2.0.\pi(\theta)\right\} + \frac{\partial}{\partial\eta}\left\{\eta^2\theta^2.0.\pi(\theta)\right\} + \frac{\partial}{\partial\theta}\left(\theta^2\left\{\frac{\partial\pi}{\partial\theta}\right.\right.$$
$$\left.\left. - \pi\left\{\theta^2 E\big((\frac{\partial\log f}{\partial\theta})(\frac{\partial^2\log f}{\partial\theta^2})\big) - \eta^2\big(E(\frac{\partial^3\log f}{\partial\beta^2\partial\theta}) + E(\frac{\partial^3\log f}{\partial\eta^2\partial\theta})\big)\right\}\right\}\right) = 0.$$

Again from Lemma 2.1, (5.4) reduces to

$$\frac{\partial}{\partial\theta}[\theta^2\{\frac{\partial\pi}{\partial\theta} + \frac{2}{\theta}\pi + \pi\frac{2}{\theta}\}] = 0,$$

i.e.,

$$\frac{\partial}{\partial\theta}[\theta^2\frac{\partial\pi}{\partial\theta} + 4\theta\pi] = 0$$

which holds for $\pi \propto (\theta\eta)^{-1}$.

Finally, when η is the parameter of interest, the LR matching prior is obtained by solving

$$\frac{\partial}{\partial\beta}\left\{\pi\{\eta^4\eta^{-3}.0\}\right\} + \frac{\partial}{\partial\theta}\left\{\eta^2\theta^2\frac{1}{\theta\eta^2}\pi\right\}$$
$$(5.5) \qquad + \frac{\partial}{\partial\eta}\left(\eta^2\left\{\frac{\partial\pi}{\partial\eta} - \pi\left\{\eta^2 E\big((\frac{\partial\log f}{\partial\eta})(\frac{\partial^2\log f}{\partial\eta^2})\big) - \eta^2\frac{\theta/\eta}{\theta\eta^2} - 0\right\}\right\}\right) = 0.$$

Once again, using Lemma 2.1, (5.5) reduces to

$$\frac{\partial}{\partial\theta}(\theta\pi) + \frac{\partial}{\partial\eta}[\eta^2(\frac{\partial\pi}{\partial\eta}) - \pi(\eta^2(-\frac{2}{\eta})] = 0,$$

and the prior $\pi \propto (\theta\eta)^{-1}$ provides a solution.

6. Posteriors and Numerically Computed Coverage

The prior $\pi(\mu_1, \mu_2, \beta, \theta, \eta) \propto (\theta\eta)^{-1}$ is improper. In this section, we write down the marginal posteriors for β, θ and η, and discuss methods for finding the HPD intervals for each one of these parameters. The analytical findings of Section 4 are strengthened with some numerical coverage probability computations.

The marginal posterior of β is given by

$$\pi(\beta|\mathbf{X}_1, \mathbf{X}_2) \propto \int_o^\infty \eta^{n-2}\left(\eta^2 + \frac{S_{22} + \beta^2 S_{11} - 2\beta S_{12}}{S_{11}}\right)^{-(n-1)} d\eta.$$

Putting $\eta = z[S_{22} + \beta^2 S_{11} - 2\beta S_{12}/S_{11}]^{-1/2}$ in the above integral, one gets after simplification,

$$\pi(\beta|\mathbf{X}_1, \mathbf{X}_2) \propto \left(1 + \frac{(\beta - S_{12}/S_{11})^2}{S_{22.1}}\right)^{-\frac{n-1}{2}}, \tag{6.1}$$

where $S_{22.1} = S_{22} - {S_{12}}^2/S_{11}$. This posterior is a t-distribution with location parameter S_{12}/S_{11}, scale parameter $\{S_{22.1}/(n-2)\}^{1/2}$ and degrees of freedom $n-2$.

The posterior of θ is given by

$$\pi(\theta|\mathbf{X}_1, \mathbf{X}_2) \propto \theta^{-(n-1)} \exp(-{S_{11}}^{1/2}{S_{22.1}}^{1/2}/\theta) I_{[\theta>0]}, \tag{6.2}$$

so that θ^{-1} has a Gamma distribution with shape parameter $n-2$ and scale parameter $(S_{11}S_{22.1})^{-\frac{1}{2}}$.

Finally, the marginal posterior of η is given by

$$\begin{aligned}\pi(\eta|\mathbf{X}_1, \mathbf{X}_2) \propto & \eta^{-1/2}(\frac{S_{22.1}}{\eta} + \eta S_{11})^{-(n-3/2)} \\ & \propto \eta^{n-2}(\eta^2 + \frac{S_{22.1}}{S_{11}})^{-(n-3/2)}.\end{aligned} \tag{6.3}$$

The construction of HPD credible intervals is fairly simple. The posterior of β being a univariate-t (thus symmetric and unimodal), from (6.1), the $100(1-\alpha)\,\%$ HPD credible interval for β is given by $S_{12}/S_{11} \pm \{S_{22.1}/(n-2)\}^{1/2} t_{n-2;\alpha/2}$, where $t_{n-2;\alpha/2}$ denotes the upper $100(\alpha/2)\%$ point of a Student's t-distribution with $n-2$ degrees of freedom.

Observing that the posterior of θ is log-concave, the $100(1-\alpha)\,\%$ region for θ is given by $[\theta_1, \theta_2]$, where θ_1 and θ_2 satisfy

$${\theta_1}^{-(n-1)} \exp(-{S_{11}}^{1/2}{S_{22.1}}^{1/2}/\theta_1) = {\theta_2}^{-(n-1)} \exp(-{S_{11}}^{1/2}{S_{22.1}}^{1/2}/\theta_2) \tag{6.4}$$

and

$$\int_{\theta_1}^{\theta_2} \theta^{-(n-1)} \exp(-{S_{11}}^{1/2}{S_{22.1}}^{1/2}/\theta)(S_{11}S_{22.1})^{\frac{n-2}{2}} d\theta = 1 - \alpha. \tag{6.5}$$

It is important to note that if $w = \theta^{-1}$, then the posterior pdf of w is given by

$$\pi(w|\mathbf{X_1}, \mathbf{X_2}) \propto w^{n-3} \exp(-wS_{11}^{1/2}S_{22.1}^{1/2}).$$

Noting the log-concavity of this pdf as well, the HPD region $[w_1, w_2]$ for w is obtained by solving

$${w_1}^{n-3} \exp(-w_1{S_{11}}^{1/2}{S_{22.1}}^{1/2}) = {w_2}^{n-3} \exp(-w_2{S_{11}}^{1/2}{S_{22.1}}^{1/2}) \tag{6.6}$$

and

$$\int_{w_1}^{w_2} \frac{w^{n-3}}{\Gamma(n-2)} \exp(-w{S_{11}}^{1/2}{S_{22.1}}^{1/2})(S_{11}S_{22.1})^{\frac{n-2}{2}} \, dw = 1 - \alpha. \tag{6.7}$$

Clearly the solution $[w_1, w_2]$ of (6.6) and (6.7) is different from the solution $[\theta_2^{-1}, \theta_1^{-1}]$ of (6.4) and (6.5).

Table 1

Frequentist coverage probabilities of 95% HPD intervals for β, θ and η when $\sigma_1^2 = 1$ and $\sigma_2^2 = 1$

ρ	n	β	θ	η
0.25	4	0.952	0.947	0.949
	8	0.946	0.955	0.950
	12	0.954	0.952	0.948
	16	0.952	0.954	0.950
	20	0.945	0.948	0.950
0.50	4	0.950	0.952	0.949
	8	0.944	0.952	0.948
	12	0.954	0.953	0.944
	16	0.946	0.950	0.949
	20	0.952	0.948	0.949
0.75	4	0.955	0.952	0.953
	8	0.953	0.948	0.949
	12	0.950	0.946	0.947
	16	0.948	0.946	0.951
	20	0.956	0.946	0.951

Finally observing that the posterior of η in (6.3) is log-concave, the $100(1-\alpha)\,\%$ HPD interval $[\eta_1, \eta_2]$ for η is obtained by solving

$$\eta_1^{n-2}(\eta_1^2 + \frac{S_{22.1}}{S_{11}})^{-(n-3/2)} = \eta_2^{n-2}(\eta_2^2 + \frac{S_{22.1}}{S_{11}})^{-(n-3/2)},$$

where

$$c\int_{\eta_1}^{\eta_2} \eta^{n-2}(\eta^2 + \frac{S_{22.1}}{S_{11}})^{-(n-3/2)}\,d\eta = 1-\alpha,$$

c being the normalizing constant.

Now we evaluate the frequentist coverage probability by investigating the HPD credible interval of the marginal posterior densities of β, θ and η under our probability matching prior for several ρ and n. That is to say, the frequentist coverage of a $100(1-\alpha)\%$ HPD interval should be close to $1-\alpha$. This is done numerically. The results were fairly insensitive to the choice of σ_1 and σ_2. Table 1 gives numerical values of the frequentist coverage probabilites of 95% HPD intervals for β, θ and η for $\sigma_1 = \sigma_2 = 1$.

The computation of these numerical values is based on simulation. In particular, for fixed $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$ and n, we take $5,000$ independent random samples of $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ from the bivariate normal model. In our simulation study, we take $\mu_1 = \mu_2 = 0$ without loss of generality. Under the prior π, the frequentist coverage probability can be estimated by the relative frequency of HPD intervals containing the true parameter value. An inspection of Table 1 reveals that the agreement between the frequentist and posterior coverage probabilities of HPD intervals is quite good for the probability matching prior even if n is small.

7. Summary

The paper considers several probability matching criteria, and develops a prior that meets all the matching criteria individually for several parameters of the bivariate normal distribution including the regression coefficient and the generalized variance. Future work will address development of matching priors when the parameter of interest is the correlation coefficient. Possible multivariate extensions will also be considered.

Acknowledgements

The paper has benefitted much from the comments of Bertrand Clarke and a reviewer. Ghosh's research was partially supported by the NSF Grant SES-0631426.

References

[1] Berger, J. and Sun, D. (2006). Objective priors for a bivariate normal model with multivariate generalizations. ISDS Technical Report, Duke University.

[2] Berger, J. and Sun, D. (2007). Objective priors for the bivariate normal model. To appear in the *Annals of Statistics*.

[3] Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with Discussion). *J. R. Statist. Soc. B***53**,79-109.

[4] Datta, G. S. and Ghosh, J. K. (1995). On priors providing frequentist validity of Bayesian inference. *Biometrika*, **82**, 37-45.

[5] Datta, G. S. and Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics.* Lecture notes in Statistics. Springer, New York.

[6] Datta, G. S. and Sweeting, T. J. (2005). *Probability matching priors.* Handbook of Statistics, Vol 25: Bayesian Thinking: Modeling and Computation, Eds.: D. Dey and C.R. Rao, pp. 91–114. Elsevier.

[7] Ghosh, M. and Mukerjee, R. (1998). Recent Developments on probability matching priors. In: S. E. Ahmed, M. Ahsanullah and B. K. Sinha, eds., *Applied Statistical Science, III,* pp. 227-52, Nova Science Publishers, New York.

[8] Huzurbazar, V. S. (1950). Probability distributions and orthogonal parameters. *Proc. Camb. Phil. Soc* **46**, 281-284.

[9] Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Am. Statist. Assoc.* **91**, 1343-70.

[10] Mukerjee, R. and Ghosh, M. (1997). Second order probability matching priors. *Biometrika*, **84**, 970-975.

[11] Peers, H. W. (1965). Confidence properties of Bayesian interval estimates. *J. R. Statist. Soc. B* **30**, 535-44.

[12] Stein, C, (1985). On the coverage probability of confidence sets based on a prior distribution. In *Sequential Methods in Statistics*, Banach center publications, **16**. Warsaw: PWN-Polish scientific publishers.

[13] Tibshirani, R. J. (1989). Noninformative priors for one parameter of many. *Biometrika* **76**, 604-8.

[14] Welch, B. L. and Peers, H. W. (1963). On formulae for confidence points based on integrals of weighted likelihoods. *J. R. Statist. Soc. B* **35**, 318-29.

[15] Yin, M. and Ghosh, M. (1997). A note on the probability difference between matching priors based on posterior quantiles and on inversion of conditional likelihood ratio statistics. *Calcutta Statist. Assoc. Bull.* **47**, 59-65.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 82

Fuzzy Set Representation of a Prior Distribution

Glen Meeden

University of Minnesota

Abstract: In the subjective Bayesian approach uncertainty is described by a prior distribution chosen by the statistician. Fuzzy set theory is another way to representing uncertainty. Here we give a decision theoretic approach which allows a Bayesian to convert their prior distribution into a fuzzy set membership function. This yields a formal relationship between these two different methods of expressing uncertainty.

Contents

1. Introduction

For a subjective Bayesian uncertainty about the unknown parameter or state of nature can be expressed through a prior distribution. If θ denotes a typical parameter value and Θ the set of all possible parameter values then the prior distribution over Θ summarizes their knowledge and beliefs about the parameter.

Fuzzy set theory, introduced in Zadeh [7], is another approach to representing uncertainty. A fuzzy set A, a subset of Θ, is characterized by its membership function. This is a function defined on Θ whose range is contained in the unit interval. At a point θ the value of the membership function is a measure of how much we think θ belongs to the set A. Statisticians have been slow to embrace fuzzy set theory. Taheri [6] gives a review of applications of fuzzy set theory concepts to statistical methodology. Bayesians have shown less interest in fuzzy ideas than frequentists. Singpurwalla and Booker [5] have proposed a model which incorporates fuzzy membership functions into a subjective Bayesian setup. However, they do not give membership functions a probabilistic interpretation. In the imprecise or vague approach to Bayesian statistics a decision maker selects a family of possible prior distributions to represent their prior beliefs. de Cooman [2] presents an uncertainty model for vague probability assessments that is closely related to Zadeh's approach [7].

[1] School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street, SE, Minneapolis, MN 55455

AMS 2000 subject classifications: Primary 62F15; secondary 62C05
Keywords and phrases: prior distribution, fuzzy sets, Bayesian inference

The concept of confidence intervals is a frequentist approach to expressing uncertainty about an unknown parameter given data. It has long been recognized that naive users have difficulty interpreting confidence intervals. They have a tendency to give a probabilistic interpretation to the observed confidence interval.

It has also long been known that for discrete data conventional confidence intervals, which we will also call 'crisp' confidence intervals, using a term from fuzzy set theory, can preform poorly. A recent article [1] reviews the problems with crisp confidence intervals for binomial models. Because of the inherent flaws in crisp confidence intervals for discrete problems a new confidence interval notion has been suggested called fuzzy confidence intervals [3]. Given the data a fuzzy confidence interval is just the membership function of the set of plausible or reasonable values for θ. One way to think about such membership functions is that they are generalizations of randomized intervals where no randomization is ever implemented. They argued that fuzzy confidence intervals overcome the difficulties of the usual crisp intervals for discrete probability models.

In terms of frequency of coverage discrete data Bayesian credible intervals will suffer from the same problem that conventional intervals do. This should be of concern to objective Bayesians who want their intervals to have good frequentist properties. One way to approach this problem is to find a method that allows them to use their posterior to get a sensible fuzzy interval instead of the usual Bayesian credible interval.

Here we consider a no data statistical decision problem where the set of possible decisions is the class of all membership functions defined on Θ. We then define a family of loss functions. These functions measure the loss incurred when a probability distribution is replaced by a fuzzy membership function. For any loss function in the family and a given prior distribution we solve the resulting no data decision problem. This gives a method for converting a prior or posterior into a fuzzy membership function. For a given fuzzy membership function we also study the problem of identifying the family of prior distributions whose common solution to the no data decision problem is this function. This sets up a formal relationship between the two theories.

2. Fuzzy Set Theory

We will only use some of the basic concepts and terminology of fuzzy set theory, which can be found in the most elementary of introductions to the subject [4].

A *fuzzy set* A in a space Θ is characterized by its *membership function*, which is a map $I_A : \Theta \to [0, 1]$. The value $I_A(\theta)$ is the 'degree of membership' of the point θ in the fuzzy set A or the "degree of compatibility ... with the concept represented by the fuzzy set". See ([4], p. 75). The idea is that we are uncertain about whether θ is in or out of the set A. The value $I_A(\theta)$ represents how much we think θ is in the fuzzy set A. The closer $I_A(\theta)$ is to 1.0, the more we think θ is in A. The closer $I_A(\theta)$ is to 0.0, the more we think θ is not in A.

A fuzzy set whose membership function only takes on the values zero or one is called *crisp*. For a crisp set, the membership function I_A is the same thing as the indicator function of an ordinary set A. Thus 'crisp' is just the fuzzy set theory way of saying 'ordinary', and 'membership function' is the fuzzy set theory way of saying 'indicator function'. The *complement* of a fuzzy set A having membership function I_A is the fuzzy set B having membership function $I_B = 1 - I_A$.

If I_A is the membership function of a fuzzy set A, the γ-*cut* of A ([4], Section 5.1) is the crisp set

$$^{\gamma}I_A = \{\theta \,:\, I_A(\theta) \geq \gamma\}.$$

Clearly, knowing all the γ-cuts for $0 \leq \gamma \leq 1$ tells us everything there is to know about the fuzzy set A. The 1-cut is also called the *core* of A, denoted core(A) and the set

$$\text{supp}(A) = \bigcup_{\gamma>0} {}^{\gamma}I_A = \{\theta \,:\, I_A(\theta) > 0\}$$

is called the *support* of A ([4], p. 100). A fuzzy set is said to be *convex* if each γ-cut is convex ([4], pp. 104–105).

3. A Decision Problem

For simplicity we assume that Θ is an interval of real numbers and the prior π is a continuous probability density function defined on it.

Let $\mathcal{A}$ be the class of all measurable membership functions defined on Θ. Then $\mathcal{A}$ is the space of possible decisions or actions with a typical member denoted by A. Given a prior density π on Θ we want to find the membership function or fuzzy set A which best represents π. We do this by defining a loss function and then solving the no data decision problem.

Our loss function will depend on four known parameters which are specified by the statistician. They are $a_1 \geq 0$, $a_2 \geq 0$, $b_1 \geq 0$ and $b_2 \geq 0$ where at least one of the a_i's and at least one of the b_i's must be strictly positive. Then the loss incurred when action A is taken and θ is the true state of nature is given by

$$(1)\quad L(A,\theta) = a_1\{1 - I_A(\theta)\} + \frac{a_2}{2}\{1 - I_A(\theta)\}^2 + \int_{\Theta}\left\{b_1 I_A(\theta) + \frac{b_2}{2}(I_A(\theta))^2\right\} d\theta.$$

To understand this loss function remember that we want to find the fuzzy set or membership function A which best represents the set of sensible or reasonable parameter values under our prior π. Hence if θ is the true parameter point we want $I_A(\theta)$ to be close to 1. This explains the presence of the first two terms in equation 1. But on the other hand we do not want the fuzzy set to be too large. This is controlled by the last term in the equation which is a measure of the overall size of the fuzzy set.

We now find the solution for this no data decision problem.

Theorem 1. *Let $\pi(\theta)$ be a prior density on Θ. Then for the loss function of equation (1) the fuzzy set membership A which satisfies*

$$\int_{\Theta} L(A,\theta)\pi(\theta)\, d\theta = \inf_{A' \in \mathcal{A}} \int_{\Theta} L(A',\theta)\pi(\theta)\, d\theta$$

is given by

$$(2)\qquad I_A(\theta) = \begin{cases} 0 & \textit{for} \quad 0 \leq \pi(\theta) < b_1/(a_1+a_2) \\ \frac{(a_1+a_2)\pi(\theta)-b_1}{a_2\pi(\theta)+b_2} & \textit{for} \quad b_1/(a_1+a_2) \leq \pi(\theta) \leq (b_1+b_2)/a_1 \\ 1 & \textit{for} \quad \pi(\theta) > (b_1+b_2)/a_1. \end{cases}$$

Proof. Note that we can write

$$\int_{\Theta} L(A',\theta)\pi(\theta)\, d\theta = \int_{\Theta}\left\{a_1\{1 - I_A(\theta)\} + \frac{a_2}{2}\{1 - I_A(\theta)\}^2\right\}\pi(\theta) + b_1 I_A(\theta) + \frac{b_2}{2}(I_A(\theta))^2\right\} d\theta$$

so that to find the solution it is enough to minimize the integrand of the previous equation for each fixed value of θ. But for a fixed θ the integrand is just a quadratic function of $I_{A'}(\theta)$ and a simple calculus argument completes the proof. □

The theorem remains true when $a_1 = 0$ if we assume dividing by zero yields infinity.

Note that the solution is unchanged if the loss function is multiplied by a positive number. Without loss of generality we could set one of the four parameters defining the loss function equal to one but having four parameters will be convenient in the following discussion.

As with any decision problem the solution depends strongly on the loss function. We believe our family of loss functions is flexible and captures some of the important aspects of the problem. Finding a good fuzzy set to summarize our information about a parameter is much like finding a good credible set. We want it to include the likely values but without it getting to large. The loss function in equation (1) is essentially the sum of two quadratic functions. The first part is quadratic in non-membership in the set of likely values while the second part is quadratic in a measure of the size of the set. If we just include the linear terms in each part then the optimal solution will always be a crisp set. It is necessary to include the quadratic terms to get a true fuzzy set as a solution.

We see from equation (2) that the optimal membership function is related to the prior π in a sensible fashion. The solution is 1 where the prior is large, 0 where the prior is small and a rescaling between the two cases. Note that for a given bounded π if b_1 is chosen large enough then the solution to our decision problem is the membership function which is identically zero. On the other hand if π is bounded away from zero and a_1 is chosen large enough then the solution to our decision problem is the membership function which is identically one.

4. Relating Priors and Fuzzy Sets

We have considered the problem of converting a prior distribution into a fuzzy membership function. In some situations it could be of interest to be able to move in the other direction. That is, transform the uncertainty expressed in a fuzzy membership function into the Bayesian paradigm. One way to do this would be to find a loss function and prior for which the solution to our decision problem is the fuzzy membership function in hand. This suggests the following three questions.

- For a specified fuzzy membership function, I_A, and a specified loss function does there exist a prior density function for which the solution to our decision problem is I_A?
- For a specified fuzzy membership function, I_A, does there exist a loss function and a prior density function for which the solution to our decision problem is I_A?
- If a solution does exist for question 1 is it unique?

We see from equation (2) that for I_A to be a solution for π we must have

$$(3) \qquad \pi(\theta) = \frac{b_1 + b_2 I_A(\theta)}{a_1 + a_2(1 - I_A(\theta))} \qquad \text{for } \theta \text{ where} \qquad 0 < I_A(\theta) < 1$$

From this we see that the answer to our first question is no. This is because when Θ is unbounded π in the previous equation need not be integrable and even when it is it need not integrate to one. The answer to the second question is yes whenever $I_A(\theta)$ is integrable. Since in this case we can always select $b_1 \geq 0$ and $b_2 > 0$ to make $\pi(\theta)$ of equation (3) a density. When a solution exists it need not be unique.

For a simple example we set $a_2 = 0$ and let the other three parameters be positive. Consider the special case where Θ is bounded. If we set

$$r_1 = b_1/a_1 \quad \text{and} \quad r_2 = (b_1 + b_2)/a_1 \tag{4}$$

we find that

$$a_1 = b_2/(r_2 - r_1) \quad \text{and} \quad b_1 = r_1/(r_2 - r_1) \tag{5}$$

and the solution from equation (2) has the form

$$I_A(\theta) = \begin{cases} 0 & \text{for} \quad 0 \leq \pi(\theta) < r_1 \\ (\pi(\theta) - r_1)/(r_2 - r_1) & \text{for} \quad r_1 \leq \pi(\theta) \leq r_2 \\ 1 & \text{for} \quad \pi(\theta) > r_2. \end{cases} \tag{6}$$

Now let I_A be given and assume that the length of Θ is ℓ. If $r_1 < 1/\ell$ then there exist a unique $r_2 > r_1$ such that

$$\pi_{A,r_1}(\theta) = (r_2 - r_1)I_A(\theta) + r_1 \quad \text{for } \theta \in \Theta \tag{7}$$

is a prior distribution over Θ. Moreover we can find values for a_1, b_1 and b_2 which satisfy equation 4. With this loss function I_A will be the solution to our decision problem when the prior is π_{A,r_1}. Furthermore if the sets where $I_A(\theta) = 0$ and $I_A(\theta) = 1$ each have positive Lebesgue measure then it will not be the unique prior with this property. Any prior density π satisfying

$$\begin{aligned} &\pi(\theta) \leq r_1 \quad \text{when} \quad I_A(\theta) = 0 \\ &\pi(\theta) = \pi_{A,r_1}(\theta) \quad \text{when} \quad 0 < I_A(\theta) < 1 \\ &\pi(\theta) \geq r_2 \quad \text{when} \quad I_A(\theta) = 1 \end{aligned} \tag{8}$$

will also be a solution for our decision problem.

Among the set of possible solutions the one in equation (7) has two nice properties. First of all it is continuous whenever $I_A(\theta)$ is continuous. Secondly it treats the members of $\{\theta : I_A(\theta) = 1\}$ similarly and the members of $\{\theta : I_A(\theta) = 0\}$ similarly. More importantly, this identification of a fuzzy membership function with a class of prior distributions demonstrates that we can give roughly equivalent expressions of uncertainty in the Bayesian and fuzzy paradigms.

Finally, we address the question of uniqueness. The previous discussion indicates that if we want uniqueness we should consider membership functions which never take on zero or one as a possible value. Let I_A be such a membership function and let $a_1 > 0$ and $a_2 > 0$ be fixed and suppose Θ is the unit interval. Then integrating equation 3 we have

$$\begin{aligned} \int_0^1 \pi(\theta)\, d\theta &= b_1 \int_0^1 \frac{1}{a_1 + a_2(1 - I_A(\theta))}\, d\theta + b_2 \int_0^1 \frac{I_A(\theta)}{a_1 + a_2(1 - I_A(\theta))}\, d\theta \\ &= b_1 c_1(a_1, a_2) + b_2 c_2(a_1, a_2). \end{aligned}$$

Hence π will be a probability density function whenever

$$b_1 \in [0, 1/c_1(a_1,a_2)] \quad \text{and} \quad b_2 = (1 - b_1 c_1(a_1,a_2))/c_2(a_1,a_2).$$

To better understand the relationship between I_A and its corresponding prior we consider a simple example. Let

$$I_A(\theta) = 6.075\,\theta^2(1-\theta) \quad \text{for} \quad \theta \in [0,1]. \tag{9}$$

We consider two different choices of the a_i's and for each case two different choices of b_1. For the first case $a_1 = 1$ and $a_2 = 7$. The maximum possible value for b_1 is 3.40 and our two choices for the b_i's are $b_1 = 0.01$, $b_2 = 5.15$ and $b_1 = 3.35$, $b_2 = 0.072$. In the second case $a_1 = 4$ and $a_2 = 2$. The maximum possible value for b_1 is 4.91 and our two choices for the b_i's are $b_1 = 0.01$, $b_2 = 9.02$ and $b_1 = 4.50, b_2 = 0.76$. For each of the four combinations we found the unique prior whose solution to the decision problem yields the fuzzy membership function of equation (9). The membership function along with the four priors are shown in the figure.

The membership function is the solid curve. The two curves with the two largest maximums are the solutions for the first case where $a_1 = 1$ and $a_2 = 7$. Of the two solutions the one with $b_1 = 0.01$ has the largest maximum. The other two curves are the solutions for the second case. Again the solution for $b_1 = 0.01$ has the largest of the two maximums. These curves demonstrate what a closer inspection of equation (3) yields. For a fixed a_1 and a_2 the solution becomes less concentrated about its mode as b_1 increases from zero to its maximum value. Also the solution becomes less concentrated about its mode as we increase a_1 and decrease a_2. But in all cases the priors do reflect the shape of their common solution.

An interesting consequence of this unique correspondence is that it gives a way to update a large class of fuzzy membership functions given data. Suppose an expert has selected a fuzzy membership function to represent their uncertainty. The statistician then selects appropriate values of the a_i's and the b_i's and uses equation 3 to transform it into a prior. Then given the data they find the posterior distribution which is then converted back to a fuzzy membership function using the theorem with the a_i and b_i values.

This result is somewhat surprising since a fuzzy membership function must satisfy less conditions then a probability density function since it need not be integrable. At first glance the previous example where a membership function corresponded to a family of priors seems more reasonable. To get the unique correspondence, however, we made two fairly strict assumptions. The function in equation (3) needed to be integrable and the range of the membership function had to lay in the open unit interval. Both these conditions on the membership function seem not so surprising if we hope to convert it to a probability density function.

5. Some Final Remarks

Mainline statistics has shown little interest in fuzzy set theory. This is especially true for most Bayesians since they believe that they already have a good way to express uncertainty. Here we have argued that Bayesians should be more interested in fuzzy set theory. For discrete data, just as for frequentists, there are certain advantages to considering interval estimates as fuzzy sets. We noted that our scheme for converting a prior density into a fuzzy membership function could also be used to relate some fuzzy membership functions to prior densities. In some cases a fuzzy

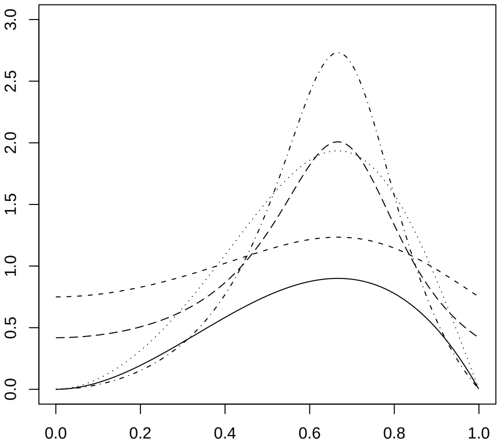


Fig 1. *A plot of the fuzzy membership function (the solid line) in equation 9 and four priors whose common solution for four loss functions is the fuzzy membership function. The two priors for the $a_1 = 1$ and $a_2 = 7$ case are the ones with the two largest maximums. The other two priors are for the two $a_1 = 4$ and $a_2 = 2$ cases.*

membership function will correspond to a family of densities while under more restricted conditions it will correspond to a unique density. The relationship seems intuitively sensible and as far as we know it is the first simple formal correspondence between the two theories which until now have lived in different worlds.

A copy of Geyer and Meeden [3] and related material can be found at

`http//:www.stat.umn.edu/~glen/papers/`

Acknowledgements

Research supported in part by NSF Grant DMS 0406169. The author would like to thank a referee for a helpful comment.

References

[1] Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion (with discusion). *Statistical Science 16*, 101–133.

[2] de Cooman, G. (2005). A behavioural model for vague probability models. *Fuzzy Sets and Systems 154*, 305–358.

[3] Geyer, C. and Meeden, G. (2005). Fuzzy and randomized confidence intervals and p-values (with discussion). *Statistical Science 20*, 358–387.

[4] Klir, G. J., St. Clair, U. H., and Yuan, B. (1997). *Fuzzy Set Theory: Foundations and Applications.* Prentice Hall PTR, Upper Saddle River, NJ.

[5] Singpurwalla, N. D. and Booker, J. M. (2004). Membership functions and probability measures of fuzzy sets (with discussion. *Journal of the American Statistical Association 99*, 867–889.

[6] Taheri, S. M. (2003). Trends in fuzzy sets. *Austrian Journal of Statistics 32*, 239–257.

[7] Zadeh, L. (1965). Fuzzy sets. *Information and Control 8*, 338–359.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 90
© Institute of Mathematical Statistics, 0000
arXiv: math.PR/0000000

Fuzzy Sets In Nonparametric Bayes Regression

Jean-François Angers[1] **and Mohan Delampady**[2]

Université de Montréal and Indian Statistical Institute, Bangalore

Abstract: A simple Bayesian approach to nonparametric regression is described using fuzzy sets and membership functions. Membership functions are interpreted as likelihood functions for the unknown regression function, so that with the help of a reference prior they can be transformed to prior density functions. The unknown regression function is decomposed into wavelets and a hierarchical Bayesian approach is employed for making inferences on the resulting wavelet coefficients.

Contents

1. Introduction

Consider the model

$$y_i = g(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \text{ and } x_i \in \mathcal{T}, \tag{1.1}$$

where $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)' \sim N(0, \sigma^2 I)$, σ^2 is unknown and $g(\cdot)$ is a function defined on some index set $\mathcal{T} \subset \mathcal{R}^1$. Inferences about g such as its estimation and estimation error as well as model checking are of interest.

Without parametric assumptions such as those leading to linear regression, this is a nonparametric regression problem. A Bayesian approach to (fully) nonparametric regression problems typically requires specifying prior distributions on function spaces which is rather difficult to handle. The extent of the complexity of this

[1] Département de mathématiques et de statistique, Université de Montréal, Montréal H3C 3J7, Canada e-mail: angers@dms.umontreal.ca

[2] Statistics and Mathematics Unit, Indian Statistical Institute, Bangalore 560 059, India e-mail: mohan@isibang.ac.in

AMS 2000 subject classifications: Primary 62G08; secondary 62A15, 62F15

Keywords and phrases: function estimation, hierarchical Bayes, membership function, model choice, wavelet

approach can be gauged from sources such as Ghosh and Ramamoorthi [9], Lenk [10], and so on. Furthermore, quantifying useful prior information such as 'g is close to (a specified function) g_0' is difficult probabilistically, whereas this seems quite straightforward if instead an appropriate metric on the concerned function space is used. This is where fuzzy sets or membership functions can be made use of. Before proceeding to this problem, let us recall the following details on fuzzy sets and membership functions.

Definition 1.1. *A fuzzy subset A of a space $\mathcal{G}$ (or just a fuzzy set A) is defined by a membership function $h_A : \mathcal{G} \longrightarrow [0, 1]$.*

The membership function, $h_A(g)$, is supposed to express the degree of compatibility of g with A. For example, if $\mathcal{G}$ is the real line and A is the set of points 'close to 0', then $h_A(0) = 1$ indicates that 0 is certainly included in A, but $h_A(.05) = .01$ says that .05 is not really 'close' to 0 in this context. Similarly, if $\mathcal{G}$ is a set of functions and $A \subset \mathcal{G}$ is a set of functions 'close' to a given function g_0, then $h_A(g_0) = 1$ indicates that g_0 is certainly included in A; however, if $h_A(g_1) = .01$ with $g_1(x) = 10g_0 + 100$ then g_1 is not really 'close' to g_0 in this case.

Note that even when $\mathcal{G} = \Theta$ is the parameter space, a membership function $h_A(\theta)$ is not a probability density or mass function defined on Θ, and hence cannot be used to obtain a prior distribution directly. Instead, as we have done in [4], we propose that a reasonable interpretation for a fuzzy subset A of Θ is that it is a likelihood function for θ given A. (See also [12, 13]). This interpretation seems to be able to answer some of the questions regarding how appropriate the concept of *fuzziness* is in modeling our perception of imprecision. French [7] discusses a few of these questions. First of all, likelihood is an accepted means for modeling imprecision. Another important question is how to define $h_{A\cap B}$ from h_A and h_B for incorporating h_A and h_B in Bayesian inference. If A and B are independent, then interpreting h_A and h_B as likelihood functions leads to the result that $h_{A\cap B} = h_A h_B$, for this purpose. Further, the qualitative ordering that underlies a membership function can also be investigated with this interpretation, in conjunction with a prior distribution, as we study later. See [4] for an application to hierarchical and robust Bayes inference.

This paper is closely related to Angers and Delampady [3, 4]. In the latter, we used fuzzy sets to help specify a prior density on a finite parameter space, whereas in the former a nonparametric function estimator using wavelet decomposition was presented. It is therefore natural to explore whether a combination of these two ideas can be fruitfully employed. Towards this end, we adopt the semi-parametric approach of using the wavelet decomposition of a regression function along with a remainder, thus effecting a drastic reduction of the dimension of the parameter space. Next we propose to incorporate the imprecise prior information of the kind 'the true regression function g is close to g_0' using membership functions. Wavelet decomposition of g_0 provides the wavelet coefficients θ_0 which we take as the prior mean of the wavelet coefficients θ corresponding to g. Precision of this prior mean estimate is unclear, so we treat the prior variance of these coefficients as a hyperparameter with a second stage (reference) prior. This approach is also tied to the question of Bayesian robustness. If a membership function stating 'we think g is close to g_0' is the only prior input that we intend to incorporate, how should we then proceed with the Bayesian inference related to g? We claim that the only natural approach is to study the robustness of the inferences resulting from the class of prior densities compatible with this membership function (see Section 5.1). This approach is further explained in the following sections.

This paper is organized as follows. In Sections 2 and 3, a brief summary of our previous work [3, 4] is presented. In Section 4, details on the posterior calculations are given. The main focus of this paper, namely, model checking using Bayes factors and fuzzy models, is discussed in Section 5. Simulations and practical examples are presented in the last section to illustrate this theme.

2. Nonparametric Regression and Wavelets

To proceed further, we assume that the regression function g and its prior guess g_0 are in $\mathcal{L}_2$ and impose a wavelet structure on both of them. To do this, we consider a compactly supported wavelet function $\psi \in \mathcal{C}^s$, the set of real-valued functions with continuous derivatives up to order s. Thus, $g(x)$ can be written as

$$\begin{aligned} g(x) &= \sum_{|k| \le K_0} \alpha_k \phi_k(x) + \sum_{j=0}^{\infty} \sum_{|k| \le K_j} \beta_{j,k} \psi_{j,k}(x) \\ &= g_J(x) + R_J(x), \end{aligned}$$

where

$$\begin{aligned} g_J(x) &= \sum_{|k| \le K_0} \alpha_k \phi_k(x) + \sum_{j=0}^{J} \sum_{|k| \le K_j} \beta_{j,k} \psi_{j,k}(x), \\ R_J(x) &= \sum_{j=J+1}^{\infty} \sum_{|k| \le K_j} \beta_{j,k} \psi_{j,k}(x), \\ \phi_k(x) &= \phi(x-k), \\ \psi_{j,k}(x) &= 2^{j/2}\psi(2^j x - k), \\ \alpha_k &= \int_{\mathcal{T}} \phi(x-k) g(x) dx, \text{ and} \\ \beta_{j,k} &= \int_{\mathcal{T}} 2^{j/2}\psi(2^j x - k) g(x) dx. \end{aligned} \tag{2.1}$$

Here K_j is such that $\phi_k(x)$ and $\psi_{j,k}(x)$ vanish on $\mathcal{T}$ whenever $|k| > K_j$, and ϕ is the scaling function ('father wavelet') corresponding to the 'mother wavelet' ψ. Such K_j's exist (and are finite) since the wavelet function that we have chosen has compact support. (See [3] or [8] for details.) Since the number of observations is finite, only a finite number of parameters can be estimated. Therefore, as suggested in [3], the resolution level J should be chosen such that the total number of unknown parameters which need to be estimated is no larger than n, the number of observations. Since the total number of α and β parameters is bounded by

$$l_X 2^{J+1} + J(l_\psi + 1) + (l_\phi + l_\psi + 2)$$

where l_X, l_ψ and l_ϕ represent the length of the support of $\mathcal{T}$, $\psi(\cdot)$ and $\phi(\cdot)$ respectively, we require this upper bound to be less than n. The optimal resolution level J, however, should be chosen using a Bayesian model selection technique, for example [3]. Consequently, we propose that the function $g_J(\cdot)$ be estimated from the data and the function $R_J(\cdot)$ be considered as a 'nuisance parameter' to be eliminated by integrating it out.

Further we assume that

$$g_0(x) = \sum_{|k| \le K_0} \alpha_{0,k} \phi_k(x) + \sum_{j=0}^{\infty} \sum_{|k| \le K_j} \beta_{0,j,k} \psi_{j,k}(x),$$

where

$$\alpha_{0,k} = \int_{\mathcal{T}} \phi(x-k) g_0(x) dx$$
$$\beta_{0,j,k} = \int_{\mathcal{T}} 2^{j/2} \psi(2^j x - k) g_0(x) dx.$$

Since g_0 is given, the wavelet coefficients can be assumed known.

From [1], it is known that $\beta_{j,k} = \mathrm{O}(2^{-js})$ where s denotes the degree of 'smoothness' of ψ, that is $\psi(\cdot) \in \mathcal{C}^s$ $(s > 1/2)$. Further, some of the *a priori* information wavelet coefficients can be translated into

$$(2.2) \qquad E[\alpha_k] = \alpha_{0,k}; \qquad E[\beta_{j,k}] = \beta_{0,j,k};$$
$$(2.3) \qquad Var(\alpha_k) = \tau^2; \qquad Var(\beta_{j,k}) = \tau^2/2^{2js};$$

(see [3] for details). Here, τ^2 is a first stage hyper-parameter, a suitable prior on which will be specified later. (It is also supposed that $E[\beta_{j,k}] = \beta_{0,j,k} = 0$ for $j \geq J+1$.) However, the prior probability distribution on the coefficients α_k and $\beta_{j,k}$ itself will be specified by a membership function to be introduced later.

Since there are only a finite number of $\beta_{j,k}$ that can be estimated, we cannot expect to have a very informative prior distribution on the remainder term, R_J. However, to be able to proceed with the Bayesian approach, we need to assume that $R_J(\cdot)$ is a stochastic process. For simplicity and computational ease, we consider a Gaussian process with zero mean (compatible with $E[\beta_{j,k}] = \beta_{0,j,k} = 0$ for $j \geq J+1$). Specifically, following [3], we now assume that $R_J(x)$ is a Gaussian process with mean function 0 and covariance kernel $\tau^2 Q(x,y)$ where

$$(2.4) \qquad Q(x,y) = \sum_{j=J+1}^{\infty} \frac{1}{2^{2js}} \sum_{|k| \leq K_j} \psi_{j,k}(x) \psi_{j,k}(y).$$

Note that a prior assumption such as $\beta_{j,k}$ being independent normal random quantities will naturally lead to this prior distribution as can be seen from the Karhunen–Loeve expansion [2]. We, however, make the stronger assumption that the remainder $R_J(\cdot)$ itself is a zero-mean Gaussian but with an unknown prior variance component τ^2. Our reason for this assumption is that, once the optimal resolution level J is chosen using a powerful mechanism such as a Bayes factor, as we suggest later, the wavelet coefficients at higher resolutions are not expected to have substantial influence on the wavelet smoother. Hence $R_J(\cdot)$ which is made up of these higher-order wavelet coefficients will also have negligible influence.

3. Prior Information and Membership Functions

We have explained in the previous section that we would like to make use of imprecise prior information such as 'g is close to g_0' by using a membership function which translates this into a measure of distance between the corresponding wavelet coefficients. Let us examine the implications of assuming that the available prior information is quantified in terms of a membership function

$$(3.1) \qquad h_A(g) = \xi(\rho(g, g_0)),$$

where ρ is a measure of distance. Due to the wavelet decomposition assumed on g as well as g_0 (see Section 2), a natural choice for ρ is the $\mathcal{L}_2$ distance given by

$$\rho^2(g, g_0) = ||g - g_0||^2 = \sum(\theta_{j,k} - \theta^0_{j,k})^2, \tag{3.2}$$

where $\theta_{j,k}$ and $\theta^0_{j,k}$ are, respectively, the wavelet coefficients of g and g_0. Using Parseval's identity, note that

$$\rho^2(g, g_0) = \|g - g_0\|^2 = \sum_{|k| \le K_0} (\alpha_k - \alpha_{0,k})^2 + \sum_{j=0}^{\infty} \sum_{|k| \le K_j} (\beta_{j,k} - \beta_{0,j,k})^2 .$$

Since we cannot estimate $\beta_{j,k}$ for $j = J + 1, ...,$ and because $\beta_{j,k} = \mathrm{O}(2^{-js})$ (see [1]) we then have

$$\begin{aligned}
\rho^2(g, g_0) &= \sum_{|k| \le K_0} (\alpha_k - \alpha_{0,k})^2 + \sum_{j=0}^{J} \sum_{|k| \le K_j} (\beta_{j,k} - \beta_{0,j,k})^2 \\
&\qquad + \sum_{j=J+1}^{\infty} \sum_{|k| \le K_j} (\beta_{j,k} - \beta_{0,j,k})^2 \\
&= \sum_{|k| \le K_0} (\alpha_k - \alpha_{0,k})^2 + \sum_{j=0}^{J} \sum_{|k| \le K_j} (\beta_{j,k} - \beta_{0,j,k})^2 + \sum_{j=J+1}^{\infty} \sum_{|k| \le K_j} \mathrm{O}(2^{-2js}) \\
&= \sum_{|k| \le K_0} (\alpha_k - \alpha_{0,k})^2 + \sum_{j=0}^{J} \sum_{|k| \le K_j} (\beta_{j,k} - \beta_{0,j,k})^2 + \mathrm{O}(2^{-2(J+1)s}) \\
&= \rho^2_J(g, g_0) + O(2^{-2(J+1)s}).
\end{aligned} \tag{3.3}$$

(3.4)

For this reason, to quantify the imprecise prior information, we will use a membership function that will depend only on $\rho^2_J(g, g_0)$. Some possibilities for h_A are the following:
(i) The Gaussian membership function given by

$$h_A(g) = \exp(-\rho^2_J(g, g_0)) = \exp(-\alpha||\theta - \theta^0||^2). \tag{3.5}$$

This membership function can be explained as follows. Suppose we have available some past data of the form

$$y^*_i = g(x^*_i) + \varepsilon_i, \quad i = 1, \ldots, n^*,$$

with ε_i denoting i.i.d. normal errors, and suppose g is estimated from this data by $\hat{g}$. Then the information in this data may be quantified using a membership function of the type

$$h_A(g) = \exp(-\alpha||g - \hat{g}||^2) = \exp(-c\sum(\theta_j - \hat{\theta}_j)^2).$$

g_0 may then be identified with $\hat{g}$. If we have multiple past data sets, we may then have available $h_{A_1}(g) = \exp(-\alpha_1||g - \hat{g}_1||^2)$, $h_{A_2}(g) = \exp(-\alpha_2||g - \hat{g}_2||^2)$, and so on, which may be combined into

$$h_A(g) = h_{A_1 \cap A_2}(g) = h_{A_1}(g)h_{A_2}(g) = \exp(-\left\{\alpha_1||g - \hat{g}_1||^2 + \alpha_2||g - \hat{g}_2||^2\right\}).$$

As an example one could consider fitting regression lines to two (or more) sets of past data with possibly different error variances and use the fitted regression lines along with the estimated variances for constructing the membership functions. This justifies to some extent our previous suggestion that membership functions quantify prior information in the sense of the likelihood. The constants α_1 and α_2 provide additional scope for assigning different weights to the two sources of information, which is another appealing feature of this approach.
(ii) The multivariate t membership function

$$(3.6) \quad h_A(g) = \left(1+\rho_J^2(g,g_0)\right)^{-(p+q)/2} = \left(1+(\theta-\theta^0)'V^{-1}(\theta-\theta^0)/q\right)^{-(p+q)/2},$$

where $q > 2$ is the degrees of freedom and p denotes the dimension of θ. This is a continuous scale mixture of Gaussian membership functions with the same g_0 for each of the membership functions. Since this vanishes more slowly than (3.5), one could expect better robustness with this.
(iii) The uniform function

$$h_A(g) = \begin{cases} 1 & \text{if } \rho_J(g,g_0) \le \delta; \\ 0 & \text{otherwise.} \end{cases}$$

This is an extreme case where g is restricted to a neighborhood of g_0.

In order to proceed with Bayesian inference on g, we need to convert the membership function into a prior density. This is done as in [4] with the aid of a reference prior density π_0. Thus we obtain the prior density

$$\pi(g) \quad \propto \quad h_A(g)\pi_0(g),$$

or, upon utilizing the wavelet decomposition for g, we have an equivalent prior density

$$\pi(\theta,\sigma^2) \quad \propto \quad h_A(\theta)\pi_0(\theta,\sigma^2).$$

4. Posterior Calculations

As in [3], let $Q_n = (Q_n)_{il}$, $1 \le i,l \le n$, where $(Q_n)_{il} = Q(x_i,x_l)$, which was introduced in (2.4). Note that

$$(Q_n)_{il} = \sum_{j\ge J+1}\sum_{|k|\le K_j} 2^{-2js}\psi_{jk}(x_i)\psi_{jk}(x_l).$$

Let $X = (\Phi', S')$ with the ith row of Φ' being $\{\phi_k(x_i)\}'_{|k|\le K_0}$ and the ith row of S' being $\{\psi_{jk}(x_i)\}'_{|k|\le K_j, 0\le j\le J}$. Then we have the model

$$(4.1) \qquad \mathbf{y}|\theta,\sigma^2,\tau^2 \sim N(X\theta, \sigma^2 I_n + \tau^2 Q_n).$$

Unless θ has a normal prior distribution or a hierarchical prior with a conditionally normal prior distribution, analytical simplifications in the computation of posterior quantities are not expected. For such cases, we have the joint posterior density of the wavelet coefficients θ and the error variances σ^2 and τ^2 given by the expression

$$\pi(\theta,\sigma^2,\tau^2|\mathbf{y}) \quad \propto \quad f(\mathbf{y}|\theta,\sigma^2,\tau^2)h_A(\theta)\pi_0(\theta,\sigma^2,\tau^2),$$

where f is the likelihood. From (4.1), f can be expressed as

$$f(\mathbf{y}|\theta,\sigma^2,\tau^2) \quad \propto \quad |\sigma^2 I_n + \tau^2 Q_n|^{-1/2} \exp(-\frac{1}{2}\{(\mathbf{y}-X\theta)'(\sigma^2 I_n + \tau^2 Q_n)^{-1}(\mathbf{y}-X\theta)\}).$$

Proceeding further, suppose π_0 of the form

$$\pi_0(\theta,\sigma^2,\tau^2) = \pi_1(\sigma^2,\tau^2), \tag{4.2}$$

which is constant in θ, is chosen.

MCMC based approaches to posterior computations are now readily available. For example, Gibbs sampling is straightforward. Note that the conditional posterior densities are given by

$$\pi(\theta|\mathbf{y},\sigma^2,\tau^2) \propto \exp(-\frac{1}{2}\{(\mathbf{y}-X\theta)'(\sigma^2 I_n + \tau^2 Q_n)^{-1}(\mathbf{y}-X\theta)\})h_A(\theta), \tag{4.3}$$

(4.4)
$$\begin{aligned}\pi(\sigma^2|\mathbf{y},\theta,\tau^2) &\propto |\sigma^2 I_n + \tau^2 Q_n|^{-1/2}\\ &\qquad \exp(-\frac{1}{2}\{(\mathbf{y}-X\theta)'(\sigma^2 I_n + \tau^2 Q_n)^{-1}(\mathbf{y}-X\theta)\})\pi_1(\sigma^2,\tau^2),\end{aligned}$$

(4.5)
$$\begin{aligned}\pi(\tau^2|\mathbf{y},\theta,\sigma^2) &\propto |\sigma^2 I_n + \tau^2 Q_n|^{-1/2}\\ &\qquad \exp(-\frac{1}{2}\{(\mathbf{y}-X\theta)'(\sigma^2 I_n + \tau^2 Q_n)^{-1}(\mathbf{y}-X\theta)\})\pi_1(\sigma^2,\tau^2).\end{aligned}$$

However, major simplifications are possible with the Gaussian h_A as in (i). In this case, the posterior analysis can proceed along the lines of [3]. Specifically, assuming that $h_A(\theta)$ is proportional to the density of $N(\theta_0,\tau^2\Gamma)$ with

$$\Gamma = \begin{pmatrix} I_{2K_0+1} & 0 \\ 0 & \Delta_{M_\beta} \end{pmatrix},$$

where $M_\beta = \sum_{j=0}^{J}(2K_j+1)$ and with $\tau^2\Delta$ being the variance-covariance matrix of β (which is also diagonal, having the diagonal entries specified by $Var(\beta_{jk}) = \tau^2/2^{2js}$), we obtain

$$\begin{aligned}\mathbf{y}|\theta,\sigma^2,\tau^2 &\sim N(X\theta,\sigma^2 I_n + \tau^2 Q_n),\\ \theta|\tau^2 &\sim N(\theta_0,\tau^2\Gamma).\end{aligned} \tag{4.6}$$

Therefore, it follows that

$$\mathbf{y}|\sigma^2,\tau^2 \sim N(X\theta_0,\sigma^2 I_n + \tau^2(X\Gamma X' + Q_n)), \tag{4.7}$$

$$\theta|\mathbf{y},\sigma^2,\tau^2 \sim N(\theta_0 + A(\mathbf{y}-X\theta_0), B), \tag{4.8}$$

where

$$\begin{aligned} A &= \tau^2\Gamma X'\left(\sigma^2 I_n + \tau^2(X\Gamma X' + Q_n)\right)^{-1},\\ B &= \tau^2\Gamma - \tau^4\Gamma X'\left(\sigma^2 I_n + \tau^2(X\Gamma X' + Q_n)\right)^{-1} X\Gamma.\end{aligned}$$

Now proceeding as in [3], we employ spectral decomposition to obtain $X\Gamma X' + Q_n = HDH'$, where $D = \text{diag}(d_1, d_2, \ldots, d_n)$ is the matrix of eigenvalues and H is the orthogonal matrix of eigenvectors. Thus,

$$\sigma^2 I_n + \tau^2(X\Gamma X' + Q_n) = H\left(\sigma^2 I_n + \tau^2 D\right)H' = \sigma^2 H\left(I_n + uD\right)H',$$

where $u = \tau^2/\sigma^2$. Then, the first stage (conditional) marginal density of $\mathbf{y}$ given σ^2 and u can be written as

$$\begin{aligned} m(\mathbf{y} \mid \sigma^2, u) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \frac{1}{\det(I_n + uD)^{1/2}} \\ &\quad \times \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - X\theta_0)' H (I_n + uD)^{-1} H' (\mathbf{y} - X\theta_0) \right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \frac{1}{\prod_{i=1}^n (1 + ud_i)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n \frac{s_i^2}{1 + ud_i} \right\}, \end{aligned} \tag{4.9}$$

where $\mathbf{s} = (s_1, \ldots, s_n)' = H'(\mathbf{y} - X\theta_0)$. We choose the prior on σ^2 and $u = \tau^2/\sigma^2$ qualitatively similar to that used in [3]. Specifically, we take $\pi_1(\sigma^2, u)$ to be proportional to the product of an inverse gamma density $\{k^{c-1}/\Gamma(c-1)\} \exp(-k/\sigma^2)(\sigma^2)^{-c}$ for σ^2 and the density of a $F(b, a)$ distribution for u (for suitable choice of k, c, a and b). Conditions apply on a and b as indicated in [3].

Once $\pi_1(\sigma^2, u)$ is chosen as above, we obtain the posterior mean and covariance matrix of θ as in the following result.

Theorem 4.1.

$$E(\theta|\mathbf{y}) = \theta_0 + \Gamma X' H E\left[(I_n + uD)^{-1} \mid \mathbf{y} \right] s, \tag{4.10}$$

where the expectation is taken with respect to

$$\pi_{22}(u \mid \mathbf{y}) \propto \frac{u^{b/2}}{(a + bu)^{(a+b)/2}} \left(\prod_{i=1}^n (1 + ud_i) \right)^{-1/2} \left(2k + \sum_{i=1}^n \frac{s_i^2}{1 + ud_i} \right)^{-(n+2c)/2}. \tag{4.11}$$

and

$$\begin{aligned} Var(\theta \mid \mathbf{y}) &= \frac{1}{n + 2c} E\left[2k + \sum_{i=1}^n \frac{s_i^2}{1 + ud_i} \mid \mathbf{y} \right] \Gamma \\ &\quad - \frac{1}{n + 2c} \Gamma X' H E\left[\left(2k + \sum_{i=1}^n \frac{s_i^2}{1 + ud_i} \right) (I_n + uD)^{-1} \mid \mathbf{y} \right] H' X \Gamma \\ &\quad + E\left[M(u) M(u)' \mid \mathbf{y} \right], \end{aligned} \tag{4.12}$$

where $M(u) = \Gamma X' H (I_n + uD)^{-1} s$.

The proof of Theorem 4.1 readily follows upon using standard hierarchical Bayesian model techniques (see [8], Section 9.1). The second part of Equation (4.10) can be viewed as a correction term added to the prior guess after observing the data $\mathbf{y}$.

Coming to the multivariate t form of h_A as in (ii), we note that the multivariate t density is a continuous scale mixture of multivariate normal densities as, for example, shown in [11], i.e., θ has the multivariate t distribution with location θ_0 and scale matrix V with density of the form (3.6) if and only if

$$\theta \mid \delta^2 \sim N(\theta_0, q\delta^2 V), \qquad (\delta^2)^{-1} \sim \chi_q^2.$$

This implies that with a π_0 as in (4.2), we obtain a hierarchical normal prior structure for θ with an additional hyper-parameter δ^2. Consequently, we can proceed with the posterior computations exactly as with the Gaussian h_A, except that we will have a two-dimensional integration after we simplify our calculations as above. However, MCMC techniques can easily handle this case.

Computations with the h_A given in (iii) are more difficult. Analytical simplifications as shown for the cases (i) and (ii) are not available here. Therefore, we utilize the MCMC computational scheme outlined in (4.3)-(4.5) above. Alternatively, the Metropolis–Hastings (M-H) algorithm may be employed.

5. Model Checking and Bayes Factors

An important and useful model checking problem in the present setup is checking the two models

$$M_0 : g = g_0 \qquad \text{versus} \qquad M_1 : g \neq g_0.$$

Under M_1, $(g = g(\theta), \tau^2, \sigma^2)$ is given the prior $h_A(\theta)\pi_0(\theta, \tau^2, \sigma^2)I(g \neq g_0)$, whereas under M_0, $\pi_0(\sigma^2)$ induced by $\pi_0(\theta, \tau^2, \sigma^2)$ is the only part needed. In order to conduct the model checking, we compute the Bayes factor, B_{01}, of M_0 relative to M_1:

$$B_{01}(\mathbf{y}) = \frac{m(\mathbf{y}|M_0)}{m(\mathbf{y}|M_1)}, \tag{5.1}$$

where $m(\mathbf{y}|M_i)$ is the predictive (marginal) density of $\mathbf{y}$ under model M_i, $i = 0, 1$. We have

$$m(\mathbf{y}|M_0) = \int f(\mathbf{y}|g_0, \sigma^2)\pi_0(\sigma^2)\, d\sigma^2,$$

and

$$m(\mathbf{y}|M_1) = \int f(\mathbf{y}|\theta, \tau^2, \sigma^2)h_A(\theta)\pi_0(\theta, \tau^2, \sigma^2)\, d\theta\, d\tau^2\, d\sigma^2.$$

Since improper priors can lead to difficulties in model checking problems, here we must employ proper priors. (Note that for estimation purposes, the noninformative, improper prior $(\sigma^2)^{-c}$ corresponding to $k = 0$ would have worked in the previous section.) We will develop the methodology here for the Gaussian membership function only; the other cases are similar but computationally more intensive.

Recall from the previous section that in the case of a Gaussian membership function h_A, the posterior analysis is similar to that discussed in [3], and for the same reason, the computation of the Bayes factor is also similar to what was discussed there. As in the previous section $\pi_0(\theta, \tau^2, \sigma^2)$ will be constant in θ, while σ^2 is inverse gamma and is independent of $v = \sigma^2/\tau^2$ which is given the $F_{a,b}$ prior distribution. (Equivalently, $u = 1/v = \tau^2/\sigma^2$ is given the $F_{b,a}$ prior as before.) Specifically, $\pi_0(\sigma^2) = \{k^{c-1}/\Gamma(c-1)\}\exp(-k/\sigma^2)(\sigma^2)^{-c}$, where c and k (small) are suitably chosen. Therefore,

$$m(\mathbf{y}|M_0) = \int f(\mathbf{y}|g_0, \sigma^2)\pi_0(\sigma^2)\, d\sigma^2 = (2\pi)^{-n/2}\frac{k^{c-1}}{\Gamma(c-1)}$$
$$\Gamma(n/2 + c - 1)\left\{k + \frac{1}{2}\sum_{i=1}^{n}(y_i - g_0(x_i))^2\right\}^{-(n/2+c-1)}.$$

Further, using (4.7), it follows that

$$m(\mathbf{y}|M_1, \sigma^2, u) = (2\pi\sigma^2)^{-n/2}\prod_{i=1}^{n}(1 + ud_i)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{s_i^2}{1 + ud_i}\right\}, \tag{5.2}$$

where $\mathbf{s} = (s_1, \ldots, s_n)' = H'(\mathbf{y} - X\theta_0)$ as before. Therefore,

$$\begin{aligned} m(\mathbf{y}|M_1) &= \int m(\mathbf{y}|M_1, \sigma^2, u)\pi_0(\sigma^2, u)\, d\sigma^2\, du \\ &= (2\pi)^{-n/2}\frac{k^{c-1}}{\Gamma(c-1)}\int \prod_{i=1}^{n}(1+ud_i)^{-1/2}\pi_0(u) \\ &\quad \left\{\int \exp\left\{-\frac{1}{\sigma^2}\left(k+\frac{1}{2}\sum_{i=1}^{n}\frac{s_i^2}{1+ud_i}\right)\right\}(\sigma^2)^{-(n/2+c)}\, d\sigma^2\right\} du \end{aligned}$$

$$\begin{aligned} (5.3) \qquad &= (2\pi)^{-n/2}\frac{k^{c-1}}{\Gamma(c-1)}\Gamma(n/2+c-1) \\ &\quad \int \left\{k+\frac{1}{2}\sum_{i=1}^{n}\frac{s_i^2}{1+ud_i}\right\}^{-(n/2+c-1)} \prod_{i=1}^{n}(1+ud_i)^{-1/2}\pi_0(u)\, du, \end{aligned}$$

where $\pi_0(u)$ denotes the $F_{b,a}$ density of u. Note that this involves only a straightforward single-dimensional integration. The resulting Bayes factor is illustrated later in our examples.

5.1. Prior Robustness of Bayes Factors

Note that the most informative part of the prior density that we have used is contained in the membership function h_A. Since a membership function $h_A(\theta)$ is to be treated only as a likelihood for θ, any constant multiple $ch_A(\theta)$ also contributes the same prior information about θ. Therefore, a study of the robustness of the Bayes factor that we obtained above with respect to a class of priors compatible with h_A is of interest. Here we consider a sensitivity study using the *density ratio class* defined as follows. Since the prior π that we use has the form $\pi(\theta, \tau^2, \sigma^2) \propto h_A(\theta)\pi_0(\theta, \tau^2, \sigma^2)$, we consider the class of priors

$$\mathcal{C}_A = \left\{\pi : c_1 h_A(\theta)\pi_0(\theta, \tau^2, \sigma^2) \leq \alpha\pi(\theta, \tau^2, \sigma^2) \leq c_2 h_A(\theta)\pi_0(\theta, \tau^2, \sigma^2), \alpha > 0\right\},$$

for specified $0 < c_1 < c_2$. We would like to investigate how the Bayes factor (5.1) behaves as the prior π varies in $\mathcal{C}_A$. We note that for any $\pi \in \mathcal{C}_A$, the Bayes factor B_{01} has the form

$$B_{01} = \frac{\int f(\mathbf{y}|g_0, \sigma^2)\pi(\theta, \tau^2, \sigma^2)\, d\theta\, d\tau^2\, d\sigma^2}{\int f(\mathbf{y}|\theta, \tau^2, \sigma^2)\pi(\theta, \tau^2, \sigma^2)\, d\theta\, d\tau^2\, d\sigma^2}.$$

Even though the integration in the numerator above need not involve θ and τ^2, we do so to apply the following result (see [8], Theorem 3.9, or [4], Theorem 4.1).

Consider the density-ratio class

$$\Gamma_{DR} = \{\pi : L(\eta) \leq \alpha\pi(\eta) \leq U(\eta) \text{ for some } \alpha > 0\},$$

for specified non-negative functions L and U. Further, let $q \equiv q^+ + q^-$ be the usual decomposition of q into its positive and negative parts, i.e., $q^+(u) = \max\{q(u), 0\}$ and $q^-(u) = -\max\{-q(u), 0\}$. Then we have the following theorem (see [6]).

Theorem 5.1. *For functions q_1 and q_2 such that $\int |q_i(\eta)|U(\eta)\,d\eta < \infty$, for $i = 1, 2$, and with q_2 positive a.s. with respect to all $\pi \in \Gamma_{DR}$,*

$$\inf_{\pi\in\Gamma_{DR}} \frac{\int q_1(\eta)\pi(\eta)\,d\eta}{\int q_2(\eta)\pi(\eta)\,d\eta} \text{ is the unique solution } \lambda \text{ of}$$

$$\int (q_1(\eta) - \lambda q_2(\eta))^- U(\eta)\,d\eta + \int (q_1(\eta) - \lambda q_2(\eta))^+ L(\eta)\,d\eta = 0, \tag{5.4}$$

$$\sup_{\pi\in\Gamma_{DR}} \frac{\int q_1(\eta)\pi(\eta)\,d\eta}{\int q_2(\eta)\pi(\eta)\,d\eta} \text{ is the unique solution } \lambda \text{ of}$$

$$\int (q_1(\eta) - \lambda q_2(\eta))^+ U(\eta)\,d\eta + \int (q_1(\eta) - \lambda q_2(\eta))^- L(\eta)\,d\eta = 0. \tag{5.5}$$

We shall discuss this result for the Gaussian membership function only. Then, since the prior π that we use has the form $\pi(\theta, \tau^2, \sigma^2) \propto h_A(\theta)\pi_0(\tau^2, \sigma^2)$, and we don't intend to vary $\pi_0(\tau^2, \sigma^2)$ in our analysis, we redefine $\mathcal{C}_A$ as

$$\mathcal{C}_A = \{\pi(\theta) : c_1 h_A(\theta) \le \alpha\pi(\theta) \le c_2 h_A(\theta), \alpha > 0\},$$

for specified $0 < c_1 < c_2$. Now, we re-express B_{01} as

$$B_{01}(\pi) = \frac{\int \left\{\int f(\mathbf{y}|g_0, \sigma^2)\pi_0(\sigma^2)d\sigma^2\right\} \pi(\theta)\,d\theta}{\int \left\{\int f(\mathbf{y}|\theta, \tau^2, \sigma^2)\pi_0(\tau^2, \sigma^2)\,d\tau^2\,d\sigma^2\right\} \pi(\theta)\,d\theta} = \frac{\int q_1(\theta)\pi(\theta)\,d\theta}{\int q_2(\theta)\pi(\theta)\,d\theta},$$

where

$$\begin{aligned} q_1(\theta) &\equiv \int f(\mathbf{y}|g_0, \sigma^2)\pi_0(\sigma^2)d\sigma^2 = m(\mathbf{y}|M_0), \\ q_2(\theta) &= \int f(\mathbf{y}|\theta, \tau^2, \sigma^2)\pi_0(\tau^2, \sigma^2)\,d\tau^2\,d\sigma^2. \end{aligned}$$

Then, Theorem 5.1 is readily applicable, and we obtain

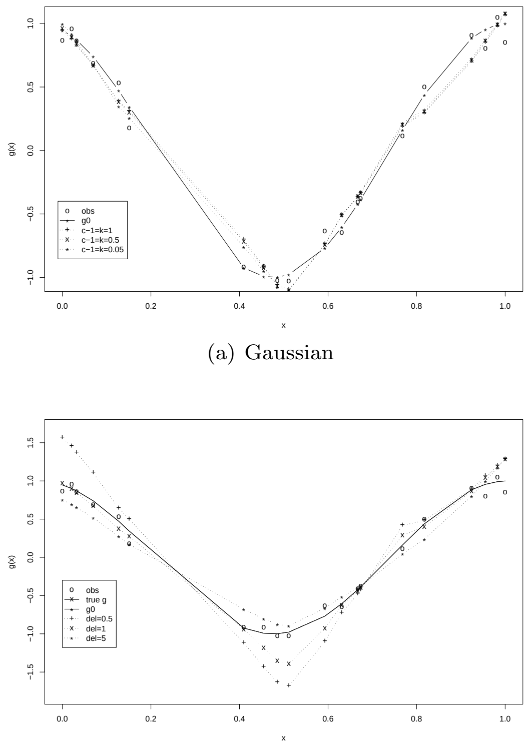
Theorem 5.2.

$$\inf_{\pi\in\mathcal{C}_A} B_{01}(\pi) \text{ is the unique solution } \lambda \text{ of}$$

$$c_2 \int (q_1(\theta) - \lambda q_2(\theta))^- h_A(\theta)\,d\theta + c_1 \int (q_1(\theta) - \lambda q_2(\theta))^+ h_A(\theta)\,d\theta = 0, \tag{5.6}$$

$$\sup_{\pi\in\mathcal{C}_A} B_{01}(\pi) \text{ is the unique solution } \lambda \text{ of}$$

$$c_2 \int (q_1(\theta) - \lambda q_2(\theta))^+ h_A(\theta)\,d\theta + c_1 \int (q_1(\theta) - \lambda q_2(\theta))^- h_A(\theta)\,d\theta = 0. \tag{5.7}$$



(a) Gaussian

(b) Ellipsoid

FIG 1. *Wavelet smoother for* $g(x) = \cos(2\pi x)$ *and* $g_0(x) = \cos(2\pi x)$

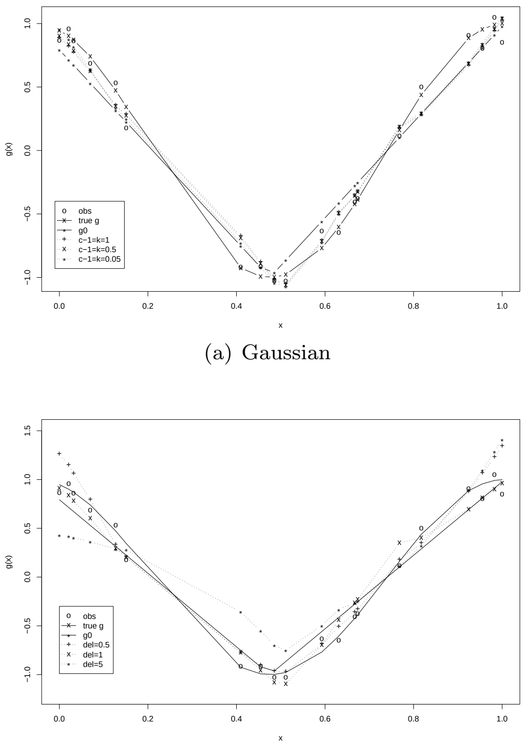
6. Examples, Simulations and Illustrations

Illustrative examples, simulated as well as those involving real-life data, are discussed below. In all examples we have used two membership functions:

1. Gaussian membership function: $h_A(g)$ proportional to the density of $N(\theta_0, \tau^2\Gamma)$, where θ_0 is obtained from the wavelet decomposition of g_0. The hyper-parameters a and b (see (4.11) and (4.12)) are $b = 3$ and $a = 8(b+2)/(b-2)$. Sensitivity analysis shows that the values of these hyper-parameters do not influence the results very much. To show that the other hyper-parameters c and k (see (4.11) and (4.12)) have some effect, though not substantial, we have displayed the wavelet smoothers (see Figures 1(a), 2(a) and 3(a)) for $(c, k) = (2, 1.5), (1.5, 0.5)$ and $(1.05, 0.05)$.
2. Uniform on the ellipsoid (see (iii) in Section 3): $h_A(g) = I_{\{\rho_J(g,g_0)\le\delta\}}$. We have used three different values (0.5, 1 and 5) for δ.

For the simulated examples, we generated observations from the model (1.1) with the regression function $g(x) = \cos(2\pi x)$ where x is drawn from a uniform density on the unit interval. Then we considered three different prior guesses for g_0: (i) $g_0(x) = \cos(2\pi x)$ (see Figure 1), (ii) $g_0(x) = 4|x - 0.5| - 1$ (see Figure 2), and (iii) $g_0 \equiv 0$ (see Figure 3).

Note that in (i) the chosen g_0 is the best possible prior guess. Further, since the normal prior is very informative, as expected the smoother (see Figure 1 (a))



(a) Gaussian

(b) Ellipsoid

FIG 2. *Wavelet smoother for* $g(x) = \cos(2\pi x)$ *and* $g_0(x) = 4|x - 0.5| - 1$

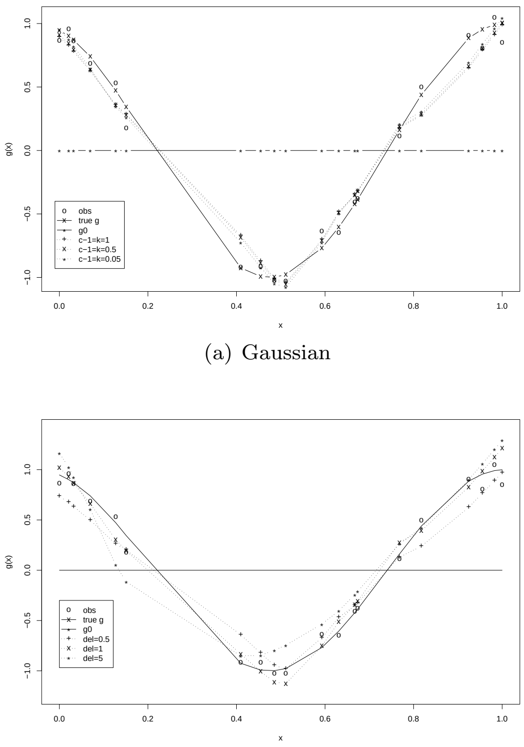
does an excellent job of extracting the true regression function. The behavior of the smoother obtained from the ellipsoid membership function (see Figure 1 (b)) is similar to that seen in Figure 1 (a), even though the prior is different.

The smoother presented in Figure 2 behaves very similar to what was seen in Figure 1. The prior guess, g_0, is slightly different from the true g.

The behavior seen in Figure 3 emphasizes our comment following Figure 2. In fact, the smoother here looks better than the one in Figure 2. This is perhaps because the prior is less informative (concentrated) than the normal prior used there, so that the smoother can follow the data more closely than the prior.

We next applied our wavelet smoother to the 'Humidity data' example from (see [3]; also [8], Example 10.2). The variable of interest y that we have chosen from the data set is the weekly average humidity level. The observations were made from June 1, 1995 to December 13, 1998. We have chosen time (day of recording the observation) as the covariate x. Since we have 185 observations here, the maximum possible value for J is 6.

In this data (see Figure 4 (a)) a seasonality effect is present, so we have chosen $g_0(x) = 22.5\cos(2\pi(x+0.1)/0.2)+62.5$, where $x = (\text{day}-\text{June 1, 1995})/(\text{December 13, 1998 - June 1, 1995})$. This choice of g_0 is rather arbitrary but it seems to follow the seasonal variations. For the analysis which led to Figure 4 (a) the Gaussian membership function was used while in Figure 4 (b), it was the ellipsoid one. In this latter figure, we have also added the lower and upper envelopes obtained from the



(a) Gaussian

(b) Ellipsoid

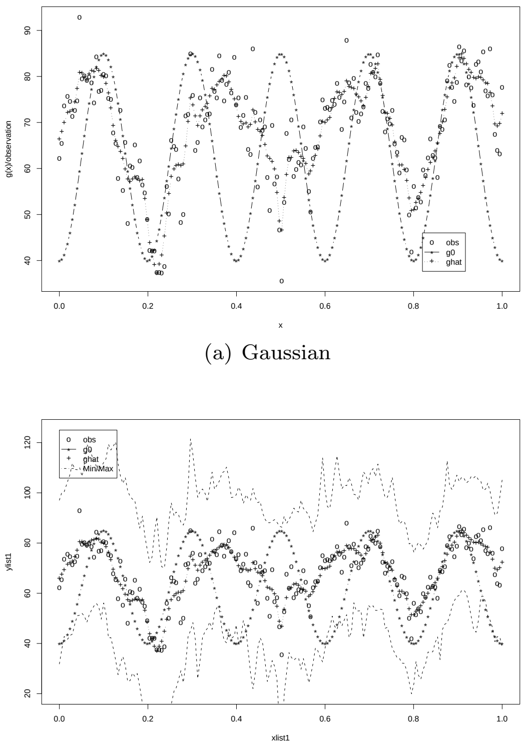
FIG 3. *Wavelet smoother for* $g(x) = \cos(2\pi x)$ *and* $g_0 \equiv 0$

prior (labeled Min/Max). From these two figures, it can be seen that the proposed estimator fits the data well and the final result does not depend much on the membership function used.

6.1. Model Checking

The model checking approach based on Bayes factors developed in the previous section has been tested on simulated examples. For this, we generated observations form Equation (1.1) with the true function $g(x_i) = \cos(2\pi x_i)$ where x_i, $i = 1, 2, \ldots, 20$ were sampled from $U(0, 1)$. For the error term, ϵ_i, we used $\sigma^2 = 0.1$. Then, for illustration purposes, we considered three different g_0 functions in $(M_0 : g = g_0)$:
(i) $g_0(x) = \cos(2\pi x)$,
(ii) $g_0(x) = 4|x - 0.5| - 1$, and
(iii) $g_0 \equiv 0$.

Note that, the g_0 function in (i) corresponds to the true function. The g_0 functions in both (i) and (ii) are similar while the last one is very different from the true function g. Therefore, it is fair to assume that the Bayes factor (see Equation (5.1)) for the first two cases should not provide evidence against the model $M_0 : g = g_0$ while we can expect strong evidence in the case of the third function. These Bayes factors are given in Table 1. From this table, it can be seen that the model



(a) Gaussian

(b) Ellipsoid

Fig 4. *Wavelet smoother for the 'Humidity data'*

Table 1
Bayes factor for $M_0 : g = g_0$ vs $M_1 : g \neq g_0$

g_0	$B_{01}(\mathbf{y})$	Evidence
$\cos(2\pi x)$	933.4275	very strongly favors M_0
$4\|x - 0.5\| - 1$	57.4735	strongly favors M_0
0	7.2845×10^{-6}	very strongly favors M_1

corresponding to the correct function ($g_0(x) = \cos(2\pi x)$) obtains the largest Bayes factor followed by that for $g_0(x) = 4|x-0.5|-1$. Moreover, if we test $M_0 : g_0(x) \equiv 0$ against $M_1 : g_0(x) \neq 0$, the Bayes factor favors M_1 with strong evidence.

7. Conclusions

In this paper we suggest a simple approach to nonparametric regression by proposing an alternative to dealing with complicated analyses on function spaces. The proposed technique uses fuzzy sets to quantify the available prior information on a function space by starting with a 'prior guess' baseline regression function g_0. First, wavelet decomposition is used to represent both the unknown regression function g as well as the prior guess g_0. Then the prior uncertainty of g relative to its distance from g_0 is specified in the form of a membership function which translates this distance into a measure of distance between the corresponding wavelet coefficients.

Furthermore, a Bayesian test is proposed to check whether the baseline function g_0 is compatible with the data or not.

Acknowledgements

We thank the editors and two referees for providing critical comments which have brought significant improvements to our presentation.

References

[1] Abramovich, F. and Sapatinas, T. (1999). Bayesian approach to wavelet decomposition and shrinkage. In *Bayesian Inference in Wavelet-Based Models, Lecture Notes in Statistics*, **141**, P. Müller and B. Vidakovic (Eds.), Springer, New York.

[2] Adler, R.J. (1981). *The Geometry of Random Fields.* Wiley, New York.

[3] Angers, J.-F. and Delampady, M. (2001). Bayesian nonparametric regression using wavelets. *Sankhyä* Ser. B **63,** 287-308.

[4] Angers, J.-F. and Delampady, M. (2004). Fuzzy sets in hierarchical Bayes and robust Bayes inference. *J. Statist. Research.* **38**, 1-11.

[5] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer, New York.

[6] DeRobertis, L. and Hartigan, J.A. (1981). Bayesian inference using intervals of measures. *Ann. Statist.* **9**, 235-244.

[7] French, S. (1986). *Decision Theory.* Ellis Horwood, Chichester, England.

[8] Ghosh, J.K., Delampady, M. and Samanta, T. (2006). *Introduction to Bayesian Analysis: Theory and Methods.* Springer, New York.

[9] Ghosh, J.K. and Ramamoorthi, R.V. (2003). *Bayesian Nonparametrics.* Springer, New York.

[10] Lenk, P.J. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *J. Amer. Statist. Assoc.*, **83**, 509-516.

[11] Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory.* Wiley, New York.

[12] Singpurwalla, N., Booker, J.M. and Bement, T.R. (2002). Probability theory. In *Fuzzy Logic and Probability Applications*, T.J. Ross et al. (Eds.), *Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA & American Statistical Association, Alexandria.*

[13] Singpurwalla, N. and Booker, J.M. (2004). Membership functions and probability measures of fuzzy sets (with discussion). *J. Amer. Statist. Assoc.* **99**, 867-889.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 106
© Institute of Mathematical Statistics, 0000
arXiv: math.PR/0000000

Objective Bayes Testing of Poisson versus Inflated Poisson Models

M. J. Bayarri[1], James O. Berger[2] and Gauri S. Datta[3,*]

University of Valencia, Duke University and SAMSI, and University of Georgia

Abstract: The Poisson distribution is often used as a standard model for count data. Quite often, however, such data sets are not well fit by a Poisson model because they have more zeros than are compatible with this model. For these situations, a zero-inflated Poisson (ZIP) distribution is often proposed. This article addresses testing a Poisson versus a ZIP model, using Bayesian methodology based on suitable objective priors. Specific choices of objective priors are justified and their properties investigated. The methodology is extended to include covariates in regression models. Several applications are given.

Contents

1. Introduction

The Poisson distribution is often used as a standard probability model for count data. For example, a production engineer may count the number of defects in items randomly selected from a production process. Quite often, however, such data sets are not well fit by a Poisson model because they contain more zero counts than are

[*] Corresponding author.

[1] Dept. of Statistics and O.R., University of Valencia, Av. Dr. Moliner 50 46100 Burjassot, Valencia, Spain, e-mail: susie.bayarri@uv.es

[2] ISDS, Box 90251, Durham, NC 27708-0251, and 19 T.W. Alexander Dr., P.O. Box 14006, Research Triangle Park, NC 27709-4006, USA, e-mail: berger@samsi.info

[3] Dept. of Statistics, University of Georgia, Athens, GA 30602-1952, USA, e-mail: gaurisdatta@gmail.com

AMS 2000 subject classifications: Primary 62F15, 62F03

Keywords and phrases: Bayes factor, Jeffreys prior, Model selection

compatible with the Poisson model. An example is again provided by the production process; indeed, according to Ghosh et al. [14], when some production processes are in a near perfect state, zero defects will occur with a high probability. However, random changes in the manufacturing environment can lead the process to an imperfect state, producing items with defects. The production process can move randomly back and forth between the perfect and the imperfect states. For this type of production process many items will be produced with zero defects, and this excess might be better modeled by a ZIP distribution than a Poisson distribution.

For $0 \leq p \leq 1, \lambda > 0$, the ZIP$(\lambda, p)$ distribution has the probability function

$$f_1(x \mid \lambda, p) = p\, I(x=0) + (1-p)\, f_0(x \mid \lambda), \quad x = 0, 1, 2, \dots \,, \tag{1.1}$$

where $I(\cdot)$ is the indicator function, and $f_0(x|\lambda)$ is the Poisson probability function

$$f_0(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!},\; x = 0, 1, 2, \dots \,. \tag{1.2}$$

The parameter p is referred to as the *zero-inflation parameter*.

Many authors used the ZIP distribution with and without covariates to model count data. In a ZIP regression model, Lambert [18] used a frequentist approach and Ghosh et al. [14] used a Bayesian approach to analyze industrial data sets.

While the aforementioned authors used the ZIP model to analyze their data, a number of authors have addressed the problem of checking whether a ZIP model is needed to model the data. From the frequentist perspective, score tests have been developed for testing the hypothesis $\mathcal{H}_0 : p = 0$ vs. $\mathcal{H}_1 : p \neq 0$ in a ZIP regression model ([10],[12]). From the Bayesian perspective, Bhattacharya et al. [9] presented a Bayesian method to test $p \leq 0$ versus the alternative $p > 0$ by computing a certain posterior probability of the alternative hypothesis. As in ([10],[12]), p is allowed to be negative in their model [9], as long as $p + (1-p)e^{-\lambda} \geq 0$.

In this paper, we consider Bayesian testing of M_0 versus M_1 given by

$$M_0 : \qquad X_i \overset{i.i.d.}{\sim} f_0(\cdot \mid \lambda),\; i = 1, \dots, n, \tag{1.3}$$

$$M_1 : \qquad X_i \overset{i.i.d.}{\sim} f_1(\cdot \mid \lambda, p),\; i = 1, \dots, n, \tag{1.4}$$

where f_0, f_1 are given in (1.1) and (1.2), respectively. Note that, as opposed to the situations in the papers mentioned above, $p < 0$ is not possible here.Indeed, we can alternatively formulate the problem as that of testing, within the ZIP model,

$$\mathcal{H}_0 : p = 0 \quad \text{versus} \quad \mathcal{H}_1 : p > 0.$$

Unlike the analysis in [9], $p = 0$ (i.e., the Poisson model) is assumed to have a priori believability (e.g., prior probability 1/2).

In Section 2 we develop the suggested objective testing of Poisson versus ZIP models when not all counts are zeros. For all zeros, the ZIP distribution is not identifiable, and a proper prior is required for all parameters; we address this in Section 5. Section 3 is devoted to some comparative examples. We consider inclusion of covariates in Section 4, where we address the testing of Poisson versus ZIP regression models and give an example involving AIDS related deaths in men. In the regression case, in order for the objective Bayesian model selection to be successful we need enough positive counts so that the design matrix based on the positive counts is full column rank. When this condition does not hold we suggest in Section 5 a partially proper prior on the regression parameters to be used for model selection. Proofs and technical details are relegated to an Appendix.

2. Formulation of the Problem

The Bayesian methodology for choosing between two models for some data is conceptually very simple (see, e.g., [3]). One assesses the prior probabilities of each model, the prior distributions for the model parameters, and computes the posterior probabilities of each model. These posterior probabilities can be computed directly from the prior probabilities and the *Bayes Factor*, an (integrated) likelihood ratio for the models which is very popular in Bayesian testing and model selection.

Often it is not possible (for lack of time or resources) to carefully assess in a subjective manner all the needed priors. In these situations, very satisfactory answers are provided by *objective Bayesian analyses* that do not use external information other than that required to formulate the problem (see [4]). First we review below some difficulties of model selection via objective Bayesian analysis. Then we justify the objective prior we chose for our problem, derive the corresponding Bayes Factor and study properties of the prior and the Bayes factor.

2.1. Bayesian Model Selection and Bayes Factors

To compare two models, M_0 and M_1, for the data $\boldsymbol{X} = (X_1, \ldots, X_n)$, the Bayesian approach is based on the *Bayes factor* B_{10} of M_1 to M_0 given by

$$B_{10} = \frac{m_1(\boldsymbol{x})}{m_0(\boldsymbol{x})} = \frac{\int f_1(\boldsymbol{x} \mid \boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1)d\boldsymbol{\theta}_1}{\int f_0(\boldsymbol{x} \mid \boldsymbol{\theta}_0)\pi_0(\boldsymbol{\theta}_0)d\boldsymbol{\theta}_0} \ , \tag{2.1}$$

where, under model M_i, $\boldsymbol{X}$ has density $f_i(\boldsymbol{x} \mid \boldsymbol{\theta}_i)$ and the unknown parameters $\boldsymbol{\theta}_i$ in M_i are assigned a prior density $\pi_i(\boldsymbol{\theta}_i), i = 0, 1$. For given prior model probabilities $Pr(M_0)$ and $Pr(M_1) = 1 - Pr(M_0)$, the posterior probability of, say, M_0 is

$$Pr(M_0 \mid \boldsymbol{x}) = \left[1 + B_{10}\ \frac{Pr(M_1)}{Pr(M_0)}\right]^{-1} \ . \tag{2.2}$$

In objective Bayesian analyses $\pi_i(\boldsymbol{\theta}_i)$ is chosen in an objective or conventional fashion and the hypotheses would be assumed to be equally likely a priori.

Use of objective priors has a long history in Bayesian inference (see, for example, [8] and [17] for justifications and references). They are, however, typically improper and are only defined up to an arbitrary multiplicative constant. This is not a problem in the posterior distribution, since the same constant appears in both the numerator and the denominator of Bayes theorem and so cancels. In model selection and hypothesis testing, however, it can be seen from (2.1) that when at least one of the priors $\pi_i(\boldsymbol{\theta}_i)$ is improper, the arbitrary constant does not cancel, so that the Bayes factor is then arbitrary and undefined. An important exception to this arises in invariant situations for parameters occurring in all of the models; Berger et al. [7] show that use of the (improper) right Haar invariant prior is then permissible.

One of the ways to address this difficulty is to try to directly 'fix' the Bayes factor by appropriately choosing the multiplicative constant, as in [13]. Popular methods (the *intrinsic Bayes factor* [5] and the *fractional Bayes factor* [20]) for fixing this constant arise as a consequence of 'training' the improper priors into proper priors based on part of the data or of the likelihood. We refer to Berger and Pericchi [6] for a review, references and comparisons. Another possibility is to directly derive appropriate 'objective' but proper distributions $\pi_i(\boldsymbol{\theta}_i)$ to use in model selection; see [2] and [15] for methods and references. This is the approach taken in this paper (with a slight exception in Section 5).

2.2. Specification and Justification of the Objective Priors

Returning to the testing of the Poisson (M_0) *vs.* the ZIP (M_1) models, i.e., testing

$$M_0 : \ \boldsymbol{X} \sim f_0(\boldsymbol{x} \mid \lambda) \ \ vs. \ \ M_1 : \ \boldsymbol{X} \sim f_1(\boldsymbol{x} \mid \lambda, p), \tag{2.3}$$

the key issue is the choice of the priors $\pi_0(\lambda)$ and $\pi_1(\lambda, p) = \pi_1(\lambda)\,\pi_1(p \mid \lambda)$.

A frequent simplifying procedure (both for subjective and objective methods) is to take $\pi_0(\lambda)$ equal to $\pi_1(\lambda)$, that is, to give the same prior to the parameters occurring in all models under consideration. This, however, may be inappropriate, since λ might have entirely different meanings under model M_0 and under model M_1; the fact that we have used the same label does not imply that they have the same meanings. This frequent mistake is discussed, for example, in [7].

It has been argued that, if the common parameters are *orthogonal* to the remaining parameters in each model (that is, the Fisher information matrix is block diagonal), then they can be assigned the same prior distribution ([15], [16]). In this case, improper priors can be used, since the arbitrary constant would cancel in the Bayes factor.

Unfortunately, p and λ in the ZIP model are not orthogonal. We first reparameterize the original model. With $p^* = p + (1-p)e^{-\lambda}$, we rewrite $f_1(x \mid \lambda, p)$ as

$$f_1^*(x \mid \lambda, p^*) = p^* I(x=0) + (1-p^*) f^T(x \mid \lambda), \quad x = 0, 1, 2, \ldots, \tag{2.4}$$

where $f^T(x \mid \lambda)$ is the zero-truncated Poisson distribution with parameter λ. Note that $p^* \geq e^{-\lambda}$. We can trivially express the Poisson (M_0) model as:

$$f_0^*(x \mid \lambda) = e^{-\lambda} I(x=0) + (1-e^{-\lambda}) f^T(x \mid \lambda), \quad x = 0, 1, 2, \ldots, \tag{2.5}$$

and now it can intuitively be seen that λ has the same meaning in both f_1^* and f_0^*. Indeed the Fisher Information matrix for p^* and λ can be checked to be diagonal.

With an orthogonal reparameterization, Jeffreys (1961) recommended using (i) *Jeffreys prior* (the square root of Fisher information) for the 'common' parameters; and (ii) a reasonable *proper* prior for the extra parameters in the more complex model.

The situation here is very unusual, however, in that the Jeffreys prior for the 'common' λ is different for each model. The *Jeffreys prior* for λ in the Poisson model is well known to be $\pi_J^0 = 1/\sqrt{\lambda}$, whereas the Jeffreys prior for the orthogonalized ZIP model is easily shown to be the same as the Jeffreys prior for the truncated distribution $f^T(x \mid \lambda)$, which is

$$\pi_J^1(\lambda) = \frac{k(\lambda)}{\sqrt{\lambda}} \ , \quad \text{where} \quad k(\lambda) = \frac{\{1-(\lambda+1)e^{-\lambda}\}^{1/2}}{1-e^{-\lambda}} \ .$$

That these priors are different after orthogonalization is highly unusual and can be traced to the fact that λ also enters into the definition of the nested model, through $p^* = e^{-\lambda}$. In any case, we are left without clear guidance as to whether π_J^0 or π_J^1 should be used as the prior for λ. (Note that, in computing the Bayes factor, the same prior for λ must be used in both the numerator and the denominator; otherwise one is facing the indeterminacy issues discussed earlier.)

Under the orthogonalized ZIP model, we also need to specify a proper prior for p^* given λ, which we propose to take uniform over the interval $(e^{-\lambda}, 1)$, that is,

$$\pi_1(p^* \mid \lambda) = \frac{I(e^{-\lambda} < p^* \leq 1)}{1 - e^{-\lambda}} \ .$$

We can thus write the overall priors being considered for the two models $f_0^*(x \mid \lambda)$ and $f_1^*(x \mid \lambda, p^*)$ as, respectively,

$$\pi_0^l(\lambda) = \frac{k(\lambda)^l}{\sqrt{\lambda}}, \quad \pi_1^l(\lambda, p^*) = \frac{k(\lambda)^l}{\sqrt{\lambda}} \frac{I(e^{-\lambda} < p^* \leq 1)}{1 - e^{-\lambda}} \ ,$$

where l is 0 or 1 as we utilize one or the other of the two Jeffreys priors for λ.

It is computationally more convenient to work in the original (p, λ) parameterization. A change of variables above then results in the priors

$$\pi_0^l(\lambda) = \frac{k(\lambda)^l}{\sqrt{\lambda}}, \quad \pi_1^l(\lambda, p) = \frac{k(\lambda)^l}{\sqrt{\lambda}} \ I(0 < p \leq 1) \ , \tag{2.6}$$

which we will henceforth consider (for l equal to 0 or 1).

We are not aware of any desiderata that would suggest a preference for either the $l = 0$ prior or the $l = 1$ prior, but luckily the two yield almost the same answers. Indeed, simple algebra shows that $k(\lambda)$ is a strictly increasing function of λ and that

$$\inf \ k(\lambda) = \frac{1}{\sqrt{2}} = 0.71 \ , \quad \text{and} \quad \sup \ k(\lambda) = 1. \tag{2.7}$$

Thus $k(\lambda)$ is quite flat as a function of λ, so that $k(\lambda)^1$ and $k(\lambda)^0 = 1$ are very similar. An immediate consequence for the Bayes factors B_{10}^l, $l = 0, 1$ is that

$$B_{10}^0/\sqrt{2} \ \leq B_{10}^1 \leq \sqrt{2}\, B_{10}^0 \ ,$$

so that the two Bayes factors can only differ by a modest amount (and in practice the difference is much smaller than this).

It is obviously a bit simpler to work with the $l = 0$ prior, so we drop the l superscript and henceforth utilize the prior

$$\pi_0(\lambda) = \frac{1}{\sqrt{\lambda}}, \quad \pi_1(p, \lambda) = \frac{1}{\sqrt{\lambda}} \ I(0 < p \leq 1) \ . \tag{2.8}$$

2.3. Objective Bayes Factor for Poisson versus ZIP models

Recall that the model M_0 is the standard Poisson model and the model M_1 is the ZIP model. For a sample of n counts $X_1, \ldots, X_n$, let $\boldsymbol{X}$ denote the sample, $k = \sum_{i=1}^n I(X_i = 0)$ be the number of zero counts, and $s = \sum_{i=1}^n X_i$ be the total count. Note that $k = n$ is equivalent to $s = 0$. For given data $\boldsymbol{x}$, the densities $f_0(\boldsymbol{x} \mid \lambda)$ and $f_1(\boldsymbol{x} \mid \lambda, p)$ under the two models are given by

$$f_0(\boldsymbol{x} \mid \lambda) = \frac{e^{-n\lambda}\lambda^s}{\prod_{i=1}^n x_i!}, \quad f_1(\boldsymbol{x} \mid \lambda, p) = \frac{[p + (1-p)e^{-\lambda}]^k (1-p)^{n-k} e^{-(n-k)\lambda}\lambda^s}{\prod_{i=1}^n x_i!} \ .$$

For $s > 0$ (i.e., the counts are not all zero),

$$m_0(\boldsymbol{x}) = \int f_0(\boldsymbol{x} \mid \lambda)\pi_0(\lambda) d\lambda = \frac{\Gamma(s + \frac{1}{2})}{n^{s+\frac{1}{2}} \prod x_i!} \ .$$

Using the binomial expansion of $[p + (1-p)e^{-\lambda}]^k$,

$$\begin{aligned} m_1(\boldsymbol{x}) &= \int f_1(\boldsymbol{x} \mid \lambda, p)\pi_1(p,\lambda)dp\ d\lambda \\ &= \frac{1}{\prod x_i!}\sum_{j=0}^{k}\frac{k!}{j!(k-j)!}\int_0^{\infty}\int_0^1 p^j(1-p)^{n-j}e^{-(n-j)\lambda}\lambda^{s-\frac{1}{2}}dpd\lambda \\ &= \frac{k!}{(n+1)!\prod x_i!}\sum_{j=0}^{k}\frac{(n-j)!}{(k-j)!}\Gamma(s+\frac{1}{2})(n-j)^{-(s+\frac{1}{2})}. \end{aligned}$$

Both $m_0(\boldsymbol{x})$ and $m_1(\boldsymbol{x})$ are finite and the Bayes factor $B_{10}(\boldsymbol{x}) = m_1(\boldsymbol{x})/m_0(\boldsymbol{x})$ is

$$B_{10}(\boldsymbol{x}) = \frac{k!}{(n+1)!}\sum_{j=0}^{k}\frac{(n-j)!}{(k-j)!}(1-\frac{j}{n})^{-(s+1/2)}\,. \tag{2.9}$$

Note that, as intuitively expected, for any given n the Bayes factor is increasing in s (total count) for any fixed k (the number of zero's), and is increasing in k for any fixed s. We use (2.9) to calculate the Bayes factors for the examples in Section 3.

When $s = 0$ or equivalently all counts are zero ($\boldsymbol{x} = \boldsymbol{0}$), there is a problem. While $m_0(\boldsymbol{0}) = \Gamma(1/2)/\sqrt{n}$ remains finite, it is easy to see that $m_1(\boldsymbol{0})$ is infinite. Indeed for *any* prior of the form $h(p)\pi(\lambda)$, where $\pi(\lambda)$ is improper and $h(p)$ is a proper density (as is required for testing), the marginal density $m_1(\boldsymbol{0})$ will be infinite. This is because, for $\boldsymbol{x} = \boldsymbol{0}$, the density $f_1(\boldsymbol{x} \mid \lambda, p) \geq p^n$ implying $m_1(\boldsymbol{0}) \geq \int_0^1 p^n h(p)dp \int_0^{\infty}\pi(\lambda)d\lambda = \infty$. We discuss what to do for this case in Section 5.

3. Applications

In this section we apply our methodology to two datasets to detect if zero-inflation is present in the data. These examples have been analyzed for zero-inflation previously using both frequentist and Bayesian procedures. Since there are non- zero counts in both examples, the Bayes factors are computed using (2.9).

Example 3.1. The first dataset is the Urinary Tract Infection (UTI) data used in Broek [10], which used a score test to detect zero-inflation in a Poisson model. The data are collected from 98 HIV-infected men treated at the Department of Internal Medicine at the Utrecht University hospital. The number of times they had a urinary tract infection was recorded as X. The data are recorded in Table 1. Merely by looking at the data it is apparent that zero-inflation is present.

X	0	1	2	3	Total
Frequency	81	9	7	1	98

Table 1
UTI Data

Equation (2.9) yields a Bayes factor $B_{10} = 223.13$ in favor of model M_1 versus model M_0; if the models were believed to be equally likely a priori, the resulting posterior model probabilities would be $Pr(M_1 \mid \boldsymbol{x}) = 0.995$ and $Pr(M_0 \mid \boldsymbol{x}) = 0.005$. This is indeed strong evidence in favor of the ZIP model.

In Bayesian testing of $\mathcal{H}_0 : p \leq 0$ versus $\mathcal{H}_1 : p > 0$, Bhattacharya et al. [9] obtained $Pr(p > 0 \mid \boldsymbol{x}) = .999$. The observed value of the score statistic was

X	0	1	2	3	4	Total
Frequency	38	26	8	2	1	75

Table 2
Terror Data

reported as 15.34 [10], yielding a $p-$value of 0.0001. All three analyses present strong evidence in favor of the ZIP model, but notice that the p-value seems to suggest stronger evidence against the Poisson null than the Bayesian analysis, and the point null Bayesian analysis suggests weaker evidence than the interval Bayesian test.

Example 3.2. The next dataset we consider is the Terrorism data from [11]. Table 2 gives the number of incidents of international terrorism per month (X) in the United States between 1968 and 1974. It is not intuitively clear whether or not there is zero-inflation in this data set.

The Bayes factor here is $B_{10} = 0.28$, yielding an objective posterior probability $Pr(M_1 \mid \boldsymbol{x}) = 0.219$, which actually supports the Poisson model. A previous analysis found $Pr(p > 0 \mid \boldsymbol{x}) = 0.507$, an indeterminate value [9]. The observed value of the score statistic is 0.04, with a $p-$value of 0.83. Conigliani et al [11] test a Poisson null model against a nonparametric alternative, finding a fractional Bayes factor B_{10}^F of 0.0089 of the nonparametric alternative to the Poisson; the apparent strength of this conclusion, compared with the other results, is rather puzzling.

4. Model Selection in ZIP Regression

Many applications involve count data where covariate information is available; see, for example, [14] and [18]. In this section we consider selecting between Poisson regression and ZIP regression models given by

$$M_0^R: \quad X_i \overset{ind}{\sim} Poisson(\lambda_i),\ i = 1, \ldots, n, \tag{4.1}$$

$$M_1^R: \quad X_i \overset{ind}{\sim} ZIP(\lambda_i, p),\ i = 1, \ldots, n. \tag{4.2}$$

For a known offset variable a_{0i}, a $q \times 1$ vector of covariates $\boldsymbol{a}_i$ and regression parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$, suppose the λ_i follow the log-linear relationship

$$\log(\lambda_i) = a_{0i} + \boldsymbol{a}_i^T \boldsymbol{\beta}.$$

We assume that the matrix $\boldsymbol{A}^T = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)$ is of rank q. Let k denote the number of zero counts in the data. For simplicity of notation, we index the observations in such a way that all the zeros are given by the first k counts.

4.1. Objective Priors for Model Selection

Generalizing the argument in Section 2.2 to the regression case is easy in one case, but difficult in the other. If we choose to base the analysis on the Jeffreys prior for $\boldsymbol{\beta}$ under the Poisson regression model M_0^R, the generalization is straightforward: the Jeffreys prior is easily computed as

$$\pi_0^R(\boldsymbol{\beta}) = |\sum_{i=1}^n \lambda_i \boldsymbol{a}_i \boldsymbol{a}_i^T|^{1/2}. \tag{4.3}$$

Note that this prior is positive since the rank of $\boldsymbol{A}$ is q. Also, utilizing this prior for $\boldsymbol{\beta}$ under model M_1^R, along with the independent uniform prior for p, results in the following priors to be utilized to compute B_{10}:

$$\text{(4.4)} \qquad \pi_0^0(\boldsymbol{\beta}) = |\sum_{i=1}^n \lambda_i \boldsymbol{a}_i \boldsymbol{a}_i^T|^{1/2}, \quad \pi_1^0(\boldsymbol{\beta}, p) = |\sum_{i=1}^n \lambda_i \boldsymbol{a}_i \boldsymbol{a}_i^T|^{1/2} I(0 < p \leq 1)\,.$$

The generalization to the regression case of the second prior considered in Section 2.2 is much more difficult, because the Jeffreys prior under the ZIP regression model is very complicated. In Section 2.2, the derivation of the corresponding Jeffreys prior was essentially done by ignoring the non-zero counts, utilizing only the truncated Poisson distribution. This suggests modifying (4.3) by removing the terms corresponding to the zero counts, resulting in

$$\text{(4.5)} \qquad \pi_1^R(\boldsymbol{\beta}) = |\sum_{i=k+1}^n \lambda_i \boldsymbol{a}_i \boldsymbol{a}_i^T|^{1/2}.$$

From another intuitive perspective, the zero counts arising from the inflation factor are clearly irrelevant in fitting the log linear model to the λ_i and, since we do not know which zero counts arise from the inflation factor, dropping them all from the Jeffreys prior has an appeal. Let $\boldsymbol{A}_+ = (\boldsymbol{a}_{k+1}, \ldots, \boldsymbol{a}_n)^T$. The prior (4.5) can only be used provided it is positive, which is ensured if the rank of $\boldsymbol{A}_+$ is q.

The resulting overall prior for use in computing B_{10} is then

$$\text{(4.6)} \quad \pi_0^1(\boldsymbol{\beta}) = |\sum_{i=k+1}^n \lambda_i \boldsymbol{a}_i \boldsymbol{a}_i^T|^{1/2}, \quad \pi_1^1(\boldsymbol{\beta}, p) = |\sum_{i=k+1}^n \lambda_i \boldsymbol{a}_i \boldsymbol{a}_i^T|^{1/2} I(0 < p \leq 1)\,.$$

The first basic issue in use of these priors is whether or not they yield finite marginal distributions. This is addressed in the following theorems, the first of which deals with the marginal density under the Poisson regression model.

Theorem 4.1. *For the Poisson regression model and either the Jeffreys prior* ($j = 0$) *or the modified Jeffreys prior* ($j = 1$),

$$\text{(4.7)} \qquad m_0^R(\boldsymbol{x}) = \int_{R^q} \prod_{i=1}^n \{\frac{e^{-\lambda_i}\lambda_i^{x_i}}{x_i!}\} \pi_j^R(\boldsymbol{\beta}) d\boldsymbol{\beta} < \infty.$$

Proof. See the Appendix. □

Note that with more than one covariate there is typically no closed-form expression for $m_0^R(\boldsymbol{x})$. Hence $m_0^R(\boldsymbol{x})$ needs to be evaluated by numerical or Monte Carlo integration.

For the ZIP regression model, the marginal density $m_1^R(\boldsymbol{x})$, under an arbitrary improper prior $\pi(\boldsymbol{\beta})$ for $\boldsymbol{\beta}$ and an independent uniform prior for p, is given by

$$\text{(4.8)} \qquad m_1^R(\boldsymbol{x}) = \int_{R^q} \int_0^1 f_1(\boldsymbol{x} \mid \boldsymbol{\beta}, p)\, \pi(\boldsymbol{\beta})\, dp\; d\boldsymbol{\beta},$$

where the density of $\boldsymbol{x}$, under model M_1^R, is given by

$$f_1(\boldsymbol{x} \mid \boldsymbol{\beta}, p) = \prod_{i=1}^k \{p + (1-p)e^{-\lambda_i}\}(1-p)^{n-k} \prod_{i=k+1}^n \frac{e^{-\lambda_i}\lambda_i^{x_i}}{x_i!}.$$

Again, as for $m_0^R(\boldsymbol{x})$, there is usually no closed-form expression for $m_1^R(\boldsymbol{x})$ and the marginal needs to be computed via numerical or Monte Carlo integration.

To investigate the finiteness of $m_1^R(\boldsymbol{x})$, note first that

$$p^k(1-p)^{n-k}\prod_{i=k+1}^{n}\frac{e^{-\lambda_i}\lambda_i^{x_i}}{x_i!}\le f_1(\boldsymbol{x}\mid\boldsymbol{\beta},p)\le\prod_{i=k+1}^{n}\frac{e^{-\lambda_i}\lambda_i^{x_i}}{x_i!}\,. \tag{4.9}$$

In view of this inequality and the independent uniform prior for p, the marginal $m_1^R(\boldsymbol{x})$ is finite if and only if

$$\int_{R^q}\prod_{i=k+1}^{n}\frac{e^{-\lambda_i}\;\lambda_i^{x_i}}{x_i!}\;\pi(\boldsymbol{\beta})\;d\boldsymbol{\beta}<\infty\,. \tag{4.10}$$

Theorem 4.2 below gives sufficient conditions for this to be finite under the priors (4.3) and (4.5) respectively. Recall that the k zeros in the sample are labeled to correspond to the first k observations. A key condition will be that the matrix $\boldsymbol{A}_+$ has rank q which implies that $n\ge k+q$ (analogous to the condition of at least one positive count for the case of no covariate treated in Section 2).

Theorem 4.2. Using $\pi_0^R(\boldsymbol{\beta})$: *Suppose that, for the observation $X_j, j=1,\ldots,k$, corresponding to the zero counts, the corresponding covariate vector $\boldsymbol{a}_j$ is such that*

$$\boldsymbol{a}_j=\sum_{m=k+1}^{n}c_{mj}\,\boldsymbol{a}_m\quad \textit{with}\quad c_{mj}\ge 0, j=1,\ldots,k, m=k+1,\ldots,n. \tag{4.11}$$

Then the marginal $m_1^R(\boldsymbol{x})$ is finite.
Using $\pi_1^R(\boldsymbol{\beta})$: *If $\boldsymbol{A}_+$ has rank q, the marginal $m_1^R(\boldsymbol{x})$ is finite.*

Proof. See the Appendix. □

Clearly the condition under which $m_1^R(\boldsymbol{x})$ is finite is more general and much easier to check for $\pi_1^R(\boldsymbol{\beta})$ than for $\pi_0^R(\boldsymbol{\beta})$. This, together with the intuitive appeal of $\pi_1^R(\boldsymbol{\beta})$, leads us to recommend its use in practice. (Note that either of the two priors reduces to the prior recommended in Section 2 for the non-regression case.)

Remark 4.1. If the condition (4.11) fails, the marginal density $m_1^R(\boldsymbol{x})$ based on the Jeffreys prior may be infinite. For example, consider $n=3$ and $q=2$, with $\lambda_1=\lambda_2^{c_1}\lambda_3^{c_2}$, $\lambda_2=\exp(\beta_1)$, $\lambda_3=\exp(\beta_2)$ for suitable nonzero c_1,c_2 to be chosen later. Then the determinant of information matrix for $\boldsymbol{\beta}$ is given by

$$|\boldsymbol{I}(\boldsymbol{\beta})|=\lambda_2\lambda_3+c_1^2\lambda_2^{c_1}\lambda_3^{c_2+1}+c_2^2\lambda_2^{c_1+1}\lambda_3^{c_2}\,,$$

so that $|\boldsymbol{I}(\boldsymbol{\beta})|^{1/2}\ge|c_1|\lambda_2^{c_1/2}\lambda_3^{(c_2+1)/2}$. If $X_1=0$, $X_2=x_2$ and $X_3=x_3$, then

$$\begin{aligned} m_1^R(\boldsymbol{x}) &\ge \frac{|c_1|}{2}\int_{R^2}\frac{e^{-\lambda_2}\lambda_2^{x_2}}{x_2!}\frac{e^{-\lambda_3}\lambda_3^{x_3}}{x_3!}\lambda_2^{c_1/2}\lambda_3^{(c_2+1)/2}d\boldsymbol{\beta}\\ &= \frac{|c_1|}{x_2!x_3!2}\int_0^\infty e^{-\lambda_2}\lambda_2^{x_2-1+.5c_1}d\lambda_2\int_0^\infty e^{-\lambda_3}\lambda_3^{x_3-1+.5c_2+.5}d\lambda_3=\infty\,, \end{aligned}$$

providing that $x_2\le-.5c_1$ or that $x_3\le-.5-.5c_2$. For example, if $c_1=-5$ and a sample produces $x_2=2$, then $m_1^R(\boldsymbol{x})=\infty$. Note that here $\boldsymbol{a}_1=-5\boldsymbol{a}_2+c_2\boldsymbol{a}_3$, with $\boldsymbol{a}_2=(1,0)^T$ and $\boldsymbol{a}_3=(0,1)^T$, so that the condition (4.11) does not hold.

4.2. An Illustrative Application

We apply the methodology recommended in Section 4.1 to a dataset involving the number of AIDS-related deaths in men. The data provides the number of deaths for 598 census tracts in a large city of Spain over a period of eight years. The dataset, which was supplied to us by Dr. M.A.M. Beneyto, has a large number of tracts with zero deaths (actually, 303, which is k in our notation). Along with the number of deaths, the dataset also provides, for each census tract, the expected number of deaths E from AIDS (adjusting for the population and the distribution of ages in each tract) and an auxiliary variable W (continuous in nature) measuring the social status of each census tract.

In our application and for the ith census tract, we take $\log(E_i)$ as the offset a_{0i} and propose a log-linear regression for λ_i with $q = 2$ and $\boldsymbol{a}_i = (1, W_i)^T$. First, we will ignore the covariate W and compute the Bayes factor taking $q = 1$ and $\boldsymbol{a}_i = 1$ based on the Jeffreys prior. This model modifies the common mean model of Section 2.2 by incorporating the offset variable in the mean, which is here given by $E_i\lambda$ with $\lambda = \beta_1$. The marginal $m_1(\boldsymbol{x})$ is computed by one-dimensional numerical integration. Although it has a closed-form expression, it is rather complicated and omitted here to save space. This expression is given in the Appendix in [1]. For the specific data here, $B_{10} = 22,975$ which gives overwhelming evidence in favor of the ZIP model.

Epidemiologists who are knowledgeable about this study believed that the large number of zero counts in the data could be explained by the covariate measuring the social status and, indeed, suspected that a ZIP regression model would not be needed if the covariate were incorporated into the analysis. The Bayes factor in favor of the ZIP regression model versus the Poisson regression model (with $q = 2$) is given by 7.25. While this Bayes factor provides a moderate amount of evidence in favor of the ZIP regression model, it is much smaller than $22,975$, indicating that, indeed, the covariate can explain most of the excess zero counts.

In this example, it is possible that the same inflation parameter p may not be appropriate for all individuals. Just like using the log-linear models for λ_i, we can treat each p_i differently (as p may change according to the covariates) and fit a logistic regression model for p_i. But it is highly likely that there would be severe confounding between the two regressions, which is particularly problematical with objective Bayesian analysis (since there is not a proper subjective prior to overcome the confounding).

5. Analysis with Insufficient Positive Counts

As noted in Section 2, the marginal density under model M_1 based on an improper prior for λ is not finite when all counts are zeros, and hence the Bayes factor is not well-defined. This is not a difficulty of only model selection; in this situation, it is also not possible to make inferences about the parameters of the ZIP model, since the joint posterior of the parameters (under the ZIP model) is improper. Indeed, when all counts are zero, the ZIP model parameters are not identifiable, and the data do not provide enough information to estimate the parameters. Since objective Bayes methods are typically based on information from the data alone, it is not surprising that problems are encountered.

We could simply invoke this argument and refrain from considering the case when all counts are zero. However, it is interesting to explore several methodologies that have been proposed for difficult testing situations, partly to judge the success of the methodologies and partly to try to provide a reasonable answer to this case. We continue, throughout the section, to assume that $p \sim Un(0,1)$.

5.1. All Zero Counts in the Non-Regression Case

We mentioned that to resolve the identifiability issue in the ZIP model for the data with all zeros we need a proper prior on λ. This can be done by either subjectively specifying a proper prior for λ or by 'training' the improper priors into proper priors based on part of the data or of the likelihood. In particular, the intrinsic Bayes factor approach [5] utilizes a part of the data as a training sample to train the improper prior to get a proper posterior. Although this approach works successfully in many examples, it is not successful in the present problem. Our investigation of this approach [1] is omitted here to save space. We discuss below the case where a subjective proper prior on λ is specified based on certain considerations.

If a proper prior is needed to define the Bayes factor for the situation of all zero counts, the most direct approach is to find a proper prior that seems compatible with certain behaviors that we expect of the Bayes factor in this situation. A natural proper prior to consider for λ is a Gamma ($Ga(a,b)$) conjugate prior under the Poisson model (M_0) given by the Gamma $g(\lambda \mid a,b)$ density

$$g(\lambda \mid a,b) = \frac{b^a e^{-b\lambda} \lambda^{a-1}}{\Gamma(a)},$$

where a, b are suitably chosen positive constants. Of course, one is welcome to simply make subjective choices here, but we will argue for a certain choice (or choices) based on rather neutral thinking.

First, we assume that the *same* gamma prior is appropriate for λ, both under the Poisson and the ZIP models. This can be justified by the orthogonalization argument used in Section 2.2. With the uniform density for p and the $Ga(a,b)$ prior for λ, the resulting Bayes factor for arbitrary data $\boldsymbol{x}$ can be computed to be

$$B_{10}(\boldsymbol{x}) = \frac{k!}{(n+1)!} \sum_{j=0}^{k} \frac{(n-j)!}{(k-j)!} \left(1 - \frac{j}{n+b}\right)^{-(s+a)}, \tag{5.1}$$

by a similar argument to that leading to (2.9). This Bayes factor includes as a special case the objective Bayes factor in (2.9); indeed the Jeffreys prior used there was a limiting case of the $g(\lambda \mid a,b)$ for $a = 1/2$ and $b = 0$. Note that the Bayes factor (5.1) is increasing in s, k and a, and decreasing in b.

For the special case $\boldsymbol{x} = \mathbf{0}$ (that is $s = 0$ and $k = n$), note that $f_1(\mathbf{0}|\lambda, p) \geq f_0(\mathbf{0}|\lambda)$. Hence, using the same proper prior for λ with both the Poisson and the ZIP models, it follows that $m_1(\mathbf{0}) \geq m_0(\mathbf{0})$, and hence, $B_{10}(\mathbf{0}) \geq 1$. In particular, for the $Un(0,1)$ prior for p and $Ga(a,b)$ prior for λ, it can be checked that

$$B_{10}(\mathbf{0}) = \frac{(n+b)^a}{n+1} \sum_{j=0}^{n} \frac{1}{(j+b)^a} \geq 1\,. \tag{5.2}$$

This is reasonable: when a long stream of *only* zeros is observed, it is entirely natural to say that the data favor the ZIP model. But the degree of favoritism depends on a and b, and we turn to rather speculative desiderata to narrow the choice. Recall that the mean of the $Ga(a,b)$ distribution for λ is ab^{-1} and the variance is ab^{-2}.

In order for the prior not to be too sharp, it is reasonable to require the prior standard deviation to be no less than the prior mean. This implies that $a \leq 1$. It also seems reasonable to require the prior mean to be at least 1, so that small values of λ do not have excessive prior probability. This leads to $b \leq a$. Since the Bayes factor is decreasing in b, the smallest Bayes factor satisfying the above constraints (that is, the one lending the most support for the Poisson model M_0) is then obtained by taking $b = a$ (this gives a prior mean of 1). It is not unreasonable to select this prior as it belongs to a reasonable class which is most favorable to the null model. Finally, one might judge it to be unappealing to utilize a prior for λ which is not bounded near zero (for $a < 1$ the gamma density is decreasing with an asymptote at $\lambda = 0$) which implies that a should be at least 1. Thus we end up with the choice $a = b = 1$. Note that $a = 1$ is the upper limit of $a \leq 1$ and the choice $a = 1$ now counterbalances the Bayes factor in favor of M_1 (whereas $b = a$ in the range $b \leq a$ tilts the Bayes factor in favor of M_0). This reasoning is all rather speculative and, of course, the result is a particular prior, which may not reflect actual prior beliefs. Nevertheless it is instructive to study the behavior of the Bayes factor when this prior is used.

For $a = b = 1$, that is, the Exponential(1) distribution, it can be checked that $B_{10} = \sum_{j=0}^{n}(j+1)^{-1}$,which is thus our recommended default Bayes factor when observing only zero counts. Note that $B_{10}(\mathbf{0}) \approx \log(n+1)$ for large n. So a *large* string of all zero counts in a sample will lead to a Bayes factor approaching infinity at the slow rate of $\log(n)$. The large sample behavior of the Bayes factor for this type of sample seems intuitively reasonable.

5.2. Insufficient Positive Counts in the Regression Case

In the regression situation of Section 4, it was necessary to have sufficient positive counts so that the conditions of Theorem 4.2 were satisfied. We will restrict discussion here to the situation involving the prior specifications in (4.6), for which the key condition needed for the marginal to be finite was that the matrix $\boldsymbol{A}_{+}((n-k) \times q)$ should be of rank q. If the number of positive counts $n - k$ is insufficient so that t, the rank of $\boldsymbol{A}_{+}$, is less than q, this solution will not work.

Remark 5.1. Indeed, neither the prior for $\boldsymbol{\beta}$ given by (4.3) nor by (4.5) guarantees a finite positive marginal density. We omit the proof to save space. A proof may be found in the Appendix in [1].

We call this situation one of rank deficiency, with the rank deficiency of $\boldsymbol{A}_{+}$ equal to $q - t$. The situation is analogous to the case of all zero counts without covariates discussed in Subsection 5.1. (In the setup of that section, $q = 1$ and rank $\boldsymbol{A}_{+}$ less than 1 means that $k = n$, i.e., no positive counts.) We could again merely recognize that this type of data is just not informative enough to allow for objective Bayes analysis. We shall however propose a prior that yields finite marginal densities, following similar reasoning to that used in Section 5.1.

We continue to use a $Un(0, 1)$ prior for p and focus on proposing suitable priors for $\boldsymbol{\beta}$. A discussion similar to that in subsection 5.1 shows that this prior has to be at least, partially proper.

Note that, instead of specifying a prior on $\boldsymbol{\beta}$, we can specify a prior on q independent parametric functions of $\boldsymbol{\beta}$; our specific proposal is to carefully choose these functions such that t of them are well identified by the data with positive counts while the remaining $q-t$ are not. We then propose to use a version of Jeffreys prior on the former t functions, and a proper prior on the latter $q-t$ functions.

Specifically, let $\boldsymbol{A}_0$ denote the $k \times q$ matrix whose k rows are $\boldsymbol{a}_1^T, \ldots, \boldsymbol{a}_k^T$. A rank of $\boldsymbol{A} = q$ and a rank of $\boldsymbol{A}_+ = t$ imply a rank of $\boldsymbol{A}_0 \geq q-t$. Let $V_+ \subseteq R^q$ denote the vector space of dimension t formed by the columns of $\boldsymbol{A}_+^T$. Suppose $\boldsymbol{a}_{i_1}, \ldots, \boldsymbol{a}_{i_r}$ are all of the vectors from $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k$ corresponding to the zero counts which are in V_+. Note that $0 \leq r \leq k-(q-t)$. These vectors are linear combinations of the vectors $\boldsymbol{a}_{j_1}, \ldots, \boldsymbol{a}_{j_t}$ and the corresponding $\lambda_{i_1}, \ldots, \lambda_{i_r}$ are functions of $\lambda_{j_1}, \ldots, \lambda_{j_t}$. From the set of $\{\lambda_j : j \in \{1, \ldots, k\} - \{i_1, \ldots, i_r\}\}$ we select $q-t$ λ's, $\lambda_{l_1}, \ldots, \lambda_{l_{q-t}}$ such that $\{\boldsymbol{a}_{j_1}, \ldots, \boldsymbol{a}_{j_t}, \boldsymbol{a}_{l_1}, \ldots, \boldsymbol{a}_{l_{q-t}}\}$ is linearly independent.

Note that there is an $(n-k) \times t$ matrix $\boldsymbol{C}$ of rank t such that

$$(\boldsymbol{a}_{k+1}, \ldots, \boldsymbol{a}_n) = (\boldsymbol{a}_{j_1}, \ldots, \boldsymbol{a}_{j_t})\boldsymbol{C}^T.$$

Let $\boldsymbol{D} \equiv \boldsymbol{D}(\lambda_{j_1}, \ldots, \lambda_{j_t})$. Then, the information matrix for $\lambda_{j_1}, \ldots, \lambda_{j_t}$ based on the Poisson model for the observations $k+1, \ldots, n$ is given by

$$\boldsymbol{I}(\lambda_{j_1}, \ldots, \lambda_{j_t}) = \boldsymbol{D}^{-1}\boldsymbol{C}^T Diag(\lambda_{k+1}, \ldots, \lambda_n)\boldsymbol{C}\boldsymbol{D}^{-1}. \tag{5.3}$$

We define a partial Jeffreys prior for $\lambda_{j_1}, \ldots, \lambda_{j_t}$ by

$$\pi_{PJ}(\lambda_{j_1}, \ldots, \lambda_{j_t}) = \{\prod_{i=1}^{t} \lambda_{j_i}^{-1}\}|\boldsymbol{C}^T Diag(\lambda_{k+1}, \ldots, \lambda_n)\boldsymbol{C}|^{1/2}. \tag{5.4}$$

Let $\{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{q-t}\}$ denote an orthonormal basis of the space spanned by $\boldsymbol{a}_{l_1}, \ldots, \boldsymbol{a}_{l_{q-t}}$. Define $\xi_w = e^{\boldsymbol{b}_w^T\boldsymbol{\beta}}$, $w = 1, \ldots, q-t$. Note that $\lambda_{l_w}, w = 1, \ldots, q-t$ can be expressed in terms of $\xi_1, \ldots, \xi_{q-t}$. Indeed,

$$\log(\lambda_{l_w}) = a_{0l_w} + \sum_{h=1}^{q-t} d_{wh} \log(\xi_h), \quad w = 1, \ldots, q-t,$$

where $d_{wh} = \boldsymbol{b}_h^T \boldsymbol{a}_{l_w}$. Finally, we assign independent exponential distributions with mean 1 to each of $\xi_1, \ldots, \xi_{q-t}$. This prior will induce a proper distribution on $\lambda_{l_w}, w = 1, \ldots, q-t$ with a density which we denote by $\pi_{prop}(\lambda_{l_1}, \ldots, \lambda_{l_{q-t}})$. The final prior used to calculate the marginal density under model M_1^R is then given by

$$\pi(\lambda_{j_1}, \ldots, \lambda_{j_t}, \lambda_{l_1}, \ldots, \lambda_{l_{q-t}}) = \pi_{PJ}(\lambda_{j_1}, \ldots, \lambda_{j_t})\pi_{prop}(\lambda_{l_1}, \ldots, \lambda_{l_{q-t}})\,;$$

this is partially Jeffreys prior and partially proper. The corresponding prior density on $\boldsymbol{\beta}$ is, of course, obtained through transformation. Further, along the line of the proof of Theorem 4.2, it can be checked that the marginal density $m_1^R(\boldsymbol{x})$ will be finite. We omit the details to save space.

While there is arbitrariness in the specific choice of $\lambda_{l_1}, \ldots, \lambda_{l_{q-t}}$ to assign a subjective prior distribution based on exponential distributions, the partial Jeffreys prior in (5.4) remains invariant to the choice of t independent λ's from $\lambda_{k+1}, \ldots, \lambda_n$. This solution thus seems reasonable for small $q-t$.

To avoid the arbitrariness, we could consider all possible selections of $(q-t)$ of the λ's from $\lambda_1, \ldots, \lambda_k$ so that these $q-t$ and t of the λ's from $\lambda_{k+1}, \ldots, \lambda_n$ define a reparameterization of $\boldsymbol{\beta}$. For each selection we can calculate the Bayes factor, and in the spirit of IBF we can take a suitable average over all these Bayes factors. If the rank deficiency of $\boldsymbol{A}_+$ is 1, we will have $k-r$ Bayes factors to average.

Acknowledgments

The authors would like to thank the Community of Valencia Group in the Project *Desigualdades socioeconómicas y medioambientales en ciudades en España, Proyecto MEDEA* for the data used in Section 4, a referee for valuable comments, and Archan Bhattacharya for computing help. Part of this research was conducted while Bayarri and Datta were visiting SAMSI/Duke University, whose support is gratefully acknowledged. This research is supported in part by NSF under grants DMS-0103265, SES-0241651, and AST-0507481, and by the Spanish Ministry of Education and Science, under Grant MTM2007-61554.

Appendix

Proof of Theorem 4.1: From (4.3) and (4.5) it is immediate that $\pi_1^R(\boldsymbol{\beta}) \leq \pi_0^R(\boldsymbol{\beta})$. Thus it is enough to prove (4.7) for $j = 0$. Let $\boldsymbol{i}$ denote the indices $(i_1, \dots, i_q)$ and $\boldsymbol{A}(\boldsymbol{i})$ denote a $q \times q$ submatrix of $\boldsymbol{A}$ based on rows $i_1, \dots, i_q$. Then by Binet-Cauchy expansion of determinant (cf. Noble, 1969, p. 226) it can be shown that

$$|\sum_{i=1}^{n} \lambda_i \boldsymbol{a}_i \boldsymbol{a}_i^T| = \sum (\lambda_{i_1} \dots \lambda_{i_q}) |\boldsymbol{A}(\boldsymbol{i}) \boldsymbol{A}(\boldsymbol{i})^T|, \tag{A1}$$

where the summation is over all submatrices of order $q \times q$. Dropping the terms from the above summation for which $|\boldsymbol{A}(\boldsymbol{i})\boldsymbol{A}(\boldsymbol{i})^T| = 0$ we get from (4.3) that

$$\pi_0^R(\boldsymbol{\beta}) \leq \sum^{*} (\lambda_{i_1} \dots \lambda_{i_q})^{1/2} |\boldsymbol{A}(\boldsymbol{i}) \boldsymbol{A}(\boldsymbol{i})^T|^{1/2}, \tag{A2}$$

where $\sum^*$ denotes summation over all $q \times q$ matrices for which $|\boldsymbol{A}(\boldsymbol{i})\boldsymbol{A}(\boldsymbol{i})^T| > 0$.

Since $e^{-\lambda_i} \lambda_i^{x_i} / x_i! < 1$, from (4.7) and (A2) we get

$$m_0^R(\boldsymbol{x}) \leq \sum^{*} \int_{R^q} \prod_{j=1}^{q} \{ \frac{e^{-\lambda_{i_j}} \lambda_{i_j}^{x_{i_j}}}{x_{i_j}!} \} (\lambda_{i_1} \dots \lambda_{i_q})^{1/2} |\boldsymbol{A}(\boldsymbol{i}) \boldsymbol{A}(\boldsymbol{i})^T|^{1/2} d\boldsymbol{\beta}. \tag{A3}$$

Recall that $\log(\lambda_i) = a_{0i} + \boldsymbol{a}_i^T \boldsymbol{\beta}$. Now transforming $\boldsymbol{\beta}$ to $(\lambda_{i_1}, \dots, \lambda_{i_q})$ and using the Jacobian of transformation $(\lambda_{i_1} \dots \lambda_{i_q})^{-1} |\boldsymbol{A}(\boldsymbol{i})\boldsymbol{A}(\boldsymbol{i})^T|^{-1/2}$, we get from (A3) that

$$m_0^R(\boldsymbol{x}) \leq \sum^{*} \prod_{j=1}^{q} \int_0^\infty \frac{e^{-\lambda_{i_j}} \lambda_{i_j}^{x_{i_j} - .5}}{x_{i_j}!} d\lambda_{i_j} < \infty, \tag{A4}$$

since each of the integrals in the right hand side of (A4) is finite. This completes the proof of Theorem 4.1.

Proof of Theorem 4.2: First, as in (A1) and (A2), it can be shown that for some positive c (not depending on parameters) less than 1

$$c \sum^{*} (\lambda_{i_1} \dots \lambda_{i_q})^{1/2} |\boldsymbol{A}(\boldsymbol{i}) \boldsymbol{A}(\boldsymbol{i})^T|^{1/2} \leq \pi_0^R(\boldsymbol{\beta}) \leq \sum^{*} (\lambda_{i_1} \dots \lambda_{i_q})^{1/2} |\boldsymbol{A}(\boldsymbol{i}) \boldsymbol{A}(\boldsymbol{i})^T|^{1/2}. \tag{A5}$$

In view of this inequality and (4.10), the marginal $m_1^R(\boldsymbol{x})$ is finite if and only if

$$\int_{R^q} \prod_{i=k+1}^{n} \frac{e^{-\lambda_i} \lambda_i^{x_i}}{x_i!} (\lambda_{i_1} \dots \lambda_{i_q})^{1/2} |\boldsymbol{A}(\boldsymbol{i}) \boldsymbol{A}(\boldsymbol{i})^T|^{1/2} d\boldsymbol{\beta} < \infty \tag{A6}$$

for each $\boldsymbol{i} = (i_1, \dots, i_q)$ for which $|\boldsymbol{A}(\boldsymbol{i})\boldsymbol{A}(\boldsymbol{i})^T| > 0$.

Note that the sufficient condition stated in the theorem and the condition that rank of $\boldsymbol{A}$ is q imply that the regression matrix $\boldsymbol{A}_+^T = (\boldsymbol{a}_{k+1}, \dots, \boldsymbol{a}_n)$ corresponding to the set of positive counts has rank q.

Suppose, with no loss of generality, $i_1 < \cdots < i_q$ in (A6). Also, suppose $i_1 < \cdots < i_u \leq k < i_{u+1} < \cdots < i_q$. It is possible that u may be 0 or may be q. By the assumed condition that for $j = 1, \dots, k$, $\boldsymbol{a}_j$ can be expressed as a linear combination of $\boldsymbol{a}_{k+1}, \dots, \boldsymbol{a}_n$ with nonnegative coefficients, it follows that

$$\lambda_{i_j} = h_{i_j} \prod_{m=k+1}^{n} \lambda_m^{c_{mi_j}}, \quad j = 1, \dots, u,$$

where $c_{mi_j} \geq 0$ and $h_{i_j} > 0$. Then

$$\prod_{j=1}^{u} \lambda_{i_j} = f \prod_{m=k+1}^{n} \lambda_m^{b_m},$$

where $b_m = \sum_{j=1}^{u} c_{mi_j} \geq 0$ and $f > 0$ are free from parameters.

Then the integrand (without $|\boldsymbol{A}(\boldsymbol{i})\boldsymbol{A}(\boldsymbol{i})^T|^{1/2}$) in (A6) can be simplified as

$$\begin{aligned} \prod_{i=k+1}^{n} \frac{e^{-\lambda_i}\lambda_i^{x_i}}{x_i!}(\lambda_{i_1}\dots\lambda_{i_q})^{1/2} &= \prod_{i=k+1}^{n} \frac{e^{-\lambda_i}\lambda_i^{x_i+\frac{1}{2}b_i}}{x_i!}(\lambda_{i_{u+1}}\dots\lambda_{i_q})^{1/2} \\ \text{(A7)} \qquad &= [\prod_{j=u+1}^{q} \frac{e^{-\lambda_{i_j}}\lambda_{i_j}^{x_{i_j}+\frac{1}{2}b_{i_j}+\frac{1}{2}}}{x_{i_j}!}][\prod_{l=1}^{n+u-k-q} \frac{e^{-\lambda_{\alpha_l}}\lambda_{\alpha_l}^{x_{\alpha_l}+\frac{1}{2}b_{\alpha_l}}}{x_{\alpha_l}!}], \end{aligned}$$

where $\{\alpha_1, \dots, \alpha_{n+u-k-q}\} = \{k+1, \dots, n\} - \{i_{u+1}, \dots, i_q\}$.

Suppose $\{s_1, \dots, s_q\} \subset \{k+1, \dots, n\}$ is such that $\{\boldsymbol{a}_{s_1}, \dots, \boldsymbol{a}_{s_q}\}$ is a linearly independent set (such a set exists since $\boldsymbol{A}_+$ is of rank q). Note that for $y > 0$ the function $g(u) = e^{-u}u^y$ is maximized at $u = y$ implying

$$\text{(A8)} \qquad e^{-u}u^y \leq e^{-y}y^y \quad \text{for all } u > 0.$$

By (A8) we get from (A7) that

$$\text{(A9)} \qquad \prod_{i=k+1}^{n} \frac{e^{-\lambda_i}\lambda_i^{x_i}}{x_i!}(\lambda_{i_1}\dots\lambda_{i_q})^{1/2} \leq D(\prod_{j=1}^{q} e^{-\lambda_{s_j}}\lambda_{s_j}^{d_{s_j}}),$$

where $D > 0$ is a constant independent of the parameters and $d_{s_j} = x_{s_j} + \frac{1}{2}b_{s_j} + \frac{1}{2}$ if $s_j \in \{i_{u+1}, \dots, i_q\}$, and $d_{s_j} = x_{s_j} + \frac{1}{2}b_{s_j}$ if $s_j \in \{\alpha_1, \dots, \alpha_{n+u-k-q}\}$.

The Jacobian of transformation from $\boldsymbol{\beta}$ to $\lambda_{s_1}, \dots, \lambda_{s_q}$ is $H/(\lambda_{s_1}\dots\lambda_{s_q})$ for some $H > 0$ constant. Then since $d_{s_j} \geq 1$ for $j = 1, \dots, q$, by (A9) we have

$$\text{(A10)} \quad \int_{R^q} \prod_{i=k+1}^{n} \frac{e^{-\lambda_i}\lambda_i^{x_i}}{x_i!}(\lambda_{i_1}\dots\lambda_{i_q})^{1/2} d\boldsymbol{\beta} \leq HD \prod_{j=1}^{q} \int_0^{\infty} e^{-\lambda_{s_j}}\lambda_{s_j}^{d_{s_j}-1} d\lambda_{s_j} < \infty \, .$$

By (A10) and (A6) we conclude that $m_1^R(\boldsymbol{x})$ corresponding to $\pi_0^R(\boldsymbol{\beta})$ is finite. To prove finiteness of $m_1^R(\boldsymbol{x})$ corresponding to $\pi_1^R(\boldsymbol{\beta})$ note that by (4.10)

$$m_1^R(\boldsymbol{x}) \leq \int_{R^q} (\prod_{i=k+1}^{n} \frac{e^{-\lambda_i}\lambda_i^{x_i}}{x_i!})\pi_1^R(\boldsymbol{\beta}) d\boldsymbol{\beta} \, .$$

Finiteness of the right hand quantity in the last display follows from a version of Theorem 4.1 corresponding to the prior $\pi_0^R(\boldsymbol{\beta})$ by replacing n observations from the Poisson by $n-k$ observations from Poisson. This completes the proof.

References

[1] Bayarri, M.J., Berger, J.O. and Datta, G.S. (2007). Objective Bayes testing of Poisson versus inflated Poisson models. Technical Report. Department of Statistics, University of Georgia, Athens, GA 30602, USA.

[2] Bayarri, M.J. and García-Donato, G. (2007). Extending conventional priors for testing general hypotheses in linear models. *Biometrika*, **94**, 135-152.

[3] Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag.

[4] Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, **1**, 385-402.

[5] Berger, J.O. and Pericchi, L.R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109-122.

[6] Berger, J.O. and Pericchi, L.R. (2001). Objective Bayesian methods for model selection: introduction and comparison (with discussion). In *Model Selection, Institute of Mathematical Statistics Lecture Notes- Monograph Series*, **38**, Ed. P. Lahiri, pp. 135-207, Beachwood Ohio: Institute of Mathematical Statistics.

[7] Berger, J., Pericchi, L. and Varshavsky, J. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhya A*, **60**, 307-321.

[8] Berger, J. and Sun, D. (2008). Objective priors for a bivariate normal model with multivariate generalizations. To appear in *Annals of Statistics.*

[9] Bhattacharya, A., Clarke, B.S. and Datta, G.S. (2007). A Bayesian test for excess zeros in a zero-inflated power series distribution. In *Beyond Parametrics in Interdisciplinary Research: Festschrift in honour of Professor Pranab K. Sen.* **IMS Lecture Notes and Monographs Series**. Eds.: N. Balakrishnan, E. Peña and M. Silvapulle. Institute of Mathematical Statistics.

[10] Broek, J.V.D. (1995). A score test for zero inflation on a Poisson distribution. *Biometrics*, **51**, 738-743.

[11] Conigliani, C., Castro, J. I. and O'Hagan, A. (2000). Bayesian assessment of goodness of fit against nonparametric alternatives. *Canadian Journal of Statistics*, **28**, 327-342.

[12] Deng, D. and Paul, S.R. (2000). Score test for zero inflation in generalized linear models. *Canadian Journal of Statistics*, **28**, 563-570.

[13] Ghosh, J.K. and Samanta, T. (2002). Nonsubjective Bayes testing - an overview. *Journal of Statistical Planning and Inference*, **103**, 205-223.

[14] Ghosh, S.K., Mukhopadhyay, P. and Lu, J.C. (2006). Bayesian analysis of zero-inflated regression models. *Journal of Statistical Planning and Inference*, **136**, 1360-1375.

[15] Jeffreys, H. (1961). *Theory of Probability, 3rd ed.* London: Oxford University Press.

[16] Kass, R.E. and Vaidyanathan, S. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society B*, **54**, 12944.

[17] Kass, R.E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343-1370.

[18] Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1-14.

[19] Noble, B. (1969). *Applied Linear Algebra.* Prentice-Hall, New York.

[20] O'Hagan, A. (1995). Fractional Bayes factors for model comparisons. *Journal of the Royal Statistical Society, Ser. B*, **57**, 99-138.

[21] Pérez, J.M. and Berger, J.(2001). Analysis of mixture models using expected posterior priors, with application to classification of gamma ray bursts. In *Bayesian Methods, with applications to science, policy and official statistics*, E. George and P. Nanopoulos, eds., Official Publications of the European Communities, Luxembourg, 401–410.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 123

Consistent selection via the Lasso for high dimensional approximating regression models

Florentina Bunea [1]

Florida State University

Abstract: In this article we investigate consistency of selection in regression models via the popular Lasso method. Here we depart from the traditional linear regression assumption and consider approximations of the regression function f with elements of a given dictionary of M functions. The target for consistency is the index set of those functions from this dictionary that realize the most parsimonious approximation to f among all linear combinations belonging to an L_2 ball centered at f and of radius $r^2_{n,M}$. In this framework we show that a consistent estimate of this index set can be derived via ℓ_1 penalized least squares, with a data dependent penalty and with tuning sequence $r_{n,M} > \sqrt{\log(Mn)/n}$, where n is the sample size. Our results hold for any $1 \leq M \leq n^\gamma$, for any $\gamma > 0$.

Contents

1. Introduction

In this paper we show that the popular Lasso technique can be used for consistent feature selection in high dimensional approximating regression models. We consider the following framework. Given a random pair (X, Y), we let $f(x) = E(Y|X = x)$ be the conditional mean function, henceforth called the regression function. We aim to reconstruct consistently a sparse approximation of f via linear combinations of elements of a given dictionary of functions $\mathcal{F} = \{f_1, \ldots, f_M\}$. This reconstruction will be based on $(X_1, Y_1), \ldots, (X_n, Y_n)$, a sample of independent random pairs distributed as $(X, Y) \in (\mathcal{X}, \Re)$, where $\mathcal{X}$ is a Borel subset of $\Re^d$; all functions f_j are defined on $\mathcal{X}$. Our aim expresses the belief that, in many instances, even if M is large, only a subset of $\mathcal{F}$ may be needed to approximate f well. If that is the case, it

[1] Department of Statistics, Florida State University, Tallahassee, Florida, e-mail: flori@stat.fsu.edu

AMS 2000 subject classifications: Primary 62G08; Secondary 62C20, 62G05, 62G20

Keywords and phrases: Lasso, l_1 regularization, penalty, consistency, selection, regression, high dimension.

may be of interest to determine whether this set can be estimated consistently via a computationally efficient method. The focus of this work is on consistent selection via the Lasso when the size of $\mathcal{F}$ grows polynomially with the sample size n, that is $M = n^\gamma$, for any $\gamma > 0$.

We begin by giving a number of examples of dictionaries $\mathcal{F}$ and associated consistency issues.

1. If $d = M$ and $f_j(X) = X_j$ for all j, one may be interested in identifying the subset of variables with linear combinations close to f. A familiar particular case is linear regression, where one assumes that $f(X) = \lambda' X$, with $\lambda \in \Re^M$ having non-zero components in positions corresponding to a set $J^* \subseteq \{1, \ldots, M\}$. Here we depart from this traditional equality assumption and consider the more realistic case where f is not equal to, but can be well approximated by a linear combination of the given variables. We discuss this in detail in the next section.
2. Another problem of interest is that of finding consistently a sparse linear approximation of f realized with elements from a large list of M possibly competing estimators. These estimates may correspond to M different methods of estimation, may be computed from M different samples with the same mean function, or may correspond to M different values of a tuning parameter of the same method. Instances of the latter arise in kernel based methods that require the choice of a grid of values for the bandwidth parameter or in Bayesian methods, where the specification of a grid of values for hyperparameters is needed. A consistent identification of a subset of the estimates in these examples would validate the use of a particular restriction on an initially large grid. In such situations, when the elements of $\mathcal{F}$ are estimators, we will assume that they have been computed on samples independent of the one used for subset selection and treat them here as fixed functions.
3. A last example is the nonparametric estimation of f from a collection of M given basis functions, where only a subset may realize a good approximation of f, as described in the following subsection.

There exist a number of model selection methods that yield consistent subset selection in regression models. In discussing them a number of distinctions are needed.

The first one pertains to the evolution of the literature on model selection techniques in regression. One important cut-off point in this evolution seems to be the computational complexity of a particular method and, within this, the size of M relative to n plays a crucial role. If $M \le n$, procedures based on various information criteria occupy an important place. They are referred to now as the BIC/AIC-type methods; we mention here the seminal works of ([1], [15]), the unifying theory of [2], and, various generalizations of these methods ([4], [7]). Such procedures can be easily implemented for small to moderate M. For larger values of M multiple testing procedures, in particular of the FDR type (e.g., [3], [9]), or cross-validation with all its variants (holdout validation, K-fold) [21], are popular, but become more computationally complex as M increases. If $M > n$ these techniques may become computationally intractable, unless they are used as part of a multiple-stage scheme. For a further overview on computational aspects in model selection, from a Bayesian perspective, see [11].

Whereas the above mentioned methods can still be used in very particular regression models when $M > n$, for instance, for sequence-space models, where model selection via BIC is equivalent to hard thresholding, they typically fail, computationally, when M is large. A standard solution in this case is to seek estimates that solve a certain class of convex optimization problems. Among the most popular estimates of this type in regression is the penalized least squares estimate with an ℓ_1-type penalty (Lasso), which we describe in detail in the next section. In a Bayesian framework it can be derived from a Gaussian likelihood with a Laplace prior. Two important aspects set the ℓ_1 regularized (Lasso) type estimators apart: they are easy and fast to compute; see [8], [13], [14], [18], among others, for efficient algorithms; and, if $M > n$, some components of the estimate will be set to zero, in finite samples, see, e.g., [13]. Therefore, via a one-step easily implementable procedure, one obtains subset selection even if $M > n$. To date, this method (or its variants) is the most widely used in regression problems of very high dimension, especially when dimension reduction is of interest.

The second distinction in discussing consistency of selection in regression is related to the target for consistency. Consistency of selection has been studied for all aforementioned techniques *only* in the following context, which we term parametric: the target for selection is typically an index set J^* corresponding to the non-zero true regression coefficients, whereas the remaining coefficients are assumed to be *exactly* zero. An estimation method that uses the data and all M elements f_j to yield a subset $\hat{I}$ of indices such that $P(\hat{I} = J^*) \to 1$ for large n is called a consistent method of selection.

In light of these two distinctions we give below a summary of the existing results on consistency of selection. They have all been established for the traditional parametric target J^*.

If $M \leq n$ and under appropriate assumptions all the above methods, or close variants, yield consistent subset selection for the parametric target J^*. References include those for AIC/BIC-type methods ([4], [10], and [22], among others), multiple testing procedures [5], cross-validation procedures [16], and Lasso-type procedures [24].

If $M > n$ consistency of selection has only been studied for Lasso-type estimators. Again, in the existing literature, the target is the standard target J^*. The results are limited. Meinshausen and Buhlmann [12] showed that $P(\hat{I} = J^*) \to 1$ in Gaussian graphical models, under assumptions that are tailored to models for which, in our notations, $(Y, X_1, \ldots, X_M) \sim N(0, \Sigma)$. Consistency of selection has been established when $M > n$, for fixed design linear regression models and a target set J^* that corresponds to coefficients λ_j^* that are assumed to be lower bounded by a sequence of order $O(n^{-\delta/2})$, for $0 < \delta < 1$ [23]. Similar results, under slightly different assumptions, have also been obtained for a three stage procedure [20]: in the first stage Lasso estimates are computed for a number of values of the tuning parameter, in the second step cross-validation is performed to select one Lasso estimate, and in the third one the model is refitted on the variables present in the selected Lasso estimate. We also refer to a related notion of consistency, in fixed design regression with Gaussian errors [19].

If $M > n$ consistent subset selection via the Lasso has not been investigated, to the best of our knowledge, in the general framework we describe in detail below. Within this framework, we extend the existing results to more general regression models on a random design and a more general target index set.

1.1. Beyond Linear Regression

Despite its practical appeal, the study of selection procedures that are consistent for target sets other than the classical one has received very little attention. Our target set will be defined relative to linear approximations of f with elements of $\mathcal{F}$ with respect to the $L_2(\nu)$ norm $\| \ \|$, where we denote the probability measure of X by ν.

Formally, define

$$\Lambda = \left\{ \lambda \in \Re^M : \ \| \sum_{j=1}^{M} \lambda_j f_j - f \|^2 \leq C_f r_{n,M}^2 \right\}, \tag{1.1}$$

where $C_f > 0$ is a constant depending only on f and $r_{n,M}$ is a positive sequence that converges to zero and which will be specified in the next section. In what follows we assume that Λ is not void. For any $\lambda \in \Re^M$ we let $J(\lambda)$ denote the index set corresponding to the non-zero components of λ and denote by $M(\lambda)$ its cardinality. Let $k^* = \min\{M(\lambda) : \ \lambda \in \Lambda\}$. We define our target vector

$$\lambda^* = \text{argmin} \left\{ \| \sum_{j=1}^{M} \lambda_j f_j - f \|^2 : \ \lambda \in \mathbb{R}^M, \ M(\lambda) = k^* \right\}. \tag{1.2}$$

Let $I^* = J(\lambda^*)$ denote the index set corresponding to the non-zero elements of λ^* and note that I^* has cardinality k^*. Thus $f^* = \sum_{j \in I^*} \lambda_j^* f_j$ provides the sparsest approximation to f that can be realized with $\lambda \in \Lambda$ and, in particular,

$$\|f^* - f\|^2 \leq C_f r_{n,M}^2. \tag{1.3}$$

This motivates our treating I^* as the target index set.

We note that if one assumes, as in standard linear regression models, that $f(x) = \sum_{j=1}^{M} \lambda_j x_j = \sum_{j \in I^*} \lambda_j^* x_j = f^*(x)$, where λ_j^* denotes the non-zero components of λ, then (1.3) is trivially satisfied for any positive sequence $r_{n,M}$. Therefore, the classical target J^* is a particular case of ours.

In order to ensure that λ^* captures the essential features of f in a parsimonious way we require that its components not be unnecessarily small, otherwise we can place their indices outside I^*. Formally, we will require that the following condition holds.

Condition (C): There exists $B > 0$, independent of n or M, such that

$$\min_{j \in I^*} |\lambda_j^*| > B r_{n,M}.$$

We show below that ℓ_1 penalized least squares can be used to estimate consistently the new target I^*, even if M is larger than n, in particular if it grows as n^γ, for any $\gamma > 0$, under minimal assumptions on the dictionary $\mathcal{F}$ and appropriate choices for $r_{n,M}$. In Section 2 below we introduce the estimate and discuss these choices. Section 2.1 contains our main result, Theorem 2.1, together with a discussion of the assumptions under which it holds. The proof of the main result is given in Section 2.2 and intermediate results are proved in the Appendix.

2. Consistent Selection via ℓ_1 Penalized Least Squares

We estimate the set I^* of the previous section via ℓ_1 penalized least squares. We first compute

$$\widehat{\lambda} = \mathop{\arg\min}_{\lambda \in \Re^M} \left\{ \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \sum_{j=1}^{M} \lambda_j f_j(X_i)\}^2 + \mathrm{pen}(\lambda) \right\}, \tag{2.4}$$

where

$$\mathrm{pen}(\lambda) = 2 \sum_{j=1}^{M} \omega_{n,j} |\lambda_j| \quad \text{with} \quad \omega_{n,j} = r_{n,M} \|f_j\|_n, \tag{2.5}$$

for a sequence $r_{n,M}$ given below, where we write $\|g\|_n^2 = n^{-1} \sum_{i=1}^n g^2(X_i)$ for any function $g : \mathcal{X} \to \Re$. We note that each λ_j in the penalty term has a different, data-dependent, weight. The estimate $\hat{\lambda}$ thus obtained is in one-to-one correspondence with the following estimate. For each $1 \leq j \leq M$ define $\theta_j = 2\omega_{n,j}\lambda_j$ and let A be the $M \times M$ diagonal matrix with diagonal entries $2\omega_{n,j}$. Next observe that $F\lambda = F_1\theta$, where F is the $n \times M$ matrix with entries $f_j(X_i)$, $F_1 = FA^{-1}$ and $\theta = A\lambda$. Thus, denoting by Y the n dimensional vector with entries Y_i, the problem reduces to calculating

$$\widehat{\theta} = \mathop{\arg\min}_{\theta \in \Re^M} \frac{1}{n} (Y - F_1\theta)'(Y - F_1\theta) + \sum_{j=1}^{M} |\theta_j|,$$

for which the aforementioned fast algorithms can be used. Then, we compute our sought solution $\hat{\lambda} = A^{-1}\hat{\theta}$.

We let $\hat{I}$ denote the index set corresponding to the non-zero components of $\hat{\lambda}$. We show in the next subsection that $P(\hat{I} = I^*) \to 1$ when $n \to \infty$. We begin by noticing that we always have

$$P(\hat{I} = I^*) \geq 1 - P(I^* \not\subseteq \hat{I}) - P(\hat{I} \not\subseteq I^*).$$

Therefore, proving that $\hat{I}$ is consistent reduces to showing that each of the probabilities in the right-hand side of the inequality above converge to zero. In what follows we motivate choices for the sequence $r_{n,M}$ that stem from sufficient conditions under which this convergence is achieved. The proofs are presented in the next section.

We begin by noticing that if $\hat{\lambda} \to \lambda^*$, with probability converging to one, then $I^* \not\subseteq \hat{I}$ with probability converging to zero. To see this, further note that if component-wise consistency of $\hat{\lambda}$ holds, we will estimate *all* non-zero elements of λ^* by non-zero sequences, but we may also estimate some of its zero components by some small, but non-zero sequences. In light of this fact, a first set of restrictions on $r_{n,M}$ will be such that $\hat{\lambda}$ is close to λ^*, in the sense below. It follows immediately (by [5], Theorem 2.3; see the Appendix below for a full formulation) that, with high probability

$$r_{n,M} |\hat{\lambda} - \lambda^*|_1 \leq D\{\|f - f^*\|^2 + k^* r_{n,M}^2\},$$

for some positive constant D, and where $|a|_1 = \sum_{j=1}^M |a_j|$ denotes the ℓ_1 norm of any vector in $\Re^M$. Next, notice that the optimal parametric rate of convergence for a component $\hat{\lambda}_j$ of $\hat{\lambda}$ is of order $1/\sqrt{n}$, and it can be achieved if we knew I^* of cardinality $k^* < M$ in advance. However, this is not known, so the best we can do is mimic this behavior in our context. We can do this by choosing $r_{n,M}$ of order $1/\sqrt{n}$, where we recall that we have assumed that $\|f - f^*\|^2 \leq r_{n,M}^2$. Notice further that this choice is optimal for the rate of convergence of $\hat{\lambda}$, which is not the focus here. Indeed, more modest rates of convergence of $\hat{\lambda}$ can be considered when consistency of selection is of main importance. We discuss in detail two concrete choices, and defer a complete analysis for future work.

One can consider $r_{n,M} = A\sqrt{\log(Mn)/n}$, for an appropriately large constant $A > 0$. Notice that this choice differs from the one that yields the optimal rate only by logarithmic factors, which are needed to accommodate dictionaries with $M > n$. With this choice, the target set I^* corresponds to linear combinations of the elements of $\mathcal{F}$ that belong to, up to logarithmic factors, a $1/\sqrt{n}$ neighborhood of f, with respect to the $L_2(\nu)$ norm. This provides only a slight departure from the standard linear model assumption and standard target index set J^*. It is therefore not surprising that, in this case, our tuning sequence $r_{n,M}$ is also comparable to the one considered in parametric models ([12], [23]), where a sequence of the order of $1/n^{1/2-\theta}$, $\theta \in (0, 1/2)$, is employed. We note that this choice is slightly conservative, and can be relaxed to $O(\sqrt{\log(Mn)/n})$ in our framework, and therefore, as a particular case, in theirs.

In order to accommodate consistent selection in a purely nonparametric framework we need to increase the size of $r_{n,M}$. For instance, if all f_j are estimates of f, and $r_{n,M}$ is as before, the set Λ defined in (1.1) may be empty, as non-parametric estimates of f have typically slower rates than $1/\sqrt{n}$. We therefore consider target sets I^* corresponding to $L_2(\nu)$ neighborhoods around f of radius $r_{n,M}^2$, now with $r_{n,M} = O\left((\log(Mn)/n)^{1/4}\right)$. In this case, the set Λ given in (1.1) above is not empty if at least one of the estimators f_j has, up to logarithmic factors, a rate of the order $n^{-1/4}$, which is a modest rate to require. Of course, if $f_j(X) = X_j$, as in linear regression, this choice means that we may be content with a coarser approximation than before. However, note that this approximation has the benefit of being realized with a smaller number of variables and that this may increase the interpretability of that particular model and be a desirable property in practical situations.

The results presented below hold for either of these choice, in particular for any $r_{n,M} \geq A\sqrt{\log(Mn)/n}$, and we will therefore not distinguish between them.

2.1. Main Result: Consistent Subset Selection

We begin by listing and commenting on the assumptions under which our result holds. The first assumption refers to the error terms $W_i = Y_i - f(X_i)$. We recall that $f(X) = E(Y|X)$.

Assumption (A1). *The random variables $X_1, \ldots, X_n$ are independent, identically distributed random variables with probability measure μ. The random variables W_i are independently distributed with*

$$E\{W_i \,|\, X_1, \ldots, X_n\} = 0$$

and

$$E\left\{\exp(|W_i|) \,|\, X_1, \ldots, X_n\right\} \leq b \quad \textit{for some finite } b > 0 \textit{ and } i = 1, \ldots, n.$$

We also impose mild conditions on f and on the functions f_j. Let $\|g\|_\infty = \sup_{x\in\mathcal{X}} |g(x)|$ for any function g on $\mathcal{X}$.

Assumption (A2).

(a) There exists $0 < L < \infty$ such that $\|f_j\|_\infty \le L$ for all $1 \le j \le M$.
(b) There exists $c_0 > 0$ such that $\|f_j\| \ge c_0$ for all $1 \le j \le M$.
(c) There exists $L_0 < \infty$ such that $E[f_i^2(X)f_j^2(X)] \le L_0$ for all $1 \le i, j \le M$.
(d) There exists $L_1 < \infty$ such that $\|f\|_\infty \le L_1 < \infty$.
(e) There exists $L^* < \infty$ such that $\|f - f^*\|_\infty \le L^*$.

Remark 2.1 We note that (a) trivially implies (c). However, as the implied bound may be too large, we opted for stating (c) separately. Note also that (a) and (d) imply the following: for any fixed $\lambda \in \Re^M$, there exists a positive constant $L(\lambda)$, depending on λ, such that $\|f - \sum_{j=1}^M \lambda_j f_j\|_\infty = L(\lambda)$. Inspection of the proof of Theorem 2.1 below shows that we can allow L^* to grow very slowly with n. However, for sake of clarity in presentation we opted for treating it as fixed.

Assumption (A3). *Let*

$$\rho_M(i,j) = \frac{< f_i, f_j >}{\|f_i\|\|f_j\|},$$

where $< f_i, f_j >= Ef_i(X)f_j(X)$ and $\|f_i\| = E^{1/2} f_i^2(X)$. Assume that

$$\max_{i\in I^*}\max_{j\neq i} |\rho_M(i,j)| \le \frac{C}{k^*},$$

for some constant $C > 0$.

Remark 2.2 Following [6], $C = 1/45$ is an allowable choice. Other choices are possible, but improvement of constants is beyond the scope of this paper.

Remark 2.3 Assumption (A3) reflects the belief that the correlations between functions f_j with $j \in I^*$ and functions f_j with $j \notin I^*$ should be small. However, we allow the correlations outside I^* to be arbitrary. We note that this assumption replaces the standard orthonormality assumption on the design matrix: it is given in terms of theoretical quantities and it can hold even if $M > n$. It can be checked in practice by replacing the theoretical correlations by sample correlations.

We denote by G the event that the $n \times M$ matrix F with entries $f_j(X_i)$ has full rank. To avoid additional technicalities, the results of this paper can be regarded as conditional on G. Otherwise, all the results can be re-derived by intersecting all the relevant events with G and G^c, under the additional assumption that $P(G^c)$ is appropriately small.

We can now state our main result which we prove in the next subsection.

Theorem 2.1. *If assumptions (A1)-(A3) and condition (C) hold, and $k^* r_{n,M} \to 0$ then $P(\hat{I} = I^*) \to 0$.*

Remark 2.4 The convergence above holds either if M is fixed and $n \to \infty$ or if both $M, n \to \infty$, if $r_{n,M} \ge A\sqrt{\log(Mn)/n}$ for an appropriately large constant A. Therefore we obtain consistency for both choices of $r_{n,M}$ discussed above. In our derivations we require that M does not grow faster than a power of n.

Remark 2.5 The condition $r_{n,M}k^* \to 0$ imposes restrictions on the size of k^*. If $r_{n,M} = O(\sqrt{\log(Mn)/n})$ the theorem above shows that we can recover consistently subsets of size $k^* = O(\sqrt{n}/\log n)$, up to other logarithmic factors. The choice $r_{n,M} = O(\log(Mn)/n)^{1/4}$ corresponds to a coarser approximation of f than before, and the restriction on the number of approximating functions is now $k^* = O(n^{1/4}/\log n)$.

2.2. Proof of Theorem 2.1

Recall that

$$P(\hat{I} = I^*) \geq 1 - P(I^* \not\subseteq \hat{I}) - P(\hat{I} \not\subseteq I^*).$$

Therefore, proving that $\hat{I}$ is consistent reduces to showing that each of the probabilities in the right hand side of the inequality above converge to zero. We present this in the following two propositions. We defer the proof of the intermediate results to the Appendix.

Proposition 2.2. *If assumptions (A1)-(A3) and condition (C) hold, and $r_{n,M}k^* \to 0$, then $P(I^* \not\subseteq \hat{I}) \to 0$ as $n \to \infty$, for any $r_{n,M} \geq A\sqrt{\log(Mn)/n}$, with $A > 0$ large enough.*

Proof. We follow the same reasoning as [4]. Let $c_n = \min_{k \in I^*} |\lambda_k^*|$ and recall that $c_n > Br_{n,M}$, by condition (C). Therefore

$$\begin{aligned} P(I^* \not\subseteq \hat{I}) &\leq P(j \notin \hat{I} \text{ for some } j \in I^*) \\ &\leq P(|\widehat{\lambda}_j - \lambda_j^*| = |\lambda_j^*|) \\ &\leq P(|\widehat{\lambda}_j - \lambda_j^*| > c_n) \to 0, \text{ as } n \to \infty \end{aligned}$$

where, in the second inequality, we used that $\widehat{\lambda}_j = 0$ for $j \notin \hat{I}$, by the definition of $\hat{I}$. The last inequality follows from Corollary 1 presented in the Appendix below. ■

Proposition 2.3. *If assumptions (A1)-(A3) hold and $r_{n,M}k^* \to 0$, then $P(\hat{I} \not\subseteq I^*) \to 0$, as $n \to \infty$, for any $r_{n,M} \geq A\sqrt{\log(Mn)/n}$, with $A > 0$ large enough.*

Proof. Let

$$h(\mu) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \sum_{j \in I^*} \mu_j f_j(X_i)\}^2 + 2r_{n,M} \sum_{j \in I^*} ||f_j||_n |\mu_j|,$$

and define

$$\tilde{\mu} = \arg\min_{\mu \in \Re^{k^*}} h(\mu). \tag{2.6}$$

Let

$$\mathcal{B} = \bigcap_{k \notin I^*} \left\{ |\frac{2}{n}\sum_{i=1}^{n}[Y_i - \sum_{j \in I^*} \tilde{\mu}_j f_j(X_i)] f_k(X_i)| < 2r_{n,M}||f_k||_n \right\}.$$

Let $\tilde{\lambda} \in \Re^M$ be the vector that has the components of $\tilde{\mu}$ in positions corresponding to the index set I^* and components equal to zero otherwise. Thus, by abuse of notation, $\tilde{\lambda} = (\tilde{\mu}, 0)$. From Lemma 3.4 in the Appendix it follows that, on the set $\mathcal{B}$, $\tilde{\lambda}$ is a solution of (2.4). Recall that $\hat{\lambda}$ is a solution of (2.4) by construction. Then, by arguments similar to those used in ([13], Theorems 3.1 and 3.2) regarding the closeness of two solutions it follows that, on the set $\mathcal{B}$, $\hat{\lambda}_k = 0$ for $k \in {I^*}^c$. Therefore $\hat{I} \subseteq I^*$ on the set $\mathcal{B}$. Hence

$$\begin{aligned} & P(\hat{I} \not\subseteq I^*) \leq P(\mathcal{B}^c) \\ = \; & P\left(\bigcup_{k \in \{1,\dots,M\} \setminus I^*} \left\{ |\frac{2}{n}\sum_{i=1}^n [Y_i - \sum_{j \in I^*} \tilde{\mu}_j f_j(X_i)] f_k(X_i)| \geq 2r_{n,M} ||f_k||_n \right\} \right) \\ \leq \; & \sum_{k \in \{1,\dots,M\} \setminus I^*} P\left(\left\{ |\frac{2}{n}\sum_{i=1}^n [Y_i - \sum_{j \in I^*} \tilde{\mu}_j f_j(X_i)] f_k(X_i)| \geq 2r_{n,M} ||f_k||_n \right\} \right). \end{aligned}$$

Let $k \in \{1,\dots,M\} \setminus I^*$ be fixed. Define the sets

$$\begin{aligned} E_1(k) &= \left\{ \frac{1}{n} |\sum_{i=1}^n W_i f_k(X_i)| < r_{n,M} \|f_k\|_n / 2 \right\}, \\ E_2(k) &= \left\{ \|f_k\|_n^2 \geq \frac{1}{4} \|f\|^2 \right\}, \\ E_3(k) &= \left\{ |\frac{1}{n}\sum_{i=1}^n f_j(X_i) f_k(X_i)| \leq 2|\langle f_j, f_k \rangle| + \delta_{n,M}, \ j \in I^* \right\}, \end{aligned}$$

where $\delta_{n,M} = 2CL^2 r_{n,M}$ will be specified below. The choice of $\delta_{n,M}$ is purely technical and does not affect the overall results.

Let $\tilde{f} = \sum_{j \in I^*} \tilde{\mu}_j f_j$. Recall that $\lambda^* \in R^M$ given by (1.2) has zero components in positions corresponding to indices in ${I^*}^c$, by definition. Let μ^* be the vector in $\Re^{k^*}$ obtained from λ^* by deleting these zeros. Therefore $f^* = \sum_{j=1}^M \lambda_j^* f_j = \sum_{j \in I^*} \mu_j^* f_j$. By successive applications of the triangle inequality and since $\|f_k\|_n \leq L$, for all $k \in {I^*}^c$, by assumption (A2) (a), we obtain:

$$\begin{aligned} & P\left(\frac{1}{n} |\sum_{i=1}^n [Y_i - \sum_{j \in I^*} \tilde{\mu}_j f_j(X_i)] f_k(X_i)| \geq r_{n,M} \|f_k\|_n \right) \\ \leq \; & P\left(\frac{1}{n} |\sum_{i=1}^n W_i f_k(X_i)| \geq r_{n,M} \|f_k\|_n / 2 \right) \\ & + P\left(\frac{1}{n} |\sum_{i=1}^n (f(X_i) - \tilde{f}(X_i)) f_k(X_i)| \geq r_{n,M} \|f_k\|_n / 2 \right) \\ \leq \; & P(E_1^c(k)) + P\left(\frac{1}{n} |\sum_{i=1}^n (f^*(X_i) - \tilde{f}(X_i)) f_k(X_i)| \geq r_{n,M} \|f_k\|_n / 4 \right) \\ & \qquad + P\left(\frac{1}{n} \sum_{i=1}^n |(f(X_i) - f^*(X_i))| \geq r_{n,M} \|f_k\|_n / 4L \right) \\ \leq \; & P(E_1^c(k)) + P\left(|(\sum_{j \in I^*} (\tilde{\mu}_j - \mu_j^*) \frac{1}{n} \sum_{i=1}^n f_j(X_i)) f_k(X_i)| \geq r_{n,M} \|f_k\|_n / 4 \right) \\ & \qquad + P\left(\frac{1}{n} \sum_{i=1}^n |(f(X_i) - f^*(X_i))| \geq r_{n,M} \|f_k\|_n / 4L \right). \end{aligned} \tag{2.7}$$

To bound the second term in the last inequality above we first notice that on the set $E_3(k)$ and under assumptions (A2) (a) and (A3) we have

$$\begin{aligned}
&|\sum_{j\in I^*}(\tilde{\mu}_j-\mu_j^*)\frac{1}{n}\sum_{i=1}^{n}f_j(X_i))f_k(X_i)| \\
&\qquad\qquad\leq 2\sum_{j\in I^*}|\tilde{\mu}_j-\mu_j^*||\langle f_j,f_k\rangle| + \delta_{n,M}\sum_{j\in I^*}|\tilde{\mu}_j-\mu_j^*| \\
&\qquad\qquad\leq \frac{2CL^2}{k^*}|\tilde{\mu}-\mu^*|_1+\delta_{n,M}|\tilde{\mu}-\mu^*|_1.
\end{aligned}$$

Therefore, on $E_2(k)\cap E_3(k)$, and under assumptions (A2), (a) and (b), and (A3) we have

$$\begin{aligned}
&P(|(\sum_{j\in I^*}(\tilde{\mu}_j-\mu_j^*)\frac{1}{n}\sum_{i=1}^{n}f_j(X_i))f_k(X_i)|\geq r_{n,M}\|f_k\|_n/4) \\
\leq\ & P(|\tilde{\mu}-\mu^*|_1\geq\frac{c_0}{32CL^2}k^*r_{n,M})+P(|\tilde{\mu}-\mu^*|_1\geq\frac{c_0}{16}r_{n,M}\delta_{n,M}^{-1}) \\
\leq\ & 2P(|\tilde{\mu}-\mu^*|_1\geq\frac{c_0}{32CL^2}k^*r_{n,M}), \qquad (2.8)
\end{aligned}$$

for n large enough, since the assumption $k^*r_{n,M}\to 0$ implies that $k^*r_{n,M}\leq 1$ for large n, and we recall that we defined $\delta_{n,M}=2CL^2r_{n,M}$.

Lastly, notice that on the set $E_2(k)$ and under assumption (A2) (b) and (e) the third term of the last inequality in display (2.7) can be bounded by

$$P(\frac{1}{n}\sum_{i=1}^{n}|(f(X_i)-f^*(X_i))|\geq\frac{c_0}{8L}r_{n,M}). \tag{2.9}$$

To complete the proof we need to show that $P(E_1^c(k)), P(E_2^c(k))$ and $P(E_3^c(k))$ and the probabilities in (2.8) and (2.9), when summed over $k\in\{1.,\ldots,M\}\setminus I^*$, converge to zero as $n\to\infty$. We show this in Lemma 3.5, Corollary 2 and Lemma 3.6, respectively, in the Appendix below. This completes the proof of this result. ■

Appendix

In order to show Proposition 2.2 and to bound (2.8) above we will use twice ([6], Theorem 2.3 page 177) and we begin by stating it here, for completeness. For any $\lambda\in\Re^M$ we let $J(\lambda)$ denote the index set corresponding to the non-zero components of λ and denote by $M(\lambda)$ its cardinality. Let $\rho(\lambda)=\max_{i\in J(\lambda)}\max_{j\neq i}|\rho_M(i,j)|$. With Λ given by (1.1) in Section 1.1, let $\Lambda_1=\{\lambda\in\Lambda:\rho(\lambda)\leq C/M(\lambda)\}$.

Theorem 2.3 [6]. *Assume that (A1) and (A2) hold. Then the ℓ_1 penalized least squares estimator $\hat{\lambda}$ given by (2.4) satisfies, for any $\lambda\in\Lambda_1$*

$$P\left\{|\widehat{\lambda}-\lambda|_1\leq B_1r_{n,M}M(\lambda)\right\}\geq 1-\pi_{n,M}(\lambda), \tag{3.10}$$

where

$$\begin{aligned}
\pi_{n,M}(\lambda)\quad\leq\quad & 14M^2\exp\left(-c_1n\,\min\left\{\frac{r_{n,M}^2}{L_0},\frac{r_{n,M}}{L^2},\frac{1}{L_0M^2(\lambda)},\frac{1}{L^2M(\lambda)}\right\}\right) \\
& \exp\left(-c_2\frac{M(\lambda)}{L^2(\lambda)}nr_{n,M}^2\right)
\end{aligned}$$

for some positive constants c_1, c_2 depending on c_0, C_f and b only, and a constant B_1 depending on c_0 and C_f.

Notice now that by (1.3) and under assumption $(A3)$, $\lambda^* \in \Lambda_1$. We therefore have the following corollary.

Corollary 1. *Assume that (A1) - (A3) hold. Then*

$$P\left\{|\widehat{\lambda}_j - \lambda^*_j| > B_1 r_{n,M}\right\} \le \pi^*,$$

for all $1 \le j \le M$, where $\pi^* = \pi_{n,M}(\lambda^*)$.

Proof. From ([6], Theorem 2.3) we obtain

$$1 - \pi^* \le P\left\{|\widehat{\lambda} - \lambda^*|_1 \le B_1 k^* r_{n,M}\right\} \le P\left\{\min_{1\le j\le M} |\widehat{\lambda}_j - \lambda^*_j| \le B_1 r_{n,M}\right\}.$$

This immediately implies the result. ∎

Remark 3.1 Notice that $\pi^* \to 0$ as $n \to 0$ for any $r_{n,M} \ge A\sqrt{\log(Mn)/n}$, and for $B = B_1$, as needed in Proposition 2.2 in Section 2.2 above.

In order to control the probability (2.8) we first define U and U_1, the analogues of the sets Λ and Λ_1 defined above.

$$U = \left\{\mu \in \Re^{k^*} : \ \|f - \sum_{j\in I^*} \mu_j f_j\|^2 \le C_f r^2_{n,M}\right\}, \quad U_1 = \{\mu \in U : \rho(\mu)M(\mu) \le C\}.$$

Recall that μ^* is the vector in $\Re^{k^*}$ obtained from λ^* by deleting the zero entries. Then, since assumption (A3) implies $\max_{i\in I^*}\max_{j\in I^*, j\ne i} |\rho_M(i,j)| \le C/k^*$ and $\|f - \sum_{j=1}^M \lambda^*_j f_j\| = \|f - \sum_{j\in I^*} \mu_j f_j\|$ we deduce that $\mu^* \in U_1$. Therefore, using again ([6], Theorem 2.3) applied now to the dictionary $\{f_j\}_{j\in I^*}$ and quantity $\tilde{\mu}$ defined in (2.6) above, we obtain the following corollary:

Corollary 2. *Assume that (A1) - (A3) hold. Then*

$$P\{|\tilde{\mu} - \mu^*|_1 \le B_2 k^* r_{n,M}\} \ge 1 - p^*, \tag{3.11}$$

where

$$\begin{aligned} p^* \quad &\le \quad 14{k^*}^2 \exp\left(-c_1 n \min\left\{\frac{r^2_{n,M}}{L_0}, \frac{r_{n,M}}{L^2}, \frac{1}{L_0 {k^*}^2}, \frac{1}{k^* L^2}\right\}\right) \\ &\quad + \exp\left(-c_2 \frac{k^*}{L^2(\lambda^*)} n r^2_{n,M}\right), \end{aligned}$$

for some positive constants c_1, c_2 as above and a constant $B_2 > 0$ that only depends on C_f and c_0.

Remark 3.2 If $r_{n,M} \ge A\sqrt{\log(Mn)/n}$, then $Mp^* \to 0$ as $n \to \infty$, for $A > 0$ large enough. Hence, the probability given by (2.8), summed over k, converges to zero for both choices of $r_{n,M}$ introduced in Section 2, adjusting the value of B_2 if needed.

The following Lemma is needed in the beginning of the proof of Proposition 2.3.

Lemma 3.4. $\tilde{\lambda} = (\tilde{\mu}, 0)$ *is a solution of (2.4) on the set*

$$\mathcal{B} = \bigcap_{k \notin I^*} \left\{ \left| \frac{2}{n} \sum_{i=1}^{n} [Y_i - \sum_{j \in I^*} \tilde{\mu}_j f_j(X_i)] f_k(X_i) \right| < 2r_{n,M} ||f_k||_n \right\}.$$

Proof. We recall that for any convex function $g : \Re^M \to \Re$ the subdifferential of g at a point λ is the set $D_\lambda = \{w \in \Re^M : \ g(u) - g(\lambda) \geq \langle w, u - \lambda \rangle\}$. Let $g(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \sum_{j=1}^{M} \lambda_j f(X_i)\}^2 + \text{pen}(\lambda)$, where we recall that our penalty term is $\text{pen}(\lambda) = 2r_{n,M}\sum_{j=1}^{M} \|f_j\|_n |\lambda_j|$. Then (e.g., [13]) we have

$$D_\lambda = \{w \in \Re^M : \ w = -\frac{2}{n} F'(Y - F\lambda) + 2r_{n,M} v\},$$

where $v \in \Re^M$ is such that

$$\begin{array}{rcll} v_k & = & ||f_k||_n, & if \ \lambda_k > 0 \\ v_k & = & -||f_k||_n, & if \ \lambda_k < 0 \\ v_k & \in & [-||f_k||_n, ||f_k||_n], & if \ \lambda_k = 0, \end{array}$$

and where we recall that $Y = (Y_1, \ldots, Y_n)$ and F is the $n \times M$ matrix with elements $f_j(X_i)$. By standard results in convex analysis, $\bar{\lambda} \in \Re^M$ is a point of local minimum for a convex function g if and only if $0 \in D_{\bar{\lambda}}$, where $0 \in \Re^M$. Therefore, $\bar{\lambda}$ minimizes our $g(\lambda)$ if and only if $0 \in D_{\bar{\lambda}}$ if and only if

$$\left| \left(\frac{2}{n} F'(Y - F\bar{\lambda}) \right)_k \right| = 2r_{n,M}|v_k| \text{ for all } k \in \{1, \ldots, M\},$$

where $(\ )_k$ above denotes the k-th component of the vector in paranthesis. Equivalently, $\bar{\lambda}$ minimizes $g(\lambda)$ if and only if, for all $1 \leq k \leq M$

$$\begin{array}{lrcl} (3.12) & \left| \frac{2}{n} \sum_{i=1}^{n} [Y_i - \sum_{j=1}^{M} \bar{\lambda}_j f_j(X_i)] f_k(X_i) \right| & = & 2r_{n,M}||f_k||_n, \ \ if \ \bar{\lambda}_k \neq 0 \\ & \left| \frac{2}{n} \sum_{i=1}^{n} [Y_i - \sum_{j=1}^{M} \bar{\lambda}_j f_j(X_i)] f_k(X_i) \right| & \leq & 2r_{n,M}||f_k||_n, \ \ if \ \bar{\lambda}_k = 0. \end{array}$$

In what follows we find conditions under which $\tilde{\lambda} = (\tilde{\mu}, 0)$, with $\tilde{\mu}$ given in (2.6) above, satisfies (3.12). First notice that, by definition, $\sum_{i=1}^{n}[Y_i - \sum_{j=1}^{M} \tilde{\lambda}_j f_j(X_i)] = \sum_{i=1}^{n}[Y_i - \sum_{j \in I^*} \tilde{\mu}_j f_j(X_i)]$. Since $\tilde{\mu}$ is a solution of (2.6) then, by the above standard results in convex analysis, applied now to the function $h(\lambda)$ defined in the proof of Proposition 2.3, the following hold

$$\begin{array}{rcl} \left| \frac{2}{n} \sum_{i=1}^{n} [Y_i - \sum_{j \in I^*} \tilde{\mu}_j f_j(X_i)] f_k(X_i) \right| & = & 2r_{n,M}||f_k||_n, \ if \ \tilde{\lambda}_k = \tilde{\mu}_k \neq 0, \ \ k \in I^* \\ \left| \frac{2}{n} \sum_{i=1}^{n} [Y_i - \sum_{j \in I^*} \tilde{\mu}_j f_j(X_i)] f_k(X_i) \right| & \leq & 2r_{n,M}||f_k||_n, \ if \ \tilde{\lambda}_k = \tilde{\mu}_k = 0, \ \ k \in I^*. \end{array}$$

Notice now that on the set $\mathcal{B}$ we also have

$$\left| \frac{2}{n} \sum_{i=1}^{n} [Y_i - \sum_{j \in I^*} \tilde{\mu}_j f_j(X_i)] f_k(X_i) \right| \leq 2 r_{n,M} ||f_k||_n, \text{ i}f \ k \notin I^* \text{ (for which } \tilde{\mu}_k = 0).$$

The above displays show that $\tilde{\lambda}$ satisfies condition (3.12) and is therefore a solution of (2.4) on $\mathcal{B}$. ∎

Remark 3.3 The observation that constitutes the statement of the above lemma has also been made elsewhere [12] for a slightly different penalty term. We have included here a full derivation of it for completeness and clarity.

To complete the proof of Proposition 2.3 we will make repeated use of Bernstein's inequality, which we state here for completeness.

Bernstein's inequality *Let $\zeta_1, \ldots, \zeta_n$ be independent random variables such that*

$$\frac{1}{n} \sum_{i=1}^{n} E|\zeta_i|^m \leq \frac{m!}{2} w^2 d^{m-2}$$

for some positive constants w and d and for all integers $m \geq 2$. Then, for any $\varepsilon > 0$ we have

$$P\left\{ \sum_{i=1}^{n} (\zeta_i - E\zeta_i) \geq n\varepsilon \right\} \leq \exp\left(-\frac{n\varepsilon^2}{2(w^2 + d\varepsilon)} \right). \tag{3.13}$$

Lemma 3.5. *Let assumptions (A1) and (A2) hold. Then*

$$\sum_{k \in \{1,\ldots,M\} \setminus I^*} P(E_1^c(k)) \to 0, \quad \sum_{k \in \{1,\ldots,M\} \setminus I^*} P(E_2^c(k)) \to 0, \textit{ and}$$

$$\sum_{k \in \{1,\ldots,M\} \setminus I^*} P(E_3^c(k)) \to 0, \textit{ as } n \to \infty.$$

Proof. To show $\sum_{k \in \{1,\ldots,M\} \setminus I^*} P(E_1^c(k)) \to 0$ it is enough to show that $(I) = \sum_{k \in \{1,\ldots,M\} \setminus I^*} P(E_1^c(k) \cap E_2(k)) \to 0$ and that $(II) = \sum_{k \in \{1,\ldots,M\} \setminus I^*} P(E_2^c(k)) \to 0$. The proofs follow immediately from Bernstein's inequality and the union bound. They are the same as ([6], proofs of Lemmas 4 and 5, page 186). We include here the derived probability bounds, for completeness.

$$(I) \leq 2M^2 \exp\left(-\frac{n r_{n,M}^2}{16b} \right) + 2M^2 \exp\left(-\frac{n r_{n,M} c_0}{8\sqrt{2}L} \right) + 2M^2 \exp\left(-\frac{n c_0^2}{12L^2} \right),$$

and

$$(II) \leq M^2 \exp\left(-\frac{n c_0^2}{12L^2} \right).$$

To bound the last quantity in the statement of the Lemma notice first that

$$\begin{aligned}
P(E_3^c(k)) &\leq 2\sum_{j\in I^*} P\left(\frac{1}{n}\sum_{i=1}^n f_j(X_i)f_k(X_i) > 2|\langle f_j, f_k\rangle| + \delta_{n,M}\right)\\
&\leq 2\sum_{j\in I^*}\exp\left\{-\frac{n}{4L_0}\left(|\langle f_j, f_k\rangle| + \delta_{n,M}\right)^2\right\}\\
&\quad + 2\sum_{j\in I^*}\exp\left\{-\frac{n}{4L}\left(|\langle f_j, f_k\rangle| + \delta_{n,M}\right)\right\}\\
&\leq M\exp\left\{-\frac{n\delta_{n,M}^2}{4L_0}\right\} + M\exp\left\{-\frac{n\delta_{n,M}}{4L}\right\}.
\end{aligned}$$

The second inequality of the display above follows from Bernstein's inequality with $\zeta_i = f_j(X_i)f_k(X_i)$, for every fixed j, and k and with $w^2 = L_0$, $d = L^2$, for $\epsilon = |\langle f_j, f_k\rangle| + \delta_{n,m}$, used together with the inequality $e^{x/a+b} \leq e^{x/2a} + e^{x/2b}$ for all x, a and b. Therefore, for $\delta_{n,M} = 2CL^2 r_{n,M}$ we obtain

$$(III) = \sum_{k\in\{1,\ldots,M\}\setminus I^*} P(E_2^3(k)) \leq M^2\exp\left\{-\frac{C^2L^4nr_{n,M}^2}{L_0}\right\} + M^2\exp\left\{-\frac{CLnr_{n,M}}{2}\right\}.$$

Thus, the quantities $(I), (II)$ and (III) converge to zero for any $r_{n,M} \geq A\sqrt{\log(M)n/n}$. ∎

Lemma 3.6. *Let assumptions (A1) and (A2) hold. Then*

$$(IV) = \sum_{k\in\{1,\ldots,M\}\setminus I^*} P\left(\frac{1}{n}\sum_{i=1}^n |(f(X_i) - f^*(X_i))| \geq \frac{c_0}{8L} r_{n,M}\right) \to 0.$$

Proof. By the Cauchy-Schwartz inequality we have

$$\begin{aligned}
(3.14)\ & P\left(\frac{1}{n}\sum_{i=1}^n |(f(X_i) - f^*(X_i))| \geq \frac{c_0}{8L} r_{n,M}\right)\\
&\leq P\left(\frac{1}{n}\sum_{i=1}^n (f(X_i) - f^*(X_i))^2 \geq \frac{c_0^2}{64L^2} r_{n,M}^2\right)\\
&\leq P\left(\sum_{i=1}^n \{(f(X_i) - f^*(X_i))^2 - \|f - f^*\|^2\} \geq n(\frac{c_0^2}{64L^2} r_{n,M}^2 - \|f - f^*\|^2)\right)\\
&\leq P\left(\sum_{i=1}^n \{(f(X_i) - f^*(X_i))^2 - \|f - f^*\|^2\} \geq C_1 n r_{n,M}^2\right),
\end{aligned}$$

where we recall that $\|f - f^*\|^2 \leq C_f r_{n,M}^2$, by definition and $C_1 = c_0^2/64L^2 - C_f$, where we assume that we have already adjusted C_f to have $C_1 > 0$, by taking an appropriate constant A in the definition of $r_{n,M}$, if needed. The proof follows immediately from Bernstein's inequality applied to $\zeta_i = (f(X_i) - f^*(X_i))^2$, with $w = \sqrt{C_f} r_{n,M}$ and $d = L^*$, and for $\epsilon = C_1 r_{n,M}^2$. Therefore

$$(IV) \leq M\exp\{-\frac{C_f C_1^2}{4} n r_{n,M}^2\} + M\exp\{-\frac{C_1}{4L^*} n r_{n,M}^2\},$$

and both terms converge to zero for either choice of $r_{n,M}$. ∎

Acknowledgments

Research partially supported by NSF grants DMS 0406049 and DMS 0706829.

References

[1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **19**, 716–723.

[2] Barron, A., Birgé, L., Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory and Relat. Fields* **113**, 301–413.

[3] Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Hypothesis Testing. *J. R. Statist. Soc.* B **57**, 289 – 300.

[4] Bunea, F. (2004). Consistent covariate selection and post model selection inference in semiparametric regression. *Annals of Statistics* **32**, 898–927.

[5] Bunea, F., Wegkamp, M. H. and Auguste, A. (2006). Consistent variable selection in high dimensional regression via multiple testing. *Journal of Statistical Planning and Inference* **136**, 4349 – 4364.

[6] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007). Sparsity oracle inequalities for the Lasso. *The Electronic Journal of Statistics* **1**, 169 - 194.

[7] Chakrabarti, A. and Ghosh, J.K. (2006). A generalization of BIC for the general exponential families. *Journal of Statistical Planning and Inference* **136**, 2847 - 2872.

[8] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32**, 407–451.

[9] Genovese, C., and Wasserman, L. (2004). A Stochastic Process Approach to False Discovery Rates. *Annals of Statististics* **32**, 1035 - 1061.

[10] Guyon, X. and Yao, J. (1999). On the underfitting and overfitting sets of models chosen by order selection criteria. *Journal of Multivariate Analysis* **70**, 221 - 315.

[11] Lahiri, P., Editor (2001). *Model Selection. Institute of Mathematical Statistics, Lecture Notes-Monograph Series* **38**.

[12] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics* **34** (3), 1436–1462.

[13] Osborne, M.R., Presnell, B. and Turlach, B.A (2000a). On the lasso and its dual. *Journal of Computational and Graphical Statistics* **9**, 319 – 337.

[14] Osborne, M.R., Presnell, B. and Turlach, B.A (2000b). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis* **20**(3), 389 – 404.

[15] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

[16] Shao, J. (1993). Linear model selection by cross validation. *J. Amer. Stat. Assoc.* **888**, 486 - 494.

[17] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B.* **58**, 267–288.

[18] Turlach, B.A. (2005). On algorithms for solving least squares problems under an L1 penalty or an L1 constraint. *2004 Proceedings of the American Statistical Association, Statistical Computing Section [CD-ROM]*, American Statistical Association, Alexandria, VA, pp. 2572-2577.

[19] Wainwright, M. J. (2007). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *Technical Report, UC Berkeley, Department of Statistics.*

[20] Wasserman, L. and Roeder, K. (2007). High dimensional variable selection. *Technical Report, Carnegie Mellon University, Department of Statistics.*

[21] Wegkamp, M.H. (2003). Model Selection in nonparametric regression. *Annals of Statistics* **31**, 252 - 273.

[22] Woodroofe, M. (1982). On model selection and the arcsine laws. *Annals of Statistics* **10**, 1182 − 1194.

[23] Zhao, P. and Yu, B. (2007). On model selection consistency of Lasso. *Journal of Machine Learning Research* **7**, 2541–2567.

[24] Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418-1429.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 139
© Institute of Mathematical Statistics, 0000
arXiv: math.PR/0000000

Asymptotic optimality of a cross-validatory predictive approach to linear model selection

Arijit Chakrabarti[1] **and Tapas Samanta**

Indian Statistical Institute

Abstract: In this article we study the asymptotic predictive optimality of a model selection criterion based on the cross-validatory predictive density, already available in the literature. For a dependent variable and associated explanatory variables, we consider a class of linear models as approximations to the true regression function. One selects a model among these using the criterion under study and predicts a future replicate of the dependent variable by an optimal predictor under the chosen model. We show that for squared error prediction loss, this scheme of prediction performs asymptotically as well as an oracle, where the oracle here refers to a model selection rule which minimizes this loss if the true regression were known.

Contents

1. Introduction

The ultimate goal of modeling in any scientific or sociological investigation is to discover the underlying regular pattern or phenomenon, if any, which controls the data generating mechanism. Although it is almost impossible to imagine that a single model or combinations of a handful will fully capture the intricate functioning of nature or sociological issues, one can always hope to be able to come close. Given a choice of several models and a set of data, a popular method is to choose the model which explains or fits the given data best (in some well-defined sense). However, it is of prime importance that any model that is chosen should be able to predict future observations from the same experiment or process reasonably well and that it does not merely fit the observed data. This is the purpose of predictive model selection.

One of the most prominent approaches to predictive model selection is cross-validation (see [17]) and variants thereof. As the name cross-validation suggests,

[1] Applied Statistics Division, Indian Statistical Institute, 203 B.T. Road, Kolkata-700108, India, e-mail: arc@isical.ac.in; tapas@isical.ac.in

AMS 2000 subject classifications: Primary 62J05; secondary 62F15 parameters of the population are estimated under each model by using a part of the data (the ‘estimation set’), while the rest of the data (the ‘validation set’) are predicted using the estimates based on the first group. This is done repeatedly by using ‘validation sets’ comprising different parts of the data, e.g., the whole data could simply be divided into 10 disjoint parts, each part consisting of an equal number of observations and predicted using the rest. If, for a particular model, such predictions match best with the actually observed values, i.e, if the average prediction error is the smallest for it among all the candidate models, it is selected. Optimality properties of classical cross-validatory techniques have been studied, e.g., in [12] and [16].

In the Bayesian literature, several approaches to model selection have been studied with the predictive aspect in mind; see, e.g, [1, 4, 5, 8–10, 13, 14]. The purpose of this paper is to study the predictive properties of a model selection criterion (see (1.2) below) based on the average of the (log) cross-validatory predictive densities (see (1.1) below) and already available in the literature. Different types of averages (e.g. arithmetic mean, (log) geometric mean) of cross-validatory predictive densities have been studied by several authors ([2], [3], [5], [9], and [14]). Chakrabarti and Ghosh [5] considered an average with respect to disjoint validation sets and studied what should be the optimal proportion of the sample kept for validation in large sample sizes, for the selection of a model closest to the true model (in terms of Kullback-Leibler divergence), and for the selection of the more parsimonious model if two models are equidistant from the truth. Using squared error prediction loss, we show that model selection using criterion (1.2) has an optimality property in predicting a future replicate of the dependent variable (for fixed values of the independent variables), when the true regression is being approximated by a class of candidate linear models. The proofs of the optimality results partly use some general techniques of Li [12] which were later adopted in [16].

In the Bayesian setup, the ordinary predictive density under a model is defined as the integral of the likelihood function of the observed data with respect to the prior distribution of the parameters under the model. Between two competing models, the one having a larger predictive density for the given data seems to be the more appropriate description of the unknown data generating process. In non-subjective Bayesian analysis, it is common to use noninformative priors for the parameters which are typically improper and defined only up to unknown multiplicative constants. In such situations, use of the ordinary predictive density as a model selection criterion will be inappropriate. To get rid of this difficulty, one updates the improper prior by getting a proper posterior based on part of the data (called the training sample) and then integrates the likelihood function of the rest of the data with respect to this posterior, thus giving the cross-validatory predictive density. This is like getting the predictive distribution of part of the data using information obtained from the rest of it. This method of obtaining a cross-validatory predictive density can also be used when one puts a proper prior on the parameters of the model. The cross-validatory predictive density can then be used to get pseudo-Bayes Factors, after appropriate averaging with respect to the different possible choices of the training sample. This line of thought owes its origin to Geisser [7] and Geisser and Eddy [8] and came to prominence through what are referred to as partial Bayes Factors or Intrinsic Bayes Factors ([2], [3], [9], [11], and [15]).

In the next few paragraphs, we describe our setup and the model selection criterion we study. We follow the notations of Shao [16].

Let $\boldsymbol{y}_n = (y_1, \ldots, y_n)'$ be a vector of observations on the dependent (response) variable and let $\boldsymbol{X}_n = (\boldsymbol{x}_1', \ldots, \boldsymbol{x}_n')'$ be an $n \times p_n$ matrix of explanatory variables (which are potentially responsible for the variability in the y's), with $\boldsymbol{x}_i$ associated with y_i. Let $\boldsymbol{\mu}_n$ denote $E(\boldsymbol{y}_n|\boldsymbol{X}_n)$, the (unknown) average value of the response variable given the values of the explanatory variables. We further assume that given $\boldsymbol{X}_n$, $\boldsymbol{e}_n = \boldsymbol{y}_n - \boldsymbol{\mu}_n$ has mean vector $\mathbf{0}$ and the components of e_i are independent with common variance σ^2, which could be known or unknown. We are interested in capturing the functional relationship, if any, between $\boldsymbol{\mu}_n$ and $\boldsymbol{X}_n$ which will be most suitable for predictive purposes. We restrict our search within a class of normal linear models. Our model space, denoted $\mathcal{A}_n$, is indexed by α, where each α consists of a subset of size $p_n(\alpha)$ ($1 \leq p_n(\alpha) \leq p_n$) of $\{1, 2, \ldots, p_n\}$ and the true mean $\boldsymbol{\mu}_n$ is assumed to be linearly related to the corresponding explanatory variables. More specifically, under model $\alpha \in \mathcal{A}_n$, $\boldsymbol{y}_n \sim N(\boldsymbol{\mu}_n(\alpha) = X_n(\alpha)\boldsymbol{\beta}_n(\alpha), \sigma^2 I_n)$ where $\boldsymbol{X}_n(\alpha)$ is the submatrix of $\boldsymbol{X}_n$ consisting of the $p_n(\alpha)$ columns specified by α and $\boldsymbol{\beta}_n(\alpha) \in \Re^{p_n(\alpha)}$. A Bayesian puts a prior on the unknown parameters within each model. We consider standard non-subjective priors (see e.g., [1]) given by

$$\begin{aligned} \pi_\alpha(\boldsymbol{\beta}_n(\alpha)) &\propto 1 \qquad \text{if } \sigma^2 \text{ is known, and} \\ \pi_\alpha(\boldsymbol{\beta}_n(\alpha), \sigma^2) &\propto \frac{1}{\sigma^2} \quad \text{if } \sigma^2 \text{ is unknown.} \end{aligned}$$

Consider, for example, the case with σ^2 unknown. Let $\pi_\alpha((\boldsymbol{\beta}_n(\alpha), \sigma^2)|y_{k+1}, \ldots, y_n)$ denote the posterior distribution of the parameters under the model given the observations $(y_{k+1}, \ldots, y_n)$. The cross-validatory predictive density of $(y_1, \ldots, y_k)$ given $(y_{k+1}, \ldots, y_n)$ under model α, denoted by the expression $f_\alpha(y_1, \ldots, y_k|y_{k+1}, \ldots, y_n)$, is given by

$$\int f_{\boldsymbol{\beta}_n(\alpha),\sigma^2}(y_1, \ldots, y_k)\pi_\alpha((\boldsymbol{\beta}_n(\alpha), \sigma^2)|y_{k+1}, \ldots, y_n)\, d\boldsymbol{\beta}_n(\alpha)\, d\sigma^2, \tag{1.1}$$

where $f_{\boldsymbol{\beta}_n(\alpha),\sigma^2}(y_1, \ldots, y_k)$ denotes the density of the k dimensional normal vector, with mean vector given by the first k components of $\boldsymbol{\mu}_n(\alpha)$ and variance-covariance matrix $\sigma^2 I_k$, evaluated at $y_1, \ldots, y_k$. Similarly, the predictive density of any subset $(y_{t_1}, \ldots, y_{t_k})$ of $\boldsymbol{y}$, given the rest of the components of $\boldsymbol{y}$ under this model can be calculated, where $(t_1, \ldots, t_k)$ denotes a subset of $(1, \ldots, n)$. Since a good criterion should not depend too much on the choice of the training sample, we consider the geometric mean of the cross-validatory predictive densities thus obtained by varying the choice of the training sample. The ratio of such geometric means for two models is precisely the Geometric Intrinsic Bayes Factor ([2], [3]). For model α, the criterion which we intend to study equals the logarithm of this geometric mean. Thus if we consider a total of r training samples, this logarithm is given by

$$\mathrm{CV}(\alpha) = \frac{1}{r}\sum_{i=1}^{r} \log f_\alpha(y_{t_{1i}}, \ldots, y_{t_{ki}}|\{y_t : t \notin (t_{1i}, \ldots, t_{ki})\}), \tag{1.2}$$

where $(y_{t_{1i}}, \ldots, y_{t_{ki}})$ is the set of y observations *not* included in the i-th training sample. One selects the model $\hat{\alpha}_n \in \mathcal{A}_n$ which maximizes $\mathrm{CV}(\alpha)$.

Once a model is thus selected, we use the mean of the predictive distribution of $\boldsymbol{y}_n^{\text{new}}$, given the observed $\boldsymbol{y}_n$ under the selected model, as the predictor for a future replicate $\boldsymbol{y}_n^{\text{new}}$ of the response variable for the same value $\boldsymbol{X}_n$ of the explanatory variables. An easy calculation shows that this turns out to be the least squares estimate $X(\hat{\alpha}_n)\hat{\boldsymbol{\beta}}_n(\hat{\alpha}_n)$ where $\hat{\boldsymbol{\beta}}_n(\alpha) = P_n(\alpha)\boldsymbol{y}_n$ and $P_n(\alpha) = X_n(\alpha)(X_n(\alpha)'X_n(\alpha))^{-1}X_n(\alpha)$ is the usual projection matrix.

Our goal is to evaluate this prediction scheme under the true regression using squared error prediction loss. Under the true $\boldsymbol{\mu}_n$, the future replicate $\boldsymbol{y}_n^{\text{new}}$

will be independent of the original observations $\boldsymbol{y}_n$. The quality of any predictor $\delta(\boldsymbol{y}_n)$ of $\boldsymbol{y}_n^{\text{new}}$ based on $\boldsymbol{y}_n$ can be evaluated by the average prediction error $E_{\boldsymbol{\mu}_n}(\frac{1}{n}||\boldsymbol{y}_n^{\text{new}} - \delta(\boldsymbol{y}_n)||^2)$, where $E_{\boldsymbol{\mu}_n}$ denotes expectation with respect to the joint distribution of $(\boldsymbol{y}_n^{\text{new}}, \boldsymbol{y}_n)$ when $\boldsymbol{\mu}_n$ is the true unknown mean. This expectation will be small if, for any fixed $\boldsymbol{y}_n$, $E_{\boldsymbol{\mu}_n}(\frac{1}{n}||\boldsymbol{y}_n^{\text{new}} - \delta(\boldsymbol{y}_n)||^2|\boldsymbol{y}_n)$ is also small. As observed before, the predictor $\delta(\boldsymbol{y}_n)$ we want to evaluate is the same as the least squares predictive estimate of $\boldsymbol{y}_n^{\text{new}}$ under the chosen model $\hat{\alpha}_n$. Now note that for any given fixed model α, the least squares predictive estimate is given by $\delta(\boldsymbol{y}_n) = \delta(\boldsymbol{y}_n)(\alpha) = \hat{\boldsymbol{\mu}}_n(\alpha) = X_n(\alpha)\hat{\boldsymbol{\beta}}_n(\alpha)$. A simple algebra shows that the above conditional expectation is, up to a constant which does not depend on α, equal to

$$L_n(\alpha) = \frac{||\boldsymbol{\mu}_n - \hat{\boldsymbol{\mu}}_n(\alpha)||^2}{n}. \tag{1.3}$$

Hence the conditional expectation will be minimized for a certain α if $L_n(\alpha)$ is minimized. If we knew the true $\boldsymbol{\mu}_n$, we could find the model which minimizes this $L_n(\alpha)$ for each $\boldsymbol{y}_n$. We shall call this the oracle model, denoted α_n^L. The best any procedure can achieve is to do as well as the oracle in the limit in terms of the loss as the sample size grows to infinity.

We show in the following sections of this article that under certain conditions, minimizing CV(α) with respect to α is asymptotically equivalent to minimizing $L_n(\alpha)$. Using this fact it is shown that the ratio of $L_n(\alpha_n^L)$ to $L_n(\hat{\alpha}_n)$ tends to 1 in probability, whereby establishing the optimum asymptotic behavior of criterion (1.2) in the problem of prediction of a set of future observations.

In Sections 2 and 3 we consider the case where the true model is not in the model space – the proposed models are only approximations to the truth. In Section 2 we consider the case when σ^2 is known. We show that under certain assumptions, the model selection procedure under study performs as well as the oracle asymptotically in the sense that the ratio of their losses tends to one in probability. In Section 3, we consider the more realistic situation when σ^2 is unknown. Under appropriate conditions it is shown that this procedure also achieves the oracle asymptotically in this case. As a validation of this method, we next consider in Section 4 the question of whether, under the assumption that the true model is indeed included in the model space, we do equally well in terms of hitting the oracle loss asymptotically. It is shown that this model selection procedure chooses the correct model with smallest dimension with probability tending to one in addition to being asymptotically optimal in terms of hitting the oracle. Some concluding remarks are made in Section 5. Technical proofs of most of the results are given in the Appendix.

For notational simplicity we write $\boldsymbol{y}$, $\boldsymbol{\mu}$, $\boldsymbol{e}$, $X(\alpha)$, $\boldsymbol{\beta}(\alpha)$ and $P(\alpha)$ in place of $\boldsymbol{y}_n$, $\boldsymbol{\mu}_n$, $\boldsymbol{e}_n$, $X_n(\alpha)$, $\boldsymbol{\beta}_n(\alpha)$ and $P_n(\alpha)$ respectively, dropping the suffix n for the rest of the paper.

2. Basic Results – Case with σ^2 Known

In this section we take the 'model false' point of view that the models are only approximations to the truth but none of them is actually true. We show that under certain conditions, the model selection procedure under study is asymptotically optimal in the sense of performing as well as the oracle defined above.

As described in the introduction, the model selection criterion under consideration is an average of the cross-validatory predictive density

$$f_\alpha(y_1, \ldots, y_k|y_{k+1}, \ldots, y_n)$$

under model α, over suitable choices of the 'training sample' $\{y_{k+1}, \dots, y_n\}$. *We do not recommend here any particular choice of the training samples; our results hold as long as each y_i, $1 \le i \le n$, appears in the same number of training samples chosen (which will be assumed throughout the paper).*

Let $\boldsymbol{y}_i, i = 1, \dots, r$ be the r training samples (each of size $n-k$). For each $\boldsymbol{y}_i$, let $\boldsymbol{\mu}_i$ and $\boldsymbol{e}_i$ be the subvector of $\boldsymbol{\mu}$ and $\boldsymbol{e}$ corresponding to the labels of the components of $\boldsymbol{y}_i$ and $X_i(\alpha)$ be the submatrix of $X(\alpha)$ consisting of the corresponding rows of it. Also, let $\hat{\boldsymbol{\beta}}_i(\alpha) = [X_i'(\alpha)X_i(\alpha)]^{-1}X_i'(\alpha)\boldsymbol{y}_i$, $P_i(\alpha) = X_i(\alpha)[X_i'(\alpha)X_i(\alpha)]^{-1}X_i'(\alpha)$, $i = 1, \dots, r$. It will be assumed throughout that $(n-k) \to \infty$ and $X_i'(\alpha)X_i(\alpha)$ is nonsingular for each i and α. With the standard non-subjective prior $\pi(\boldsymbol{\beta}(\alpha)) =$ constant, we have a closed form expression for the cross-validatory predictive density. An alternative equivalent criterion, which is to be minimized with respect to α, is

$$\begin{aligned}\Gamma(\alpha) &= \frac{1}{n}(\boldsymbol{y} - X(\alpha)\hat{\boldsymbol{\beta}}(\alpha))'(\boldsymbol{y} - X(\alpha)\hat{\boldsymbol{\beta}}(\alpha)) \\ &- \frac{1}{r}\sum_{i=1}^{r}\frac{1}{n}(\boldsymbol{y}_i - X_i(\alpha)\hat{\boldsymbol{\beta}}_i(\alpha))'(\boldsymbol{y}_i - X_i(\alpha)\hat{\boldsymbol{\beta}}_i(\alpha)) \\ &+ \frac{1}{r}\sum_{i=1}^{r}\frac{\sigma^2}{n}\log\left(\frac{|X'(\alpha)X(\alpha)|}{|X_i'(\alpha)X_i(\alpha)|}\right). \end{aligned} \tag{2.1}$$

Note that $\Gamma(\alpha)$ is equal to the negative of the criterion (1.2) up to an additive constant. We will prove that minimization of $\Gamma(\alpha)$ is equivalent to minimization of the loss $L_n(\alpha)$ (defined in (1.3)) in an appropriate asymptotic sense and this will lead to the desired asymptotic (predictive) optimality of the criterion under consideration.

Note that the loss $L_n(\alpha)$ defined in (1.3) can be written as

$$nL_n(\alpha) = n\Delta_n(\alpha) + \boldsymbol{e}'P(\alpha)\boldsymbol{e}$$

where $n\Delta_n(\alpha) = \boldsymbol{\mu}'(I - P(\alpha))\boldsymbol{\mu}$ and let

$$nR_n(\alpha) = E(nL_n(\alpha)) = n\Delta_n(\alpha) + \sigma^2 p_n(\alpha).$$

One of the key assumptions under which we prove our results is the following condition ([12], [16]):

$$\sum_{\alpha \in \mathcal{A}_n} \frac{1}{[nR_n(\alpha)]^m} \to 0 \tag{2.2}$$

for some positive integer m for which $E(e_1^{4m}) < \infty$. We also assume

$$\frac{p_n\lambda_n}{\min\limits_{\alpha \in \mathcal{A}_n} nR_n(\alpha)} \to 0, \tag{2.3}$$

where $\lambda_n = \log(n/(n-k))$.

For certain remarks justifying these assumptions, see [12] and [16]. In particular, it is argued in these papers using several concrete examples, that condition (2.2) is a natural one when the dimension p_n of the largest model grows with sample size. Also, if p_n remains bounded, $nR_n(\alpha)$ is expected to go to ∞ for all α as the sample size increases, if the candidate models are separated from the truth.

That $\min_{\alpha} nR_n(\alpha) \to \infty$ is assumption A.3$'$ of Li [12] and as remarked therein, it is a quite reasonable assumption if p_n grows with n. Condition (2.3) requires that $\min_{\alpha} nR_n(\alpha) \to \infty$ at a suitable rate. Under condition (3.3) below ([16], condition (2.5)), (2.3) holds if $(p_n\lambda_n)/n \to 0$.

It is important to note that we also need to assume $(n-k)/n \to 0$ to prove our results (see e.g. (6.10)). This addresses an important question about the required size of the training sample. We, however, do not claim that it is a necessary condition for asymptotic predictive optimality.

We now consider the criterion $\Gamma(\alpha)$ as defined in (2.1). Since $X(\alpha)\hat{\boldsymbol{\beta}}(\alpha) = P(\alpha)\boldsymbol{y}$,

$$\begin{aligned} & \frac{1}{n}(\boldsymbol{y} - X(\alpha)\hat{\boldsymbol{\beta}}(\alpha))'(\boldsymbol{y} - X(\alpha)\hat{\boldsymbol{\beta}}(\alpha)) \\ = & \frac{1}{n}\boldsymbol{y}'(I - P(\alpha))\boldsymbol{y} \\ = & \frac{1}{n}\boldsymbol{e}'\boldsymbol{e} + L_n(\alpha) - \frac{2}{n}\boldsymbol{e}'P(\alpha)\boldsymbol{e} + \frac{2}{n}\boldsymbol{e}'(I - P(\alpha))\boldsymbol{\mu}. \end{aligned} \tag{2.4}$$

Similarly,

$$\begin{aligned} & \frac{1}{r}\sum_{i=1}^{r}\frac{1}{n}(\boldsymbol{y}_i - X_i(\alpha)\hat{\boldsymbol{\beta}}_i(\alpha))'(\boldsymbol{y}_i - X_i(\alpha)\hat{\boldsymbol{\beta}}_i(\alpha)) \\ = & \frac{n-k}{n^2}\boldsymbol{e}'\boldsymbol{e} + \frac{1}{nr}\sum_{i=1}^{r}\boldsymbol{\mu}_i'(I - P_i(\alpha))\boldsymbol{\mu}_i - \frac{1}{nr}\sum_{i=1}^{r}\boldsymbol{e}_i'P_i(\alpha)\boldsymbol{e}_i \\ + & \frac{2}{nr}\sum_{i=1}^{r}\boldsymbol{e}_i'(I - P_i(\alpha))\boldsymbol{\mu}_i. \end{aligned} \tag{2.5}$$

We first state two auxiliary results.

Lemma 2.1. *Under conditions (2.2) and (2.3),*

$$\frac{1}{n}(\boldsymbol{y} - X(\alpha)\hat{\boldsymbol{\beta}}(\alpha))'(\boldsymbol{y} - X(\alpha)\hat{\boldsymbol{\beta}}(\alpha)) = \frac{1}{n}\boldsymbol{e}'\boldsymbol{e} + L_n(\alpha) + o_p(L_n(\alpha))$$

uniformly in $\alpha \in \mathcal{A}_n$.

By saying $Z_n(\alpha) = o_p(L_n(\alpha))$ uniformly in α, we mean $\max_{\alpha} |Z_n(\alpha)|/L_n(\alpha) \xrightarrow{p} 0$.

Lemma 2.2. *Suppose that conditions (2.2) and (2.3) hold and* $(n-k)/n \to 0$. *Then*

$$\frac{1}{r}\sum_{i=1}^{r}\frac{1}{n}(\boldsymbol{y}_i - X_i(\alpha)\hat{\boldsymbol{\beta}}_i(\alpha))'(\boldsymbol{y}_i - X_i(\alpha)\hat{\boldsymbol{\beta}}_i(\alpha)) = \frac{n-k}{n^2}\boldsymbol{e}'\boldsymbol{e} + o_p(L_n(\alpha)),$$

uniformly in $\alpha \in \mathcal{A}_n$.

Proofs of Lemma 2.1 and Lemma 2.2 are given in the Appendix.

In order to prove the main result of this section we need to assume another condition which is given below.

Let

$$a_{in}(\alpha) = \log\left\{\frac{(n-k)^{p_n(\alpha)}|X'(\alpha)X(\alpha)|}{n^{p_n(\alpha)}|X_i'(\alpha)X_i(\alpha)|}\right\}. \tag{2.6}$$

We assume

$$\max_{\alpha\in\mathcal{A}_n} \frac{\frac{1}{r}\sum\limits_{i=1}^{r} a_{in}(\alpha)}{nR_n(\alpha)} \to 0. \tag{2.7}$$

Remark 2.1. Let $\boldsymbol{x}_1'(\alpha), \dots, \boldsymbol{x}_n'(\alpha)$ be the n rows of $X(\alpha)$. If these n rows are 'similar', e.g., if they can be thought of as (independent) realizations of a random vector $\boldsymbol{x}$ and p_n is small compared to both $n-k$ and n, then

$$\left|\frac{X'(\alpha)X(\alpha)}{n}\right| = \left|\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x}_j(\alpha)\boldsymbol{x}_j'(\alpha)\right| \approx |E(\boldsymbol{x}\boldsymbol{x}')|$$
$$\text{and similarly } \left|\frac{X_i'(\alpha)X_i(\alpha)}{n-k}\right| \approx |E(\boldsymbol{x}\boldsymbol{x}')|.$$

In this case, it follows that $a_{in}(\alpha) \approx 0$. In such a situation, assumption (2.7) seems to be quite reasonable.

Now note that (2.3) and (2.7) will imply that the third term in the right hand side of (2.1) is also of the order $o_p(L_n(\alpha))$ uniformly in $\alpha \in \mathcal{A}_n$. Thus

$$\Gamma(\alpha) = \text{constant} + L_n(\alpha) + o_p(L_n(\alpha)) \text{ uniformly in } \alpha \in \mathcal{A}_n$$

which implies minimization of $\Gamma(\alpha)$ is essentially equivalent to minimization of $L_n(\alpha)$ in an appropriate asymptotic sense and we have the following result.

Theorem 2.1. *Suppose that conditions (2.2), (2.3) and (2.7) hold and $(n-k)/n \to 0$. Then we have the following results.*

(a) $\Gamma(\alpha) = \frac{k}{n^2}\boldsymbol{e}'\boldsymbol{e} + L_n(\alpha) + o_p(L_n(\alpha))$ *uniformly in* $\alpha \in \mathcal{A}_n$.

(b) The model selection rule under study is asymptotically optimal in the sense that

$$\frac{L_n(\hat{\alpha}_n)}{\min\limits_{\alpha\in\mathcal{A}_n} L_n(\alpha)} \xrightarrow{p} 1$$

where $\hat{\alpha}_n$ is as defined in Section 1.

Proof of Theorem 2.1 is given in the Appendix.

3. Case with σ^2 Unknown

We now consider the more realistic situation when the variance σ^2 is unknown. The standard non-subjective prior in this case is $\pi(\boldsymbol{\beta}(\alpha), \sigma^2) \propto \frac{1}{\sigma^2}$ under model α. Interestingly, the results in this case follow from the basic results obtained in Section 2. We consider here the ('model false') setup and assumptions of Section 2.

Let $\boldsymbol{y}_i, i = 1, \dots, r$ be the r training samples chosen. The cross-validatory predictive density under model α for a training sample $\boldsymbol{y}_i$ is given by

$$\frac{|X_i'(\alpha)X_i(\alpha)|^{\frac{1}{2}}}{|X'(\alpha)X(\alpha)|^{\frac{1}{2}}} \times \frac{[(\boldsymbol{y} - X(\alpha)\hat{\boldsymbol{\beta}}(\alpha))'(\boldsymbol{y} - X(\alpha)\hat{\boldsymbol{\beta}}(\alpha))]^{-\frac{n}{2}}}{[(\boldsymbol{y}_i - X_i(\alpha)\hat{\boldsymbol{\beta}}_i(\alpha))'(\boldsymbol{y}_i - X_i(\alpha)\hat{\boldsymbol{\beta}}_i(\alpha))]^{-\frac{n-k}{2}}}$$

up to a multiplicative constant.

Our criterion (to be minimized with respect to α), which is an average over the r training samples, is given by

$$\Gamma(\alpha) = \log[S(\alpha)] - \frac{n-k}{nr}\sum_{i=1}^{r}\log[S_i(\alpha)] + \frac{1}{nr}\sum_{i=1}^{r}\log\frac{|X'(\alpha)X(\alpha)|}{|X_i'(\alpha)X_i(\alpha)|} \tag{3.1}$$

where $S(\alpha) = (\boldsymbol{y} - X(\alpha)\hat{\boldsymbol{\beta}}(\alpha))'(\boldsymbol{y} - X(\alpha)\hat{\boldsymbol{\beta}}(\alpha))$ and $S_i(\alpha) = (\boldsymbol{y}_i - X_i(\alpha)\hat{\boldsymbol{\beta}}_i(\alpha))'(\boldsymbol{y}_i - X_i(\alpha)\hat{\boldsymbol{\beta}}_i(\alpha))$.

Note that $\Gamma(\alpha) = (k/n)\log(n\sigma^2) + \Gamma_1(\alpha)$ where

$$\Gamma_1(\alpha) = \log\left[\frac{S(\alpha)}{n\sigma^2}\right] - \frac{n-k}{nr}\sum_{i=1}^{r}\log\left[\frac{S_i(\alpha)}{n\sigma^2}\right] + \frac{1}{nr}\sum_{i=1}^{r}a_{in}(\alpha) + \frac{1}{n}p_n(\alpha)\lambda_n, \tag{3.2}$$

$a_{in}(\alpha)$ is as defined in (2.6) and $\lambda_n = \log(n/(n-k))$. Therefore, minimizing $\Gamma(\alpha)$ (with respect to α) is equivalent to minimizing $\Gamma_1(\alpha)$ for all σ. Let

$$u_n(\alpha) = \log\left[\frac{\boldsymbol{e}'\boldsymbol{e}}{n\sigma^2} + \frac{1}{\sigma^2}L_n(\alpha)\right].$$

In order to prove the asymptotic optimality of this method, we first note in Lemma 3.1 below that $\Gamma_1(\alpha)$ is asymptotically equivalent to $u_n(\alpha)$ and this in turn implies the desired conclusion as stated in Theorem 3.1. We prove these results by invoking certain conditions which we describe below.

We first make the following assumption (see [16], condition (2.5)):

$$\liminf_{n\to\infty}\min_{\alpha}\Delta_n(\alpha) > 0 \tag{3.3}$$

where $\Delta_n(\alpha)$ is as defined in Section 2. This may be thought of as an identifiability condition on the models in the model space, as appears in the discussion of Mervyn Stone on [16]. We further assume that

$$\frac{n-k}{n}\log n \to 0, \ \frac{p_n\lambda_n}{n} \to 0 \text{ and } \frac{1}{n}\sum_{i=1}^{n}\mu_i^2 \text{ is bounded,} \tag{3.4}$$

$$\frac{1}{nr}\sum_{i=1}^{r}a_{in}(\alpha) \to 0, \tag{3.5}$$

and

$$\sum_{i=1}^{r}\log(S_i) > 0 \tag{3.6}$$

with probability tending to 1, where S_i is equal to $S_i(\alpha)$ with α as the full model, i.e, $\alpha = \{1, \ldots, p_n\}$. One can give sufficient conditions for (3.6) based on the relative magnitude of r and $(n-k)$ as $n \to \infty$, to the effect that r is not too large compared with $n-k$ which is the case for most practically implementable schemes. We, however, do not record the details here. The final results of this section are now stated below.

Lemma 3.1. *Under conditions (3.3)-(3.6),*

$$\Gamma_1(\alpha) = u_n(\alpha) + o_p(u_n(\alpha)) \textit{ uniformly in } \alpha. \tag{3.7}$$

Theorem 3.1. *Under conditions (3.3)-(3.6),*

$$\frac{L_n(\hat{\alpha}_n)}{L_n(\alpha_n^L)} \xrightarrow{p} 1. \tag{3.8}$$

Both Lemma 3.1 and Theorem 3.1 are proved in the Appendix.

4. The 'Model True' Case and Consistency

We now show that if some model in the model space is true, the model selection procedure under study chooses the correct model of the smallest dimension in addition to being asymptotically optimal. Thus this procedure not only captures the truth but at the same time is as parsimonious as possible. Although the assumption of a true model may not seem to be very realistic, our result in this section provides a validation of the method. We, however, consider only the simpler case when σ^2 is known.

As in [16], let $\mathcal{A}_n^c \subset \mathcal{A}_n$ denote all the proposed models that are actually correct. Thus for $\alpha \in \mathcal{A}_n^c$, $\boldsymbol{\mu} = X(\alpha)\boldsymbol{\beta}(\alpha)$ for some $\boldsymbol{\beta}(\alpha) \in \Re^{p_n(\alpha)}$. In Section 2 we assumed that $\mathcal{A}_n^c$ is empty. It is important to note that all the results of Section 2 with $\mathcal{A}_n$ replaced by $\mathcal{A}_n - \mathcal{A}_n^c$ hold under the corresponding assumptions with $\mathcal{A}_n$ replaced by $\mathcal{A}_n - \mathcal{A}_n^c$. In particular, if

$$\sum_{\alpha \in \mathcal{A}_n - \mathcal{A}_n^c} \frac{1}{[nR_n(\alpha)]^m} \to 0 \tag{4.1}$$

for some positive integer m for which $E(e_1^{4m}) < \infty$ and

$$\frac{p_n \lambda_n}{\min\limits_{\alpha \in \mathcal{A}_n - \mathcal{A}_n^c} nR_n(\alpha)} \to 0, \tag{4.2}$$

with $\lambda_n = \log(n/(n-k))$, then

$$\Gamma(\alpha) = \frac{k}{n^2}\boldsymbol{e}'\boldsymbol{e} + L_n(\alpha) + o_p(L_n(\alpha)) \tag{4.3}$$

uniformly in $\alpha \in \mathcal{A}_n - \mathcal{A}_n^c$.

For $\alpha \in \mathcal{A}_n^c$, $(I - P(\alpha))\boldsymbol{\mu} = \mathbf{0}$ and $(I - P_i(\alpha))\boldsymbol{\mu}_i = 0 \ \forall \ i$. Therefore, from (2.1), (2.4) and (2.5) we have for $\alpha \in \mathcal{A}_n^c$

$$\Gamma(\alpha) = \frac{k}{n^2}\boldsymbol{e}'\boldsymbol{e} - \frac{1}{n}\boldsymbol{e}'P(\alpha)\boldsymbol{e} + \frac{1}{nr}\sum_{i=1}^{r}\boldsymbol{e}_i'P_i(\alpha)\boldsymbol{e}_i + \frac{\sigma^2}{nr}\sum_{i=1}^{r}\log\left(\frac{|X'(\alpha)X(\alpha)|}{|X_i'(\alpha)X_i(\alpha)|}\right). \tag{4.4}$$

Also $L_n(\alpha) = \frac{1}{n}\boldsymbol{e}'P(\alpha)\boldsymbol{e}$ for $\alpha \in \mathcal{A}_n^c$.

We now assume that

$$\limsup_{n\to\infty} \sum_{\alpha \in \mathcal{A}_n^c} \frac{1}{[p_n(\alpha)]^m} < \infty. \tag{4.5}$$

for some positive integer m such that $E(e_1^{4m}) < \infty$ (condition (3.10) of Shao [16]), and

$$\max_{\alpha \in \mathcal{A}_n^c} \frac{\frac{1}{r}\sum_{i=1}^{r} a_{in}(\alpha)}{p_n(\alpha)\lambda_n} \to 0 \tag{4.6}$$

with $\lambda_n = \log(\frac{n}{n-k})$ and $a_{in}(\alpha)$ as defined in (2.6). See Remark 2.1 in this context. Let α_n^c be the model α in $\mathcal{A}_n^c$ with smallest dimension. Using the above, we now have

Proposition 4.1. *Under conditions (4.1), (4.2), (4.5) and (4.6)*

$$\Gamma(\alpha) = \frac{k}{n^2}\boldsymbol{e}'\boldsymbol{e} + \frac{1}{n}\lambda_n\sigma^2 p_n(\alpha) + o_p(\frac{1}{n}\lambda_n\sigma^2 p_n(\alpha)) \tag{4.7}$$

uniformly in $\alpha \in \mathcal{A}_n^c$*, and*

$$\Gamma(\alpha_n^c) = \frac{k}{n^2}\boldsymbol{e}'\boldsymbol{e} + o_p(L_n(\alpha)) \textit{ uniformly in } \alpha \in \mathcal{A}_n - \mathcal{A}_n^c. \tag{4.8}$$

Proof of Proposition 4.1 is given in the Appendix.

Keeping in mind the above facts, we now proceed towards proving that this model selection rule chooses the most parsimonious correct model as claimed in Theorem 4.1 below. Towards this we first observe that (4.3) and (4.8) imply

$$\max_{\alpha \in \mathcal{A}_n - \mathcal{A}_n^c} (\Gamma(\alpha_n^c) - \frac{k}{n^2}\boldsymbol{e}'\boldsymbol{e})/(\Gamma(\alpha) - \frac{k}{n^2}\boldsymbol{e}'\boldsymbol{e}) < 1$$

with probability tending to 1. It then follows that

$$P[\Gamma(\alpha_n^c) \leq \Gamma(\alpha)\ \forall \alpha \in \mathcal{A}_n - \mathcal{A}_n^c] \to 1. \tag{4.9}$$

We now try to find some conditions under which

$$P[\Gamma(\alpha_n^c) \leq \Gamma(\alpha)\ \forall \alpha \in \mathcal{A}_n^c] \to 1. \tag{4.10}$$

Let $n[\Gamma(\alpha) - \Gamma(\alpha_n^c)] = Z_n(\alpha)$. It is enough to show that

$$P[Z_n(\alpha) \geq 0\ \forall \alpha \in \mathcal{A}_n^c] \to 1. \tag{4.11}$$

Now,

$$\begin{aligned} & P[Z_n(\alpha) < 0 \text{ for some } \alpha \in \mathcal{A}_n^c] \\ \leq & \sum_{\alpha \in \mathcal{A}_n^c} P[Z_n(\alpha) < 0] \\ \leq & \sum_{\alpha \in \mathcal{A}_n^c} P[|Z_n(\alpha) - E(Z_n(\alpha))| > E(Z_n(\alpha))] \\ \leq & \sum_{\alpha \in \mathcal{A}_n^c} \frac{E|Z_n(\alpha) - E(Z_n(\alpha))|^{2m}}{[E(Z_n(\alpha))]^{2m}}. \end{aligned} \tag{4.12}$$

From (4.4)

$$Z_n(\alpha) - E(Z_n(\alpha)) = \frac{1}{r}\sum_{i=1}^{r} \boldsymbol{e}_i'[P_i(\alpha) - P_i(\alpha_n^c)]\boldsymbol{e}_i - \boldsymbol{e}'[P(\alpha) - P(\alpha_n^c)]\boldsymbol{e} \tag{4.13}$$

and $E(Z_n(\alpha))$ can be written as

$$\frac{1}{\sigma^2}E(Z_n(\alpha)) = [p_n(\alpha) - p_n(\alpha_n^c)]\lambda_n + \frac{1}{r}\sum_{i=1}^{r}[a_{in}(\alpha) - a_{in}(\alpha_n^c)]$$

where $a_{in}(\alpha)$ is as defined in (2.6). If we assume

$$\frac{1}{r}\sum_{i=1}^{r}[a_{in}(\alpha) - a_{in}(\alpha_n^c)] = o_p([p_n(\alpha) - p_n(\alpha_n^c)]\lambda_n) \tag{4.14}$$

uniformly in $\alpha \in \mathcal{A}_n^c$, then

$$\frac{1}{\sigma^2}E(Z_n(\alpha)) = [p_n(\alpha) - p_n(\alpha_n^c)]\lambda_n + o_p([p_n(\alpha) - p_n(\alpha_n^c)]\lambda_n) \tag{4.15}$$

uniformly in $\alpha \in \mathcal{A}_n^c$. Noting that $P(\alpha) - P(\alpha_n^c)$ and $P_i(\alpha) - P_i(\alpha_n^c)$ are projection matrices and the first term on the right hand side of (4.13) can be expressed as $\boldsymbol{e}'M\boldsymbol{e}$ for some matrix M, and using Theorem 2 of Whittle [18] or inequality (6.2) of the Appendix we have

$$E|Z_n(\alpha) - E(Z_n(\alpha))|^{2m} \leq \text{constant}[p_n(\alpha) - p_n(\alpha_n^c)]^m.$$

It then follows from (4.12) and (4.15) that (4.11) holds if

$$\sum_{\alpha \in \mathcal{A}_n^c} \frac{1}{\lambda_n^{2m}[p_n(\alpha) - p_n(\alpha_n^c)]^m} \to 0. \tag{4.16}$$

Thus we finally have the following.

Theorem 4.1. *Under conditions (4.1), (4.2), (4.5), (4.6), (4.14) and (4.16),*

$$P[\hat{\alpha}_n = \alpha_n^c] \to 1. \tag{4.17}$$

It is proved in the Appendix that under (4.1) and (4.2)

$$\max_{\alpha \in \mathcal{A}_n - \mathcal{A}_n^c} \frac{L_n(\alpha_n^c)}{L_n(\alpha)} \xrightarrow{p} 0. \tag{4.18}$$

Since $L_n(\alpha_n^c) \leq L_n(\alpha)\ \forall \alpha \in \mathcal{A}_n^c$, Theorem 4.1 and (4.18) imply the following.

Theorem 4.2. *Under the conditions of Theorem 4.1, one has*

$$L_n(\hat{\alpha}_n)/L_n(\alpha_n^L) \xrightarrow{p} 1. \tag{4.19}$$

5. Concluding Remarks

In this article we have studied predictive optimality of a cross-validatory Bayesian approach to model selection in the context of selecting from among a set of linear models. It has been shown that this method predicts as well as the oracle as the sample size grows. In addition, it has been shown that in case the space of candidate models contains at least one correct model, this method chooses the correct model with the smallest dimension with probability tending to one as sample size grows. Thus the method has two important facets – one of an optimal predictor and the other of a selection criterion which does not unnecessarily choose a complex model when simpler ones are apt.

Needless to say, this article has not addressed some interesting related issues. First, it will be interesting to see how this method works when it is applied in the setup of generalized linear models, through theoretical investigation and simulation. Another focus of recent research is the case when the number of potential parameters in the models is very large, e.g., when it is of the same order as the number of observations. Asymptotic optimality studies in such setup, even for the normal linear models will be a really challenging task. Also, we have not touched upon the computational aspect of this method, which becomes important if the number of potential regressors and number of models in the model space get large. We, however, emphasize that one rarely considers the set of all 2^p possible models if p regressors are available. For example, one can use expert knowledge about the problem under study and start with a pruned list of models or one can take a nested sequence of models (thereby restricting the total number of models to at most p). Li ([12], Example 1) considered a situation where the p regressors are arranged in decreasing order of importance. He then considered p models, the α-th model consisting of the first α regressors in this ordered arrangement. See in this context Examples 1 and 2 of [16] where the number of models under consideration is fixed although the number of parameters may grow with sample size. Last but not the least, as we commented before, the requirement that $k/n \to 1$ is only a sufficient condition; a careful study of the necessity of this condition is in order. In some examples, we have observed that $k/n \to c$ for any $c \in (0,1)$ is also sufficient to achieve good optimality results similar to ones we have obtained in this paper. Some theoretical investigations and simulation studies will hopefully prove conclusive to find the optimal k. It is worth mentioning that in a related problem Chakrabarti and Ghosh [5] made interesting observations regarding this issue which can be a starting point for such investigation.

Appendix

We present in this section proofs of some of the results of the earlier sections. We will need bounds for the moments of linear and quadratic forms in $\boldsymbol{e}$. Let $A = (a_{ij})$ be a non-random $n \times n$ matrix and $\boldsymbol{b}$ be a non-random n-vector. Then by Theorem 2 of Whittle [18],

$$E(|\boldsymbol{e}'\boldsymbol{b}|^{2m}) \leq C_1(||\boldsymbol{b}||^2)^m, \text{ and} \tag{6.1}$$

$$E|\boldsymbol{e}'A\boldsymbol{e} - E(\boldsymbol{e}'A\boldsymbol{e})|^{2m} \leq C_2(\sum_i \sum_j a_{ij}^2)^m \tag{6.2}$$

for some constants $C_1, C_2 > 0$ and for positive integer m for which $E(e_1^{4m}) < \infty$.

Below $\max\limits_{\alpha}$ will mean maximum over $\alpha \in \mathcal{A}_n$.

Proof of Lemma 2.1. As shown in Li ([12], p.970), using Theorem 2 of Whittle [18] or inequalities (6.1) and (6.2) stated above, and condition (2.2), we have

$$\max_{\alpha} \frac{|\boldsymbol{e}'P(\alpha)\boldsymbol{e} - \sigma^2 p_n(\alpha)|}{nR_n(\alpha)} \xrightarrow{p} 0, \text{ and} \tag{6.3}$$

$$\max_{\alpha} \frac{|\boldsymbol{e}'(I - P(\alpha))\boldsymbol{\mu}|}{nR_n(\alpha)} \xrightarrow{p} 0. \tag{6.4}$$

Also, from (6.3)

$$\max_\alpha |\frac{L_n(\alpha)}{R_n(\alpha)} - 1| = \max_\alpha \frac{|\boldsymbol{e}'P(\alpha)\boldsymbol{e} - \sigma^2 p_n(\alpha)|}{nR_n(\alpha)} \xrightarrow{p} 0. \tag{6.5}$$

Lemma 2.1 now follows from (2.3), (2.4), (6.3), (6.4) and (6.5). □

Proof of Lemma 2.2. Let

$$T_1 = \frac{1}{r}\sum_{i=1}^r \boldsymbol{\mu}_i'(I - P_i(\alpha))\boldsymbol{\mu}_i,\ T_2 = \frac{1}{r}\sum_{i=1}^r \boldsymbol{e}_i'P_i(\alpha)\boldsymbol{e}_i \text{ and } T_3 = \frac{1}{r}\sum_{i=1}^r \boldsymbol{e}_i'(I - P_i(\alpha))\boldsymbol{\mu}_i.$$

Then, in view of (2.5), the left hand side of the equality claimed in Lemma 2.2 can be written as

$$\frac{n-k}{n^2}\boldsymbol{e}'\boldsymbol{e} + \frac{1}{n}(T_1 - T_2 + 2T_3).$$

We shall prove that

$$T_j/n = o_p(L_n(\alpha)) \text{ uniformly in } \alpha \tag{6.6}$$

for $j = 1, 2, 3$.

We fix a training sample $\boldsymbol{y}_1 = (y_1, y_2, \ldots, y_{n-k})'$. Let

$$X(\alpha) = \begin{pmatrix} X_1 \\ X_{1c} \end{pmatrix} \text{ and } I - P(\alpha) = \begin{pmatrix} A \\ B \end{pmatrix}$$

where X_1 and X_{1c} are the submatrices consisting of the first $n-k$ rows and the last k rows of X, respectively, and A and B are analogous submatrices of $I - P(\alpha)$. Then

$$\begin{aligned} \boldsymbol{\mu}'(I - P(\alpha))\boldsymbol{\mu} &= \boldsymbol{\mu}'B'B\boldsymbol{\mu} + \boldsymbol{\mu}'A'A\boldsymbol{\mu}, \quad \text{and} & (6.7)\\ \boldsymbol{\mu}'(I - P(\alpha))\boldsymbol{\mu} - \boldsymbol{\mu}_1'(I - P_1(\alpha))\boldsymbol{\mu}_1 &= \boldsymbol{\mu}'B'(I - P_c)^{-1}B\boldsymbol{\mu}, & (6.8)\end{aligned}$$

where $P_c = X_{1c}(X'(\alpha)X(\alpha))^{-1}X_{1c}'$ (see, e.g., Result (5.4) of Chatterjee and Hadi [6], p. 189). One can now check that $(I - P_c)^{-1} = I + X_{1c}(X_1'X_1)^{-1}X_{1c}'$ and

$$\boldsymbol{\mu}'B'(I - P_c)^{-1}B\boldsymbol{\mu} - \boldsymbol{\mu}'B'B\boldsymbol{\mu} = \boldsymbol{\mu}'B'X_{1c}(X_1'X_1)^{-1}X_{1c}'B\boldsymbol{\mu} \geq 0 \tag{6.9}$$

as $(X_1'X_1)^{-1}$ is positive definite. From (6.7)-(6.9)

$$\frac{\boldsymbol{\mu}_1'(I - P_1(\alpha))\boldsymbol{\mu}_1}{nL_n(\alpha)} \leq \frac{\boldsymbol{\mu}_1'(I - P_1(\alpha))\boldsymbol{\mu}_1}{\boldsymbol{\mu}'(I - P(\alpha))\boldsymbol{\mu}} \leq \frac{||A\boldsymbol{\mu}||^2}{||A\boldsymbol{\mu}||^2 + ||B\boldsymbol{\mu}||^2}.$$

We now consider average over the r training samples. Since each y_i $(1 \leq i \leq n)$ appears in the same number of training samples, we have

$$\frac{T_1}{nL_n(\alpha)} \leq \frac{\frac{1}{r}\sum\limits_{i=1}^r \boldsymbol{\mu}_i'(I - P_i(\alpha))\boldsymbol{\mu}_i}{\boldsymbol{\mu}'(I - P(\alpha))\boldsymbol{\mu}} \leq \frac{n-k}{n} \tag{6.10}$$

which converges to zero.

To prove (6.6) for $j = 2$ we note that T_2 can be expressed as $\boldsymbol{e}'M(\alpha)\boldsymbol{e}$ for some matrix $M(\alpha) = (m_{ij})$, which is a sum of r matrices corresponding to the r choices of the $n-k$ indices from $\{1,2,\ldots,n\}$ ($n-k$ rows of $X(\alpha)$). For example, for the training sample $\boldsymbol{y}_1=(y_1,\ldots,y_{n-k})'$, $\boldsymbol{e}_1'P_1(\alpha)\boldsymbol{e}_1$ may be written as $\boldsymbol{e}'M_1(\alpha)\boldsymbol{e}$ where

$$M_1(\alpha)=\begin{pmatrix} P_1(\alpha) & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{pmatrix},\quad \text{thus}\quad M(\alpha)=(1/r)\sum_{i=1}^{r}M_i(\alpha).$$

As $P_i(\alpha)$'s are all idempotent matrices, one can show that $\sum\limits_i\sum\limits_j m_{ij}^2\le p_n(\alpha)$. Then proceeding as in the proof of (6.3) given in Li ([12], p.970) one can prove the result using (6.2), (2.2), (2.3) and (6.5). Indeed, by (6.2),

$$\begin{aligned} & P\left[\max_\alpha\frac{|\boldsymbol{e}'M(\alpha)\boldsymbol{e}-\sigma^2p_n(\alpha)|}{nR_n(\alpha)}>\epsilon\right] \\ \le\ & C\sum_\alpha\frac{[p_n(\alpha)]^m}{\epsilon^{2m}[nR_n(\alpha)]^{2m}}\end{aligned}$$

for some constant $C>0$. The result follows from (2.2), (2.3) and (6.5).

The proof of (6.6) for $j=3$ is similar. We note that $T_3=\boldsymbol{e}'\boldsymbol{b}$ with $\boldsymbol{b}=(1/r)\sum\limits_{i=1}^{r}(I-P_i(\alpha))\boldsymbol{\mu}_i$ and $||b||^2\le(1/r)\sum\limits_{i=1}^{r}\boldsymbol{\mu}_i'(I-P_i(\alpha))\boldsymbol{\mu}_i$. By (6.1) and (6.10)

$$\begin{aligned} & P\left[\max_\alpha\frac{|\boldsymbol{e}'\boldsymbol{b}|}{nR_n(\alpha)}>\epsilon\right] \\ \le\ & C\left(\frac{n-k}{n}\right)^m\sum_\alpha\frac{1}{[nR_n(\alpha)]^m}\end{aligned}$$

for some constant $C>0$. The result follows from (2.2) and (6.5). Thus (6.6) is proved and hence the lemma. □

Remark 6.1. Indeed, to prove Lemma 2.1 and Lemma 2.2, we need to assume

$$\frac{p_n}{\min\limits_{\alpha\in\mathcal{A}_n} nR_n(\alpha)}\to 0$$

instead of the stronger condition (2.3). We, however, need (2.3) to prove our final result.

Proof of Theorem 2.1. Since (2.3) and (2.7) imply that the third term in the right hand side of (2.1) is of the order $o_p(L_n(\alpha))$ uniformly in $\alpha\in\mathcal{A}_n$, part (a) follows from (2.1), Lemma 2.1 and Lemma 2.2. From part (a), $\Gamma(\alpha)$ can be written as

$$\Gamma(\alpha)=\frac{k}{n^2}\boldsymbol{e}'\boldsymbol{e}+L_n(\alpha)(1+\zeta_n(\alpha)),\ \alpha\in\mathcal{A}_n,$$

where $\max\limits_\alpha|\zeta_n(\alpha)|\xrightarrow{p}0$. Now $\Gamma(\hat\alpha_n)\le\Gamma(\alpha)\ \forall\ \alpha$ implies

$$\frac{L_n(\hat\alpha_n)}{L_n(\alpha)}\le\frac{1+\zeta_n(\alpha)}{1+\zeta_n(\hat\alpha_n)}\le\frac{1+\max\limits_\alpha|\zeta_n(\alpha)|}{1-\max\limits_\alpha|\zeta_n(\alpha)|}\quad\forall\ \alpha.$$

Part (b) follows from the above. □

Proof of Lemma 3.1. We first note that under suitable conditions there exist $0 < \delta < \Delta$ such that

$$\log(1+\delta) < u_n(\alpha) < \log(1+\Delta) \quad \forall\alpha \tag{6.11}$$

with probability tending to 1. This follows from (3.3), (3.4) and the fact that $\boldsymbol{e}'\boldsymbol{e}/n\sigma^2 \xrightarrow{p} 1$, noting that $\max_\alpha \boldsymbol{e}'P(\alpha)\boldsymbol{e}/n \le \boldsymbol{e}'P\boldsymbol{e}/n \xrightarrow{p} 0$ and $L_n(\alpha)$ is uniformly (in α) bounded with probability tending to 1. Here P is the projection matrix corresponding to the full model.

Consider now the expression in (3.2). By Lemma 2.1 of Section 2 and (6.11),

$$\begin{aligned} \log[S(\alpha)/n\sigma^2] &= \log[\boldsymbol{e}'\boldsymbol{e}/n\sigma^2 + L_n(\alpha)/\sigma^2 + o_p(L_n(\alpha)/\sigma^2)] \\ &= \log[\boldsymbol{e}'\boldsymbol{e}/n\sigma^2 + L_n(\alpha)/\sigma^2 + o_p(\boldsymbol{e}'\boldsymbol{e}/n\sigma^2 + L_n(\alpha)/\sigma^2)] \\ &= \log[e^{u_n(\alpha)}(1+o_p(1))] \\ &= u_n(\alpha) + o_p(1) \\ &= u_n(\alpha) + o_p(u_n(\alpha)) \end{aligned} \tag{6.12}$$

uniformly in α. In view of (3.5), to prove (3.7), it remains to show

$$\frac{n-k}{nr}\sum_{i=1}^{r}\log[S_i(\alpha)/n\sigma^2] = o_p(1). \tag{6.13}$$

Note that we are also using (3.4) and (6.11). Since $S_i(\alpha) \ge S_i$ for all α and all i, we have for all α

$$0 < \frac{1}{r}\sum_{i=1}^{r}\log[S_i(\alpha)] = \log[\prod_{i=1}^{r} S_i(\alpha)]^{1/r} \le \log[\frac{1}{r}\sum_{i=1}^{r} S_i(\alpha)]$$

implying

$$-\frac{n-k}{n}\log(n\sigma^2) < \frac{n-k}{nr}\sum_{i=1}^{r}\log\left[\frac{S_i(\alpha)}{n\sigma^2}\right] \le \frac{n-k}{n}\log[\frac{1}{r}\sum_{i=1}^{r} S_i(\alpha)] - \frac{n-k}{n}\log(n\sigma^2).$$

Then (6.13) follows from Lemma 2.2 of Section 2, condition (3.4) and the fact that $L_n(\alpha)$ is uniformly (in α) bounded with probability tending to 1 (as noted earlier in the argument for (6.11)). □

Proof of Theorem 3.1. Let $\hat{\alpha}_n$ be the model which minimizes $\Gamma(\alpha)$. Proceeding as in the proof of part (b) of Theorem 2.1, and using (3.7) we can prove that

$$\frac{u_n(\hat{\alpha}_n)}{u_n(\alpha_n^L)} \xrightarrow{p} 1.$$

This, together with (6.11), imply that

$$\begin{aligned} &u_n(\hat{\alpha}_n) - u_n(\alpha_n^L) \xrightarrow{p} 0 \\ \text{i.e.,}\quad &\frac{\boldsymbol{e}'\boldsymbol{e} + nL_n(\hat{\alpha}_n)}{\boldsymbol{e}'\boldsymbol{e} + nL_n(\alpha_n^L)} \xrightarrow{p} 1. \end{aligned}$$

Since $\frac{\boldsymbol{e}'\boldsymbol{e}}{n} \xrightarrow{p} \sigma^2$ and $L_n(\alpha_n^L) \ge \min_\alpha \Delta_n(\alpha)$, using (3.3) we have

$$\frac{L_n(\hat{\alpha}_n)}{L_n(\alpha_n^L)} \xrightarrow{p} 1.$$

□

Proof of Proposition 4.1. We first prove equation (4.7). Below, by $\max\limits_{\alpha}$ we mean maximum over $\alpha \in \mathcal{A}_n^c$. Let $Z_n(\alpha) = (\boldsymbol{e}'P(\alpha)\boldsymbol{e})/(\sigma^2 p_n(\alpha))$. We first show that $\max\limits_{\alpha}|Z_n(\alpha)| = O_p(1)$. By (6.2)

$$\begin{aligned} & P[\max_{\alpha}|Z_n(\alpha)-1| > M] \\ \leq & \sum_{\alpha} E|Z_n(\alpha)-1|^{2m}/M^{2m} \\ \leq & \frac{C}{M^{2m}}\sum_{\alpha}\frac{1}{[p_n(\alpha)]^m} \end{aligned}$$

for some constant $C > 0$ and by (4.5) this can be made arbitrarily small by choosing suitable $M > 0$. Thus $\max\limits_{\alpha}|Z_n(\alpha) - 1| = O_p(1)$ implying $\max\limits_{\alpha}|Z_n(\alpha)| = O_p(1)$. This implies $(1/n)\boldsymbol{e}'P(\alpha)\boldsymbol{e} = o_p(\frac{1}{n}\lambda_n\sigma^2 p_n(\alpha))$ uniformly in $\alpha \in \mathcal{A}_n^c$ as $\lambda_n \to \infty$. Proceeding in a similar manner and noting that $(1/r)\sum\limits_{i=1}^{r}\boldsymbol{e}_i'P_i(\alpha)\boldsymbol{e}_i$ can be written as $\boldsymbol{e}'M(\alpha)\boldsymbol{e}$ (see proof of Lemma 2.2) one can prove

$$\frac{1}{nr}\sum_{i=1}^{r}\boldsymbol{e}_i'P_i(\alpha)\boldsymbol{e}_i = o_p(\frac{1}{n}\lambda_n\sigma^2 p_n(\alpha)) \text{ uniformly in } \alpha \in \mathcal{A}_n^c.$$

The result now follows from (4.2), (4.4) and (4.6).
In order to complete the proof of Proposition 4.1, we now prove equation (4.8). From (4.7),

$$\Gamma(\alpha_n^c) = \frac{k}{n^2}\boldsymbol{e}'\boldsymbol{e} + \frac{1}{n}\lambda_n\sigma^2 p_n(\alpha_n^c) + o_p\left(\frac{1}{n}\lambda_n\sigma^2 p_n(\alpha_n^c)\right).$$

The result follows from (4.1) and (4.2) noting that (4.1) implies (6.5) with $\max\limits_{\alpha\in\mathcal{A}_n}$ replaced by $\max\limits_{\alpha\in\mathcal{A}_n-\mathcal{A}_n^c}$. □

Proof of (4.18). Note that

$$\max_{\alpha\in\mathcal{A}_n-\mathcal{A}_n^c}\frac{L_n(\alpha_n^c)}{L_n(\alpha)} = \max_{\alpha}\frac{\boldsymbol{e}'P(\alpha_n^c)\boldsymbol{e}}{nL_n(\alpha)}.$$

By (6.2) and by arguments used earlier

$$P\left[\max_{\alpha\in\mathcal{A}_n-\mathcal{A}_n^c}\left|\frac{\boldsymbol{e}'P(\alpha_n^c)\boldsymbol{e} - \sigma^2 p_n(\alpha_n^c)}{nR_n(\alpha)}\right| > \epsilon\right] \leq C\left[\frac{p_n}{\min\limits_{\alpha} nR_n(\alpha)}\right]^m \sum_{\alpha\in\mathcal{A}_n-\mathcal{A}_n^c}\frac{1}{[nR_n(\alpha)]^m}$$

for some constant C. The result follows from (4.1) and (4.2). □

Acknowledgements

We would first like to thank Professors Bertrand Clarke and Subhashis Ghosal for inviting us to contribute to this volume. We consider it a great privilege to be able to write an article as an expression of our deep regard for our mentor and teacher Professor Jayanta K. Ghosh. We cannot express in words how much we learned in numerous academic discussions we had with him over the years which immensely influenced our thought process. Needless to say, this work owes much to the way we learned to think about model selection through our association with him. Finally, we thank an anonymous referee and the editors for helpful comments and suggestions towards improvement of the paper.

References

[1] Barbieri, M.M. and Berger, J.O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870-897.

[2] Berger, J.O. and Pericchi, L.R. (1996a). The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.* **91** 109-122.

[3] Berger, J.O. and Pericchi, L.R. (1996b). The intrinsic Bayes factor for linear models (with discussion). In: Bernardo, J.M. et al. (eds) *Bayesian Statistics* **5** 25-44. Oxford Univ. Press, London.

[4] Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory.* Wiley, Chichester, England.

[5] Chakrabarti, A. and Ghosh, J.K. (2007). Some aspects of Bayesian model selection for prediction (with discussion). In: Bernardo, J.M. et al. (eds) *Bayesian Statistics* **8** (To appear).

[6] Chatterjee, S. and Hadi, A.S. (1988). *Sensitivity Analysis in Linear Regression.* Wiley, New York.

[7] Geisser, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.* **70** 320-328.

[8] Geisser, S. and Eddy, W.F. (1979). A predictive approach to model selection. *J. Amer. Statist. Assoc.* **74** 153-160.

[9] Gelfand, A.E. and Dey, D.K. (1994). Bayesian model choice: Asymptotics and exact calculations. *J. Roy. Statist. Soc.* (Ser. B) **56** 501-514.

[10] Gelfand, A.E. and Ghosh, S.K. (1998). Model choice: a minimum posterior predictive loss approach. *Biometrika* **85** 1-11.

[11] Ghosh, J.K. and Samanta, T. (2002). Nonsubjective Bayes testing – an overview. *J. Statist. Plann. Inference* **103** 205-223.

[12] Li, K.-C. (1987). Asymptotic optimality for C_p, CL, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.* **15** 958-975.

[13] Mukhopadhyay, N. (2000). Bayesian model selection for high dimensional models with prediction error loss and $0-1$ loss. Ph.D. thesis, Purdue Univ.

[14] Mukhopadhyay, N., Ghosh, J.K. and Berger, J.O. (2005). Some Bayesian predictive approaches to model selection. *Stat. Prob. Lett.* **73** 369-379.

[15] O'Hagan, A. (1995). Fractional Bayes factors for model comparisons. *J. Roy. Statist. Soc.* (Ser. B) **57** 99-138.

[16] Shao, J. (1997). An asymptotic theory for linear model selection (with discussion). *Statistica Sinica* **7** 221-264.

[17] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statist. Soc.* (Ser. B) **36** 111-147.

[18] Whittle (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab. Appl.* **5** 302-305.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 156

Risk and resampling under model uncertainty

Snigdhansu Chatterjee[1], **and Nitai D. Mukhopadhyay**[2]

University of Minnesota and Virginia Commonwealth University

Abstract:

In statistical exercises where there are several candidate models, the traditional approach is to select one model using some data driven criterion and use that model for estimation, testing and other purposes, ignoring the variability of the model selection process. We discuss some problems associated with this approach. An alternative scheme is to use a model-averaged estimator, that is, a weighted average of estimators obtained under different models, as an estimator of a parameter. We show that the risk associated with a Bayesian model-averaged estimator is bounded as a function of the sample size, when parameter values are fixed. We establish conditions which ensure that a model-averaged estimator's distribution can be consistently approximated using the bootstrap. A new, data-adaptive, model averaging scheme is proposed that balances efficiency of estimation without compromising applicability of the bootstrap. This paper illustrates that certain desirable risk and resampling properties of model-averaged estimators are obtainable when parameters are fixed but unknown; this complements several studies on minimaxity and other properties of post-model-selected and model-averaged estimators, where parameters are allowed to vary.

Contents

1. Introduction

In typical statistical applications, it is rare that a precise model is available to fit to the data. Selecting one model from several competing models, is often the first step in the process. However, in the subsequent analysis, it is common to ignore the variability in the initial model selection. Two of the many consequences of ignoring modeling variability are (i) under-estimation of the variability of estimators and

[1] School of Statistics, 313 Ford Hall, 224 Church Street SE, Minneapolis, MN 55455e-mail: `chatterjee@stat.umn.edu`

[2] Department of Biostatistics, 730 E. Broad Street, Richmond, Virginia 23298 e-mail: `ndmukhopadhy@vcu.edu`

AMS 2000 subject classifications: Primary 60F12; secondary 60J05, 62C10, 62F40

Keywords and phrases: Model selection, model averaging, bounded risk, bootstrap, linear regression.

predictors, and (*ii*) erroneous inference and prediction, resulting from incorrectly computing the distributions of estimators and predictors. An alternative to selecting a model first and then computing an estimator under that model is to consider several models and appropriately average the estimators computed under these models.

Several studies have been published recently on the properties of post-model-selected and model-averaged estimators; see for example, [8], [23] and [24]. These studies are discouraging as they show that many nice properties associated with estimators under a known model vanish when there is model uncertainty. For example, Yang [23] shows that consistent model selection/averaging, and minimax-rate optimality cannot be simultaneously obtained. The review of Leeb and Pötscher [8] contains a discussion of several other problems with inference after model selection.

In view of these negative results, it seems desirable to scale down our expectations while working under model uncertainty, and strive for positive, if weaker, results. This may be achieved in one of two ways: we may either impose less stringent conditions on our estimators, or we may relax the criterion by which an estimator is evaluated. The latter is the goal of the present study.

The computation of an estimator is generally one of the early steps in a statistical exercise. Estimators of parameters are used for various purposes, notably for quantifying evidence for or against scientific hypotheses, obtaining interval estimates for the parameter under consideration, for prediction and forecasting, and for quantifying the accuracy of predictions and forecasts. These applications require knowledge about the distribution of the estimator, and knowledge about the risk associated with the usage of such estimators. In this paper, we concentrate on the risk behavior of a model-averaged estimator, and on approximating the distribution of a model-averaged estimator using the bootstrap.

In the first part of our study we show that under the traditional frequentist assumption that the parameters are fixed but unknown constants, the mean squared error in regression estimation under consistent model selection/averaging is bounded as a function of sample size. This complements Yang [23], where it was shown that a similar quantity cannot achieve minimax-rate optimality. Several of the negative results, including those of Yang [23], arise when a parameter is a known constant in a smaller model, while it is allowed to vary in a local neighborhood of that constant in a larger model. Recently, Hjort and Claeskens [5] studied model averaged estimators under a local parameter framework. Local parameters are ideal for mathematical development, but they are not reflective of statistical reality; see [17]. Indeed, as Hjort and Claeskens themselves remark in the rejoinder to the discussion of their paper, "a too literal belief in sample-size-dependent parameters would clash with Kolmogorov consistency and other requirements of natural statistical models." [5]. In view of this, it is meaningful to verify that estimators have reasonable risk behavior under consistent model selection/averaging when parameters are fixed constants. Our result also implies that *integrated risks* under consistent model selection/averaging are bounded, when integrals are taken with respect to any probability measure on the parameter space that does not depend on sample size.

In the second part of our study, in addition to the assumption that the parameters are fixed but unknown constants, we also weaken the consistency requirement of the model averaging procedure. In the terminology of Yang [23], a model selection/averaging scheme is *consistent* if it is asymptotically degenerate at the true model, when the true model is one of the candidate models. When the models are nested and several of them can correctly describe the data generation process, the most parsimonious correct model is taken as the true model. We call this *strong consistency.* We define a model selection/averaging scheme as *weakly consistent* if it selects or averages over all candidate models that correctly describe the data generation process. When only one model is correct, the strong and weak consistency requirements are identical; but if models are nested and several of them are correct, a weakly consistent scheme may distribute weights among all of them while a strongly consistent one is asymptotically degenerate at the smallest one. Recently, Leung and Barron [11] proposed a scheme of model averaging that results in nice risk behavior. Their scheme is an example of a weakly consistent procedure. We show that a particular choice of a weakly consistent model-averaged estimator has a distribution that can be approximated using the bootstrap.

In Section 2 we propose a simple linear regression model framework to study model uncertainty. We also discuss some of the properties of post-model-selection estimators that make them unsuitable for further applications, and also some properties of model-averaged estimators. This is followed in Section 3 with a discussion of mean squared error of the Bayesian model-averaged estimator. In Section 4 we propose a new adaptive, model-averaged estimator whose distribution may be consistently approximated using the bootstrap. A simulation example is discussed in Section 5. Finally, in Section 6 we discuss some aspects of our results, and point to some open issues relating to model uncertainty.

2. Issues with Model Selection or Averaging

We select a simple regression framework for our study, which is the same as that used by [8], and similar to that of [24]. The observed data $\{(Y_t, \mathbf{x}_t = (x_{t1},\ x_{t2})^T), t = 1, \dots, n\}$, are modeled as

$$Y_t = \alpha x_{t1} + \beta x_{t2} + e_t, \tag{2.1}$$

where the e_t's are independent, identically distributed $N(0, \sigma^2)$, σ^2 known. The design matrix $\mathbf{X}$ with rows given by $\mathbf{x}_t^T = (x_{t1}, x_{t2})$ is non-random. We denote the two columns of $\mathbf{X}$ as X_1 and X_2, the vector of errors as $\mathbf{e}$, and the vector of observations as $\mathbf{Y}$. The inner products and norms used below are the usual Euclidean ones. The notation D is used for the determinant of the design matrix, thus $D = ||X_1||^2||X_2||^2 - < X_1, X_2 >^2$. The unknown parameters in this model are (α, β). Model uncertainty surrounds the issue of whether or not $\beta = 0$. In this paper, for ease in presentation, we consider the problem of estimation of α.

We make the standard assumption that $n^{-1}\mathbf{X}^T\mathbf{X} \to Q$ for a positive definite matrix Q. This, in particular, implies the standard design conditions

$$\begin{aligned} ||X_1||^2 &= O(n), \quad ||X_2||^2 = O(n), && (2.2)\\ < X_1, X_2 > &= O(n), \quad D = ||X_1||^2||X_2||^2 - < X_1, X_2 >^2 = O(n^2). && (2.3) \end{aligned}$$

We also assume that $n^{-1} < X_1, X_2 > \not\longrightarrow 0$ as $n \to \infty$, since without this restriction the effect of model uncertainty vanishes in this framework.

The true model, called M_0, may be described as

$$M_0 = \begin{cases} U \text{ (unrestricted)} & \text{if } \beta \neq 0; \\ R \text{ (restricted)} & \text{if } \beta = 0. \end{cases}$$

Under U, we adopt the ordinary least squares or maximum likelihood estimators $\widehat{(\alpha, \beta)} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Our notation for these are $(\hat{\alpha}(U), \hat{\beta}(U))$. Under R, $\hat{\beta}(R) \equiv 0$, and the ordinary least squares or maximum likelihood estimator for α is $\hat{\alpha}(R) = [\sum_i x_{1i}^2]^{-1} \sum x_{1i}y_i$. Define $V_1 = \sigma^{-1}||X_1||^{-1} < X_1, \mathbf{e} >

and $V_2 = \sigma^{-1} D^{-1/2} ||X_1|| \left\{ < X_2, \mathbf{e} > - ||X_1||^{-2} < X_1, X_2 >< X_1, \mathbf{e} > \right\}$, thus $\mathbf{V} = (V_1, \;\; V_2)^T \sim N(0, \mathbf{I}_2)$. In terms of $\mathbf{V}$, the estimators are

$$\begin{pmatrix} \hat{\alpha}(R) \\ \hat{\alpha}(U) \\ \hat{\beta}(U) \end{pmatrix} = \begin{bmatrix} \alpha + \beta ||X_1||^{-2} < X_1, X_2 > + \sigma ||X_1||^{-1} V_1, \\ \alpha + \sigma ||X_1||^{-1} V_1 - \sigma ||X_1||^{-1} D^{-1/2} < X_1, X_2 > V_2, \\ \beta + \sigma ||X_1|| D^{-1/2} V_2. \end{bmatrix}$$

The dichotomy between the *bias* of the restricted model R and the *variance* of the unrestricted model U can be clearly seen in the above formula. The restricted model estimator $\hat{\alpha}(R)$ has a bias factor $\beta ||X_1||^{-2} < X_1, X_2 >$, which vanishes under R, while $\hat{\alpha}(U)$ has an extra factor of $\sigma ||X_1||^{-1} D^{-1/2} < X_1, X_2 > V_2$ that inflates its variance relative to $\hat{\alpha}(R)$. Hence, model selection or model averaging is essentially a process of balancing bias and variance; see [20].

Let σ_β be the standard deviation of $\hat{\beta}(U)$. This is a non-random, known number depending on σ^2 and $\mathbf{X}$. The following model selection criterion is used:

$$\hat{M} = \begin{cases} U & \text{if } |n^{-1/2} \sigma_\beta^{-1} \hat{\beta}(U)| > c; \\ R & \text{if } |n^{-1/2} \sigma_\beta^{-1} \hat{\beta}(U)| \le c. \end{cases}$$

The above criterion may be identified as representative of standard model selection tools, in the simple regression model. In particular, the above criterion is the traditional pre-test procedure based on the likelihood ratio, coincides with the *Akaike Information Criterion* (AIC) if $c = \sqrt{2}$, and coincides with the *Bayesian Information Criterion* (BIC) if $c = \sqrt{\log n}$. The post-model-selection estimator of α is

$$\tilde{\alpha} = \hat{\alpha}(R) I_{\{\hat{M}=R\}} + \hat{\alpha}(U) I_{\{\hat{M}=U\}}. \tag{2.4}$$

Several nice properties are known about $\hat{M}$ and, consequently, it is generally believed that $\tilde{\alpha}$ will also have good properties. Some of the important properties include that for all β and as $c \to \infty, n^{-1/2} c \to 0$, $P[\hat{M} = M_0] \to 1$, $\{\hat{M} = M_0\} \subseteq \{\tilde{\alpha} = \hat{\alpha}(M_0)\}$ and thus $P[\tilde{\alpha} = \hat{\alpha}(M_0)] \to 1$ (see [15]). Note that $\hat{\alpha}(M_0)$ is the 'oracle's guess' about α, and is not a statistic, since it is based on the knowledge of β. The above properties tend to give the impression that $\tilde{\alpha}$ is a very good estimator.

However, there are some major problems since the above results are asymptotic in nature, and the asymptotics can take a long time to kick in, as well as be dependent on the value of β. Our primary reference for this model and its basic properties [8] identifies this as a problem of *non-uniformity* in β of the convergence of $\hat{M}$ and $\tilde{\alpha}$. It can be immediately seen that the estimator $\tilde{\alpha}$ is super-efficient when $c \to \infty$, $c/\sqrt{n} \to 0$, as with BIC. The major repercussions of super-efficiency of $\tilde{\alpha}$ and the non-uniformity of its asymptotics is in its risk performance, and in its finite sample behavior. The mean squared error of $\tilde{\alpha}$ is unbounded and depends on β, while that of $\hat{\alpha}(M_0)$ is a constant. As a consequence, the finite sample behavior of $\tilde{\alpha}$ is erratic and can be quite unlike its asymptotic approximation. Available simulations confirm this; see [8]. Several other studies conducted by Leeb, Pötscher, Yang and others reveal how and why the properties of $\tilde{\alpha}$ and $\hat{\alpha}(M_0)$ differ. For further information see, for example, [6–10, 22, 24, 25].

The super-efficiency of $\tilde{\alpha}$ results in most variations of the bootstrap being inapplicable. Only subsampling ([14]) and the m-out-of-n bootstrap with $m/n \to 0$ would yield consistent approximations of the distribution of $\tilde{\alpha}$. Unfortunately, these methods have problems of their own, some details of which can be found in [18] and

[1]. Specifically, although subsampling is asymptotically consistent, it can perform miserably in finite samples. For any $\alpha \in (0, 1)$, the actual asymptotic coverage of a standard level $(1 - \alpha)$ subsampling confidence interval can be zero; see [1] for details. The finite sample properties of subsampling based methods can be improved sometimes by considering hybrid techniques, calibrations and other modifications, as documented by [2]. However, the asymptotic zero coverage of subsampling intervals for $\tilde{\alpha}$ cannot be reversed by, for example, size correction, since technical conditions that allow for such correction to work are not satisfied by $\tilde{\alpha}$.

The above issues with post-model-selection estimators lead to model-averaged estimators. A model-averaged estimator of α is of the form

$$\breve{\alpha} = \hat{\alpha}(R)p_R + \hat{\alpha}(U)p_U, \tag{2.5}$$

where p_R and p_U are two *weights* associated with the models R and U. Yang and his co-authors have extensively studied aggregation across models for several statistical procedures like estimators and forecasts, in both their algorithmic as well as theoretical aspects (see [22–25]). In particular, a result of [23] implies that when the model averaging technique is strongly consistent, the supremum of the mean squared error of $n^{1/2}(\breve{\alpha} - \alpha)$ over values of $(\alpha,\ \beta)$ tends to infinity. Thus, strongly consistent model averaging does not attain the minimax rate. Our result in Section 3 shows that, up to constant terms, it is no worse than the post-model-selection estimator when $(\alpha,\ \beta)$ are held fixed.

Recently, [5] studied several forms of model averaging and showed that a typical model-averaged estimator converges weakly to a mixture of normal laws, when the parameters of the true model are in a $O(n^{-1/2})$ neighborhood of the simplest candidate in a nesting of models. Since subsampling does not seem to perform well in practice, it is important to study conditions on model weights under which bootstrap approximations of finite sample distributions hold, *i.e.*, conditions under which the statistic under consideration is smooth and asymptotically normal (see [12], [13]). This is studied in Section 4.

3. Risk Profile of Model-Averaged Estimators

Several problems associated with the post-model-selection estimator can be attributed to its lack of uniformity, as discussed extensively by others [8]. One is the super-efficiency of $\tilde{\alpha}$, for example, when BIC is used for model selection. The core problem of lack of uniformity in the convergence pattern of $\tilde{\alpha}$ is unavoidable – even with model averaging – when a strongly consistent model averaging technique is used, as described by [23]. In this section we show that when parameter values are fixed, model averaging is no worse than model selection, up to constant terms.

Under the unrestricted model, U, we choose the prior on $(\alpha,\ \beta)$ to be a standard mean zero, identity covariance bivariate Normal distribution, $N(0, I)$. Under the restricted model, R, the prior on α is a standard univariate Normal distribution, $N(0, 1)$. We put equal prior weights, *i.e.*, 1/2, on the models, so the prior odds is 1. Our notation for the posterior probabilities of the two models are π_{nU} and π_{nR}. Since σ is known, without loss of generality we also assume $\sigma = 1$ in this section. Thus the Bayesian model-averaged estimator of α is

$$\hat{\alpha}_{BMA} \quad = \quad \pi_{nU}\hat{\alpha}(U) + \pi_{nR}\hat{\alpha}(R). \tag{3.6}$$

We use the pre-selected, least squares estimators $\hat{\alpha}(U)$ and $\hat{\alpha}(R)$ as constituents of $\hat{\alpha}_{BMA}$, and consider the squared error loss function. The case where a general loss function is used, with $\hat{\alpha}(U)$ and $\hat{\alpha}(R)$ taken to be the Bayes estimators under models U and R, is very similar. The following Proposition is our main result in this section.

Proposition 3.1. *The normalized risk of* $\hat{\alpha}_{BMA}$*,* $nR(\alpha) = nE(\hat{\alpha}_{BMA} - \alpha)^2$*, satisfies* $\sup_n nR(\alpha) < \infty$*, for every fixed choice of* α *and* β*. Hence, the integrated normalized risk* $\sup_n \int_{\alpha,\ \beta} nR(\alpha) d\lambda(\alpha,\ \beta) < \infty$ *for any probability measure* $\lambda(\cdot)$ *that does not depend on* n*.*

Proof of Proposition 3.1. In the following, we use C as a generic constant, not depending on the parameters α and β or the sample size n.

Note that $\hat{\alpha}(R) = \hat{\alpha}(U) + \hat{\beta}(U)||X_1||^{-2} < X_1, X_2 >$. Therefore,

$$
\begin{aligned}
nR(\alpha) &= nE\left[\pi_{nU}\hat{\alpha}(U) + \pi_{nR}\hat{\alpha}(R) - \alpha\right]^2 \\
&\leq 2nE\left(\hat{\alpha}(U) - \alpha\right)^2 + 2n||X_1||^{-4} < X_1, X_2 >^2 E\left\{\pi_{nR}^2 \hat{\beta}^2(U)\right\}. \qquad (3.7)
\end{aligned}
$$

Note that $E\left(\hat{\alpha}(U) - \alpha\right)^2 = \sigma^2 ||X_1||^{-2} E\left[V_1 - < X_1, X_2 > D^{-1/2} V_2\right]^2 = Cn^{-1}$ and $E\pi_{nR}^2 \hat{\beta}^2(U). \leq 2\beta^2 E\pi_{nR}^2 + Cn^{-1}$. Thus, we need suitable bounds for $\beta^2 E\pi_{nR}^2$. We now have

$$
p_{nR} = m_R(\mathbf{Y}) / \left(m_U(\mathbf{Y}) + m_U(\mathbf{Y})\right) = \frac{m_R(\mathbf{Y})}{m_U(\mathbf{Y})}\left(1 + \frac{m_R(\mathbf{Y})}{m_U(\mathbf{Y})}\right)^{-1} \leq \frac{m_R(\mathbf{Y})}{m_U(\mathbf{Y})}.
$$

Then, making use of the moment generating function of a χ^2 random variable, we can deduce that

$$
E\left(\frac{m_R(\mathbf{Y})}{m_U(\mathbf{Y})}\right)^2 = Cn^2 \exp\left\{-nC_0(\alpha^2 + \beta^2)\right\}
$$

for a particular constant C_0. This yields, at (3.7), that

$$
nR(\alpha) = Cn^{-1} + Cn^3 \beta^2 \exp\left\{-nC_0(\alpha^2 + \beta^2)\right\}.
$$

which is bounded for every fixed $(\alpha,\ \beta)$, as a function of n. The rest of the result follows. □

Remark 3.1 A lower bound for $nR(\alpha)$ can also be established using arguments similar to those above. With slight modification, the above approach using the moment generating function of a non-central χ^2 random variable can be used to provide an alternative proof of Theorem 2 of [23]. It can also be seen that even when $(\alpha,\ \beta)$ vary over a compact set, the supremum of $nR(\alpha)$ over (α, β) is unbounded.

4. Adaptive Model-Averaged Estimators and the Bootstrap

The results of Hjort and Claeskens [5] and Leeb and Potscher [8] indicate that the post-model-selection estimator and many model-averaged estimators cannot be consistently bootstrapped. The problems associated with the risk behavior, and those associated with bootstrap approximation, arise from two different sources. Undesirable behavior of the risk function arises from considering scenarios as parameters vary, while a major reason why the distribution of post-model-selection or model-averaged estimators cannot be approximated by bootstrap methods is because of lack of smoothness of the estimator, or lack of asymptotic normality.

In this section we study the conditions on the model weights which are required for consistent bootstrap approximation of the distribution of the resulting model-averaged estimator. Clearly, since the distribution of $\hat{\alpha}(U)$ can be approximated using the bootstrap, putting the entire weight on model U is an option. However, balancing between $\hat{\alpha}(U)$ and $\hat{\alpha}(R)$ can lead to a more efficient estimator. We propose below a data-adaptive model weighing scheme that achieves the dual goals of reasonable efficiency and bootstrap consistency.

A model-averaged estimator of α is of the form

$$\breve{\alpha} = \hat{\alpha}(R)p_{nR} + \hat{\alpha}(U)p_{nU}. \tag{4.8}$$

Notice that we have adopted a different notation (p_{nR} and p_{nU}) for the model weights in this Section, from those (π_{nR} and π_{nU}) used in Section 3. This is to emphasize that the nature of these weights may be different. We retain the condition that the parameters (α, β) are fixed but unknown.

A primary requirement for consistency is $p_{nR} + p_{nU} = 1$, as pointed out in [5]. In order to avoid pathologies, we also specify that $p_{nU} \in [0, 1]$. Note that the weights p_{nR} and p_{nU} may depend on the parameters (α, β), and the random component $\mathbf{V}$, apart from the known constants $\mathbf{X}$ and σ^2.

Replacing p_{nU} by $1 - p_{nR}$, we thus have

$$\begin{aligned} \breve{\alpha} \quad &= \quad \alpha + \sigma||X_1||^{-1}V_1 + \beta p_{nR}||X_1||^{-2} < X_1, X_2 > \\ &\qquad -\sigma||X_1||^{-1}D^{-1/2} < X_1, X_2 > (1 - p_{nR})V_2. \end{aligned}$$

A primary requirement on $\breve{\alpha}$ is that it should be consistent, and the following proposition establishes a necessary and sufficient condition for this.

Proposition 4.1. *The model-averaged estimator $\breve{\alpha}$ converges in probability to α if and only if βp_{nR} converges in probability to zero as $n \to \infty$.*

Proof of Proposition 4.1. The sufficiency part follows easily from the design conditions (2.2)-(2.3). For the necessity part, suppose that $\beta p_{nR} \xrightarrow{p} \tilde{c} \neq 0$ as $n \to \infty$. This is clearly equivalent to $p_{nR} \xrightarrow{p} c = \tilde{c}/\beta \neq 0$ as $n \to \infty$ and $\beta \neq 0$. Hence, we also have $(1 - p_{nR}) \left\{\sigma||X_1||^{-1}D^{-1/2} < X_1, X_2 > V_2\right\} \xrightarrow{p} (1 - c)0 = 0$. This implies $\breve{\alpha} \xrightarrow{p} \alpha - \tilde{c}\gamma \neq \alpha$, where $||X_1||^{-1} < X_1, X_2 > \to \gamma$ as $n \to \infty$. The case where p_{nR} does not have a limit can be treated similarly with a little more algebra. □

The next proposition is an extension of the previous one, and establishes sufficient conditions for asymptotic normality of $\breve{\alpha}$.

Proposition 4.2. *The scaled and centered model-averaged estimator $n^{1/2}(\breve{\alpha} - \alpha)$ has an asymptotic normal distribution if (i) $n^{1/2}\beta p_{nR}$ converges in probability to zero as $n \to \infty$, and (ii) p_{nR} converges in probability as $n \to \infty$ for all values of (α, β).*

Proof of Proposition 4.2. The first condition forces the bias component in $\breve{\alpha}$ to be $o(n^{-1/2})$, while the second condition allows for use of Slutsky's theorem. □

By requiring $n^{1/2}\beta p_{nR} \xrightarrow{p} 0$ as $n \to \infty$ we have ensured that, when $\beta \neq 0$, we have $n^{1/2}p_{nR} \xrightarrow{p} 0$. Thus the model-averaged estimator is close to the unrestricted model estimator $\hat{\alpha}(U)$, and has the same limiting distribution up to first order terms. However, when $\beta = 0$, the asymptotic distribution of $n^{1/2}(\check{\alpha} - \alpha)$ depends on the limit of p_{nR}, which is between zero and one. Thus, when the restricted model holds, the asymptotic variance of $\check{\alpha}$ is between that of $\hat{\alpha}(R)$ and $\hat{\alpha}(U)$. The relative strengths of different candidates for model weight p_{nR} may be evaluated by their probability limits when $\beta = 0$. We note that we consider $(\alpha,\ \beta)$ as fixed constants and do not allow them to vary with n. If, for example, we assumed $\beta = O(n^{-1/2})$, then the first condition of Proposition 4.2 would imply asymptotically zero weight on the restricted model.

In order to progress towards bootstrap consistency, apart from asymptotic normality of $\check{\alpha}$, we also need p_{nR} to be a smooth function. Thus ruling out the indicator function $p_{nR} = I_{\{|n^{-1/2}\sigma_\beta^{-1}\hat{\beta}(U)| \le c\}}$ used in $\tilde{\alpha}$. Keeping in view the nice properties of $\tilde{\alpha}$, we now develop an adaptive, data-driven model weight function p_{nR} that is a smooth version of $I_{\{|n^{-1/2}\sigma_\beta^{-1}\hat{\beta}(U)| \le c\}}$.

For any k_n, we split the event $\{-k_n \le \hat{\beta}(U) \le k_n\}$ into two events, $\{\hat{\beta}(U) - k_n \le 0\}$ and $\{\hat{\beta}(U) + k_n \ge 0\}$, and approximate the indicators of these events separately. Our approximation for $I_{\{\hat{\beta}(U) - k_n \le 0\}}$ is

$$\begin{aligned}\xi_{1n} &\equiv \xi_{1n}\left(\gamma_{1n}, \hat{\beta}(U), k_n\right) \\ &= \left(1 + \exp\left\{-\gamma_{1n}(\hat{\beta}(U) - k_n)\right\}\right)^{-1} \exp\left\{-\gamma_{1n}(\hat{\beta}(U) - k_n)\right\},\end{aligned}$$

and for $I_{\{\hat{\beta}(U) + k_n \ge 0\}}$ is

$$\xi_{2n} \equiv \xi_{2n}\left(\gamma_{1n}, \hat{\beta}(U), k_n\right) = \left(1 + \exp\left\{\gamma_{2n}(\hat{\beta}(U) + k_n)\right\}\right)^{-1} \exp\left\{\gamma_{2n}(\hat{\beta}(U) + k_n)\right\}.$$

We take the two tuning values γ_{1n} and γ_{2n} to be always positive. However, they change with n; and in a major departure from traditional model weights, they are not equal to each other, and also depend on the data. Thus, $\gamma_{1n} \equiv \gamma_{1n}(\alpha, \beta, \mathbf{V})$ and $\gamma_{2n} \equiv \gamma_{2n}(\alpha, \beta, \mathbf{V})$ are unequal, random weights.

Equipped with these functions, we define

$$p_{nR} = 0.5\xi_{1n} + 0.5\xi_{2n}.$$

We adopt the *paired bootstrap* as our resampling strategy. Thus, we draw a simple random sample with replacement of the data pairs $(Y_i^*, \mathbf{x}_i^*),\ i = 1, \ldots, n$, from the original data $(Y_i, \mathbf{x}_i),\ i = 1, \ldots, n$. The entire process of obtaining $\hat{\alpha}(R)$, $\hat{\alpha}(U)$, $\hat{\beta}(R)$, p_{nR}, and $\check{\alpha}$ is imitated with the resample $(Y_i^*, \mathbf{x}_i^*),\ i = 1, \ldots, n$, and we approximate the distribution of $n^{1/2}(\check{\alpha} - \alpha)$ with the distribution of $n^{1/2}(\check{\alpha}^* - \check{\alpha})$, conditional on $(Y_i, \mathbf{x}_i),\ i = 1, \ldots, n$. A technical condition guarantees that the design matrix from the resampled data is non-singular with high probability; see condition (1.17) of [3].

The following Theorem is our main result in this section, and establishes consistency of the bootstrap for a adaptively weighted model-averaged estimator.

Theorem 4.1. *Assume that sequence of constants $k_n \downarrow 0$ as $n \to \infty$. Suppose the tuning constants are chosen as $\gamma_{1n} = a_n\hat{\beta}(U)$ $\gamma_{2n} = -a_n\hat{\beta}(U)$ where $\{a_n\}$ is a sequence of positive constants satisfying $a_n^{-1}\log(n) \downarrow 0 \quad$ as $n \to \infty$.*

Then $n^{1/2}(\check{\alpha} - \alpha)$ has an asymptotic Normal distribution and the paired bootstrap is consistent for it.

Proof of Theorem 4.1. For the asymptotic normality we only need to check that the conditions of Proposition 4.2 are met. We illustrate the calculation for verifying $n^{1/2}\xi_{1n} \xrightarrow{p} 0$, when $\beta \neq 0$.

$$\begin{aligned}
& P\left[|n^{1/2}\xi_{1n}| > \epsilon\right] \\
= \;& P\Big[| \left(1+\exp\left\{-\gamma_{1n}(\hat{\beta}(U)-k_n)\right\}\right)^{-1} \\
& \qquad\qquad \exp\left\{-\gamma_{1n}(\hat{\beta}(U)-k_n)+0.5\log(n)\right\} | > \epsilon\Big] \\
\leq \;& P\left[\exp\left\{-\gamma_{1n}(\hat{\beta}(U)-k_n)+0.5\log(n)\right\} > \epsilon\right] \\
= \;& P\left[\hat{\beta}(U) \text{ lies between the roots of } x^2-k_nx-0.5a_n^{-1}\log(n)+a_n^{-1}\log(\epsilon)=0\right].
\end{aligned}$$

The roots of the equation $x^2-k_nx-0.5a_n^{-1}\log(n)+a_n^{-1}\log(\epsilon)=0$ are always real when $\epsilon<1$, since $k_n^2+2a_n^{-1}\log(n)-4a_n^{-1}\log(\epsilon)>0$ for all n. Note that the square of the distance between the roots is given by $\left(k_n^2+2a_n^{-1}\log(n)-4a_n^{-1}\log(\epsilon)\right)/4$. When $k_n \downarrow 0$, $k_n^2+2a_n^{-1}\log(n)-4a_n^{-1}\log(\epsilon)\downarrow 0$, hence the Lebesgue measure of the interval between the roots goes to zero as $n\to\infty$, thus ensuring

$$P\left[\hat{\beta}(U) \text{ lies between the roots of } x^2-k_nx-0.5a_n^{-1}\log(n)+a_n^{-1}\log(\epsilon)=0\right] \to 0,$$

as $n\to\infty$. Note that this result actually does not depend on the value of β, as long as it is non-zero.

Other parts of the proof for asymptotic Normality may be verified similarly. Since $\breve{\alpha}$ is a smooth function of α, β and $\mathbf{V}$, and has an asymptotic Normal distribution, the consistency of the paired bootstrap procedure follows from [12] and [13]. □

Remark 4.1. The condition $k_n \downarrow 0$ as $n\to\infty$ is a weaker restriction than typically found in literature. Since $\hat{\beta}(U)=O_p(n^{-1/2})$, the AIC criterion uses $k_n=O(n^{-1/2})$, while the BIC uses $k_n=O(n^{-1/2}\sqrt{\log(n)})$.

Remark 4.2. The assumptions of Proposition 4.1 and Proposition 4.2 cannot be weakened in general. The example of Section 10.6 of [5] provides a test case. It is a simpler version of the model described in Section 2, and simply has $Y_1,\ldots,Y_n$ independent, identically distributed as $N(\mu,1)$ random variables. Model uncertainty is about whether $\mu=0$, and the natural estimator for μ is $\bar{Y}_n=n^{-1}\sum_{i=1}^{n}Y_i$ in the unrestricted model, and 0 in the restricted model. A model-averaged estimator is $\hat{\mu}=W(n^{1/2}\bar{Y}_n)\bar{Y}_n$, for some weight $W(\cdot)\in[0,1]$. Note that under a model with contiguous alternatives $\mu_{\text{true}}=n^{-1/2}\delta$, the requirement that $\hat{\mu}$ be consistent for μ_{true} actually places no restriction on the weight $W(\cdot)$, which may take any value in $[0,1]$. However, if we want consistency under arbitrary μ, $W(n^{1/2}\bar{Y}_n)\xrightarrow{p}1$ is a requirement.

For asymptotic normality, $n^{1/2}\mu(1-W(n^{1/2}\bar{Y}_n))\xrightarrow{p}0$ and convergence in probability of $W(n^{1/2}\bar{Y}_n)$, are requirements. Under μ_{true}, this implies that $W(n^{1/2}\bar{Y}_n)\xrightarrow{p}1$ must hold, while for general μ, the stronger condition $n^{1/2}(1-W(n^{1/2}\bar{Y}_n))\xrightarrow{p}0$ must be satisfied.

Under μ_{true}, it is of interest to approximate the distribution of the standardized statistic

$$\Lambda_n=n^{1/2}(\hat{\mu}-\mu_{\text{true}})=n^{1/2}W(n^{1/2}\bar{Y}_n)\bar{Y}_n-\delta=W(\delta+Z_n)(\delta+Z_n)-\delta,$$

where $Z_n\sim N(0,1)$.

A natural question is what should be a bootstrap equivalent of Λ_n. Suppose $Y_1^*, \ldots, Y_n^*$ are a random sample from the data $Y_1, \ldots, Y_n$. We consider the bootstrap equivalent of $n^{1/2}\bar{Y}_n$ to be $n^{1/2}(\bar{Y}_n^* - \bar{Y}_n)$, and not $n^{1/2}\bar{Y}_n^*$. This is in keeping with [4], who put forth the guideline that for good power performance, resampling must be done to reflect the null hypothesis. While model selection is not in general a hypothesis test, some of the same principles are applicable.

Hence, we have $\hat{\mu}^* = W(n^{1/2}(\bar{Y}_n^* - \bar{Y}_n))\bar{Y}_n^*$. When $1 - W(n^{1/2}\bar{Y}_n) \xrightarrow{p} 0$, it can be readily seen that the distribution of $\Lambda_n^* = n^{1/2}(\hat{\mu}^* - \hat{\mu})$, conditional on $Y_1, \ldots, Y_n$, and that of Λ_n converge to the same limit law. □

Remark 4.3. We conjecture that for the model-averaged estimator proposed in this section, a result similar to [16] would hold. In the framework of this paper, the statement corresponding to the main result of [16] would be as follows: Let $F_{n,\alpha,\beta}(t) = P\left[n^{1/2}(\breve{\alpha} - \alpha) \leq t\right]$, and let $\hat{F}_n(t)$ be an estimator of $F_{n,\alpha,\beta}(t)$ satisfying for every $\delta > 0$ $P_{n,\alpha,\beta}\left[\mid \hat{F}_n(t) - F_{n,\alpha,\beta}(t) \mid > \delta\right] \to 0$, as $n \to \infty$. Then $\exists\ \delta_0 > 0$ and $\rho_0 > 0$ such that

$$\sup_{(\tilde{\alpha},\tilde{\beta}) \in B((\alpha,\beta);\rho_0/\sqrt{n})} P_{n,\tilde{\alpha},\tilde{\beta}}\left[\mid \hat{F}_n(t) - F_{n,\tilde{\alpha},\tilde{\beta}}(t) \mid > \delta_0\right] \to 1; \tag{4.9}$$

$$\text{where} \quad B((\alpha,\beta);a) = \{(\tilde{\alpha},\tilde{\beta}) : ||(\tilde{\alpha},\tilde{\beta}) - (\alpha,\beta)|| < a$$

is the open ball of radius a around $(\alpha,\ \beta)$. It can be seen that under standard conditions, if the supremum in (4.9) is taken over $B((\alpha,\ \beta); a_n)$ with $a_n = o(n^{-1/2})$ instead of $B((\alpha,\ \beta); \rho_0/\sqrt{n})$, the limit would be zero instead of 1. Thus the result of [16] may be improved to the case where the supremum is taken only over the set of parameter values that are exact order $n^{-1/2}$ away from the $(\alpha,\ \beta)$ under which the estimator $\hat{F}_n(\cdot)$ is computed. This is easily verified, for example, when $\alpha = 0$, $\sigma = 1$ and $X_{t2} \equiv 1$.

Note that from a bootstrap approximation point of view, (4.9) is not a negative result, but a very positive one. The uses of bootstrap approximation are for constructing interval estimates, testing hypotheses and so on. Equation (4.9) and other related results from [16] imply that a bootstrap approximation $\hat{F}_n(\cdot)$ constructed under the 'null' $(\alpha,\ \beta)$, has sup-norm distance of 1 from the true distributions under parameter values that are exact order $n^{-1/2}$ away from the $(\alpha,\ \beta)$. Thus $\hat{F}_n(\cdot)$ has power 1 in hypothesis testing under contiguous alternatives. This is a further confirmation of the tenet of [4], that resampling procedure ought to reflect the null hypothesis.

Remark 4.4. It is of interest to know that the asymptotic variance of $\breve{\alpha}$ depends on β, and is given by $\mathrm{Var}(n^{1/2}(\breve{\alpha} - \alpha)) - \mathrm{Var}(n^{1/2}(\hat{\alpha}(U) - \alpha)) \to 0$ if $\beta \neq 0$, while $\mathrm{Var}(n^{1/2}(\breve{\alpha}-\alpha)) - \{0.5\mathrm{Var}(n^{1/2}(\hat{\alpha}(U)-\alpha)) + 0.5\mathrm{Var}(n^{1/2}(\hat{\alpha}(R)-\alpha))\} \to 0$ if $\beta = 0$. This is established by checking that both ξ_{1n} and ξ_{2n} tend to $1/2$ as $n \to \infty$ when $\beta = 0$. Thus $\breve{\alpha}$ performs like the correct estimator $\hat{\alpha}(U)$ when model U is valid, and balances between the correct and conservative choices when the restricted model R is true.

5. A Simulation Example

We performed a small simulation experiment to illustrate some of the features of inference under model uncertainty that have been discussed in the previous sections. We took $n = 50$, $x_{i1} \equiv 1$, and generated 50 numbers from the Uniform distribution

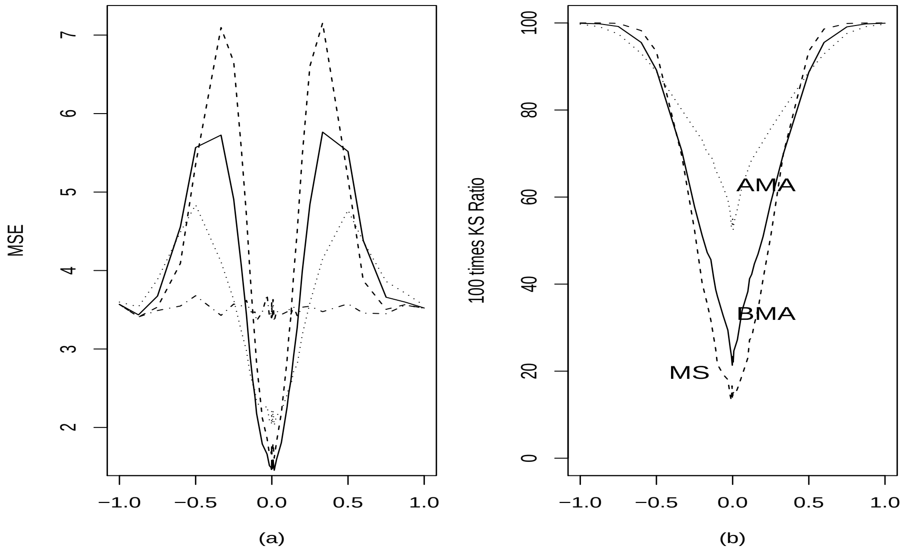


FIG 1. *Panel (a) is the mean squared error of $\hat{\alpha}_{BMA}$ (solid line), $\hat{\alpha}_{MS}$ (broken line), $\hat{\alpha}_{AMA}$ (dotted line), and $\hat{\alpha}(U)$ (dot-and-dash line). Panel (b) is the ratio of Kolmogorov Smirnov distances $KSRatio_j = KS_{jR}/(KS_{jR}+KS_{jU})$, $j = MS, BMA, AMA$, scaled by 100; between distributions of centered and scaled estimators and $\hat{\alpha}(R)$ (for KS_{jR}) and $\hat{\alpha}(U)$ (for KS_{jU}).*

supported between zero and three and fixed these as the x_{i2} values. We fixed $\alpha = 1$, and varied the β values.

For different values of $\beta \in [-1, 1]$, we obtained sampling distribution approximations of (i) the post-model-selected estimator $\hat{\alpha}_{MS}$, (ii) a version of the Bayesian model-averaged estimator $\hat{\alpha}_{BMA}$, and (iii) an adaptive model-averaged estimator $\hat{\alpha}_{AMA}$, by 5000 replications for each value of β. For the Bayesian model-averaged estimator, model R was assigned weight $q_{nR} = \exp(-BIC_R/2)/(\exp(-BIC_R/2) + \exp(-BIC_U/2))$ while model U was assigned weight $1 - q_{nR}$. We define

$$\begin{aligned} BIC_R &= \sum \left[Y_i - \hat{\alpha}_R x_{i1}\right]^2 + \log(n), \\ BIC_U &= \sum \left[Y_i - \hat{\alpha}_U x_{i1} - \hat{\beta}_U x_{i1}\right]^2 + 2\log(n). \end{aligned}$$

For the adaptive model-averaged estimator, we took $a_n = (\log(n))^2$.

The requirement that $a_n^{-1}\log(n) \downarrow 0$ suggests that a_n should be an increasing sequence, growing faster than $\log(n)$. Several choices of a_n were used initially, and it turned out that very slowly increasing sequences like $a_n = (\log(n))^2$ or very quickly increasing sequences like $a_n = n^{0.499}$ performed better than others. This is a reflection on our way of constructing the functions ξ_{1n} and ξ_{2n} using γ_{1n} and γ_{2n}. Alternative choices, like $\gamma_{1n} = a_n|\hat{\beta}(U)|\{\hat{\beta}(U)\}^{-1}$, are a subject for further research.

The first object of our study is the mean squared error of the three estimators of α, namely, $\hat{\alpha}_{MS}$, $\hat{\alpha}_{BMA}$, and $\hat{\alpha}_{AMA}$. Panel (a) in Figure 1 contains the graphs of the mean squared error (MSE) as β varies between $[-1, 1]$. In this and all subsequent figures, the solid line corresponds to $\hat{\alpha}_{BMA}$, the broken line to $\hat{\alpha}_{MS}$, and the dotted line to $\hat{\alpha}_{AMA}$. In this figure, we have also added the graph for the MSE of $\hat{\alpha}(U)$, which is the nearly horizontal dot-and-dash line. First, using model selection or av-

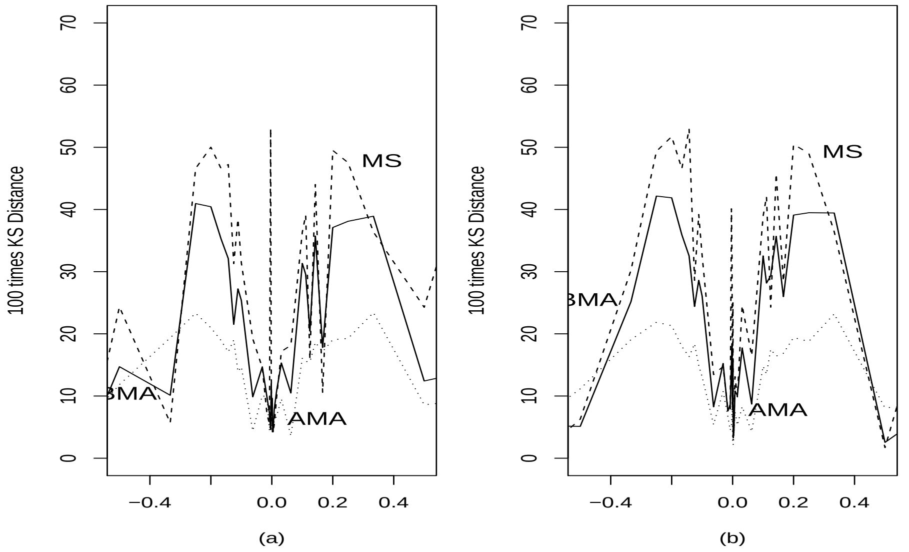


Fig 2. *Panel (a) is the subsampling approximation (subsample size* 20*) for the distribution of centered and scaled $\hat{\alpha}_{BMA}$ (solid line), $\hat{\alpha}_{MS}$ (broken line), $\hat{\alpha}_{AMA}$ (dotted line). Panel (b) is the corresponding bootstrap approximation.*

eraging is clearly better than using $\hat{\alpha}(U)$ only in the region $0 \pm 2/\sqrt{n} \approx (-0.3, 0.3)$, where MS, BMA and AMA all perform better than $\hat{\alpha}(U)$. However, in the neighboring regions $|\beta| \in (0.3, 0.8)$, $\hat{\alpha}(U)$ has smaller MSE than the three estimators. For high values of $|\beta|$, using model selection/averaging or the unrestricted model makes little difference. Thus whether model averaging/selection is useful or not depends considerably on the value of β. Also note that BMA has a lower MSE compared to MS for low values of $|\beta|$ and only marginally higher MSE otherwise, with a much lower maximum MSE value. The graph for AMA tends to stay closest to the graph for $\hat{\alpha}(U)$, and thus does better than BMA or MS in the region $|\beta| \in (0.05, 0.75)$, but is marginally poorer otherwise.

In order to study how the three estimators balance between $\hat{\alpha}(R)$ and $\hat{\alpha}(U)$, we computed the Kolmogorov–Smirnov distances KS_{jR} and KS_{jU}, between the distribution of $n^{1/2}(\hat{\alpha}_j - \alpha)$, and the distributions of $n^{1/2}(\hat{\alpha}_R - \alpha)$ and $n^{1/2}(\hat{\alpha}_U - \alpha)$, where $j = MS, BMA, AMA$ (MS: model selected, BMA=Bayesian model-averaged, AMA=adaptive model-averaged). We then computed the ratios

$$KSRatio_j = 100\frac{KS_{jR}}{KS_{jR} + KS_{jU}}, \quad j = MS, BMA, AMA.$$

Under ideal circumstances, this ratio ought to be zero at $\beta = 0$, and 100 for $\beta \neq 0$. Panel (b) in Figure 1 displays the $KSRatio_j$ values for the three estimators $j = MS, BMA, AMA$. When $\beta = 0$, MS is closest to $\hat{\alpha}_R$, while, as predicted, AMA balances between $\hat{\alpha}_R$ and $\hat{\alpha}_U$. The Bayesian model-averaged estimator BMA lies between MA and AMA, and is quite close to MS. In the region $0 \pm 2/\sqrt{n} \approx (-0.3, 0.3)$ both MS and BMA are much closer to $\hat{\alpha}_R$ than $\hat{\alpha}_U$.

Next, we studied resampling for the three estimators. Subsampling with subsample size $m = 20 = 0.4n$ and the bootstrap was studied. Note that subsampling is consistent for all three estimators, but the bootstrap is consistent only for

AMA. Panels (a) ((b)) of Figure 2, respectively, present the Kolmogorov–Smirnov distance, scaled by 100, between the distributions of $n^{1/2}(\hat{\alpha}_j - \alpha)$ and its subsampling (bootstrap) version, $j = MS, BMA, AMA$. We present the graphs for $|\beta| \leq 0.4 \approx 3/\sqrt{n}$, since there is not much difference between the three graphs for other values of β. It can be seen that the distances between the actual distribution and its subsampling/bootstrap versions are much smaller for AMA, while the resampling approximations for MS and BMA are particularly bad in the regions $\{|\beta| \in (0.1, 0.3)\}$. Also, there is little visual difference between the accuracies of the subsampling and the bootstrap approximations despite their different asymptotic behavior, which confirms some of the observations made in [1], [2] and [18].

6. Discussion and Conclusions

The problems associated with post-model-selection estimation have been discussed by several researchers. In current statistical practice, the process of selecting a model has similarities with hypothesis testing. On the other hand, estimation of parameters, some of which may be known constants in some of the models, is generally entirely separated from model selection. Estimation and testing/selection are two different paradigms of statistical analysis that are hard to integrate. The lack of uniformity across models that parameter estimators generally display, and the issues that arise subsequently, are products of the less than successful attempt to combine the two processes of estimation and selection.

In the Bayesian paradigm, model averaging seems to be a good integration of the two, since the selection step here is also an estimation exercise in spirit. The statement about integrated risks in Proposition 3.1 implies that Bayes' risks of model-averaged estimators are bounded. Thus, while minimaxity seems to be an elusive goal under model uncertainty, a fully Bayesian approach to analyzing risk behavior may be more successful.

In the context of bootstrapping model-averaged estimators, an alternative to $\breve{\alpha}$ is to estimate the bias in $\hat{\alpha}$ in all the models, and define a bias corrected average of these. As the bias of $\hat{\alpha}(R)$ is $\beta||X_1||^{-1} < X_1, X_2 >$, if we estimate this by $\hat{\beta}||X_1||^{-1} < X_1, X_2 >$, we get back $\hat{\alpha}(U)$. Nevertheless, in more complex problems the 'bias corrected model averaged' estimator may be an interesting object to study.

In Theorem 4.1 we established the consistency of the paired bootstrap for a data-adaptive model-averaged estimator. Two other kinds of bootstrap are available in the linear regression context; namely, parametric bootstrap and the residual-based bootstrap. When only one model is in use, the parametric bootstrap generates data from it using estimated values for the unknown parameters, while the residual bootstrap obtains residuals after fitting the model. The equivalents of these are not obvious under model uncertainty.

In Section 4 we remarked that the data adaptive weights p_{nR} and p_{nU} may not share the same properties as the posterior model probabilities π_{nR} and π_{nU} of Section 3. It would be interesting to study when p_{nR} and p_{nU} can be interpreted as posterior probabilities, and also under what conditions the frequentist properties of a Bayesian model-averaged estimator may be elicited using the bootstrap.

Acknowledgments

Professor Chatterjee's research was partially supported by a grant from the University of Minnesota. We thank the Referee and the editors of this monograph for some excellent comments and suggestions. Also, we would like to thank Professor Yuhong Yang, who carefully read an earlier draft of this paper and made several comments; which, along with several illuminating discussions, greatly enhanced our understanding on the scope and issues relating to model selection/averaging.

References

[1] Andrews, D. W. K. and Guggenberger, P. (2005) Hybrid and size-corrected subsample methods, *Cowles Foundation discussion paper # 1605.*

[2] Andrews, D. W. K. and Guggenberger, P. (2005) The limit of finite sample size and a problem with subsampling, *Cowles Foundation discussion paper # 1606.*

[3] Chatterjee, S. and Bose, A. (2000) Variance estimation in high dimensional regression models, *Statistica Sinica*, **10**, 497-515.

[4] Hall P., and Wilson S.R. (1991) Two guidelines for bootstrap hypothesis testing, *Biometrics*, **47**, 757-762.

[5] Hjort, N. L. and Claeskens, G. (2003) Frequentist model average estimators, *J. Amer. Statist. Assoc.*, **98**, 879-899.

[6] Leeb, H. (2006) The distribution of a linear predictor after model selection: unconditional finite sample distributions and asymptotic approximations, *IMS Lecture Notes, Monograph Series*, **49**, 291-311.

[7] Leeb, H. and Pötscher, B. M. (2003) The finite sample distribution of post-model-selection estimators and uniform versus non-uniform approximations, *Econometric Theory*, **19**, 100-142.

[8] Leeb, H. and Pötscher, B. M. (2005) Model selection and inference: facts and fiction, *Econometric Theory*, **21**, 21-59.

[9] Leeb, H. and Pötscher, B. M. (2006) Performance limits for the estimators of the risk or distribution of shrinkage type estimators, and some general lower risk bound results, *Econometric Theory*, **22**, 69-97.

[10] Leeb, H. and Pötscher, B. M. (2006) Can one estimate the conditional distribution of post-model-selection estimators?, *Ann. Statist.*, **34**, 2554-2591.

[11] Leung, G. and Barron, A. R. (2006) Information theory and mixing least squares regressions, *IEEE Transactions on Information Theory*, **52**, 3396-3410.

[12] Mammen, E. (1992a) Bootstrap, wild bootstrap and asymptotic normality, *Probability theory and related fields*, **93**, 439-455.

[13] Mammen, E. (1992b) *When does bootstrap work: asymptotic results and simulations*, Springer Lecture Notes in Statistics.

[14] Politis, D. N., Romano, J. P., and Wolf, M. (1999) *Subsampling*, Springer, New York.

[15] Pötscher, B. M. (1991) Effects of model selection on inference, *Econometric Theory*, **7**, 163-185.

[16] Pötscher, B. M. (2006) The distribution of model averaging estimators and an impossibility result regarding its estimation, *MPRA paper # 73.*

[17] Raftery, A. E. and Zheng, Y. (2003) Comment on *Frequentist model average estimators, by [5]*, *J. Amer. Statist. Assoc.*, **98**, 931-938.

[18] Samworth, R. (2003) A note on methods of restoring consistency to the bootstrap, *Biometrika*, **90**, 985-990.

[19] Sethuraman, J. (2004) Are super-efficient estimators super-powerful? *Comm. Statist.: Theory and Methods*, **33**, 2003-2013.

[20] Shen, X. and Dougherty, D. P. (2003) Discussion of *Frequentist model average estimators*, *J. Amer. Statist. Assoc.*, **98**, 917-919.

[21] Yang, Y. (2003) Regression with multiple candidate models: selecting or mixing? *Statistica Sinica*, **13**, 783-809.

[22] Yang, Y. (2004) Aggregating regression procedures to improve performance, *Bernoulli*, **10**, 25-47.

[23] Yang, Y. (2005) Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation, *Biometrika*, **92**, 937-950.

[24] Yang, Y. (2007) Prediction/Estimation with Simple Linear Model: Is It Really that Simple? *Econometric Theory*, **23**, 1-36.

[25] Yuan, Z. and Yang, Y. (2005) Combining linear regression models: when and how? *J. Amer. Statist. Assoc.*, **100**, 1202-1214.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 171

Remarks on Consistency of Posterior Distributions

Taeryon Choi[1] **and R. V. Ramamoorthi**[2]

University of Maryland, Baltimore County and Michigan State University

Abstract: In recent years, the literature in the area of Bayesian asymptotics has been rapidly growing. It is increasingly important to understand the concept of posterior consistency and validate specific Bayesian methods, in terms of consistency of posterior distributions. In this paper, we build up some conceptual issues in consistency of posterior distributions, and discuss panoramic views of them by comparing various approaches to posterior consistency that have been investigated in the literature. In addition, we provide interesting results on posterior consistency that deal with non-exponential consistency, improper priors and non i.i.d. (independent but not identically distributed) observations. We describe a few examples for illustrative purposes.

Contents

1. Introduction

Let θ be an unknown parameter and $X_1, X_2, \ldots, X_n$ be n random variables whose joint distribution is $P_\theta^{(n)}$. In order to draw inferences on θ, a Bayesian posits a prior distribution Π for θ and updates this prior to the posterior distribution given $X_1, X_2, \ldots, X_n$, which we denote by $\Pi(\cdot|X_1, X_2, \ldots, X_n)$. This paper focuses on some issues related to an asymptotic aspect of this posterior distribution, namely, consistency.

The sequence of posterior distributions $\{\Pi(\cdot|X_1, X_2, \ldots, X_n)\}$ is said to be consistent at θ_0, if the posterior converges, in a suitable sense, to the degenerate measure at θ_0.

[1] Department of Mathematics and Statistics, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD, 21250. e-mail: `tchoi@math.umbc.edu`

[2] Department of Statistics and Probability, Michigan State University, A413 Wells Hall, East Lansing, MI 48824-1027. e-mail: `ramamoor@stt.msu.edu`

Primary 62G20; secondary 62C10, 62F15

Keywords and phrases: Affinity, Doob's theorem, Exponential consistency, Posterior distribution, Strongly separated

Posterior consistency is a kind of frequentist validation of the updating method. If an oracle were to know the true value of the parameter, posterior consistency ensures that with enough observations one would get close to this true value. Posterior consistency also assures that as more and more observations accumulate, the observations have to dominate the role of the prior in inference. There are other interpretations related to merging of opinions and other concepts. We refer the reader to Diaconis and Freedman [10].

In order to set the perspective for this paper we begin with a short summary of earlier results in posterior consistency. Details and additional references can be found in Ghosh and Ramamoorthi [19]. The first posterior consistency result goes back to Laplace. In more recent times posterior consistency and asymptotic normality of the posterior were established for regular finite dimensional models. In a seminal paper, Freedman [12] gave a nonparameteric example, with integer-valued observations, where the posterior is inconsistent. In [10], Diaconis and Freedman showed that in the nonparametric case inconsistency can occur, even in location models with an Euclidean parameter. They suggested that instead of searching for priors that would be consistent at all unknown values of the parameter it would be fruitful to study natural priors and identify points of consistency.

On the positive side, Freedman [12, 13] and soon after Schwartz [25] provided conditions under which the posterior probability of a set A will go to 0. These conditions involved two parts, one on prior positivity of Kullback–Leibler neighborhoods and the other on existence of certain test functions. Under the assumption of prior positivity of Kullback–Leibler neighborhoods, Barron gave necessary and sufficient conditions for the posterior probability of A to go to 0. These results were then specialized to weak and L_1 neighborhoods by Barron *et al.* [3], Ghosal *et al.* [15] and Walker [32].

One aspect of these results was that they all established exponential consistency. In this paper we first give a quick review of these results from a slightly different perspective with a focus on the role of exponential consistency. We then give an example where there is consistency but not exponential consistency. The example also shows that the exponential aspect is not driven by the Kullback–Leibler condition.

Another early result in consistency is due to Doob [11], who showed that posterior consistency occurs for all θ in a set of prior measure one. In this paper we consider a study of the non i.i.d. case based on Doob's result, specifically, the simple linear regression model. The martingale techniques are not applied here and we discuss the connection of posterior consistency with orthogonality of product measures.

Consistency is just the beginning of Bayesian asymptotics. Issues such as rates of convergence and asymptotic normality have received quite a lot of attention. Yet it appears that even at the level of consistency there are still issues that need to be clarified. In this paper, we review some conceptual issues in consistency of posterior distributions, and discuss different approaches to posterior consistency that have been investigated in the literature; we view this as a followup of Ghosal *et al.* [14]. We have attempted to elucidate those sufficient conditions to establish posterior consistency and tie up some loose ends on diverse conceptual issues in consistency of posterior distributions. The paper also contains some new results along with a brief commentary to the subject. In general, detailed proofs are omitted and given only when they are different from standard published materials or when the result is unpublished.

Section 2 contains a summary of some background material, some of the notations and assumptions used in the paper. Section 3 largely describes known results although some of the proofs are reorganized. The criteria based on the uniform strong law of large numbers is new and so far we are not aware of any significant application. The result might still be of interest because of its similarities to results on the consistency of nonparametric maximum likelihood estimates (NPMLEs) and also because it affords a natural extension to non i.i.d. cases. Section 4 specializes the results in Section 2 to the context of consistency. After a brief discussion of Schwartz's result, we discuss the known conditions for L_1 consistency and the relationship between these. Section 5 extends the Schwartz theorem to improper priors and formal posteriors. The result is new even in the parametric case. We have not pursued conditions for stronger consistency because improper priors usually arise in finite dimensional situations where weak and strong consistency coincide. Section 6 contains an example. All the general consistency results in the literature actually establish exponential consistency. In Section 6 we give an example where consistency obtains but not exponential consistency. The example surprised us as we had believed that, at least in the i.i.d. case, consistency would always be at an exponential rate.

In the last section we study the extension of consistency results to a non i.i.d. case. We give an example to show that the analogue of Doob's theorem will not always hold, and we prove a Doob theorem for the linear regression model with nonparametric errors. We also briefly discuss an extension of the theorem of Walker.

2. Preliminaries

In the setup that we consider, Θ is the parameter space ; $\{f_\theta : \theta \in \Theta\}$ is a family of densities with respect to a σ-finite measure μ on a measurable space $\mathcal{X}$. We will use P_θ to denote the probability distribution generated by f_θ. Throughout the paper we assume that Θ and $\mathcal{X}$ are complete separable metric spaces and we also assume that $\theta \mapsto f_\theta$ is 1-1 and $(\theta, x) \mapsto f_\theta(x)$ is measurable.

The affinity, $\mathrm{Aff}(f,g)$, between any two densities is defined as $\mathrm{Aff}(f,g) = \int \sqrt{fg}d\mu$. Let Π be a prior distribution, i.e., a probability measure on Θ. Given $\theta, X_1, X_2, \ldots, X_n$ are assumed to be i.i.d P_θ. $f_\theta^{(n)}(x_1, x_2, \ldots, x_n)$ will stand for the joint density $\prod_{i=1}^n f_\theta(x_i)$.

The Kullback–Leibler (KL) divergence is denoted by $K(\theta_0, \theta) = E_{\theta_0} \log(f_\theta / f_{\theta_0})$. A KL neighborhood $K_\epsilon(\theta_0)$ of θ_0 is denoted by $\{\theta : K(\theta_0, \theta) < \epsilon\}$.

Definition 2.1. A point θ_0 is said to be in the KL support of Π if for all $\epsilon > 0, \Pi\left(K_\epsilon(\theta_0)\right) > 0$.

The posterior distribution $\Pi(A|X_1, X_2, \ldots, X_n)$, the version that we consider, is given by the following. For any measurable subset A of Θ,

$$\Pi(A|X_1, X_2, \ldots, X_n) = \frac{J_A(X_1, X_2, \ldots, X_n)}{J(X_1, X_2, \ldots, X_n)} \tag{2.1}$$

where

$$J_A(X_1, X_2, \ldots, X_n) = \int_A \frac{f_\theta^{(n)}}{f_{\theta_0}^{(n)}}(X_1, X_2, \ldots, X_n)\Pi(d\theta)$$

and

$$J(X_1, X_2, \ldots, X_n) = \int_\Theta \frac{f_\theta^{(n)}}{f_{\theta_0}^{(n)}}(X_1, X_2, \ldots, X_n)\Pi(d\theta).$$

3. Exponential Decrease to 0

We begin with a review of results that provide conditions under which, for a measurable subset A of Θ, $\Pi(A|X_1, X_2, \dots, X_n)$ goes to 0 exponentially with $P_{\theta_0}^{\infty}$ probability 1.

Definition 3.1. Let $\theta_0 \in \Theta$ and let $P_{\theta_0}^{\infty}$ stand for the joint distribution of $\{X_i\}_{i=1}^{\infty}$ when θ_0 is the true value of θ. Then $\Pi(A|X_1, X_2, \dots, X_n)$ is said to go to 0 exponentially with $P_{\theta_0}^{\infty}$ probability 1, if there exists a $\beta > 0$ such that

$$P_{\theta_0}^{\infty}\left(\{\Pi(A|X_1, X_2, \dots, X_n) > e^{-n\beta} \ i.o.\ \}\right) = 0$$

where *i.o.* stands for 'infinitely often'.

Proposition 3.2 goes back to [12] and [25]. For a proof see [19, Lemma 4.4.1].

Proposition 3.2. If θ_0 is in the KL support of Π then for all $\beta > 0$,

$$\lim_{n\to\infty} e^{n\beta} J(X_1, X_2, \dots, X_n) = \infty \text{ a.s. } P_{\theta_0}^{\infty}.$$

Proposition 3.2 shows that the Kullback–Leibler support condition takes care of the denominator in (2.1). The exponential convergence to 0 would follow if it can be established that there exists $\beta_0 > 0$ such that $e^{n\beta_0} J_A(X_1, X_2, \dots, X_n) \to 0$ *a.s.* $P_{\theta_0}^{\infty}$. We explore sufficient conditions to achieve this.

Definition 3.3. For a probability measure ν on θ, let $q_\nu^{(n)}$ be the marginal density of $X_1, \dots, X_n$,

$$q_\nu^{(n)}(x_1, x_2, \dots, x_n) = \int_\Theta f_\theta^{(n)}(x_1, x_2, \dots, x_n)\nu(d\theta).$$

Definition 3.4. Let $A \subset \Theta$ and $\delta > 0$. The set A and θ_0 are said to be *strongly δ separated* if for any probability ν on A,

$$\mathrm{Aff}(f_{\theta_0}, q_\nu^{(1)}) < \delta.$$

The relationship $H^2(f, g) = 1 - 2\mathrm{Aff}(f, g)$ between the Hellinger distance $H(f, g)$ and the Affinity $\mathrm{Aff}(f, g)$ shows that $\mathrm{Aff}(f_{\theta_0}, q_\nu^{(1)}) < \delta$ is equivalent to $H^2(f_{\theta_0}, q_\nu^{(1)}) > 1 - \delta$. Say that A and θ_0 are strongly separated if they are strongly δ separated for some $\delta > 0$.

Example 3.5. Suppose that the L_1 distance between f_{θ^*} and f_{θ_0} is larger than δ^* for some $\delta^* > 0$, $\|f_{\theta^*} - f_{\theta_0}\| > \delta^*$. Let

$$A = \left\{\theta : \|f_{\theta^*} - f_\theta\| < \frac{\delta^*}{2}\right\}.$$

It is easy to see that A is strongly separated from θ_0 for every ν on A.

We begin by isolating a useful consequence of strong separation. The underlying idea is in [32]. Note that the argument is essentially analytic and does not use Hoeffding's inequality as in [19]. Lemma 3.6, we believe, can be extended to non-i.i.d and even to non-independent cases. We do not pursue this here but will briefly return to it in Section 7.

Lemma 3.6. *If θ_0 and A are strongly δ separated then for all probability ν on A, for all n,*

$$\text{Aff}(f_{\theta_0}^{(n)}, q_\nu^{(n)}) < e^{-n\beta_0}, \text{ where } \beta_0 = -\log\delta. \tag{3.1}$$

Proof. The proof is straightforward by induction on n, combined with the definition of strong separation. □

Remark 3.1. The conclusion of Lemma 3.6 holds with $\beta_0 = -\log\delta/k$ if for all ν, for some k, $\text{Aff}(f_{\theta_0}^{(k)}, q_\nu^{(k)}) < \delta$, i.e., A and θ_0 are strongly separated for the parametrization $\theta \mapsto f^{(k)}$

The next result is the celebrated result of Schwartz [25] stated in terms of strong separation. A result of LeCam [21] shows that it is equivalent to the formulation of Schwartz involving an unbiased test for testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \in A$. LeCam's theorem is proved using the Hahn–Banach theorem so is essentially an existence result. Hence the point of view of strong separation could be an easier condition to verify in some situations.

Theorem 3.7 (Schwartz)**.** *If*

(1) θ_0 is in the KL support of Π
(2) for some k, A and θ_0 are strongly separated for the parametrization $\theta \mapsto f^{(k)}$

Then $\Pi(A|X_1, X_2, \ldots, X_n)$ goes to 0 exponentially a.e. $P_{\theta_0}^\infty$.

Proof. Let Π^* be the probability measure obtained by restricting Π to A and normalizing it. Then

$$\begin{aligned} P_{\theta_0}(\sqrt{J_A} > e^{-n\gamma}) &\leq e^{n\gamma} E_{\theta_0}(\sqrt{J_A}) \\ &= e^{n\gamma}\sqrt{\Pi(A)}\text{Aff}(f_{\theta_0}^{(n)}, q_{\Pi^*}^{(n)}) \\ &\leq \sqrt{\Pi(A)} e^{n\gamma} e^{-n\beta_0} \end{aligned}$$

Taking $\gamma = \beta_0/4$, it follows easily that

$$P_{\theta_0}(\sqrt{J_A} > e^{-n\gamma} \; i.o.) = 0$$

The proof can be completed easily using Proposition 3.2. For details see [19]. □

Proposition 3.2 and Lemma 3.6 easily give the following theorem of Walker [32].

Theorem 3.8. *If*

(1) θ_0 is in the KL support of Π
(2) If $A = \cup_{i\geq 1} A_i$ such that

(a) For some $\delta > 0$ all the A_i's are strongly δ separated from θ_0 and

(b) $\sum_{i\geq 1} \sqrt{\Pi(A_i)} < \infty$

Then $\Pi(A|X_1, X_2, \ldots, X_n)$ goes to 0 exponentially a.e. $P_{\theta_0}^\infty$

Proof. It follows by noting

$$\begin{aligned}
& P_{\theta_0}(\sqrt{J_A} > e^{-n\gamma}) \leq e^{n\gamma} E_{\theta_0}(\sqrt{J_A}) \\
\leq\ & e^{n\gamma} E_{\theta_0}\left(\sqrt{\sum_i J_{A_i}}\right) \leq e^{n\gamma} \sum_i E_{\theta_0}\left(\sqrt{J_{A_i}}\right) \\
=\ & e^{n\gamma} \sum_i \sqrt{\Pi(A_i)} \mathrm{Aff}(f_{\theta_0}^{(n)}, q_{\Pi_i^*}^{(n)}) \qquad (3.2) \\
\leq\ & e^{n\gamma} e^{-n\beta_0} \sum_i \sqrt{\Pi(A_i)},
\end{aligned}$$

where Π^* in (3.2) is the normalized restriction of Π^* to A_i. □

The next theorem gives another set of sufficient conditions, in terms of the uniform Strong Law of Large Numbers (SLLN), for the posterior probability of a set to go to 0 exponentially. The conditions are stronger than those of Schwartz [25]. They are similar in spirit to the conditions used in the study of Hellinger consistency of NPMLEs (see [30]) and suggest a parallel between consistency of NPMLEs and posterior consistency.

Theorem 3.9. *Let $A \subset \Theta$. If*

(1) θ_0 is in the KL support of Π

(2) $\mathrm{Aff}(f_{\theta_0}, f_\theta) < \delta$ for all $\theta \in A$

(3) $\sup_{\theta \in A} \left| \int \sqrt{\frac{f_\theta}{f_{\theta_0}}}(x) dP_n - \mathrm{Aff}(f_{\theta_0}, f_\theta) \right| \to 0$ a.s $P_{\theta_0}^\infty$, where P_n is the empirical distribution obtained from $X_1, X_2, \ldots, X_n$

Then $\Pi(A|X_1, X_2, \ldots, X_n)$ goes to 0 exponentially a.e. $P_{\theta_0}^\infty$

Proof. Note that $\int g(x) dP_n = (1/n)\sum_{i=1}^n g(X_i)$ for arbitrary function $g(x)$. Thus,

$$\begin{aligned}
J_A &= \int_A \prod_1^n \frac{f_\theta}{f_{\theta_0}}(X_i) \Pi(d\theta) \\
&= \int_A \exp\left\{2n \int \log \sqrt{\frac{f_\theta}{f_{\theta_0}}}(x) dP_n\right\} \Pi(d\theta) \\
& \quad \text{since } \log x \leq x - 1 \\
&\leq \int_A \exp\left\{2n \int \left(\sqrt{\frac{f_\theta}{f_{\theta_0}}}(x) - 1\right) dP_n\right\} \Pi(d\theta).
\end{aligned}$$

Take $\delta^* = 1 - \delta$. By assumptions 2 and 3, for all large n,

$$\begin{aligned}
\sup_{\theta \in A} \sqrt{\frac{f_\theta}{f_{\theta_0}}} dP_n &\leq \sup_{\theta \in A} \left\{ \left| \sqrt{\frac{f_\theta}{f_{\theta_0}}} dP_n - \mathrm{Aff}(f_{\theta_0}, f_\theta) \right| + \mathrm{Aff}(f_{\theta_0}, f_\theta) \right\} \\
&\leq \frac{\delta^*}{2} + 1 - \delta^* = 1 - \delta^*/2,
\end{aligned}$$

which in turn implies that $J_A < \Pi(A) \exp(-n\delta^*/2)$.

□

Proposition 3.10. Conditions (2) and (3) of Theorem 3.9 imply that there exists a uniformly consistent test for $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \in A$.

Proof. Choose δ_0 such that $\delta + \delta_0 = \beta_0 < 1$.

Let

$$C = \left\{ (x_1, x_2, \ldots, x_n) : \sup_{\theta \in A} \left| \int \sqrt{\frac{f_\theta}{f_{\theta_0}}}(x) dP_n - \text{Aff}(f_{\theta_0}, f_\theta) \right| < \delta_0 \right\}$$

By assumption (3) of Theorem 3.9, for any $\epsilon > 0$ and sufficiently large n, $P_{\theta_0}(C) > 1 - \epsilon$. For each $(x_1, x_2, \ldots, x_n)$ in C, for all $\theta \in A$,

$$\frac{1}{n} \sum \sqrt{\frac{f_\theta}{f_{\theta_0}}(x_i)} \leq \sup_{\theta \in A} \text{Aff}(f_{\theta_0}, f_\theta) + \delta_0$$

so that

$$\sum \left(\sqrt{\frac{f_\theta}{f_{\theta_0}}(x_i)} - 1 \right) < n(\delta + \delta_0 - 1) = -n\beta_0.$$

Therefore, for $\theta \in A$,

$$P_\theta(C) = \int_C \frac{f_\theta^{(n)}}{f_{\theta_0}^{(n)}}(x_i) f_{\theta_0}^{(n)}(x_i) \prod \mu(dx_i) < P_{\theta_0}(C) e^{-2n\beta_0}.$$

□

Remark 3.2. Salinetti [24] has used the notion of hypo convergence to study consistency of posterior and consistency of maximum likelihood estimates. A somewhat related result is due to Ghosal and van der Vaart who show that her condition is related to Schwartz's testing condition in the discussion of [24].

While the results discussed so far deal with sufficient conditions for $\Pi(A|X_1, \ldots, X_n)$ to go to 0 exponentially, the next basic result due to Barron [2] gives conditions that are both necessary and sufficient.

Theorem 3.11 (Barron). *$A \subset \Theta$. Assume that θ_0 is in the KL support of Π. Then the following are equivalent.*

(i) There exists a β_0 such that

$$P_{\theta_0}\{\Pi(A|X_1, X_2, \ldots, X_n) > e^{-n\beta_0} \text{ i.o.}\} = 0.$$

(ii) There exist subsets V_n, W_n of Θ, positive numbers $c_1, c_2, \beta_1, \beta_2$, and a sequence of tests $\{\phi_n\}$, ϕ_n based on n observations, such that

(a) $A \subset V_n \cup W_n$,

(b) $\Pi(W_n) \leq C_1 e^{-n\beta_1}$, and

(c) $P_{\theta_0}\{\phi_n > 0 \text{ i.o.}\} = 0$ and $\inf_{f \in V_n} E_f \phi_n \geq 1 - c_2 e^{-n\beta_2}$.

4. Consistency

As before, Π stands for a prior and $\{\Pi(\cdot|X_1, X_2, \ldots, X_n)\}$ denotes a sequence of posterior distributions. The sequence of posteriors is said to be consistent at θ_0 if $\{\Pi(U|X_1, X_2, \ldots, X_n)\} \to 1$ a.s.$P_{\theta_0}^\infty$ for all neighborhoods U of θ_0.

Typically the parametrization $\theta \mapsto f_\theta$ turns out to be continuous when the space of densities is endowed with weak convergence or with the L_1 or the Hellinger metric. Consequently the neighborhoods of interest are those that arise from weak or L_1 topology.

In view of the last section, what is required then is to verify that the conditions developed in the last section apply to neighborhoods.

Let $g(x)$ be a bounded measurable function and define

$$A_g = \left\{\theta : \int g(x) f_\theta(x)\mu(dx) - \int g(x) f_{\theta_0}(x)\mu(dx) \geq \epsilon\right\}. \tag{4.1}$$

Clearly A is strongly ϵ separated from θ_0 and hence if θ_0 is in the KL support of Π then by Theorem 3.7 the posterior probability of A goes to 0 exponentially. If U is a weak neighborhood then U^c is a finite union of sets of the type displayed in (4.1). This establishes exponential consistency for weak neighborhoods.

Consider the L_1 neighborhood

$$U = \{\theta : \|f_\theta - f_{\theta_0}\| < \epsilon\}.$$

In this case, in general, U^c cannot be expressed as a finite union of sets strongly separated from θ_0. Unlike the case of weak neighborhoods, in this case we need conditions beyond requiring that θ_0 is in the KL support of Π.

Theorem 3.8 can be easily adapted in this context.

Theorem 4.1. *Assume*

(1) θ_0 is in the KL support of Π
(2) For all $\delta > 0$, there exist sets $A_1, A_2, \ldots$ such that the diameter of A_i, $\operatorname{diam}(A_i) < \delta$, $\bigcup A_i = \Theta$ and $\sum \sqrt{\Pi(A_i)} < \infty$.

Then for any L_1 neighborhood U of θ_0, the posterior probability of U^c goes to 0 exponentially a.e. $P_{\theta_0}^\infty$.

The theorem follows from observing that if U is an ϵ neighborhood, then taking $\{A_i\}_{i=1}^\infty$ for $\delta = \epsilon/3$ it is easily seen that the A_i's that have non-empty intersection with U^c cover U^c. These A_i's satisfy the assumptions of Theorem 3.8.

Definition 4.2 (Bracketing entropy). Let $\Gamma \subset \Theta$. For a $\delta > 0$ define the bracketing entropy $\mathcal{H}(\Gamma, \delta)$ to be the logarithm of the minimum integer k such that, there exist non negative functions $f_1^U, f_2^U, \cdots, f_k^U$ satisfying

(1) $\int f_i^U(x)\mu(dx) < 1 + \delta$,
(2) for each θ there exists i such that $f_\theta \leq f_i^U$.

Definition 4.3 (Metric entropy). Let $\Gamma \subset \Theta$. For $\delta > 0$ the Metric entropy $J(\Gamma, \delta)$ is defined to be the logarithm of minimum of all integers k such that there exist densities $f_1^*, f_2^*, \cdots, f_k^*$ such that for each θ there exists i such that $\|f_\theta - f_i^*\| < \delta$.

If θ_0 is in the KL support of Π then each of the three conditions listed below ensures that the posterior is exponentially L_1 consistent. The first condition (W) is from Walker's theorem, Theorem 3.8, the next (BSW) is due to [3] and the third (GGR) appears in [15]. A formal statement and proof can be found in [3] and [15].

(W) For each $\delta > 0$, there exist sets $A_1, A_2, \ldots$ such that $\cup A_i = \Theta$, L_1-diameter of $\{f_\theta : \theta \in A_i\} < \delta$ and $\sum_i \sqrt{\Pi(A_i)} < \infty$.

(BSW) For each $\epsilon > 0$, there exist $\Theta_n \subset \Theta$, and C, c_1, c_2, δ all positive such that

(a) $\Pi(\Theta_n^c) < e^{-nc_2}$

(b) $\mathcal{H}(\Theta_n, \delta) \leq nc$ for $c < ([\epsilon - \sqrt{\delta}]^2 - \delta)/2, \delta < \epsilon^2/4$.

(GGR) If for each $\epsilon > 0$, there is a $0 < \delta < \epsilon, c_1, c_2, \beta < \epsilon^2/2$ and Θ_n such that

(a) $\Pi(\Theta_n^c) < c_1 e^{-n\beta}$

(b) $J(\Theta_n, \delta) \leq n\beta$.

The next theorem shows that both (W) and (BSW) imply (GGR). A proof that (W) $\Rightarrow$ (GGR) was also communicated to us by Ghosal, S. [personal communication].

Theorem 4.4. *(W)⇒(GGR) and (BSW) ⇒ (GGR)*

Proof. (W)⇒(GGR)

Assume without loss of generality that $\Pi(A_i) = \Pi_i$ is decreasing in i and let $\sum \sqrt{\Pi_i} = c < \infty$. Set

$$\Theta_n = \bigcup_1^{k_n} A_i.$$

Since the L_1-diameter of $\{f_\theta : \theta \in A_i\} < \delta$, it is easy to see that $J(\Theta_n, \delta) < \log k_n$. Thus, by taking $k_n = \exp(n\beta)$ one then obtains sieves with the properties required by (GGR).

Next, we shall argue that $\Pi(\Theta_n^c) = \Pi\left(\bigcup_{i>k_n} A_i\right) \leq 2c^2/k_n$. Note that, for any j, $j\sqrt{\Pi_j} \leq \sum_{i=1}^{j} \sqrt{\Pi_i} \leq c$ so that $j \leq c/\sqrt{\Pi_j}$. Therefore,

$$\Pi\left(\bigcup_{j>k_n} A_i\right) \leq \sum_{j>k_n} \Pi_j \leq c^2 \sum_{j>k_n} \frac{1}{j^2} \leq \frac{2c^2}{k_n}$$

(BSW)⇒(GGR)

Let $f_1, f_2, \ldots, f_k$ be functions such that $\int f_i = 1 + c_i < (1 + \delta)$ and such that for any $\theta \in \Gamma, \exists\ i$ such that $f_\theta \leq f_i$. Let $f_i^* = f_i/(1 + c_i)$. Then

$$\begin{aligned} \|f_i^* - f_\theta\| &\leq \frac{1}{1 + c_i}\|f_i - (1 + c_i)f_\theta\| \\ &\leq \|f_i - f_\theta\| + c_i \\ &\leq 2\delta \end{aligned}$$

Hence $f_1^*, f_2^*, \ldots, f_k^*$ forms a 2δ net for Γ and $J(\Gamma, 2\delta) \leq \mathcal{H}(\Gamma, \delta)$. □

5. Improper Priors and Formal Posteriors

Suppose that Π is an improper prior on Θ, that is, a σ-finite measure with $\Pi(\Theta) = \infty$. A formal posterior density given $X_1 = x_1, X_2 = x_2, \ldots X_n = x_n$ is defined as in Equation (2.1). This is of course well defined only if

$$J(x_1, x_2, \ldots, x_n) = \int_\Theta \frac{f_\theta^{(n)}}{f_{\theta_0}^{(n)}}(x_1, x_2, \ldots, x_n)\Pi(d\theta) < \infty.$$

This situation occurs widely in the context of noninformative priors (see for example, Ghosh and Ramamoorthi [18] and Kass and Wasserman [20]).

The next theorem shows that if P_0 is in the KL support of Π then the posterior is weakly consistent. Improper priors largely arise in the context of finite dimensional regular models and in these situations weak consistency and strong consistency coincide. Hence, we do not develop conditions akin to (W), (BSW) or (GGR) for improper priors. First, Lemma 5.1 states a result of the KL support of $\Pi(\cdot|x)$.

Lemma 5.1. *Let P_0 is in the KL support of Π. Denote by $A = \{x : J(x) = \int f_\theta(x)\Pi(d\theta) < \infty\}$. Then, for P_0 almost all x in A, θ_0 is in the KL support of $\Pi(\cdot|x)$.*

Proof. Fix $\epsilon > 0$. Consider $E = \{x \in A : \Pi(K_\epsilon|x) = 0\}$. We shall show that $P_{\theta_0}(E) = 0$. Note that

$$\Pi(K_\epsilon|x) = \frac{\int_\Theta I_{K_\epsilon}(\theta) f_\theta(x)\Pi(d\theta)}{\int f_\theta(x)\Pi(d\theta)}.$$

Denoting by Π^* the measure $\Pi(\cdot \cap K_\epsilon)/\Pi(K_\epsilon)$, since, for $x \in E$, $\Pi(K_\epsilon|x) = 0$, we have that

$$\Pi^*\{\theta : f_\theta(x) = 0\} = 1.$$

Consequently $\int\limits_E \int_{K_\epsilon} f_\theta(y)\Pi^*(d\theta)(E)d\mu(y) = 0$. Interchanging the integrals,

$\int_{K_\epsilon}[\int_E f_\theta(y)d\mu(y)]\Pi^*(d\theta) = 0$ and hence there exists some θ' such that $\int_E f_{\theta'}(y)d\mu(y) = 0$ so that $P_{\theta'}(E) = 0$. For every θ in K_ϵ P_θ dominates P_{θ_0}, so $P_{\theta_0}(E) = 0$. Letting ϵ run through rationals, the lemma is established. □

Theorem 5.2. *Let Π be an improper prior on Θ. $\{f_\theta : \theta \in \Theta\}$ is a family of densities. Assume that the formal posterior is defined with P_0^∞ probability one. Formally, if*

$$A_n = \{x_1, x_2, \dots, x_n : J(x_1, x_2, \dots, x_n) < \infty\} \textit{ then } P_0^\infty(\cup A_n) = 1$$

If θ_0 is in the KL support of Π then the formal posterior is weakly consistent at θ_0.

Proof. By Lemma 5.1, for each n, except for those in a set of P_{θ_0} measure 0, for all $(x_1, x_2, \dots, x_n) \in A_n$, θ_0 is in the KL support of $\Pi(\cdot|(x_1, x_2, \dots, x_n))$.

Since on A_n, $\Pi(\cdot|(x_1, x_2, \dots, x_{n+1})) = \Pi_{(x_1,x_2,\dots,x_n)}(\cdot|x_{n+1})$, the result follows. □

6. Example

All the results discussed so far are related to exponential consistency. The next example shows that, even in the context of i.i.d. observations, the posterior can be consistent at a non-exponential rate.

Consider an example where we have a prior Π, f_0 is in the KL support of Π and the posterior is not L_1 consistent, i.e., there is a set A which is a complement of a neighborhood of f_0 and whose posterior does not go to 0. Such an example appears, for instance, in Barron *et al.* [3].

Consider the prior to $\Pi^* = .5\delta_{f_0} + .5\Pi$. Then by Doob's theorem the posterior of A goes to 0. It cannot go exponentially, for if it does, by Barron's theorem (e.g. [2] and [19, Theorem 4.4.3]), there would be sieves V_n and sets U_n of exponentially small Π^* probability that cover A. These properties also carry over to Π and now the first part of Barron's result would imply that the original prior Π is itself consistent, in fact, exponentially consistent.

7. Independent but Non-Identically Distributed Models

7.1. Extension to Posterior Consistency

Here we look at the setup where, as before, Θ is a parameter space and Π is a prior on Θ. Given θ, we assume that $X_1, X_2, \ldots$ are independent with X_i distributed as $f_{i,\theta}$.

All the results discussed so far can be easily adapted, but not necessarily easily applied in the non-identically distributed case. As in Section 1, the posterior can be written as the ratio of two integrals. A stronger form of KL support (for instance, see Choudhuri *et al.* [9] and Amewou-Atisso *et al.* [1]) takes care of the denominator. It is not clear if there is a simple version of the (GGR) type of sufficient condition. Instead, those results for independent but non-identically distributed models as in Amewou-Atisso *et al.* [9], Choudhuri *et al.* [1], Ghosal and Roy [16] and Choi and Schervish [8], tried to establish the existence of uniformly consistent tests directly, which makes the numerator in the ratio of two integrals decrease to 0 exponentially.

Alternatively, LeCam [22] and Birge [4] showed that for independent non-identically distributed variables, tests with exponentially small errors exist when we use the average squared Hellinger distance to separate densities and convex sets. That is, uniformly consistent tests are always obtained if the entropy with such a distance is controlled. In the recent paper by Ghosal and van der Vaart [17], (GGR) type results have been investigated in the test construction for the convergence rates of posterior distributions for non i.i.d. observations.

On the other hand, Walker's sufficient conditions are easily adaptable in this case. Note that the proof of Lemma 3.6 does not require the assumption of the identically distributed observations; hence Theorem 3.8 easily follows to this case. We state it formally below.

Theorem 7.1. *If $A = \bigcup_{i\geq 1} A_i$ such that*

1. *For some $\delta > 0$ all the A_i's are strongly δ separated from θ_0 for the model $\theta \mapsto f_{i,\theta}$ and*
2. *$\sum_{i\geq 1} \sqrt{\Pi(A_i)} < \infty$*

Then for some $\beta_0 > 0$,

$$e^{n\beta_0} \int_A \prod_{i=1}^{n} \frac{f_{i,\theta}(x_i)}{f_{i,\theta_0}(x_i)} \Pi(d\theta) \to 0 \ a.s \prod_{i=1}^{\infty} P_{i,\theta_0}.$$

Similar results to Theorem 7.1 along with regression problems have been discussed in Walker [33].

Example 7.2 (Orthogonal series expansion)**.** Let

$$Y_i = \eta(X_i) + \epsilon_i, \ i = 1, \ldots, n \tag{7.1}$$

where the ϵ_i's are assumed to be independent $N(0,1)$ random variables, the X_i's are sampled from a known probability distribution, and $\eta(\cdot)$ is a regression function. An orthogonal series expansion for the regression function $\eta(x)$ is a representation of $\eta(x)$ by an infinite sum,

$$\eta(x) = \sum_{j=1}^{\infty} \eta_j \phi_j(x),$$

where $\{\phi_j(x)\}_{j=1}^{\infty}$ is an orthonormal basis for an L^2 space containing η. Regarding either posterior consistency or rate of convergence of posterior distributions, this model has been investigated by Shen and Wasserman [26], Walker [33] and Choi and Schervish [8].

Let $\{\phi_j(\cdot)\}_{j=1}^{\infty}$ be an orthonormal basis for $L^2[0,1]$ such that for some $C>0$, $\sup_{x\in[0,1]}|\phi_j(x)|\le C$ for all j.

In this case, we consider the following δ-covering of Ω, a union of sets of the type

$$\{\psi\ :\ n_j\delta_j<\psi_j<(n_j+1)\delta_j,\ j=1,2,\ldots\}, \tag{7.2}$$

which was also examined for Hellinger consistency in density estimation problems from infinite-dimensional exponential families in Walker [32] and regression problems in Walker [33]. Based on (7.2), the condition (b) in Theorem 7.1 can be verified as in Section 6.1 [32]. When the regression function is uniformly bounded, the L_1(or Hellinger) neighborhood of the true density f_{θ_0} becomes equivalent to the L_1 neighborhood of the true regression function η_0. Therefore, by considering a δ-covering in (7.2) and its corresponding prior probability, two conditions (a) and (b) are easily verified. Hence, the conclusion of Theorem 7.1 is achieved when A is in the L_1 neighborhood of the true density generating the regression model (7.1).

Example 7.3 (Gaussian process regression)**.** Gaussian process regression is one of the popular approaches to Bayesian nonparametric regression problems, and it is used to model the regression function $\eta(x)$ as a Gaussian process a priori. Posterior consistency based on Gaussian processes has been established in Ghosal and Roy [16] and Choi [7] for nonparametric binary regression, Tokdar and Ghosh [29] for density estimation and Choi [6] and Choi and Schervish [8] for nonparametric regression. Interestingly, all the results mentioned above have been based on constructing uniformly consistent tests rather than the condition (b) in Theorem 7.1. The challenges in the study of posterior consistency based on Gaussian processes is to find a rate that a prior probability shrinks as we consider a sequence of δ-coverings that satisfies the condition (b). In this case, the important task to be achieved is obtaining the exponentially small lower bound for small balls of Gaussian processes. There is a recent investigation in this regard (e.g. see Li and Shao [23] and van der Vaart and van Zanten [31]). It would be interesting to explore if this difficulty in verifying (b) under Gaussian process priors can be bypassed when we apply Theorem 7.1.

7.2. Doob's Theorem

Doob [11] showed that when Θ is the parameter space and given θ, $\mathbf{X}_1,\mathbf{X}_2,\ldots,$ are i.i.d. P_θ then, for any sequence of posterior distributions $\Pi(\cdot|\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_n)$, under mild set theoretic assumptions (for instance when $\mathcal{X}$ and Θ are Borel subsets of Polish spaces) for any prior Π, there is a $\Theta_\Pi\subset\Theta$ with $\Pi(\Theta_\Pi)=1$ such that the posterior is consistent at all $\theta\in\Theta_\Pi$. In what follows we explore the analogue of Doob's theorem in independent non-identically distributed models.

To change the notation a bit, given θ in Θ, let $\mathbf{Y}_1,\mathbf{Y}_2,\ldots$,be $\mathcal{Y}$ valued random variables with joint distribution $P_{\theta,\infty}$. For any prior Π on Θ, denote by λ_Π the joint distribution induced on $\Theta\times\mathcal{Y}^\infty$ by Π and $\{P_{\theta,\infty}:\theta\in\Theta\}$. We will denote the elements of $\mathcal{Y}^\infty$ by $\mathbf{y}$ and of $(\mathbf{Y}_1,\mathbf{Y}_2,\ldots,)$ by $\mathbf{Y}$. As before $\Pi(\cdot|\mathbf{Y}_1,\mathbf{Y}_2,\ldots,\mathbf{Y}_n)$ will stand for a fixed version of the posterior distribution of θ.

By going through an appropriate countable set of continuous functions g and applying the martingale convergence theorem to each posterior mean of $g(\theta)$, it can be seen that there is a conditional probability $\Pi^*(\cdot|\boldsymbol{y})$ such that for all $\boldsymbol{y}$ outside a λ_Π null set

$$\Pi(\cdot|\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n) \overset{weakly}{\to} \Pi^*(\cdot|\mathbf{y}).$$

Clearly the posterior is consistent at θ if $\Pi^*(\cdot|\boldsymbol{y}) = \delta_\theta$ a.e. $P_{\theta,\infty}$.

Proposition 7.4. Consider the following two sets of statements for a given prior Π:

1. There is a set Θ_Π with $\Pi(\Theta_\Pi) = 1$ and the posterior is consistent at all $\theta \in \Theta_\Pi$.
2. There is a set Θ_Π with $\Pi(\Theta_\Pi) = 1$ and a measurable set $E^\Pi \subset \Theta \times \mathcal{Y}^\infty$ such that
   (a) for each $\theta \in \Theta_\Pi, P_{\theta,\infty}(E^\Pi_\theta) = 1$,
   (b) $E^\Pi_\theta \cap E^\Pi_{\theta'} = \emptyset$ for $\theta \neq \theta'$.

The two sets of statements are equivalent.

Proof. Suppose (1) holds. Then it is easy to verify that the set

$$E^\Pi = \{(\theta, \boldsymbol{y}) : \Pi^*(\cdot|\boldsymbol{y}) = \delta_\theta,\ (\theta \in \Theta_\Pi)\}$$

is measurable and satisfies the conditions in (2).

On the other hand if (2) holds then define $\phi(\boldsymbol{y}) = \theta$ if $(\theta, \boldsymbol{y}) \in E^\Pi$. Then, using a result from set theory [28, Theorem 4.5.7], it can be shown that ϕ is measurable. It is easy to verify that $\tilde{\Pi}$ defined by

$$\tilde{\Pi}(\cdot|\boldsymbol{y}) = \delta_{\phi(\boldsymbol{y})}$$

is a version of the conditional distribution of θ given $\mathbf{Y}$ and hence $\Pi^*(\cdot|\boldsymbol{y}) = \tilde{\Pi}(|\boldsymbol{y})$ a.e. λ_Π. An application of Fubini's theorem yields the result.

□

Our interest is in establishing (1) for all priors Π and it is convenient to work with a stronger version of (2) by seeking a decomposition that does not depend upon Π. Formally,

Proposition 7.5. Let Π be a prior for Θ.

Suppose there exists a measurable set $E \subset \Theta \times \mathcal{Y}^\infty$ such that

1. For each $\theta \in \Theta, P_{\theta,\infty}(E_\theta) = 1$ where E_θ is the θ-section $\{\boldsymbol{y} : (\theta, \boldsymbol{y}) \in E\}$.
2. $E_\theta \cap E_{\theta'} = \emptyset$ for $\theta \neq \theta'$.

Then there is a set Θ_Π with Π measure 1, such that the posterior is consistent at all $\theta \in \Theta_\Pi$.

Thus, Doob's theorem is intimately related to uniform orthogonality of $\{P_{\theta,\infty} : \theta \in \Theta\}$. There is a wide literature on singularity and mutual absolute continuity of measures on infinite product spaces ([27] and [5]). This literature in general deals with pairwise orthogonality whereas Proposition 7.5 requires uniform orthogonality. The step from pairwise to uniform orthogonality can be formidable. Yet we feel that some of these results are likely to be useful in establishing Doob-type theorems in the non-i.i.d. set up.

Motivated by Proposition 7.5, we present an example where the Doob-type theorem fails to hold. On the positive side, Proposition 7.5 enables us to prove a theorem for linear regression models with nonparametric errors.

The case that we consider is

$$Y_i = \alpha + \beta x_i + \epsilon_i \qquad i = 1, 2, \ldots$$

where

1. $x_1, x_2, \ldots,$ are fixed non-random design points.
2. $\epsilon_1, \epsilon_2, \ldots$ are independent and identically distributed random variables with a probability density symmetric around 0.

Example 7.6. Suppose $\sum_i x_i^2 < \infty$ and $\epsilon_i \sim N(0,1)$, and let $\alpha = 0$. In this case it follows from a result of Shepp [27] that $\prod_1^\infty N(\beta x_i, 1)$ are mutually absolutely continuous. Hence the decomposition required by Proposition 7.4 fails and Doob's theorem cannot hold.

The last example we consider is semiparametric regression, the linear regression model where the distribution of the noise is assumed to be unknown and thus needs to be estimated. This example has been investigated in terms of posterior consistency, following from the generalization of the Schwartz theorem in Amewou-Atisso *et al.* [1]. We revisit this example in Theorem 7.7 and show that the Doob-type theorem holds with an assumption on the fixed non-random design points, similar to that of [1].

Let Assumption A be defined as the following: There exists $\epsilon_0 > 0$ such that the covariate values x_i's satisfy

$$\sum_i I_{(-\infty, -\epsilon_0)}(x_i) = \infty \text{ and } \sum_i I_{(\epsilon_0, \infty)}(x_i) = \infty.$$

Theorem 7.7. *Consider the model*

$$Y_i = \alpha + \beta x_i + \epsilon_i \qquad i = 1, 2, \ldots$$

where

1. *$x_1, x_2, \ldots,$ are fixed nonrandom design points*
2. *$\epsilon_1, \epsilon_2, \ldots$ are i.i.d. variables with an unknown distribution of which density f is symmetric, continuous at 0 and $f(0) > 0$.*

If Assumption A holds, then given any prior Π for (α, β, f), there is a set Θ_Π of Π measure 1 such that the posterior is consistent at all $(\alpha, \beta, f) \in \Theta_\Pi$.

Proof. Let $\mathcal{F}$ be all densities f on the real line which are symmetric, continuous at 0 and $f(0) > 0$. Formally, we have as the parameter space $\Theta = R \times R \times \mathcal{F}$ and given (α, β, f), the Y_i's are independent with $Y_i \sim f_{\alpha+\beta x_i}$, where $f_{\alpha+\beta x_i}(y) = f(y - (\alpha + \beta x_i))$.

We now construct a decomposition satisfying the conditions of Proposition 7.5.

Let $N_1 = \{n_1, n_2, \ldots\}$ be the subsequence of all i with $x_i > \epsilon$ and $M_1 = \{m_1, m_2, \ldots\}$ be the subsequence of all i with $x_i < -\epsilon$. Let t be a real number and define $A_t = (t, \infty)$.

Note that the unknown parameter θ is the triple $\theta \equiv (\alpha, \beta, f)$. Following notations in Proposition 7.5, let Π be a prior on Θ and let E^Π be the set of all $(\alpha, \beta, f, \boldsymbol{y})$ such that for any real number t,

1. $\lim_{n\to\infty} \frac{1}{n}\sum_1^n I_{A_t}(y_i - (\alpha+\beta x_i)) = P_f(A_t)$
2. $\lim_{k\to\infty} \frac{1}{k}\sum_1^k I_{A_t}(y_{n_i} - (\alpha+\beta x_{n_i})) = P_f(A_t)$
3. $\lim_{l\to\infty} \frac{1}{l}\sum_1^l I_{A_t}(y_{m_i} - (\alpha+\beta x_{m_i})) = P_f(A_t)$

Since N_1, M_1 are fixed subsequences and since it is enough to work with t - rational, E^Π is easily seen to be measurable.

Further, for each (α,β,f), $[\prod_1^\infty P_{\alpha+\beta x_i}]\,(E^\Pi_{\alpha,\beta,f}) = 1$, where $E^\Pi_{\alpha,\beta,f}$ is the (α,β,f)-section $\{\boldsymbol{y} : (\alpha,\beta,f,\boldsymbol{y}) \in E^\Pi\}$ for each (α,β,f) as defined in Proposition 7.5. This follows by noting that under $[\prod_1^\infty P_{\alpha+\beta x_i}]$, $Y_1 - (\alpha+\beta x_1), Y_2 - (\alpha+\beta x_2), \ldots$ are i.i.d. with common density f. An application of the law of large numbers proves the claim.

We next argue that if $(\alpha_1,\beta_1,f_1) \neq (\alpha_2,\beta_2,f_2)$ then $E^\Pi_{\alpha_1,\beta_1,f_1} \cap E^\Pi_{\alpha_2,\beta_2,f_2} = \emptyset$.

If $\alpha_1 = \alpha_2, \beta_1 = \beta_2$ and $f_1 \neq f_2$, and if $\boldsymbol{y} \in E^\Pi_{\alpha,\beta,f_1} \cap E^\Pi_{\alpha,\beta,f_2}$ then a contradiction is easily obtained by considering a t for which $P_{f_1}(A_t) \neq P_{f_2}(A_t)$.

Now suppose that for some $\Delta > 0$, $\alpha_1 - \alpha_2 > \Delta$ and $\beta_1 - \beta_2 > \Delta$. Clearly for every $n_i \in N_1$, $(\beta_1 - \beta_2)x_{n_i} > \Delta\epsilon$. Choose η such that $\eta < \Delta\epsilon$ and $\inf_{|x|<\eta} f_1(x) > C > 0$.

Since f is symmetric and $\eta > 0, P_{f_1}(A_\eta) < 1/2$. We will get a contradiction by showing that if $\boldsymbol{y} \in E^\Pi_{\alpha_1,\beta_1,f_1} \cap E^\Pi_{\alpha_2,\beta_2,f_2}$, then $P_{f_1}(A_\eta) \geq 1/2$.

If $\boldsymbol{y} \in E^\Pi_{\alpha_1,\beta_1,f_1} \cap E^\Pi_{\alpha_2,\beta_2,f_2}$, then for all t,

$$\frac{1}{k}\sum_1^k I_{A_t}(y_{n_i} - (\alpha_1+\beta_1 x_{n_i})) \to P_{f_1}(A_t) \tag{7.3}$$

$$\frac{1}{k}\sum_1^k I_{A_t}(y_{n_i} - (\alpha_2+\beta_2 x_{n_i})) \to P_{f_2}(A_t) \tag{7.4}$$

$$\begin{aligned} \alpha_1 + \beta_1 x_{n_i} &= \alpha_2 + \beta_2 x_{n_i} + (\alpha_1 - \alpha_2) + (\beta_1 - \beta_2)x_{n_i} \\ &\geq \alpha_2 + \beta_2 x_{n_i} + \eta \end{aligned}$$

and hence

$$I_{A_t}(y_{n_i} - (\alpha_1+\beta_1 x_{n_i})) \geq I_{A_t}(y_{n_i} - (\alpha_2+\beta_2 x_{n_i} + \eta)) = I_{A_{t-\eta}}(y_{n_i} - (\alpha_2+\beta_2 x_{n_i}))$$

In particular with $t = \eta$,

$$I_{A_\eta}(y_{n_i} - (\alpha_1+\beta_1 x_{n_i})) \geq I_{(0,\infty)}(y_{n_i} - (\alpha_2+\beta_2 x_{n_i})).$$

Consequently

$$\frac{1}{k}\sum_1^k I_{A_t}(y_{n_i} - (\alpha_1+\beta_1 x_{n_i})) \geq \frac{1}{k}\sum_1^k I_{(0,\infty)}(y_{n_i} - (\alpha_2+\beta_2 x_{n_i})) \to P_{f_2}(0,\infty) = \frac{1}{2}$$

The case when $\alpha_1 - \alpha_2 < \Delta, \beta_1 - \beta_2 > \Delta$ can be handled by considering the subsequence M_1. Similarly, the other remaining cases follow.

□

Acknowledgments

This paper, like all my research endeavours, has benefitted from J.K. Ghosh. Bayesian nonparametrics is but a small part of all that I have learned from him, and I am

but one among many who owe their intellectual development to his influence. My association with J.K. Ghosh, JKG as I call him, began more than thirty five years ago in the form of student and teacher. This equation has remained constant but over the years, on top of it, has developed a friendship that I greatly cherish. This article is dedicated to JKG with admiration, appreciation, affection and .... gratitude.

R.V. Ramamoorthi

References

[1] Amewou-Atisso, M., Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (2003). Posterior consistency for semi-parametric regression problems. *Bernoulli* **9**, 2, 291–312. MR1997031 (2004f:62075)

[2] Barron, A. (1988). The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions. Tech. Rep. 7, Dept. of Statistics, Univ. Illinois, Champaign.

[3] Barron, A., Schervish, M. J., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist. 27*, 536–561.

[4] Birgé, L. (1983). Robust testing for independent non-identically distributed variables and Markov chains. In *Practical nonparametric and semiparametric Bayesian statistics, Lecture Notes in Statistics 133*, D. Dey, P. Müller, and D. Sinha, Eds. Springer-Verlag, New York.

[5] Chatterji, S. D. and Mandrekar, V. (1978). *Equivalence and singularity of Gaussian measures and applications.* Probabilistic analysis and related topics, Vol. 1. Academic Press, New York.

[6] Choi, T. (2005). Posterior consistency in nonparametric regression problems under Gaussian process priors. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.

[7] Choi, T. (2007). Alternative posterior consistency results in nonparametric binary regression using gaussian process priors. *J. Statist. Plann. Inference 137*, 2975–2983.

[8] Choi, T. and Schervish, M. J. (2007). On posterior consistency in nonparametric regression problems. *J. Multivariate Anal. 98*, 1969–1987.

[9] Choudhuri, N., Ghosal, S., and Roy, A. (2004). Bayesian estimation of the spectral density of a time series. *J. Amer. Statist. Assoc.* **99**, 468, 1050–1059. MR2109494 (2005j:62175)

[10] Diaconis, P. and Freedman, D. A. (1986). On the consistency of bayes estimates. *Ann. Statist. 14*, 1–26.

[11] Doob, J. L. (1949). Application of the theory of martingales. *Coll. Int. du C. N. R. S. Paris.*, 23–27.

[12] Freedman, D. A. (1963). On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist. 34*, 1386–1403.

[13] Freedman, D. A. (1965). On the asymptotic behavior of Bayes' estimates in the discrete case ii. *Ann. Math. Statist. 36*, 454–456.

[14] Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999a). Consistency issues in Bayesian nonparametrics. In *Asymptotics, nonparametrics, and time series.* Statist. Textbooks Monogr., Vol. **158**. Dekker, New York, 639–667. MR1724711

[15] Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999b). Posterior

consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27**, 1, 143–158. MR1701105 (2000j:62053)

[16] Ghosal, S. and Roy, A. (2006). Posterior consistency of Gaussian process prior for nonparametric binary regression. *Ann. Statist. 34*, 2413–2429.

[17] Ghosal, S. and van der Vaart. (2007). Convergence rates of posterior distributions for noniid observations. *Ann. Statist. 35*, 192–223.

[18] Ghosh, J. K., Delampady, M., and Samanta, T. (2006). *An introduction to Bayesian analysis Theory and Methods.* Springer Texts in Statistics. Springer, New York.

[19] Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian nonparametrics.* Springer Series in Statistics. Springer, New York. MR1992245 (2004g:62004)

[20] Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc. 91*, 1343–1370.

[21] Kraft, C. (1955). Some conditions for consistency and uniform consistency of statistical procedures. *Univ. California Publ. Statist. 2*, 125–141. MR0073896 (17,505a)

[22] Le Cam, L. M. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer, New York.

[23] Li, W. and Shao, Q.-M. (2001). Gaussian processes : inequalities, small ball probabilities and applications. In *In Stochastic processes : theory and methods.* Handbook of Statist., Vol. **19**. North-Holland, Amsterdam, 533–597.

[24] Salinetti, G. (2003). New tools for consistency in bayesian nonparametrics. In *Bayesian Statistics 7.* 369–384, Oxford Univ. Press, New York.

[25] Schwartz, L. (1965). On Bayes procedures. *Z. Wahr. Verw. Gebiete 4*, 10–26.

[26] Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *Ann. Statist. 29*, 666–686.

[27] Shepp, L. A. (1965). Distingunishing a sequence of random variables from a translate of itself. *Ann. Math. Statist. 36*, 1107–1112.

[28] Srivastava, S. M. (1998). *A course on Borel sets.* Number 180 in Graduate Texts in Mathematics. Springer-Verlag, Berlin.

[29] Tokdar, S. and Ghosh, J. K. (2007). Posterior consistency of Gaussian process priors in density estimation. *J. Statist. Plann. Inference 137*, 34–42.

[30] van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist. 21*, 14–44.

[31] van der Vaart, A. W. and van Zanten, J. (2007). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*. to appear.

[32] Walker, S. (2004). New approaches to Bayesian consistency. *Ann. Statist.* **32**, 5, 2028–2043. MR2102501 (2006c:62049)

[33] Walker, S. G. (2003). Bayesian consistency for a class of regression problems. *South African Stat. Journal 37*, 149–167.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 188
© Institute of Mathematical Statistics, 0000
arXiv: `math.PR/0000000`

Large Sample Asymptotics for the Two-Parameter Poisson-Dirichlet Process

Lancelot F. James[1]

Hong Kong University of Science and Technology

Abstract: This paper explores large sample properties of the two-parameter (α, θ) Poisson-Dirichlet Process in two contexts. In a Bayesian context of estimating an unknown probability measure, viewing this process as a natural extension of the Dirichlet process, we explore the consistency and weak convergence of the the two-parameter Poisson-Dirichlet posterior process. We also establish the weak convergence of properly centered two-parameter Poisson-Dirichlet processes for large $\theta + n\alpha$. This latter result complements large θ results for the Dirichlet process and Poisson-Dirichlet sequences, and complements a recent result on large deviation principles for the two-parameter Poisson-Dirichlet process. A crucial component of our results is the use of distributional identities that may be useful in other contexts.

Contents

1. Introduction

In this work, for $0 \leq \alpha < 1$ and $\theta > -\alpha$, we are interested in the two-parameter class of random probability measures that are formed by

$$P_{\alpha,\theta}(\cdot) \overset{d}{=} \sum_{k=1}^{\infty} V_k \prod_{j=1}^{k-1} (1 - V_j) \delta_{Z_k}(\cdot) \tag{1.1}$$

where the V_k are independent beta$(1-\alpha, \theta+k\alpha)$ random variables and, independent of these, the (Z_k) are an i.i.d. sequence with values in some Polish space I with common (non-atomic) distribution H. That is to say $P_{\alpha,\theta}$ is a random probability measure taking values in $\mathcal{P}_I$, where $\mathcal{P}_I$ is the set of all probability measures on I. We will

[1] Lancelot F. James, The Hong Kong University of Science and Technology, Department of Information and Systems Management, Clear Water Bay, Kowloon, Hong Kong. e-mail: `lancelot@ust.hk`.

AMS 2000 subject classifications: Primary 62G05; secondary 62F15
Keywords and phrases: Bayesian consistency, multiplier CLT, weak convergence, two-parameter Poisson-Dirichlet process

simply say that a random probability measure P is a two-parameter (α, θ) Poisson-Dirichlet process, having law say $\Pi_{\alpha,\theta}$ on $\mathcal{P}_I$, suppressing dependence on H, if P can be represented as in (1.1). That is $P(\cdot) \overset{d}{=} P_{\alpha,\theta}(\cdot)$. For shorthand we write $P \sim \Pi_{\alpha,\theta}$. We denote the expectation operator corresponding to the law $\Pi_{\alpha,\theta}$ as $E_{\alpha,\theta}$, which is such that $E_{\alpha,\theta}[P(\cdot)] = H(\cdot)$. We note that the V_k are obtained by size-biasing a ranked sequence of probabilities known as the two-parameter Poisson-Dirichlet sequence. It follows that by permutation invariance $P_{\alpha,\theta}$ may also be represented in terms of this sequence. Many properties of the two-parameter Poisson-Dirichlet sequence, as it primarily relates to Bessel and Brownian phenomena, were discussed in [36]. This sequence has gained in importance as it is seen to arise in a number of different areas including, for instance, Bayesian statistics, population genetics and random fragmentation and coalescent theory connected to physics. See [34] for some updated references and [19] for some connections to Dirichlet means.

When $\alpha = 0$ then P is a Dirichlet process in the sense of Ferguson [9]. Pitman [35] showed that within a Bayesian context, these random probability measures can be seen as natural and quite tractable extensions of the Dirichlet process. In particular Pitman [35] showed that if random variables $X_1, \ldots, X_n$ given P are i.i.d. P and P has prior distribution $\Pi_{\alpha,\theta}$, then the posterior distribution of $P|X_1, \ldots, X_n$, denoted as $\Pi^{(n)}_{\alpha,\theta}$, corresponds to the law of the random probability measure,

$$P^{(n)}_{\alpha,\theta}(\cdot) = R_{n(\mathbf{p})} P_{\alpha,\theta+n(\mathbf{p})\alpha}(\cdot) + (1 - R_{n(\mathbf{p})}) D_n(\cdot)$$

where

$$D_n(\cdot) = \sum_{j=1}^{n(\mathbf{p})} \Delta_j \delta_{Y_j}(\cdot).$$

$(\Delta_1, \ldots, \Delta_{n(\mathbf{p})})$ is a Dirichlet$(e_1 - \alpha, \ldots, e_{n(\mathbf{p})} - \alpha)$ random vector. All the random variables appearing on the right hand side are conditionally independent given the data. $R_{n(\mathbf{p})}$ is a beta$(\theta + n(\mathbf{p})\alpha, n - n(\mathbf{p})\alpha)$ random variable and $\{Y_1, \ldots, Y_{n(\mathbf{p})}\}$ denotes the $1 \leq n(\mathbf{p}) \leq n$ unique values of $\{X_1, \ldots, X_n.\}$ Furthermore, $P_{\alpha,\theta+n(\mathbf{p})\alpha} \sim \Pi_{\alpha,\theta+n(\mathbf{p})\alpha}$. e_j is the number of X_i equivalent to Y_j for $j = 1, \ldots, n(\mathbf{p})$., When $\alpha = 0$ one obtains the posterior distribution derived in Ferguson [9]. The notation e_j and $n(\mathbf{p})$ are taken from Lo [29], as discussed in Ishwaran and James [17]. Furthermore, a generalization of the Blackwell–MacQueen [4] prediction rule is given by

$$(1.2) \qquad P(X_{n+1} \in \cdot\, |X_1, \ldots, X_n) = \frac{\theta + n(\mathbf{p})\alpha}{\theta + n} H(\cdot) + \sum_{j=1}^{n(\mathbf{p})} \frac{(e_j - \alpha)}{\theta + n} \delta_{Y_j}(\cdot).$$

Note also that

$$P(X_{n+1} \in \cdot\, |X_1, \ldots, X_n) = E[P^{(n)}_{\alpha,\theta}(\cdot)].$$

We also write

$$\tilde{F}_n(\cdot) = \sum_{j=1}^{n(\mathbf{p})} \frac{(e_j - \alpha)}{n - n(\mathbf{p})\alpha} \delta_{Y_j}(\cdot)$$

which importantly reduces to the empirical distribution when $\alpha = 0$, or when $n(\mathbf{p}) = n$.

Let P_0^∞ denote a product measure on I^∞ making X_i for $i = 1, \ldots, \infty$ independent with common (true) distribution P_0. In this paper, extending the known case of the Dirichlet process, we show that as $n \to \infty$ the posterior distribution $\Pi^{(n)}_{\alpha,\theta}$ behaves as follows. When P_0 is discrete then $\Pi^{(n)}_{\alpha,\theta}$ converges weakly to a point mass at P_0 a.s. P_0^{∞}. When P_0 is continuous, that is, non-atomic, $\Pi^{(n)}_{\alpha,\theta}$ converges weakly to a point mass at the mixture $\alpha H + (1-\alpha)P_0$. Thus when P_0 is discrete/atomic the posterior distribution is consistent. However, when P_0 is non-atomic the posterior distribution is inconsistent, unless either $\alpha = 0$ which corresponds to the case of the Dirichlet process, or more implausibly one chooses $H = P_0$. In addition to this result we establish a functional central limit theorem, for the case where P_0 is non-atomic, by showing that the process $P^{(n)}_{\alpha,\theta}$ centered at its expectation (1.2), indexed over classes of functions, converges weakly to a Gaussian process. This is in line with nonparametric Bernstein–Von Mises results of for instance, [31], [30],[28], [5] and [22]. Additionally, we note that the weak convergence of the two-parameter Poisson-Dirichlet process may be of interest in other fields. In particular, in order for us to discuss the posterior weak convergence, when P_0 is non-atomic, we will need to address the weak convergence of the centered process

$$\nu_{\alpha,\theta+n\alpha}(\cdot) = \sqrt{n}(P_{\alpha,\theta+n\alpha} - H)(\cdot),$$

as $n \to \infty$, which poses additional challenges. Note this process does not depend on the data, except through the sample size n. Furthermore, the study of $\nu_{\alpha,\theta+n\alpha}$ is more in line with the literature on the behavior of Dirichlet processes and Poisson-Dirichlet sequences when $\theta \to \infty$. See, for instance, [21], [7] and [32]. Additionally, our work is complementary to a recent result of [8] on large deviation principles for the two-parameter Poisson-Dirichlet process.

Returning to the consistency result, in terms of estimating the true P_0 in a nonparametric statistics setting, our result shows that unlike the case of the Dirichlet process, it is perhaps unwise to use $\Pi_{\alpha,\theta}$ as a prior. However we should point out that Ishwaran and James [17, 18], owing to the attractive results in [35], suggested that one could use $\Pi_{\alpha,\theta}$ in a mixture modeling setting analogous to the case of the Dirichlet process in Lo [29]. In this more formidable setting one can deduce strong consistency of the posterior density, induced by the priors $\Pi_{\alpha,\theta}$ on the mixing distribution, by using the results of Lijoi, Prünster and Walker [25]. In fact, in a work subsequent to ours, this was recently shown by Jang, Lee and Lee[20]. We also note that those authors also obtain our consistency result as a special case. For some more results on the modern treatment of Bayesian consistency in nonparametric settings one may note, for instance, the works of Ghosal, Ghosh and Ramamoorthi [11], Barron, Schervish and Wasserman [2] and Ghosal, Ghosh and van der Vaart [12]; and the book of Ghosh and Ramamoorthi [13].

2. Consistency

This section describes the consistency behavior of the posterior distribution in the case where the true distribution P_0 is either continuous or discrete. First we note the following fact;

Lemma 2.1. *Let f and g denote measurable functions on I, then for $0 \le \alpha < 1$ and $\theta > -\alpha$,*

$$E_{\alpha,\theta}[P(f)P(g)] = \frac{\theta+\alpha}{\theta+1}H(g)H(f) + \frac{1-\alpha}{\theta+1}H(fg).$$

Proof. The proof proceeds by using disintegrations. First the joint distribution of (X_1, P) can be written as,

$$P(dx_1)\Pi_{\alpha,\theta}(dP) = \Pi_{\alpha,\theta}(dP|x_1)H(dx_1)$$

where $\Pi_{\alpha,\theta}(dP|x_1) = \Pi^{(1)}_{\alpha,\theta}(dP)$. Then,

$$P(dx_2)\Pi_{\alpha,\theta}(dP|x_1) = \Pi^{(2)}_{\alpha,\theta}(dP)E[P(dx_2)|X_1 = x_1]$$

where $\Pi^{(2)}_{\alpha,\theta}(dP) = \Pi_{\alpha,\theta}(dP|x_1, x_2)$ and

$$E[P(dx_2)|X_1 = x_1] = \frac{\theta+\alpha}{\theta+1}H(dx_2) + \frac{1-\alpha}{\theta+1}\delta_{x_1}(dx_2).$$

Now this gives

$$E_{\alpha,\theta}[P(dx_1)P(dx_2)] = H(dx_1)\left[\frac{\theta+\alpha}{\theta+1}H(dx_2) + \frac{1-\alpha}{\theta+1}\delta_{x_1}(dx_2)\right].$$

which by writing $P(g)P(f) = \int_I\int_I g(x_1)f(x_2)P(dx_1)P(dx_2)$ completes the result. □

We proceed as in Diaconis and Freedman[10] by showing that the posterior distribution concentrates around the prediction rule. First using Diaconis and Freedman ([10], p.1117), we define a suitable class of semi-norms such that convergence under such norms implies weak convergence. Let $\mathcal{A} = \bigcup_{i=1}^{\infty} A_i$ be a partition of I. Then define the semi-norm between probability measures

$$\mid P - Q \mid_{\mathcal{A}} = \sqrt{\sum_{i=1}^{\infty}[P(A_i) - Q(A_i)]^2}, \tag{2.1}$$

for a suitable generating sequence of partitions $\mathcal{A}$ where, naturally for any Q, $\sum_{i=1}^{\infty} Q(A_i) = 1$. Now similar to ([10], Equation 14) for the Dirichlet process, we will show that the posterior distribution concentrates around the prediction rule (1.2).

In order to do this one only needs to evaluate the posterior expectation, expressible as,

$$E\left[\mid P^{(n)}_{\alpha,\theta} - E[P^{(n)}_{\alpha,\theta}] \mid^2_{\mathcal{A}}\right] \tag{2.2}$$

where $E[P^{(n)}_{\alpha,\theta}]$ equates with the prediction rule probability given in (1.2.) We obtain,

Lemma 2.2.

$$E\left[\mid P^{(n)}_{\alpha,\theta} - E[P^{(n)}_{\alpha,\theta}] \mid^2_{\mathcal{A}}\right] \le \frac{1}{\theta+n+1}.$$

Proof. First using basic ideas we expand, for each set A_i,

$$\left(P^{(n)}_{\alpha,\theta}(A_i) - E[P^{(n)}_{\alpha,\theta}(A_i)]\right)^2.$$

Furthermore,

$$\begin{aligned}(P^{(n)}_{\alpha,\theta}(A_i))^2 &= R^2_{n(\mathbf{p})}P^2_{\alpha,\theta+n(\mathbf{p})\alpha}(A_i)\\ &+ 2R_{n(\mathbf{p})}(1-R_{n(\mathbf{p})})P_{\alpha,\theta+n(\mathbf{p})\alpha}(A_i)D_n(A_i) + (1-R_{n(\mathbf{p})})^2D^2_n(A_i).\end{aligned}$$

Now from Lemma 2.1

$$E[P^2_{\alpha,\theta+n(\mathbf{p})\alpha}(A_i)] = \frac{\theta+n(\mathbf{p})\alpha+\alpha}{\theta+n(\mathbf{p})\alpha+1}H^2(A_i) + \frac{1-\alpha}{\theta+n(\mathbf{p})\alpha+1}H(A_i).$$

Additionally,

$$E[R^2_{n(\mathbf{p})}] = \frac{(\theta+n(\mathbf{p})\alpha)(\theta+n(\mathbf{p})\alpha+1)}{(\theta+n)(\theta+n+1)},$$

$$E[(1-R_{n(\mathbf{p})})^2] = \frac{(n-n(\mathbf{p})\alpha)(n-n(\mathbf{p})\alpha+1)}{(\theta+n)(\theta+n+1)},$$

and

$$E[\Delta_l\Delta_j] = \frac{(e_j-\alpha)(e_l-\alpha)}{(n-n(\mathbf{p})\alpha)(n-n(\mathbf{p})\alpha+1)}.$$

It follows that

$$\begin{aligned} E[(P^{(n)}_{\alpha,\theta}(A_i))^2] &= \frac{(\theta+n(\mathbf{p})\alpha)(\theta+n(\mathbf{p})\alpha+\alpha)}{(\theta+n)(\theta+n+1)}H^2(A_i) \\ &+ 2\frac{(\theta+n(\mathbf{p})\alpha)(n-n(\mathbf{p})\alpha)}{(\theta+n)(\theta+n+1)}\tilde{F}_n(A_i)H(A_i) \\ &+ \frac{(n-n(\mathbf{p})\alpha)^2}{(\theta+n)(\theta+n+1)}\tilde{F}^2_n(A_i). \end{aligned}$$

Now

$$\begin{aligned} (E[P^{(n)}_{\alpha,\theta}(A_i)])^2 &= \frac{(\theta+n(\mathbf{p})\alpha)^2}{(\theta+n)^2}H^2(A_i) \\ &+ 2\frac{(\theta+n(\mathbf{p})\alpha)(n-n(\mathbf{p})\alpha)}{(\theta+n)^2}\tilde{F}_n(A_i)H(A_i) \\ &+ \frac{(n-n(\mathbf{p})\alpha)^2}{(\theta+n)^2}\tilde{F}^2_n(A_i). \end{aligned}$$

Taking differences and using the fact that $\sum_{i=1}^{\infty} F_n(A_i)H(A_i) \le 1$ and similar arguments completes the result. □

Proposition 2.1. *If P_0 is continuous then the posterior distribution $\Pi^{(n)}_{\alpha,\theta}$ converges weakly to point mass at the distribution*

$$\alpha H(\cdot) + (1-\alpha)P_0(\cdot) \text{ a.e. } P_0^{\infty}.$$

Hence the posterior is consistent only if either P is a Dirichlet process or $H = P_0$.

Proof. From Lemma 2.2 it follows that the posterior distribution must concentrate around the prediction rule. Now, under P_0 (assuming a continuous P_0) the prediction rule becomes,

$$P(X_{n+1} \in \cdot\,|X_1,\ldots,X_n) = \frac{\theta+n\alpha}{\theta+n}H(\cdot) + \sum_{j=1}^{n}\frac{(1-\alpha)}{\theta+n}\,\delta_{X_j}(\cdot). \tag{2.3}$$

It is clear that under P_0, that $P\{X_{n+1} \in \cdot\,|X_1,\ldots,X_n\}$ converges uniformly over appropriate Glivenko–Cantelli classes to

$$\alpha H(\cdot) + (1-\alpha)P_0(\cdot)$$

for almost all sample sequences. One gets this by simple algebra and classical results about empirical processes (the second term in (2.3)) appropriately modified. □

The previous results says that the posterior distribution is inconsistent for all non-atomic P_0 unless $\alpha = 0$ or one has chosen $H = P_0$. The behavior in the case where P_0 admits ties is quite different and is summarized in the next result

Proposition 2.2. *Suppose that P_0 is a discrete law such that $n(\mathbf{p})/n \to 0$ then the posterior distribution $\Pi^{(n)}_{\alpha,\theta}$ converges weakly to point mass at P_0, a.e. P_0^∞, for all $0 \le \alpha < 1$ and $\theta > -\alpha$.*

Proof. Since, $n(\mathbf{p})/n \to 0$, it follows that

$$P(X_{n+1} \in \cdot \,|X_1, \ldots, X_n) = \frac{\theta + n(\mathbf{p})\alpha}{b+n} H(\cdot) + \sum_{j=1}^{n(\mathbf{p})} \frac{(e_j - \alpha)}{\theta + n} \delta_{Y_j}(\cdot), \tag{2.4}$$

converges uniformly to P_0 for almost all sample sequences $X_1, X_2, \ldots,$. This is true since,

$$\sum_{j=1}^{n(\mathbf{p})} \frac{(e_j - \alpha)}{\theta + n} \delta_{Y_j} = \sum_{i=1}^{n} \frac{1}{\theta + n} \delta_{X_i} - \sum_{j=1}^{n(\mathbf{p})} \frac{\alpha}{\theta + n} \delta_{Y_j}$$

where the second term on the right converges to zero. □

2.1. Some More Limits

Note that one obtains some information on the limit behavior of the random probability measure $P_{\alpha,\theta+n(\mathbf{p})\alpha}$ which has law $\Pi_{\alpha,\theta+n(\mathbf{p})\alpha}$. This type of result is more in line with large θ type asymptotics. In this case large θ is replaced by large $n(\mathbf{p})\alpha$.

Proposition 2.3. *Suppose that P_0 is such that $n(\mathbf{p}) \to \infty$, then the two-parameter Poisson-Dirichlet law $\Pi_{\alpha,\theta+n(\mathbf{p})\alpha}$ converges weakly to point mass at H.*

Proof. The proof proceeds by again utilizing the semi-norm in (2.1). We have that $E_{\alpha,\theta+n(\mathbf{p})\alpha}[P] = H$. Furthermore, from Lemma 2.1. the variance of $P(A)$ under $\Pi_{\alpha,\theta+n(\mathbf{p})\alpha}$ is, for each A, equal to,

$$\frac{1-\alpha}{\theta + n(\mathbf{p})\alpha + 1} H(A)(1 - H(A)).$$

Hence if P has law $\Pi_{\alpha,\theta+n(\mathbf{p})\alpha}$, then,

$$E_{\alpha,\theta+n(\mathbf{p})\alpha}\left[|P - H|^2_{\mathcal{A}}\right] \le \frac{1-\alpha}{\theta + n(\mathbf{p})\alpha + 1},$$

completing the result. □

3. Bernstein–von Mises and Functional Central Limit Theorems

In this section we address the more formidable problem of establishing functional central limit theorems for centered versions of the two-parameter process. We will restrict ourselves to the case where P_0 is continuous which presents some difficulties. In particular, in that setting, we are interested in the asymptotic behavior in distribution, as $n \to \infty$, of the posterior process

$$\nu^{(n)}_{\alpha,\theta}(\cdot) = \sqrt{n}(P^{(n)}_{\alpha,\theta} - E[P^{(n)}_{\alpha,\theta}])(\cdot)$$

conditional on the sequence $X_1, X_2, \ldots$, and the asymptotic behavior of the process,

$$\nu_{\alpha,\theta+n\alpha}(\cdot) := \sqrt{n}(P_{\alpha,\theta+n\alpha} - H)(\cdot),$$

uniform over classes of functions $\mathcal{F}$. For clarity, we first mention some elements of the (modern) framework of weak convergence of stochastic processes on general function indexed Banach spaces. There is a rich literature on this subject, here we use as references [39], ([38], Ch. 10) and [14]. Let $\mathcal{F}$ denote a collection of measurable functions $f : I \to \Re$ and let $\ell^\infty(\mathcal{F})$ denote the set of all uniformly bounded real functions on $\mathcal{F}$. Now for a random probability measure Q_n, and the probability measure defined as its expectation $E[Q_n] := Q$, we consider the maps from $\mathcal{F} \to \Re$ given by the linear functional

$$f \to Q_n(f) = \int_I f(x) Q_n(dx),$$

and

$$f \to Q(f) = \int_I f(x) Q(dx).$$

$\mathbb{G}_n(\cdot) = \sqrt{n}(Q_n - Q)(\cdot)$ denotes its centered process and let $\mathbb{G}_Q$ denote a Gaussian process with zero mean and covariances

$$E[\mathbb{G}_Q(f)\mathbb{G}_Q(g)] = Q(fg) - Q(f)Q(g). \tag{3.1}$$

A Gaussian process with covariance (3.1) is said to be a Q-Brownian bridge. We will assume that $\mathcal{F}$ possesses enough measurability for randomization and write, as in ([37], p. 2056), $\mathcal{F} \in M(Q)$. The notation $L_2(Q)$ represents the equivalence class of square integrable functions. Furthermore, a function $F(x)$ such that $|f(x)| \le F(x)$ for all x and $f \in \mathcal{F}$ is said to be an envelope.

We are interested in the cases where the sequence of processes $\{G_n(f) : f \in \mathcal{F}\} \in \ell^\infty(\mathcal{F})$ converges in distribution to a Gaussian process $\mathbb{G}_Q$ uniformly over $\mathcal{F}$. In that case we write

$$G_n \rightsquigarrow \mathbb{G}_Q \text{ in } \ell^\infty(\mathcal{F}).$$

Furthermore, because we are interested in the convergence of posterior distributions, if G_n depends on data $X_1, X_2 \ldots$, we will need the more delicate notion of conditional weak convergence of $G_n(\cdot)$ for almost all sample sequences $X_1, X_2 \ldots$, and we write

$$G_n \rightsquigarrow \mathbb{G}_Q \text{ in } \ell^\infty(\mathcal{F}) \text{ a.s. .}$$

More formally, one may say that the processes converge in the sense of a bounded Lipschitz metric outer almost surely (see [14], and [39]).

The rich theory of weak convergence of empirical processes addresses the cases where $Q_n = P_n = n^{-1}\sum_{i=1}^n \delta_{X_i}(\cdot)$ is the empirical measure, and $Q = E_{P_0}[P_n] = P_0$ is the true underlying distribution of the data. Hence one has,

$$\sqrt{n}(P_n - P_0) \rightsquigarrow \mathbb{G}_{P_0} \text{ in } \ell^\infty(\mathcal{F}) \tag{3.2}$$

for many classes of $\mathcal{F}$. The classes are said to be P_0-Donsker. The classical case of the empirical distribution function, $F_n(t) = \sum_{i=1}^n I_{(X_i \le t)}/n$ arises by setting $f_t(x) = \mathrm{I}_{(-\infty < x \le t)}$ for t ranging over $\Re$; see ([38], Example 19.6).

The most notable results for convergence conditionally on the data, center around the bootstrap and its exchangeably weighted extensions where

$$Q_n(\cdot) := P_W(\cdot) = \sum_{i=1}^\infty W_i \delta_{X_i}(\cdot)$$

for (W_i) an exchangeable sequence of positive weights summing to 1. In particular, we will use the following result of ([37],Theorem 2.1),

$$\sqrt{n}(P_W - P_n) \rightsquigarrow c\mathbb{G}_{P_0} \text{ in } \ell^\infty(\mathcal{F}), a.s. \tag{3.3}$$

provided that (3.2) holds, $P_0(F^2) < \infty$, and the weights satisfy certain conditions as given in [37]. The constant c is determined by the weights. The result generalizes the result of [16] for Efron's bootstrap empirical measure. In the case of Efron's bootstrap and the Bayesian bootstrap $c = 1$. For results on the real line see [3], [27] and [33].

As mentioned at the beginning of this section, we will consider the behavior of the process $\nu^{(n)}_{\alpha,\theta}$ in the case where P_0 is continuous. We shall see that we can handle part of the weak convergence of $\nu^{(n)}_{\alpha,\theta}$ by utilizing results in [37]. This is very much in the spirit of [27], [26] and [5]. However we will also need to deal with the behavior of the process $\nu_{\alpha,\theta+n\alpha}$. This process is considerably more challenging to handle as it is not obviously related to an empirical-type measure. However, in section 4, we will exploit an important distributional identity that allows us to treat $\nu_{\alpha,\theta+n\alpha}$, as a *measurelike* sum of i.i.d. processes in the sense of [1], ([39], section 2.11.1.1) and [41]. Throughout we will assume that $\mathcal{F} \in M(H)$ and that there exist an envelope $F(x)$ satisfying $H(F^2) < \infty$.

We shall also have need of the (unconditional) multiplier central limit theorem on Banach spaces for i.i.d. random variables in the Lorentz $L_{2,1}$ space, which is found in [24]. More details may be found in [39] (see also [15]and [14]). A random variable ξ (see [39], section 2.9) is said to be in $L_{2,1}$ if

$$\|\xi\|_{2,1} = \int_0^\infty \sqrt{P(|\xi| > x)}dx < \infty.$$

Finiteness of $\|\xi\|_{2,1}$ requires slightly more the a second moment but is implied by a $2+\epsilon$ absolute moment. In our case, the $L_{2,1}$ condition is easily satisfied as the variables we shall encounter are gamma random variables having all moments.

3.1. Continuous Case

We now address the case of weak convergence of $\nu^{(n)}_{\alpha,\theta}$ when P_0 is continuous. Here we will need to consider the process $\nu_{\alpha,\theta+n\alpha}$. Because, we will not be working strictly with empirical-type processes, but actually measure-like processes, we will restrict ourselves to the quite rich class of $\mathcal{F}$ that constitute a Vapnik–Chervonenkis graph class (VCGC) (see for instance ([41], p.239)). This avoids the need to otherwise state uniform entropy-type conditions. We now present the result below for $\nu^{(n)}_{\alpha,\theta}$. We will present a partial proof of this result and then address the behavior of $\nu_{\alpha,\theta+n\alpha}$, in the next section.

Theorem 3.1. *Let $\mathcal{F}$ be a VCGC subclass of $L_2(P_0)$ and $L_2(H)$ with envelope F such that $P_0(F^2) < \infty$ and $H(F^2) < \infty$. For $0 \le \alpha < 1$, and $\theta > -\alpha$, let $\nu^{(n)}_{\alpha,\theta}(\cdot)$ denote the posterior, $\Pi^{(n)}_{\alpha,\theta}$, process centered at its mean and scaled by $\sqrt{n}$. Then when P_0 is continuous, conditional on the sequence $X_1, X_2, \ldots,$*

$$\nu^{(n)}_{\alpha,\theta} \rightsquigarrow \sqrt{1-\alpha}\mathbb{G}_{P_0} + \sqrt{\alpha(1-\alpha)}\mathbb{G}_H + \sqrt{\alpha}\tilde{N}(P_0 - H) \text{ in } \ell^\infty(\mathcal{F}) \text{ a.s. .}$$

Where $\mathbb{G}_{P_0}$ and $\mathbb{G}_H$ are independent Gaussian processes, independent of $\tilde{N}$ which is a standard Normal random variable.

Proof. The process $\nu^{(n)}_{\alpha,\theta}$ is equivalent to

$$\sqrt{n}(1-R_n)(D_n-\tilde{F}_n)+\sqrt{n}(R_n-\frac{\theta+n\alpha}{\theta+n})(H-\tilde{F}_n)+R_n\nu_{\alpha,\theta+n\alpha}.$$

Now since P_0 is continuous it follows that $\tilde{F}_n$ equates with the empirical measure $P_n(\cdot)=n^{-1}\sum_{i=1}^{n}\delta_{X_i}(\cdot)$, and $D_n=P_W$ where P_W has weights

$$W_i=\frac{G_{1-\alpha,i}}{\sum_{l=1}^{n}G_{1-\alpha,l}}$$

where $G_{1-\alpha,i}$ are i.i.d. gamma$(1-\alpha,1)$ random variables. Furthermore, R_n is beta$(\theta+n\alpha,n(1-\alpha))$ and hence converges in probability to α as $n\to\infty$. So the first term is asymptotically equivalent to the process

$$\sqrt{n}(1-\alpha)(P_W-P_n)$$

and it follows that the result of ([37], Theorem 2.1) applies. Hence the process, without the $(1-\alpha)$ term, satisfies (3.3) with $c=1/\sqrt{1-\alpha}$. It is easy to see that R_n centered by its mean and scaled by $\sqrt{n}$ converges to a Normal distribution, hence the second term converges in distribution to $\sqrt{\alpha}\tilde{N}(H-P_0)$. Finally, the limit of $\nu_{\alpha,\theta+n\alpha}$ will be verified in the next section. □

4. Asymptotic Behavior of a Poisson-Dirichlet $(\alpha,\theta+n\alpha)$ Process

In this last section we establish the weak convergence of the process $\nu_{\alpha,\theta+n\alpha}$. Since $P_{\alpha,\theta+n\alpha}$ is closely associated with various properties of Brownian and Bessel processes we expect that this result will be of interest in those settings. The establishment of this result requires a few non-trivial maneuvers as $P_{\alpha,\theta+n\alpha}$ does not appear to have any similarities to an empirical process. We first establish an important distributional identity

Proposition 4.1. *Let $P_{\alpha,\theta+n\alpha}$ denote a random probability measure with law $\Pi_{\alpha,\theta+n\alpha}$, then*

$$P_{\alpha,\theta+n\alpha}(\cdot)\overset{d}{=}\frac{G_\theta}{G_{\theta+n\alpha}}P_{\alpha,\theta}(\cdot)+\sum_{i=1}^{n}\frac{G_{\alpha,i}}{G_{\theta+n\alpha}}P^{(i)}_{\alpha,\alpha}(\cdot)$$

where $P^{(i)}_{\alpha,\alpha}$ are i.i.d. $\Pi_{\alpha,\alpha}$ random probability measures independent of $G_\theta,(G_{\alpha,i})$, where $G_{\alpha,i}$ are i.i.d. gamma$(\alpha,1)$ random variables, G_θ is gamma$(\theta,1)$, independent of $(G_{\alpha,i})$ and $G_{\theta+n\alpha}=G_\theta+\sum_{i=1}^{n}G_{\alpha,i}$.

Proof. It is enough to establish this result for $P_{\alpha,\theta+n\alpha}(g)$, for an arbitrary bounded measurable function g. The result follows by noting that for any $\theta>0$,

$$E[\mathrm{e}^{-\lambda G_\theta P_{\alpha,\theta}(g)}]=\left[\int_I(1+\lambda g(x))^\alpha H(dx)\right]^{-\theta/\alpha}$$

which is equivalent to the Cauchy–Stieltjes transform of order θ of $P_{\alpha,\theta}(g)$, whose form was obtained by [40]. It follows that,

$$E[\mathrm{e}^{-\lambda G_{\theta+n\alpha}P_{\alpha,\theta+n\alpha}(g)}]=\left[\int_I(1+\lambda g(x))^\alpha H(dx)\right]^{-\theta/\alpha-n}$$

which is the same as

$$E[\mathrm{e}^{-\lambda G_\theta P_{\alpha,\theta}(g)}]\prod_{i=1}^{n} E[\mathrm{e}^{-\lambda G_{\alpha,i} P^{(i)}_{\alpha,\alpha}(g)}].$$

Thus we can conclude that

$$G_{\theta+n\alpha}P_{\alpha,\theta+n\alpha} \overset{d}{=} G_\theta P_{\alpha,\theta} + \sum_{i=1}^{n} G_{\alpha,i}P^{(i)}_{\alpha,\alpha}(\cdot).$$

Now, using the fact that $G_{\theta+n\alpha} = G_\theta + \sum_{i=1}^{n} G_{\alpha,i}$ is gamma$(\theta+n\alpha,1)$, it follows using the calculus of beta and gamma random variables that

$$G_{\theta+n\alpha}P_{\alpha,\theta+n\alpha}(\cdot) \overset{d}{=} G_{\theta+n\alpha}\left[\frac{G_\theta}{G_{\theta+n\alpha}}P_{\alpha,\theta}(\cdot) + \sum_{i=1}^{n}\frac{G_{\alpha,i}}{G_{\theta+n\alpha}}P^{(i)}_{\alpha,\alpha}(\cdot)\right]$$

where on the right hand side, $G_{\theta+n\alpha}$ is independent of the term in brackets. Now we use the fact that gamma random variables are simplifiable to conclude the result. See Chaumont and Yor ([6], sec. 1.12 and 1.13) for details on simplifiable random variables. □

Proposition 4.1 now allows us to write

$$\nu_{\alpha,\theta+n\alpha} = \sqrt{n}\frac{G_\theta}{G_{\theta+n\alpha}}P_{\alpha,\theta}(\cdot) + \sqrt{n}\frac{\sum_{i=1}^{n} G_{\alpha,i}}{G_{\theta+n\alpha}}\sum_{i=1}^{n}\frac{G_{\alpha,i}}{\sum_{i=1}^{n} G_{\alpha,i}}P^{(i)}_{\alpha,\alpha}(\cdot) - \sqrt{n}H(\cdot).$$

Additionally one has for $\theta > 0$, the covariance formula

$$cov(\frac{G_\theta}{\theta}(P_{\alpha,\theta}-H)(f), \frac{G_\theta}{\theta}(P_{\alpha,\theta}-H)(g)) = \frac{1-\alpha}{\theta}[H(fg)-H(f)H(g)],$$

which follows from Lemma 2.1. Using these points we obtain the next result.

Theorem 4.1. *Let $\mathcal{F}$ be a VCGC subclass of $L_2(H)$ with envelope F such that $H(F^2) < \infty$. Let, for $0 < \alpha < 1$ and $\theta > -\alpha$, $P_{\alpha,\theta+n\alpha}$ denote the random probability measure with Poisson-Dirichlet law $\Pi_{\alpha,\theta+n\alpha}$, and define $\nu_{\alpha,\theta+n\alpha}(\cdot) := \sqrt{n}(P_{\alpha,\theta+n\alpha} - H)(\cdot)$. Then as $n \to \infty$,*

$$\nu_{\alpha,\theta+n\alpha} \rightsquigarrow \frac{\sqrt{1-\alpha}}{\sqrt{\alpha}}\mathbb{G}_H \text{ in } \ell^\infty(\mathcal{F}).$$

Proof. The key to this result is of course Proposition 4.1, which allows us to express $\nu_{\alpha,\theta+n\alpha}$, in terms of i.i.d. processes $P^{(i)}_{\alpha,\alpha}$ having mean H and variance for each A, as

$$\frac{1-\alpha}{1+\alpha}H(A)[1-H(A)].$$

In particular, it follows that the asymptotic distributional behavior of $\nu_{\alpha,\theta+n\alpha}$ is equivalent to that of

$$\sqrt{n}\sum_{i=1}^{n}\frac{G_{\alpha,i}}{\sum_{j=1}^{n} G_{\alpha,i}}(P^{(i)}_{\alpha,\alpha} - H)(\cdot).$$

Appealing, again, to the law of large numbers we may instead use the process

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{G_{\alpha,i}}{\alpha}(P^{(i)}_{\alpha,\alpha} - H)(\cdot)$$

which decomposes into the sum of two asymptotically independent pieces,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\frac{G_{\alpha,i}}{\alpha}-1)(P^{(i)}_{\alpha,\alpha}-H)(\cdot)+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(P^{(i)}_{\alpha,\alpha}-H)(\cdot).$$

The second term is a measure-like process in the sense of [1] and ([39], section 2.11.1.1). However since we have chosen the class to be VCGC, convergence of this process follows by using arguments similar to that in ([41], section 7.8). The first term then converges by the multiplier CLT in Banach spaces ([24] or ([23], Proposition 10.4)). □

4.1. Dirichlet Process Asymptotics for $\theta \to \infty$

The next result describes weak convergence of a centered Dirichlet process as $\theta \to \infty$.

Theorem 4.2. *Let $\mathcal{F}$ be a VCGC subclass of $L_2(H)$ with envelope F such that $H(F^2) < \infty$. Let, for $\theta > 0$, $P_{0,\theta}$ denote a Dirichlet Process with law $\Pi_{0,\theta}$, having mean H, and define $\tau_\theta(\cdot) := \sqrt{\theta}(P_{0,\theta} - H)(\cdot)$. Assume without loss of generality that $\theta = n\kappa$, for κ a fixed positive number. Then as $n \to \infty$,*

$$\tau_\theta \rightsquigarrow \mathbb{G}_H \text{ in } \ell^\infty(\mathcal{F}),$$

where $\mathbb{G}_H$ is a H-Brownian bridge. Furthermore the limit does not depend on κ.

Proof. By an argument similar to Proposition 4.1, one can write

$$G_\theta P_{0,\theta} \overset{d}{=} \sum_{i=1}^{n} G_{\kappa,i}P^{(i)}_{0,\kappa}.$$

Hence τ_θ is asymptotically equivalent to

$$\frac{\sqrt{\kappa}}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{G_{\kappa,i}}{\kappa}-1\right)(P^{(i)}_{0,\kappa}-H)(\cdot)+\frac{\sqrt{\kappa}}{\sqrt{n}}\sum_{i=1}^{n}(P^{(i)}_{0,\kappa}-H)(\cdot).$$

The result then follows analogous to Theorem 4.1. □

Acknowledgments

This research was supported in part by the grant HIA05/06.BM03 of the HKSAR.

References

[1] Alexander, K. S. (1987). Central limit theorems for stochastic processes under random entropy conditions. *Probab. Theory Related Fields* **75**, 351–378.

[2] Barron, A., Schervish, M. J. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27**, 536–561.

[3] Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9**, 1196–1217.

[4] Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1** 353–355.

[5] Brunner, L. J. and Lo, A. Y. (1996). Limiting posterior distributions under mixture of conjugate priors. *Statist. Sinica* **6**, 187–197.

[6] Chaumont, L. and Yor, M. (2003). *Exercises in probability. A guided tour from measure theory to random processes, via conditioning.* Cambridge Series in Statistical and Probabilistic Mathematics, 13, Cambridge University Press.

[7] Dawson, D. A. and Feng, S. (2006). Asymptotic behavior of the Poisson-Dirichlet distribution for large mutation rate. *Ann. Appl. Probab.* **16**, 562–582.

[8] Feng, S (2007). Large deviations for Dirichlet processes and Poisson-Dirichlet distribution with two parameters. *Electron. J. Probab.* **12**, 787–807.

[9] Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.

[10] Freedman, D.A. and Diaconis, P. (1983). On inconsistent Bayes estimates in the discrete case. *Ann. Statist.* **11**, 1109–1118.

[11] Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27**, 143–158.

[12] Ghosal, S., Ghosh, J. K. and van der Vaart, A.W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28**, 500–531.

[13] Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian nonparametrics.* Springer Series in Statistics. Springer-Verlag, New York.

[14] Giné, E. (1997). *Lectures on some aspects of the bootstrap.* Lectures on probability theory and statistics (Saint-Flour, 1996), 37–151, Lecture Notes in Math., 1665, Springer, Berlin.

[15] Giné, E. and Zinn, J.(1984). Some limit theorems for empirical processes. With discussion. *Ann. Probab.* **12** 929–998.

[16] Giné, E. and Zinn, J. (1990). Bootstrapping general empirical measures. *Ann. Probab.* **18**, 851–869.

[17] Ishwaran, H. and James. L.F. (2003). Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statist. Sinica* **13** 1211–1235.

[18] Ishwaran, H. and James. L.F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96**, 161–173.

[19] James, L.F., Lijoi, A. and Prünster, I.(2006). Distributions of linear functionals of two-parameter Poisson-Dirichlet random measures. http://arxiv.org/abs/math.PR/0609488. To appear *Ann. Appl. Probab.*

[20] Jang, J., Lee, J. and Lee, S. (2007). Posterior consistency of species sampling models. Preprint.

[21] Joyce, P., Krone, S. M. and Kurtz, T. G. (2002). Gaussian limits associated with the Poisson–Dirichlet distribution and the Ewens sampling formula. *Ann. Appl. Probab.* **12** 101–124.

[22] Kim, Y. and Lee, J. (2004). A Bernstein-von Mises theorem in the nonparametric right-censoring model. *Ann. Statist.* **32**, 1492–1512.

[23] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces. Isoperimetry and processes.* Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)], 23. Springer-Verlag, Berlin.

[24] Ledoux, M. and Talagrand, M. (1986). Conditions d'intégrabilité pour les multiplicateurs dans le TLC banachique. Ann. Probab. **14** (1986), 916–921.

[25] Lijoi, A., Prünster, I. and Walker, S. G. (2005). On consistency of nonparametric normal mixtures for Bayesian density estimation. *J. Amer. Statist. Assoc.* **100**, 1292–1296.

[26] Lo, A. Y. (1993). A Bayesian bootstrap for censored data. *Ann. Statist.* **21**,

100–123.

[27] Lo, A. Y. (1987). A large sample study of the Bayesian bootstrap. *Ann. Statist.* **15**, 360–375.

[28] Lo, A. Y. (1986). A remark on the limiting posterior distribution of the multiparameter Dirichlet process. *Sankhya Ser. A* **48** 247–249.

[29] Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12**, 351–357.

[30] Lo, A. Y. (1983). Weak convergence for Dirichlet processes. *Sankhya Ser. A* **45**, 105–111.

[31] Lo, A. Y. (1982). Bayesian nonparametric statistical inference for Poisson point processes. *Z. Wahrsch. Verw. Gebiete* **59**, 55–66.

[32] Lynch, J. and Sethuraman, J. (1987). Large deviations for processes with independent increments. *Ann. Probab.* **15**, 610–627.

[33] Mason, D. M. and Newton, M. A. (1992).A rank statistics approach to the consistency of a general bootstrap. *Ann. Statist.* **20** 1611–1624.

[34] Pitman, J. (2006). *Combinatorial stochastic processes.* Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002. With a foreword by Jean Picard. Lecture Notes in Mathematics, 1875. Springer-Verlag, Berlin.

[35] Pitman, J. (1996). *Some developments of the Blackwell-MacQueen urn scheme.* Statistics, probability and game theory, 245–267, IMS Lecture Notes Monogr. Ser., 30, Inst. Math. Statist., Hayward, CA.

[36] Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855–900.

[37] Praestgaard, J. and Wellner, J.A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* **21**, 2053–2086.

[38] van der Vaart, A. W. (1998). *Asymptotic statistics.* Cambridge Series in Statistical and Probabilistic Mathematics, 3. Cambridge University Press, Cambridge.

[39] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes. With applications to statistics.* Springer Series in Statistics. Springer-Verlag, New York.

[40] Vershik, A., Yor, M. and Tsilevich, N. (2004). On the Markov–Krein identity and quasi–invariance of the gamma process. *J. Math. Sci.* **121**, 2303–2310.

[41] Ziegler, K. (1997). Functional central limit theorems for triangular arrays of function-indexed processes under uniformly integrable entropy conditions. *J. Multivariate Anal.* **62**, 233–272.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (2007) 201
© Institute of Mathematical Statistics, 2007
arXiv: math.PR/0000000

Reproducing Kernel Hilbert Spaces of Gaussian priors

A. W. van der Vaart and J. H. van Zanten

Vrije Universiteit Amsterdam

Abstract: We review definitions and properties of reproducing kernel Hilbert spaces attached to Gaussian variables and processes, with a view to applications in nonparametric Bayesian statistics using Gaussian priors. The rate of contraction of posterior distributions based on Gaussian priors can be described through a concentration function that is expressed in the reproducing Hilbert space. Absolute continuity of Gaussian measures and concentration inequalities play an important role in understanding and deriving this result. Series expansions of Gaussian variables and transformations of their reproducing kernel Hilbert spaces under linear maps are useful tools to compute the concentration function.

Contents

1. Introduction

Ghosal, Ghosh and van der Vaart considered in [4] the rate of contraction of a posterior distribution based on i.i.d. observations to the true density. Given prior probability measures Π_n defined on a set $\mathcal{P}$ of densities p relative to a given σ-finite measure on a measurable space (such that the maps $(x,p) \mapsto p(x)$ are jointly

Department of Mathematics, Vrije Universiteit, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands; e-mail: aad@cs.vu.nl; harry@cs.vu.nl

AMS 2000 subject classifications: Primary 60G15, 62G05; secondary
Keywords and phrases: rate of convergence, Bayesian inference

measurable) and observations $X_1, \dots, X_n$, they characterized the rate $\varepsilon_n \downarrow 0$ at which the posterior distribution

$$\Pi_n(B|\, X_1, \dots, X_n) = \frac{\int_B \prod_{i=1}^n p(X_i)\, d\Pi_n(p)}{\int \prod_{i=1}^n p(X_i)\, d\Pi_n(p)} \tag{1.1}$$

contracts to p_0 if the observations are an i.i.d. sample from this density, i.e. the rate for which

$$\mathrm{E}_{p_0} \Pi_n\bigl(p{:}\, d(p, p_0) > M\varepsilon_n |\, X_1, \dots, X_n\bigr) \to 0,$$

for sufficiently large M. In their results d can be the Hellinger distance, the L_1-distance, or the L_2-distance if the densities are uniformly bounded above.

The paper [15] applied these results to priors Π_n constructed from Gaussian processes. They consider a prior Π_n constructed as the distribution of p_W, for W a Gaussian random element in a Banach space $\bigl(\mathbb{B}, \|\cdot\|\bigr)$ and $w \mapsto p_w$ a map such that, for some constant C and all $v, w \in \mathbb{B}$ with $\|v - w\|$ bounded above by some fixed constant,

$$\begin{aligned} d(p_v, p_w) &\leq C\|v - w\|, \\ K(p_v, p_w) &\leq C\|v - w\|^2, \\ V(p_v, p_w) &\leq C\|v - w\|^2. \end{aligned}$$

Here $K(p, q) = \int \log(p/q)\, p\, d\mu$ is the Kullback-Leibler divergence and $V(p, q) = \int \bigl(\log(p/q)\bigr)^2\, p\, d\mu$. This setting covers, for instance, the case of density estimation on $[0, 1]$ as considered in Tokdar and Ghosh [14], with d the Hellinger distance, the Banach space equal to $\mathbb{B} = C[0, 1]$ and

$$p_w(x) = \frac{e^{w_x}}{\int_0^1 e^{w_y}\, dy}.$$

It also covers logistic or probit regression as considered in [5] with appropriate choices and several other situations, as shown in [15].

In the latter paper it is shown that if the true density takes the form $p_0 = p_{w_0}$, then the rate of posterior contraction ε_n is characterized by the pair of equations

$$\inf_{h \in \mathbb{H}: \|h - w_0\| < \varepsilon_n} \|h\|_{\mathbb{H}}^2 \leq n\varepsilon_n^2, \tag{1.2}$$

$$-\log \mathrm{P}\bigl(\|W\| < \varepsilon_n\bigr) \leq n\varepsilon_n^2. \tag{1.3}$$

Here $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ is the *reproducing kernel Hilbert space* (RKHS) of the Gaussian variable, and $\mathrm{P}\bigl(\|W\| < \varepsilon\bigr)$ is its *small ball probability* (cf. [11]). Both equations have a minimal solution ε_n, and the rate is the worse of the two solutions. The second depends only on the prior, and gives a maximal rate regardless of the true parameter w_0, whereas the first involves the true parameter.

The reproducing kernel Hilbert space arises because it determines the support and the 'geometry' of the concentration of the Gaussian measure, which are crucial for its success as a prior. Results on RKHSs of Gaussian variables are spread over many research papers, and sometimes seem to belong to what is 'well known' without clear references. Moreover, there are different definitions for stochastic processes and Borel measurable maps in a separable Banach space. In this paper we review definitions, investigate when the different definitions agree, and derive results that are useful for the construction of priors and the study of posterior distributions.

2. Definitions and Elementary Properties

In this section we give and compare two definitions of RKHS, one for stochastic processes and one for Borel measurable maps in a Banach space.

2.1. Gaussian Processes

A zero-mean Gaussian stochastic process $W = (W_t\colon t \in T)$ is a set of random variables W_t indexed by an arbitrary set T and defined on a common probability space $(\Omega, \mathcal{U}, \mathrm{P})$ such that each finite subset possesses a zero-mean multivariate normal distribution. The finite-dimensional distributions of such a process are determined by the covariance function $K\colon T \times T \to \mathcal{R}$, defined by

$$K(s,t) = \mathrm{E} W_s W_t.$$

The *reproducing kernel Hilbert space (RKHS)* attached to the Gaussian process W is the completion $\mathbb{H}$ of the linear space of all functions

$$t \mapsto \sum_{i=1}^k \alpha_i K(s_i, t), \qquad \alpha_1, \ldots, \alpha_k \in \mathcal{R}, s_1, \ldots, s_k \in T, k \in \mathbb{N}, \tag{2.1}$$

relative to the norm induced by the inner product

$$\Bigl\langle \sum_{i=1}^k \alpha_i K(s_i, \cdot), \sum_{j=1}^l \beta_j K(t_j, \cdot) \Bigr\rangle_{\mathbb{H}} = \sum_{i=1}^k \sum_{j=1}^l \alpha_i \beta_j K(s_i, t_j). \tag{2.2}$$

It can be checked that this definition is independent of the representation of the functions on the left, and that this defines a valid inner product.

The completion of the collection of functions (2.1) is an abstract metric-topological operation using the metric induced by the inner product (2.2) only. As such the completion is not a space of functions $f\colon T \to \mathcal{R}$. However, it can be identified with a space of functions $f\colon T \mapsto \mathcal{R}$, through the *reproducing formula*

$$f(t) = \langle f, K(t, \cdot) \rangle_{\mathbb{H}}.$$

For f a linear combination of the form $\sum_{i=1}^k \alpha_i K(s_i, \cdot)$ this formula follows from the definition (2.2) of the inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}}$. For general $f \in \mathbb{H}$ the (extended) inner product on the right (with the extended function $K(t, \cdot)$) is well defined through the completion operation, and can be used to define a function $f\colon T \mapsto \mathcal{R}$.

Alternatively, the function in (2.1) can be written as

$$t \mapsto \mathrm{E} W_t H, \qquad H = \sum_i \alpha_i W_{s_i}. \tag{2.3}$$

With the function in the display written as $\mathrm{E} W.H$, the inner product (2.2) is equal to

$$\bigl\langle \mathrm{E} W.H_1, \mathrm{E} W.H_2 \bigr\rangle_{\mathbb{H}} = \mathrm{E} H_1 H_2.$$

Thus the map $H \mapsto \mathrm{E} W.H$ is an isometry for the norm of the L_2-space attached to the probability space $(\Omega, \mathcal{U}, \mathrm{P})$ on which the process W is defined and the RKHS-norm. The stochastic process RKHS $\mathbb{H}$, which is defined as the completion of the set of functions (2.3), is therefore precisely the set of functions $t \mapsto \mathrm{E} W_t H$ with H ranging over the closure of the set of linear combinations $H = \sum_i \alpha_i W_{s_i}$ in $L_2(\Omega, \mathcal{U}, \mathrm{P})$ (known as the *first order chaos of* W). It follows again that we can view $\mathbb{H}$ as a Hilbert space of functions on T.

2.2. Gaussian Elements in a Banach Space

A Borel measurable random element W with values in a separable Banach space $(\mathbb{B}, \|\cdot\|)$ is called *Gaussian* if the random variable b^*W is normally distributed for any element b^* of the dual space $\mathbb{B}^*$ of $\mathbb{B}$, and it is called *zero-mean* if the mean of every such variable b^*W is zero. Henceforth we shall only consider zero-mean Gaussian variables.

It is well known that the norm $\|W\|$ of a zero-mean Gaussian variable, which is a finite random variable by the assumption that W takes its values in $\mathbb{B}$, has sub-Gaussian tails. (cf. Corollary 5.1 below, or, e.g., ([17], Propositions A.2.1 and A.2.3) for a direct proof.) In particular, all moments $\mathrm{E}\|W\|^p$ are finite. We set

$$\sigma^2(W) = \sup_{b^*\in\mathbb{B}^*:\|b^*\|=1} \mathrm{E}b^*(W)^2.$$

This is a finite number, bounded by $\mathrm{E}\|W\|^2$.

For every element $b^* \in \mathbb{B}^*$ we define $Sb^* \in \mathbb{B}$ as the Pettis integral $\mathrm{E}Wb^*(W)$ of the $\mathbb{B}$-valued random element $Wb^*(W)$. By definition, this *Pettis integral* is an element Sb^* of $\mathbb{B}$ such that $b_2^*(Sb^*) = \mathrm{E}b_2^*(W)b^*(W)$ for every $b_2^* \in \mathbb{B}^*$. The following lemma allows us to derive the existence of the Pettis integral from the fact that $\mathrm{E}\|W\|^2 < \infty$.

Lemma 2.1. *If X is a Borel measurable map in a separable Banach space $\mathbb{B}$ with $\mathrm{E}\|X\| < \infty$, then there exists an element $b \in \mathbb{B}$ such that $b^*(b) = \mathrm{E}b^*(X)$ for every $b^* \in \mathbb{B}^*$.*

Proof. Because the Banach space is assumed separable, the map X is automatically tight (e.g. [17], 1.3.2). Therefore, for any $n \in \mathbb{N}$ there exists a compact set K such that $\mathrm{E}\|X\|1_{X\notin K} < 1/n$. This compact set can be partitioned into finitely many sets B_i of diameter smaller than $1/n$. Without loss of generality these partitions can be chosen as successive refinements for increasing n. Let $X_n = \sum_i b_i 1_{X\in B_i}$ for b_i arbitrary points in the partitioning sets. Then $\mathrm{E}X_n := \sum_i b_i \mathrm{P}(X \in B_i)$ satisfies $b^*(\mathrm{E}X_n) = \mathrm{E}b^*(X_n)$ for every $b^* \in \mathbb{B}^*$. Furthermore, the sequence $\mathrm{E}X_n$ is a Cauchy sequence in $\mathbb{B}$, because $\|\mathrm{E}X_n - \mathrm{E}X_m\| = \sup_{\|b^*\|=1} |\mathrm{E}b^*(X_n - X_m)| \leq \mathrm{E}\|X_n - X_m\| \to 0$ as $n, m \to \infty$. Because $\mathrm{E}\|X_n - X\| < 2/n$, we have that $b^*(\mathrm{E}X_n) = \mathrm{E}b^*(X_n) \to \mathrm{E}b^*(X)$ for every $b^* \in \mathbb{B}$. The strong limit b of the sequence $\mathrm{E}X_n$ is of course also a weak limit, whence $b^*(b) = \mathrm{E}b^*(X)$ for every $b^* \in \mathbb{B}^*$. ∎

The *reproducing kernel Hilbert space (RKHS)* $\mathbb{H}$ attached to W is the completion of the range $S\mathbb{B}^*$ of the map $S: \mathbb{B}^* \to \mathbb{B}$ defined by $Sb^* = \mathrm{E}Wb^*(W)$ for the inner product

$$\langle Sb_1^*, Sb_2^*\rangle_{\mathbb{H}} = \mathrm{E}b_1^*(W)b_2^*(W).$$

By the Hahn–Banach theorem and the Cauchy–Schwarz inequality,

$$\begin{aligned}\|Sb^*\| &= \sup_{b_2^*\in\mathbb{B}^*:\|b_2^*\|=1} |b_2^*(Sb^*)| = \sup_{b_2^*\in\mathbb{B}^*:\|b_2^*\|=1} |\mathrm{E}b_2^*(W)b^*(W)| \\ &\leq \sigma(W)\bigl(\mathrm{E}b^*(W)^2\bigr)^{1/2} = \sigma(W)\|Sb^*\|_{\mathbb{H}}. \qquad (2.4)\end{aligned}$$

It follows that the RKHS-norm on the set $S\mathbb{B}^*$ is stronger than the original norm, so that a $\|\cdot\|_{\mathbb{H}}$-Cauchy sequence in $S\mathbb{B}^* \subset \mathbb{B}$ is a $\|\cdot\|$-Cauchy sequence in $\mathbb{B}$. Consequently, the RKHS, which is by definition the completion of the set $S\mathbb{B}^*$

under the RKHS norm, can be identified with a subset of $\mathbb{B}$. In terms of the unit balls $\mathbb{B}_1$ and $\mathbb{H}_1$ of $\mathbb{B}$ and $\mathbb{H}$ the preceding display can be written as

$$\mathbb{H}_1 \subset \sigma(W)\mathbb{B}_1. \tag{2.5}$$

In other words, the norm of the embedding $i\colon \mathbb{H} \to \mathbb{B}$ is bounded by $\sigma(W)$.

Lemma 2.2. *The map $S\colon \mathbb{B}^* \to \mathbb{H}$ is weak-* continuous.*

Proof. The unit ball $\mathbb{B}_1^*$ of the dual space is weak-* metrizable ([12], 3.16). Therefore the restricted map $S\colon \mathbb{B}_1^* \to \mathbb{H}$ is weak-* continuous if and only if weak-* convergence of a sequence b_n^* in $\mathbb{B}_1^*$ to an element b^* implies that $Sb_n^* \to Sb^*$ in $\mathbb{H}$. Now the weak-* convergence $b_n^* \to b^*$ is by definition pointwise convergence on $\mathbb{B}$. Then the sequence $(b_n^* - b^*)(W)$ tends to zero (almost) surely, and hence also in distribution. Because each of these variables is zero-mean Gaussian, this implies that the variances tend to zero, i.e. $\|Sb_n^* - Sb^*\|_{\mathbb{H}}^2 = \mathrm{E}(b_n^* - b)^2(W) \to 0$. (Alternatively, use the uniform integrability of the variables b^*W instead of the Gaussianity.)

This concludes the proof that the restriction of S to the unit ball $\mathbb{B}_1^*$ is continuous. A weak-* converging net b_n^* in $\mathbb{B}^*$ is necessarily bounded in norm, by the Banach–Steinhaus theorem ([12], 2.5), and hence is contained in a multiple of the unit ball. The continuity of the restriction then shows that $Sb_n^* \to Sb^*$, which concludes the proof. ∎

Corollary 2.1. *If $\mathbb{B}_0^*$ is a weak-* dense subset of $\mathbb{B}^*$, then $\mathbb{H}$ is the completion of $S\mathbb{B}_0^*$.*

By the definitions $\langle Sb^*, S\underline{b}^*\rangle_{\mathbb{H}} = \mathrm{E}b^*W\underline{b}^*W = b^*(S\underline{b}^*)$, for any $b^*, \underline{b}^* \in \mathbb{B}^*$. By continuity of the inner product this extends to the *reproducing formula*:

$$\langle Sb^*, h\rangle_{\mathbb{H}} = b^*(h), \tag{2.6}$$

which is valid for every $h \in \mathbb{H}$ and $b^* \in \mathbb{B}^*$.

Just as for stochastic processes there is an alternative representation of the RKHS through 'first chaos', in the present setting defined as the closed linear span of the variables b^*W in $L_2(\Omega, \mathcal{U}, \mathrm{P})$. The elements Sb^* of the RKHS can be written $Sb^* = \mathrm{E}HW$ for $H = b^*W$, and the RKHS-norm of Sb^* is by definition the $L_2(\Omega, \mathcal{U}, \mathrm{P})$-norm of this H. This immediately implies the following lemma. Note that $\mathrm{E}HW$ is well defined as a Pettis integral for every $H \in L_2(\Omega, \mathcal{U}, \mathrm{P})$, by Lemma 2.1.

Lemma 2.3. *The RKHS is the set of Pettis integrals $\mathrm{E}HW$ for H ranging over the closed linear span of the variables b^*W in $L_2(\Omega, \mathcal{U}, \mathrm{P})$ with inner product $\langle \mathrm{E}H_1W, H_2W\rangle_{\mathbb{H}} = \mathrm{E}H_1H_2$.*

It is useful to decompose the map $S\colon \mathbb{B}^* \to \mathbb{B}$ as $S = A^*A$ for $A^*\colon L_2(\Omega, \mathcal{U}, \mathrm{P}) \to \mathbb{B}$ and $A\colon \mathbb{B}^* \to L_2(\Omega, \mathcal{U}, \mathrm{P})$ given by

$$\begin{aligned} A^*H &= \mathrm{E}HW, \\ Ab^* &= b^*W. \end{aligned}$$

It may be checked that the operators A and A^* are indeed adjoints, after identifying $\mathbb{B}$ with a subset of its second dual space $\mathbb{B}^*$ under the canonical embedding ([12], 3.15, 4.5), as the notation suggests. By the preceding lemma the RKHS is the image of the first chaos space under A^*. Because $R(A)^\perp = N(A^*)$ the full range $R(A^*) = A^*\bigl(L_2(\Omega, \mathcal{U}, \mathrm{P})\bigr)$ is not bigger than the image of the first chaos, although the map $A^*: L_2(\Omega, \mathcal{U}, \mathrm{P}) \to \mathbb{H}$ is an isometry only if restricted to the first chaos space.

Recall that an operator is compact if it maps bounded sets into precompact sets, or, equivalently, maps bounded sequences into sequences that possess a converging subsequence.

Lemma 2.4. *The maps $A^*: L_2(\Omega, \mathcal{U}, \mathrm{P}) \to \mathbb{B}$ and $A: \mathbb{B}^* \to L_2(\Omega, \mathcal{U}, \mathrm{P})$ and $S: \mathbb{B}^* \to \mathbb{B}$ are compact for the norms.*

Proof. In general an operator is compact if and only if its adjoint is compact, and a composition with a compact operator is compact (see ([12], 4.19). To prove the compactness of A fix some sequence b_n^* in the unit ball $\mathbb{B}_1^*$. As the unit ball is weak-* compact by the Banach–Alaoglu theorem ([12], 4.3(c)), there exists a subsequence along which $b_{n_j}^*$ converges pointwise on $\mathbb{B}$ to a limit b^*. Consequently $b_{n_j}^*(W) \to b^*(W)$ almost surely, and hence in second mean. ∎

As a consequence we can conclude that the unit ball of the RKHS is precompact in $\mathbb{B}$. Indeed, $\mathbb{H}_1 = A^*\mathbb{U}_1$ for $\mathbb{U}_1$ the unit ball of $L_2(\Omega, \mathcal{U}, \mathrm{P})$, and hence is precompact by the compactness of A^*.

Example 2.1 (Hilbert space). *The* covariance operator *of a mean zero Gaussian random element W in a Hilbert space $\mathbb{B}$ with inner product $\langle \cdot, \cdot \rangle$ is the map $S: \mathbb{B} \to \mathbb{B}$ that satisfies $\mathrm{E}\langle W, b_1\rangle\langle W, b_2\rangle = \langle b_1, Sb_2\rangle$. It is well known that S is continuous, linear, positive, self-adjoint, and of finite trace, and hence it possesses a square root, which is another positive, self-adjoint operator $S^{1/2}: \mathbb{B} \to \mathbb{B}$ such that $S^{1/2}S^{1/2} = S$. (The square root can also be described as having the same eigenfunctions as S with eigenvalues the square roots of the eigenvalues of S.) The RKHS of W can be characterized as the range of $S^{1/2}$ equipped with the norm $\|S^{1/2}b\|_{\mathbb{H}} = \|b\|$.*

To see this note that the covariance operator S is exactly the operator S as defined previously, after the usual identification of the dual space $\mathbb{B}^*$ with $\mathbb{B}$ itself: $b \in \mathbb{B}$ corresponds to the element $b_1 \mapsto \langle b, b_1\rangle$ of $\mathbb{B}^*$. Hence the RKHS is the completion of the elements Sb under the square norm $\|Sb\|_{\mathbb{H}}^2 = \mathrm{E}\langle W, b\rangle^2 = \langle b, Sb\rangle = \|S^{1/2}b\|^2$. This is the same as the completion of the set of functions $S^{1/2}c$ (with $c = S^{1/2}b$) under the norm $\|S^{1/2}c\|_V^2 = \|c\|^2$. The latter set is of course already complete, so that completion is superfluous.

2.3. Comparison

If the sample paths $t \mapsto W_t$ of a stochastic process $W = (W_t: t \in T)$ belong to a Banach space of functions, then the process can be viewed as a map W into the Banach space. If it is a Borel measurable map, then the preceding gives two definitions of a RKHS. The two definitions will coincide provided the dual space can be appropriately related to the covariance function. In particular, if the coordinate projections $\pi_t: \mathbb{B} \to \mathcal{R}$, defined by $b \mapsto b(t)$, are elements of the dual space, then $W_t = \pi_t(W)$ and the covariance function $K(s,t) = \mathrm{E}W_sW_t$ takes the form $\mathrm{E}\pi_s(W)\pi_t(W) = \langle S\pi_s, S\pi_t\rangle_{\mathbb{H}}$. If the other elements Sb^* are determined by the elements $S\pi_t$, then the two definitions should be the same. It appears that in general some conditions are needed to make the link between the two definitions. For the Banach space $\ell^\infty(T)$ of uniformly bounded functions $z: T \to \mathcal{R}$ equipped with the uniform norm $\|z\| = \sup\{|z(t)|: t \in T\}$, this can always be done.

The following result is probably known to the experts, but we do not know a published reference.

Theorem 2.1. *If W is a Borel measurable zero-mean Gaussian random element in a complete separable subspace of $\ell^\infty(T)$ equipped with the uniform norm, then the Banach space RKHS and the stochastic process RKHS coincide. Furthermore $S\pi_t = K(t,\cdot)$.*

Proof. For a given tight Borel measurable random element W in $\ell^\infty(T)$ there exists a semimetric ρ on T under which T is totally bounded and such that W takes its values in the subspace $UC(T,\rho)$ of functions $f\colon T \to \mathcal{R}$ that are uniformly continuous relative to ρ (e.g. [17], Lemma 1.5.9). Thus we may assume without loss of generality that W takes its values in $UC(T,\rho)$ for such a semimetric ρ. The space $UC(T,\rho)$ is a Banach space under the supremum norm $\|f\| = \sup\{|f(t)|\colon t \in T\}$. Let $K(s,t) = \mathrm{E}W_sW_t$.

The coordinate projections $\pi_t\colon f \mapsto f(t)$ belong to the dual space $UC(T,\rho)^*$. The corresponding Pettis integral $S\pi_t$ is the function $K(t,\cdot)$. This follows because it is contained in $UC(T,\rho)$ and, furthermore, for every $s \in T$,

$$\pi_s\bigl(K(t,\cdot)\bigr) = K(t,s) = \mathrm{E}W_sW_t = \mathrm{E}\pi_s(W)\pi_t(W).$$

Because the coordinate projections $\pi_t f$ identify f uniquely it follows that $K(t,\cdot) = \mathrm{E}W\pi_t(W) = S\pi_t$.

Thus the stochastic process RKHS, defined as the completion of the linear combinations (2.1), is contained in the Banach space RKHS. The inner products on the two spaces agree, because

$$\langle S\pi_s, S\pi_t\rangle_{\mathbb{H}} = \mathrm{E}\pi_t(W)\pi_s(W) = K(s,t) = \bigl\langle K(s,\cdot), K(t,\cdot)\bigr\rangle_{\mathbb{H}}.$$

By the Riesz representation theorem an arbitary element of $UC(T,\rho)^*$ is a map $f \mapsto \int \bar f(t)\,d\mu(t)$ for a signed Borel measure on the completion $\bar T$ of T and $\bar f\colon \bar T \to \mathcal{R}$ is the continuous extension of f. Because T is totally bounded we can write it for each $m \in \mathbb{N}$ as a finite union of sets of diameter smaller than $1/m$. If we define μ_m as the measure obtained by concentrating the masses of μ on the partitioning sets in a fixed, single point in the partitioning set, then $\int \bar f\,d\mu_m \to \int \bar f\,d\mu$ as $m \to \infty$, for each $f \in UC(T,\rho)$. The map $f \mapsto \int \bar f\,d\mu_m$ is a linear combination of coordinate projections. It follows that for any $b^* \in UC(T,\rho)^*$ there exists a sequence b_m^* of linear combinations of coordinate projections that converges pointwise on $UC(T,\rho)$ to b^*. In other words, the linear span $\mathbb{B}_0^*$ of coordinate projections is weak-* dense in $UC(T,\rho)^*$, and hence the RKHS is the completion of $S\mathbb{B}_0^*$, by Lemma 2.1. ∎

Example 2.2. *The preceding theorem applies, for instance, to the space of continuous functions $z\colon T \to \mathcal{R}$ on a compact metric space T. For instance $C[0,1]$.*

A more general connection between the two definitions of a RKHS can be made by embedding the Banach space $\mathbb{B}$ in its second dual (see [12], 4.15). This is somewhat technical and will not be needed in the rest of the paper. The canonical embedding is, as usual, the identification of $b \in \mathbb{B}$ with the map $b^{**}\colon \mathbb{B}^* \to \mathcal{R}$ defined by $b^{**}(b^*) = b^*(b)$. A Borel measurable random element W in $\mathbb{B}$ becomes identified in this way with the stochastic process $W^{**} = \bigl(b^*(W)\colon b^* \in \mathbb{B}^*\bigr)$, which has covariance function

$$K(b_1^*, b_2^*) = \mathrm{E}b_1^*(W)b_2^*(W).$$

The stochastic process RKHS $\mathbb{H}$ attached to this process in Section 2.1 is the completion of the set of functions $K(b^*,\cdot)\colon \mathbb{B}^* \to \mathcal{R}$ relative to the inner product

$$\langle K(b_1^*,\cdot), K(b_2^*,\cdot)\rangle_{\mathbb{H}} = K(b_1^*, b_2^*) = \mathrm{E}b_1^*(W)b_2^*(W).$$

The function $K(b^*,\cdot)$ is exacly the Pettis integral $\mathrm{E}Wb^*(W)$, written Sb^* in the preceding and now viewed as an element of $\mathbb{B}^{**}$; and the inner product in the display is exactly $\langle Sb_1^*, Sb_2^*\rangle_{\mathbb{H}}$. Thus the two definitions of RKHS coincide, after identification of $\mathbb{B}$ and its image in $\mathbb{B}^{**}$ under the canonical embedding.

3. Absolute Continuity

Given a zero-mean Gaussian process $W=(W_t: t\in T)$ with covariance kernel K defined on a probability space $(\Omega,\mathcal{U},\mathrm{P})$ with RKHS $\mathbb{H}$ as defined in Section 2.1, we can define a map $U:\mathbb{H}\to L_2(\Omega,\mathcal{U},\mathrm{P})$ by defining

$$UK(t,\cdot)=W_t, \tag{3.1}$$

and extending linearly and continuously. This map is an Hilbert space isometry, since

$$\mathrm{E}UK(s,\cdot)UK(t,\cdot)=\mathrm{E}W_sW_t=K(s,t)=\bigl\langle K(s,\cdot),K(t,\cdot)\bigr\rangle_{\mathbb{H}}.$$

This isometry property also implies the existence of the extension. It follows that the process $(Uh: h\in\mathbb{H})$ is the *iso-Gaussian process* indexed by $\mathbb{H}$: a mean-zero Gaussian process with covariance function $\mathrm{E}UgUh=\langle g,h\rangle_{\mathbb{H}}$.

The process W induces a distribution P^W on the product σ-field of $\mathcal{R}^T$. For a function $f: T\mapsto\mathcal{R}$ the process $(W_t+f(t): t\in T)$ induces another distribution P^{W+f} on the same space.

Lemma 3.1. *If $f\in\mathbb{H}$, then P^{W+f} and P^W are equivalent and*

$$\frac{dP^{W+f}}{dP^W}(W)=e^{Uf-\frac12\|f\|_{\mathbb{H}}^2},\qquad \text{a.s..}$$

Proof. The process W is the 'subprocess' $W^{\mathbb{G}}=(Ug: g\in\mathbb{G})$ of the iso-Gaussian process $W^{\mathbb{H}}=(Uh: h\in\mathbb{H})$ for $\mathbb{G}$ the set of functions $K(t,\cdot)$ with t ranging over T. From the general theory of Gaussian processes

$$\frac{dP^{W^{\mathbb{H}}+(\langle h,f\rangle_{\mathbb{H}}: h\in\mathbb{H})}}{dP^{W^{\mathbb{H}}}}\bigl(W^{\mathbb{H}}\bigr)=e^{Uf-\frac12\|f\|_{\mathbb{H}}^2},\qquad \text{a.s..} \tag{3.2}$$

The process $W^{\mathbb{G}}$ arises from the iso-Gaussian process by the projection $\pi_{\mathbb{G}}:\mathcal{R}^{\mathbb{H}}\to\mathcal{R}^{\mathbb{G}}$. The corresponding Radon–Nikodym derivative can be found as the conditional expectation

$$\frac{dP^{W^{\mathbb{G}}+(\langle g,f\rangle_{\mathbb{H}}: g\in\mathbb{G})}}{dP^{W^{\mathbb{G}}}}\bigl(W^{\mathbb{G}}\bigr)=\mathrm{E}\Bigl(\frac{dP^{W^{\mathbb{H}}+(\langle h,f\rangle_{\mathbb{H}}: h\in\mathbb{H})}}{dP^{W^{\mathbb{H}}}}\bigl(W^{\mathbb{H}}\bigr)\,|\,W^{\mathbb{G}}\Bigr).$$

Because $\mathrm{lin}\,(\mathbb{G})$ is dense in $\mathbb{H}$ by construction and U is continuous, the variable Uf is the $L_2(\Omega,\mathcal{U},\mathrm{P})$-limit of a sequence Ug_n with $(g_n)\subset\mathrm{lin}\,(\mathbb{G})$ and hence is measurable relative to the completion of the σ-field generated by $W^{\mathbb{G}}$. Consequently, the right side of (3.2) is $\mathcal{W}^{\mathbb{G}}$-measurable as well and hence the conditional expectation in the preceding display is unnecessary.

Finally, note that the shift $\langle g,f\rangle_{\mathbb{H}}$ is exactly the function f after the identification $g\leftrightarrow K(t,\cdot)$, by the reproducing property: $f(t)=\langle K(t,\cdot),f\rangle_{\mathbb{H}}$ for every $t\in T$. ∎

Let $\mathbb{H}$ be the abstract RKHS attached to a zero-mean, Borel measurable, Gaussian random element W in a separable Banach space $\mathbb{B}$ defined on a probability space $(\Omega,\mathcal{U},\mathrm{P})$. Let $U:\mathbb{H}\to L_2(\Omega,\mathcal{U},\mathrm{P})$ be the isometry defined by

$$U(Sb^*)=b^*(W),\qquad b^*\in\mathbb{B}^*, \tag{3.3}$$

and extending continuously. It is the same map U as in (3.1) if we make the identification $S\pi_t = K(t,\cdot)$ of Theorem 2.1; also $US = A$ for A defined in Section 2.2. As before the map U is an isometry. The preceding lemma can be translated to the present situation.

Lemma 3.2. *If $h \in \mathbb{H}$ then the distributions P^{W+h} and P^W of $W + h$ and W on $\mathbb{B}$ are equivalent and*

$$\frac{dP^{W+h}}{dP^W}(W) = e^{Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2}, \qquad \text{a.s..}$$

Proof. The process $W^{**} = \bigl(b^*(W)\colon b^* \in \mathbb{B}^*\bigr)$ arising from W through the canonical embedding generates the same σ-field on the underlying probability space as W and can be viewed as a measurable transformation of W under the map $\phi\colon \mathbb{B} \to \mathcal{R}^{\mathbb{B}^*}$ given by $\phi(b)(b^*) = b^*(b)$. The process $W + h$ is transformed in the process $W^{**} + h^{**} = \phi(W + h)$. The result therefore follows from Lemma 3.1.

The following alternative proof is given in Proposition 2.1 in [3]. The isometry property of U shows that $\mathrm{E}(Uh)^2 = \|h\|_{\mathbb{H}}^2$. Because Uh is in the closed linear span of the zero-mean Gaussian variables $USb^* = b^*W$, it is itself zero-mean Gaussian. It follows that

$$dQ = e^{Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2}\, d\mathrm{P}$$

defines a probability measure on $(\Omega, \mathcal{U})$. For any $b_1^*, b_2^* \in \mathbb{B}^*$ the joint distribution of $(USb_1^*, USb_2^*) = (b_1^*W, b_2^*W)$ is bivariate normal with mean zero and covariance matrix $\bigl(\langle Sb_i^*, Sb_j^*\rangle_{\mathbb{H}}\bigr)_{i,j=1,2}$. By taking limits we see that for every $h \in \mathbb{H}$ the joint distribution of (b_1^*W, Uh) is bivariate normal with mean zero and covariance matrix Σ with $\Sigma_{1,1} = \|Sb_1^*\|_{\mathbb{H}}^2$, $\Sigma_{1,2} = \langle Sb_1^*, h\rangle_{\mathbb{H}}$ and $\Sigma_{2,2} = \|h\|_{\mathbb{H}}^2$. Thus

$$\mathrm{E}_Q e^{ib_1^*W} = \mathrm{E}e^{ib_1^*W} e^{Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2} = e^{\frac{1}{2}(i,1)\Sigma(i,1)^T} e^{-\frac{1}{2}\|h\|_{\mathbb{H}}^2} = e^{-\frac{1}{2}\Sigma_{1,1} + i\Sigma_{1,2}}.$$

The right side is also equal to

$$\mathrm{E}e^{ib_1^*W + i\langle Sb_1^*, h\rangle_{\mathbb{H}}} = \mathrm{E}e^{ib_1^*(W+h)}.$$

The last step follows from the reproducing formula (2.6). We conclude that the distribution of $W + h$ under P is the same as the distribution of W under Q, i.e. $\mathrm{P}(W + h \in B) = \mathrm{E}_Q 1_B(W) = \mathrm{E}1_B(W)(dQ/d\mathrm{P})$. ∎

The preceding lemma requires that the shift h is contained in the RKHS. If this is not the case, then there is no density.

Lemma 3.3. *If $b \notin \mathbb{H}$ then the distributions P^{W+b} and P^W of $W + b$ and W on $\mathbb{B}$ are orthogonal.*

Proof. By Lemma 5.1 (below) the closure $\bar{\mathbb{H}}$ of $\mathbb{H}$ in $\mathbb{B}$ is the support of W. Because the affine spaces $\bar{\mathbb{H}}$ and $\bar{\mathbb{H}} + b$ are disjoint if $b \notin \bar{\mathbb{H}}$, the assertion is clear if $b \in \mathbb{B} - \bar{\mathbb{H}}$. Therefore, it is not a loss of generality to assume that $\mathbb{B}$ is the closure of $\mathbb{H}$.

Fix a sequence $\{b_n^*\} \subset \mathbb{B}^*$ whose linear span is dense (for the norm) in $\mathbb{B}^*$ and is such that the variables b_n^*W are i.i.d. standard normal variables. We prove the existence of such a sequence at the end of the proof. We claim that $\mathbb{H} = \{b \in \mathbb{B}\colon \sum_{n=1}^\infty (b_n^*b)^2 < \infty\}$. Indeed, the sequence $h_n = Sb_n^*$ is orthonormal in $\mathbb{H}$ by the definition of the inner product in $\mathbb{H}$ and $\mathrm{lin}\,(h_n) = S\,\mathrm{lin}\,(b_n^*)$ is dense in $S\mathbb{B}^*$ by construction of the sequence b_n^* and continuity of S. By the reproducing formula $b_n^*h = \langle h, h_n\rangle_{\mathbb{H}}$ for every $h \in \mathbb{H}$, whence $\sum_n (b_n^*h)^2 < \infty$. Conversely, if $\sum_n (b_n^*b)^2 <$

∞, then $h := \sum_n (b_n^* b) h_n$ is a well-defined element of $\mathbb{H}$, with $b_m^* h = b_m^* b$ for every m because $b_m^* h_n = \langle h_n, h_m \rangle_{\mathbb{H}} = \delta_{mn}$. Because the linear span of the sequence (b_n^*) is dense in $\mathbb{B}^*$ it follows that $b^* h = b^* b$ for every b^* and hence $b = h$, which is contained in $\mathbb{H}$.

The map $\phi: \mathbb{B} \to \mathcal{R}^\infty$ defined by $b \mapsto (b_n^* b)$ is well defined and measurable. It maps W onto a sequence $(Z_n) = \phi(W)$ of standard normal variables and maps $W + b$ onto the sequence $(Z_n + b_n^* b)$ of independent shifted normal variables. By Kakutani's dichotomy the latter two laws are orthogonal if $\sum_n (b_n^* b)^2 = \infty$. This implies the orthogonality of the laws of W and $W + b$.

Finally we prove the existence of (b_n^*) as claimed. Starting with an arbitrary dense sequence (b_n^*) in $\mathbb{B}^*$, we can make this linearly independent by removing from left to right in the sequence $b_1^*, b_2^*, \ldots$ every b_n^* that can be written as a linear combination of the preceding (left-over) b_j^*. This procedure yields a linearly independent sequence (b_n^*) whose span is dense in $\mathbb{B}^*$. The random variables $b_n^* W$ are automatically linearly independent in $L_2(\Omega, \mathcal{U}, \mathrm{P})$, because $\sum_n \lambda_n b_n^* W = 0$, almost surely for a sequence λ_n with finitely many nonzero elements. This implies that $\sum_n \lambda_n b_n^*$ is zero on a set with probability one under the law of W, and hence by continuity also on the support of this law, which is $\mathbb{B}$ by assumption. Thus we can apply the Gramm–Schmidt procedure to turn the sequence $b_n^* W$ into a sequence of standard normal variables (Z_n). Then $Z_n = \sum_{i=1}^n \lambda_{i,n} b_i^* W$ for every n for a triangular array of coefficients $(\lambda_{i,n})$ with $\lambda_{n,n} \neq 0$ for every n. The sequence $\sum_{i=1}^n \lambda_{i,n} b_i^*$ has the desired properties. ∎

4. Series Representation

Suppose that the covariance kernel K of the Gaussian process $W = (W_t : t \in T)$, defined on the probability space $(\Omega, \mathcal{U}, \mathrm{P})$, can be written in the form

$$K(s,t) = \sum_{j=1}^\infty \lambda_j \phi_j(s) \phi_j(t) \tag{4.1}$$

for positive numbers $\lambda_1, \lambda_2, \ldots$ and arbitrary functions $\phi_j : T \to \mathcal{R}$, where the series is assumed to converge pointwise on $T \times T$. The convergence on the diagonal implies that $\sum_j \lambda_j \phi_j^2(t) < \infty$ for all $t \in T$. Then by the Cauchy–Schwarz inequality the series $\sum_{j=1}^\infty w_j \phi_j(t)$ converges absolutely for every sequence (w_j) of numbers with $\sum_j w_j^2/\lambda_j < \infty$, for every t, and hence defines a function from T to $\mathcal{R}$. We assume that the functions ϕ_j are linearly independent in the sense that $\sum_j w_j \phi_j(t) = 0$ for every $t \in T$ for some sequence (w_j) with $\sum_j w_j^2/\lambda_j < \infty$ implying that $w_j = 0$ for every $j \in \mathbb{N}$.

Theorem 4.1. *If the covariance function K of the mean-zero Gaussian process $W = (W_t : t \in T)$ can be represented as in (4.1) for numbers λ_j and functions $\phi_j : T \to \mathcal{R}$ which satisfy $\sum_{j=1}^\infty \lambda_j \phi_j^2(t) < \infty$ for every $t \in T$ and are linearly independent as indicated, then the RKHS of the stochastic process W is the set of all functions $\sum_{j=1}^\infty w_j \phi_j(t)$ with $\sum_{j=1}^\infty w_j^2/\lambda_j < \infty$, and the inner product is given by*

$$\Big\langle \sum_{i=1}^\infty v_i \phi_i, \sum_{j=1}^\infty w_j \phi_j \Big\rangle_{\mathbb{H}} = \sum_{j=1}^\infty \frac{v_j w_j}{\lambda_j}. \tag{4.2}$$

Proof. Under the condition that $\sum_{k=1}^\infty \lambda_k\phi_k^2(t)<\infty$ for every $t\in T$, the infinite sum defining $K(s,t)$ converges for every $(s,t)\in T\times T$, by the Cauchy–Schwarz inequality, and hence the kernel is well defined. Let H be the set of all series $\sum_{k=1}^\infty f_k\phi_k$ when (f_k) ranges over the sequences with $\sum_{k=1}^\infty f_k^2/\lambda_k<\infty$. (These series were noted to converge pointwise absolutely before the statement of the theorem.) By the assumed linear independence of the functions ϕ_j, the coefficients (f_j) are identifiable from the corresponding functions $\sum_j f_j\phi_j\in H$. Therefore we can define a bijection $i\colon H\to\ell_2$ by $i\colon\sum_k f_k\phi_k\mapsto (f_k/\sqrt{\lambda_k})$. The set H becomes a Hilbert space under the inner product induced from ℓ_2, which is given on the right side of (4.2), and which we denote by $\langle\cdot,\cdot\rangle_H$. We must prove that this inner product agrees with the inner product of $\mathbb{H}$ and that H and $\mathbb{H}$ are the same as sets.

The function $K(s,\cdot)$ has a representation $\sum_{k=1}^\infty f_k\phi_k$ for $f_k=\lambda_k\phi_k(s)$, and hence is contained in H. It also follows that

$$\bigl\langle K(s,\cdot),K(t,\cdot)\bigr\rangle_H=\sum_{k=1}^\infty\frac{\lambda_k\phi_k(s)\lambda_k\phi_k(t)}{\lambda_k}=K(s,t)=\bigl\langle K(s,\cdot),K(t,\cdot)\bigr\rangle_{\mathbb{H}},$$

where the second equality follows from the series representation of K, and the third is (2.2). Thus the inner products of H and $\mathbb{H}$ agree. We conclude that H contains $\mathbb{H}$ isometrically.

The space H has the reproducing property: $\langle f,K(t,\cdot)\rangle_H=f(t)$ for every $t\in T$ and $f\in H$. This follows from

$$\bigl\langle f,K(t,\cdot)\bigr\rangle_H=\Bigl\langle\sum_k f_k\phi_k,\sum_k\lambda_k\phi_k\Bigr\rangle_H=\sum_k\frac{f_k\lambda_k\phi_k(t)}{\lambda_k}=f(t).$$

If $f\in H$ with $f\perp\mathbb{H}$, then in particular $f\perp K(t,\cdot)$ for every $t\in T$ and hence $f(t)=0$ by the reproducing formula. Thus $H=\mathbb{H}$. ∎

Series expansions of the type (4.1) are not unique, and some may be more useful than others. They may arise as an eigenvalue expansion of the operator corresponding to the covariance function. However, this is not a requirement of the proposition, which applies to arbitrary functions ϕ_j.

Example 4.1. *Suppose that (T,Θ,ν) is a measurable space and*

$$\int\!\!\int K^2(s,t)\,d\nu(s)\,d\nu(t)<\infty.$$

Then the integral operator $K\colon L_2(T,\Theta,\nu)\to L_2(T,\Theta,\nu)$ defined by

$$Kf(t)=\int f(s)\,K(s,t)\,d\nu(t)$$

is compact and positive self-adjoint. Thus there exists a sequence of eigenvalues $\lambda_k\downarrow 0$ and an orthonormal system of eigenfunctions $\phi_k\in L_2(T,\Theta,\nu)$ (thus $K\phi_k=\lambda_k\phi_k$ for every $k\in\mathbb{N}$) such that (4.1) holds, where the series converges in $L_2(T\times T,\Theta\times\Theta,\nu\times\nu)$. The series $\sum_k f_k\phi_k$ now converges in $L_2(T,\Theta,\nu)$ for any sequence (f_k) in ℓ_2. By the orthonormality of the functions ϕ_k, they are certainly linearly independent.

If the series (4.1) also converges pointwise on $T\times T$, then in particular $K(t,t)=\sum_k\lambda_k\phi_k^2(t)<\infty$ for all $t\in T$ and Theorem 4.1 shows that the RKHS is the set of all functions $\sum_k f_k\phi_k$ for sequences (f_k) such that $(f_k/\sqrt{\lambda_k})\in\ell_2$.

If the kernel is suitably regular, then we can apply the preceding with many choices of measure ν, leading to different eigenfunction expansions.

If the process itself can be expanded as a series

$$W=\sum_{j=1}^{\infty}\mu_j Z_j\phi_j,$$

for a sequence of i.i.d. standard normal variables (Z_j) and suitable functions ϕ_j, where the series converges in $L_2(\Omega,\mathcal{U},\mathrm{P})$, then (4.1) holds with $\lambda_j=\mu_j^2$ and the stochastic process RKHS takes the form given by the preceding proposition. The following proposition gives a Banach space version of this result.

Theorem 4.2. *Let (h_i) be a sequence of elements in a separable Banach space $\mathbb{B}$ such that $\sum_{i=1}^{\infty}w_ih_i=0$ for a sequence $w\in\ell_2$, where the convergence is in $\mathbb{B}$, implying that $w=0$. Let (Z_i) be an i.i.d. sequence of standard normal variables and assume that the series $W=\sum_{i=1}^{\infty}Z_ih_i$ converges almost surely in $\mathbb{B}$. Then the RKHS of W as a map in $\mathbb{B}$ is given by $\mathbb{H}=\{\sum_{i=1}^{\infty}w_ih_i\colon w\in\ell_2<\infty\}$ with squared norm $\|\sum w_ih_i\|_{\mathbb{H}}^2=\sum_i w_i^2$.*

Proof. The almost sure convergence of the series $W=\sum_{i=1}^{\infty}Z_ih_i$ in $\mathbb{B}$ implies the almost sure convergence of the series $b^*W=\sum_{i=1}^{\infty}Z_ib^*h_i$ in $\mathcal{R}$, for any $b^*\in\mathbb{B}^*$. Because the partial sums of the last series are zero-mean Gaussian, the series converges also in $L_2(\Omega,\mathcal{U},\mathrm{P})$. Hence for any $b^*,\underline{b}^*\in\mathbb{B}^*$,

$$\mathrm{E}b^*W\underline{b}^*W=\mathrm{E}\sum_{i=1}^{\infty}Z_ib^*h_i\sum_{i=1}^{\infty}Z_i\underline{b}^*h_i=\sum_{i=1}^{\infty}b^*h_i\underline{b}^*h_i.$$

In particular, the sequence (b^*h_i) is contained in ℓ_2 for every $b^*\in\mathbb{B}^*$, with square norm $\mathrm{E}(b^*W)^2$.

For $w\in\ell_2$ and natural numbers $m<n$, by the Hahn–Banach theorem and the Cauchy–Schwarz inequality,

$$\begin{aligned}\Bigl\|\sum_{m<i\le n}w_ih_i\Bigr\|^2&=\sup_{\|b^*\|\le1}\Bigl\|\sum_{m<i\le n}w_ib^*h_i\Bigr\|^2\\&\le\sum_{m<i\le n}w_i^2\sup_{\|b^*\|\le1}\sum_{m<i\le n}(b^*h_i)^2.\end{aligned}$$

As $m,n\to\infty$ the first factor on the far right tends to zero, since $w\in\ell_2$. By the first paragraph the second factor is bounded by $\sup_{\|b^*\|\le1}\mathrm{E}(b^*W)^2\le\mathrm{E}\|W\|^2$. Hence the partial sums of the series $\sum_i w_ih_i$ form a Cauchy sequence in $\mathbb{B}$, whence the infinite series converges.

Because $\sum_i(b^*h_i)^2$ was seen to converge, it follows that $\sum_i(b^*h_i)h_i$ converges in $\mathbb{B}$, and hence $\underline{b}^*(\sum_i(b^*h_i)h_i)=\sum_i b^*h_i\underline{b}^*h_i=\mathrm{E}b^*W\underline{b}^*W$, for any $\underline{b}^*\in\mathbb{B}^*$. This shows that $Sb^*=\sum_i(b^*h_i)h_i$ and the RKHS is not bigger than the space, as claimed.

The space would be smaller than claimed if there existed $w\in\ell_2$ that is not in the closure of the linear span of the elements (b^*h_i) of ℓ_2 when b^* ranges over $\mathbb{B}^*$. We can take this w without loss of generality as orthogonal to the latter collection, i.e. $\sum_i w_ib^*h_i=0$ for every $b^*\in\mathbb{B}^*$. This is equivalent to $\sum_i w_ih_i=0$, which has been excluded for any $w\neq0$. ■

It should be noted that the sequence (h_i) in the preceding lemma consists of arbitrary elements of the Banach space, only restricted by the linear independence condition that $\sum_i w_ih_i=0$ for $w\in\ell_2$, implying that $w=0$ (and the convergence of the random sequence $\sum_i Z_i h_i$). Combined with an i.i.d. standard normal sequence as coefficients, this sequence turns into an *orthonormal* basis of the RKHS.

From the proof it can be seen that the linear independence is necessary. If it fails, then the RKHS is the set of linear combinations $\sum_i w_i h_i$ with w restricted to the closure in ℓ_2 of the set of sequences $(b^* h_i)$ when b^* ranges over $\mathbb{B}^*$ and square norm $\sum_i w_i^2$. (Taking these linear combinations for all $w \in \ell_2$ gives the same set, but the ℓ_2-norm should be computed for a projected w.)

Example 4.2. *For $Z_0, \ldots, Z_k$ i.i.d. standard normal variables consider the polynomial process $t \mapsto \sum_{i=0}^k Z_i t^i/i!$ viewed as a map in (for instance) $C[0,1]$. The RKHS of this process is equal to the set of kth degree polynomials $P_a(t) = \sum_{i=0}^k a_i t^i/i!$ with square norm $\|P_a\|_{\mathbb{H}}^2 = \sum_{i=0}^k a_i^2$, i.e., the kth degree polynomials P with square norm $\|P\|_{\mathbb{H}}^2 = \sum_{i=0}^k P^{(i)}(0)^2$.*

Conversely, any Gaussian random element W in a separable Banach space can be expanded in a series $W = \sum_{j=1}^\infty Z_j h_j$ for i.i.d. standard normal variables Z_i and any orthonormal basis (h_i) of its RKHS, where the series converges in the norm of the Banach space. Because we can rewrite this expansion as $W = \sum_j \|h_j\| Z_j \tilde{h}_j$, where $\tilde{h}_j = h_j/\|h_j\|$ is a sequence of norm one, the corresponding 'eigenvalues' λ_i are in this case the square norms $\|h_i\|^2$. To prove this result, recall the isometry $U\colon \mathbb{H} \to L_2(\Omega, \mathcal{U}, \mathrm{P})$ defined in (3.3).

Theorem 4.3. *Let (h_i) be a complete orthonormal system in the RKHS $\mathbb{H}$ of a Borel measurable, zero-mean Gaussian random element W in a separable Banach space $\mathbb{B}$. Then $Uh_1, Uh_2, \ldots$ is an i.i.d. sequence of standard normal variables and $W = \sum_{i=1}^\infty (Uh_i) h_i$, where the series converges in the norm of $\mathbb{B}$, almost surely.*

Proof. It is immediate from the definitions of U and the RKHS that $U\colon \mathbb{H} \to L_2(\Omega, U, \mathrm{P})$ is an isometry. Because U maps the subspace $S\mathbb{B}^* \subset \mathbb{H}$ into the Gaussian process $b^* W$, it maps the completion $\mathbb{H}$ of $S\mathbb{B}^*$ into the completion of the linear span of this process in $L_2(\Omega, \mathcal{U}, \mathrm{P})$, which consists of normally distributed variables. Because U retains inner products, it follows that $Uh_1, Uh_2, \ldots$ is a sequence of i.i.d. standard normal variables.

By the definition of U and its continuity, for any $b^* \in \mathbb{B}^*$,

$$b^* W = U(Sb^*) = U\Big(\sum_{i=1}^\infty \langle Sb^*, h_i\rangle_{\mathbb{H}} h_i\Big) = \sum_{i=1}^\infty \langle Sb^*, h_i\rangle_{\mathbb{H}} U h_i = \sum_{i=1}^\infty b^*(h_i) U h_i,$$

where the last equality follows from the reproducing formula (2.6) and the series converges in $L_2(\Omega, \mathcal{U}, \mathrm{P})$. In other words, for any $b^* \in \mathbb{B}^*$, $b^*\big(\sum_{i=1}^n h_i(Uh_i)\big) \equiv \sum_{i=1}^n (b^* h_i) U h_i$ converges in $L_2(\Omega, \mathcal{U}, \mathrm{P})$ to $b^* W$. We wish to strengthen this to convergence almost surely of $W_n := \sum_{i=1}^n h_i(Uh_i)$ to W in $\mathbb{B}$. This is an immediate consequence of the Lévy–Ito–Nisio theorem, as given in, e.g., ([9], Theorem 2.4), according to which convergence in distribution of all 'marginals' $b^* \sum_{i=1}^n X_i$ to the marginals $b^* W$ of some Borel measurable map W in a separable Banach space, for $b^* \in \mathbb{B}^*$, implies the almost sure convergence of the series $\sum_i X_i$.

An alternative proof based on a martingale argument is given in ([9], Proposition 3.6). Let $Z_1, Z_2, \ldots$ be an orthonormal basis of the closed linear span of the variables $b^* W$ in $L_2(\Omega, \mathcal{U}, \mathrm{P})$. Then it can be seen that, for every n, $\mathrm{E}(W \,|\, Z_1, \ldots, Z_n) = \sum_{i=1}^n Z_i h_i$ in a Banach space sense, for $h_i = \mathrm{E} Z_i W$. Convergence of the infinite series follows by a martingale convergence theorem for Banach space valued variables. ∎

5. Support and Concentration

The RKHS of a zero-mean Gaussian random element W in a separable Banach space $\mathbb{B}$ is essential for an understanding of the spread of its distribution.

To begin with, the *support* of W, the smallest closed set $\mathbb{B}_0$ in $\mathbb{B}$ with $\mathrm{P}(W \in \mathbb{B}_0) = 1$, is the closure of the RKHS.

Lemma 5.1. *The support of a mean-zero Gaussian random element W in a separable Banach space $\mathbb{B}$ is the closure of its RKHS in $\mathbb{B}$. It is also the closure of the set $S\mathbb{B}^*$ in $\mathbb{B}$.*

Proof. We first show that the probability $\mathrm{P}\bigl(\|W\| < \varepsilon\bigr)$ of an arbitrary open ball centered around 0 is positive. Let V be an independent copy of W. Because we can cover $\mathbb{B}$ with countably many balls of radius ε, there exists some ball $B(h,\varepsilon)$ with positive measure under the law of W. The difference $B(h,\varepsilon) - B(h,\varepsilon)$ is contained in the ball of radius 2ε around 0. It follows that

$$\mathrm{P}\bigl(V - W \in B(0, 2\varepsilon)\bigr) \geq \mathrm{P}\bigl(V \in B(h,\varepsilon)\bigr)\mathrm{P}\bigl(W \in B(h,\varepsilon)\bigr) > 0.$$

Now $(V-W)/\sqrt{2}$ is a zero-mean Gaussian process with the same covariance function as W, and hence has the same distribution as W. It follows that $\mathrm{P}\bigl(W \in B(0,\sqrt{2}\varepsilon)\bigr) > 0$ for every $\varepsilon > 0$.

Since the distribution of $W - h$ is equivalent to the distribution of W for any h in the RKHS, by Lemma 3.2, it follows that $\mathrm{P}\bigl(\|W-h\| < \varepsilon\bigr) > 0$ for any $\varepsilon > 0$ and $h \in \mathbb{H}$.

This remains true for an element $h \in \mathbb{B}$ that can be approximated arbitrarily closely by elements from the RKHS. Thus the support of W contains the closure of the RKHS in $\mathbb{B}$.

By the Hahn–Banach theorem this closure $\bar{\mathbb{H}}$ can be written as

$$\bar{\mathbb{H}} = \bigcap_{b^* \in \mathbb{B}^*: b^*\mathbb{H} = 0} N(b^*),$$

where $b^*\mathbb{H} = 0$ means $b^*h = 0$ for all $h \in \mathbb{H}$ and $N(b^*) = \{b \in \mathbb{B}: b^*(b) = 0\}$ is the kernel of b^*. If $b^*\mathbb{H} = 0$, then in particular $b^*(Sb^*) = \mathrm{E}(b^*(W))^2 = 0$, and hence $b^*(W) = 0$ almost surely. It follows that $\mathrm{P}\bigl(W \in N(b^*)\bigr) = 1$ for every b^* in the display. By the preceding display the complement $\mathbb{B} - \bar{\mathbb{H}}$ is a union of the open sets $N(b^*)^c$. Because an open set in a separable metric space is Lindelöf ([6], section 10) this union can be written as a union of countably many of the sets $N(b^*)^c$. Equivalently, the intersection in the preceding display can be restricted to a suitable countable subset. It follows that $\mathrm{P}(W \in \bar{\mathbb{H}}) = 1$.

The second assertion follows, because the RKHS-norm is stronger than the norm of the containing Banach space. Completing the set $S\mathbb{B}^*$ for the RKHS-norm before taking the closure in $\mathbb{B}$ does therefore not give a bigger set. ∎

An inequality of [1] gives further insight in the concentration of the distribution of W. Let $\mathbb{H}_1$ and $\mathbb{B}_1$ be the unit balls of the RKHS and the space $\mathbb{B}$, respectively. The inequality involves the (centered) *small ball probability*

$$e^{-\phi_0(\varepsilon)} = \mathrm{P}(W \in \varepsilon\mathbb{B}_1).$$

Theorem 5.1. *(Borell's inequality.) For any $\varepsilon > 0$ and $M \geq 0$,*

$$\mathrm{P}\bigl(W \in \varepsilon\mathbb{B}_1 + M\mathbb{H}_1\bigr) \geq \Phi\bigl(\Phi^{-1}(e^{-\phi_0(\varepsilon)}) + M\bigr).$$

Here Φ is the cumulative distribution function of the standard normal distribution. For fixed $\varepsilon > 0$ the right side decreases as $M \to \infty$ according to the tails of the standard normal distribution. This shows that the 'geometry of the concentration' of W is given by the unit ball of the RKHS. Summing the small ball $\varepsilon\mathbb{B}_1$ to the multiple $M\mathbb{H}_1$ can be seen as enlarging the latter set with an ε-neighbourhood. In general this is necessary to capture the mass of the W, because the support of W is the closure of the RKHS; the RKHS itself may have probability zero. For $M \to \infty$ we obtain the equality $\mathrm{P}(W \in \varepsilon\mathbb{B}_1 + \mathbb{H}) = 1$, for any $\varepsilon > 0$, which (again) shows that W is supported within the closure of $\mathbb{H}$.

Example 5.1. *For a mean-zero normal vector W in $\mathbb{B} = \mathcal{R}^k$ with covariance matrix Σ, the RKHS is the range of the covariance matrix equipped with the inner product $\langle \Sigma g, \Sigma h\rangle_{\mathbb{H}} = g^T\Sigma h$. This follows, because $\mathbb{B}^* = \mathcal{R}^k$ and, for the element $g \in \mathbb{B}^*$ given by $h \mapsto h^T g$, we have $Sg = \mathrm{E}WW^T g = \Sigma g$. The inner product of the RKHS is $\langle Sg, Sh\rangle_{\mathbb{H}} = \mathrm{E}g^T W h^T W = g^T\Sigma h$.*

The unit ball $\mathbb{H}_1$ is the set $\{\Sigma h\colon h^T\Sigma h \leq 1\}$. For nonsingular Σ this set is the ellipsoid determined by the inverse matrix Σ^{-1}, i.e., the ellipsoid determined by the level sets of the density. For singular Σ the distribution is concentrated on a lower-dimensional subspace, and we have a similar interpretation after projection on this subspace.

Borell's inequality is often quoted as only an exponential inequality on the norm $\|W\|$, but this is in fact a consequence. The distribution of the norm $\|W\|$ of a non-zero Borel measurable Gaussian map W does not have atoms (cf., [2]) and therefore has a unique median $M(W)$.

Corollary 5.1. *For any $x > 0$,*

$$\mathrm{P}\bigl(\|W\| - M(W) > x\bigr) \leq 1 - \Phi\bigl(x/\sigma(W)\bigr).$$

Proof. For $\varepsilon = M(W)$ we have $\mathrm{P}(W \in \varepsilon\mathbb{B}_1) = \mathrm{P}\bigl(\|W\| \leq M(W)\bigr) = 1/2$. Hence the choices $\varepsilon = M(W)$ and $M = x/\sigma(W)$ in Borell's inequality yield the inequality $\mathrm{P}\bigl(W \in M(W)\mathbb{B}_1 + (x/\sigma(W))\mathbb{H}_1\bigr) \geq \Phi\bigl(x/\sigma(W)\bigr)$. Because $\mathbb{H}_1 \subset \sigma(W)\mathbb{B}_1$ by (2.5), the left side is smaller than $\mathrm{P}\bigl(W \in (M(W) + x)\mathbb{B}_1\bigr)$, which is 1 minus the left side of the corollary. ∎

According to Anderson's lemma (e.g., ([9], p73), ([16], p72), or ([17], 3.11.4)) a ball of fixed radius receives maximum mass of a zero-mean Gaussian distribution if centered at the origin. The following lemma gives a lower bound on the decrease in mass if the ball is centered at an element of the RKHS. The lemma is implicit in the proof of the main result in [7], and appears explicitly as (4.16) in [8].

Lemma 5.2. *If $h \in \mathbb{H}$, then for every Borel measurable set $C \subset \mathbb{B}$ with $C = -C$,*

$$\mathrm{P}(W - h \in C) \geq e^{-\frac{1}{2}\|h\|_{\mathbb{H}}^2}\mathrm{P}(W \in C).$$

Proof. By symmetry W and $-W$ are identically distributed and hence $\mathrm{P}(W + h \in C) = \mathrm{P}(-W + h \in -C) = \mathrm{P}(W - h \in C)$. By Lemma 3.2,

$$\mathrm{P}(W + h \in C) = \mathrm{E}1_C(W + h) = \mathrm{E}e^{Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2}1_C(W).$$

This is true with $-h$ instead of h as well. Combining these facts yields that

$$\mathrm{P}(W - h \in C) \quad = \quad \tfrac{1}{2}\mathrm{E}e^{Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2}1_C(W) + \tfrac{1}{2}\mathrm{E}e^{U(-h) - \frac{1}{2}\|-h\|_{\mathbb{H}}^2}1_C(W)$$

$$= \quad e^{-\frac{1}{2}\|h\|_{\mathbb{H}}^2}\mathrm{E}\cosh(Uh)1_C(W) \geq e^{-\frac{1}{2}\|h\|_{\mathbb{H}}^2}\mathrm{P}(W \in C),$$

since $\cosh x = (e^x + e^{-x})/2 \geq 1$ for every x. ■

The lemma with C equal to the ball of radius ε around 0 refers to the *noncentered small ball probabilities* $\mathrm{P}\bigl(\|W - w\| < \varepsilon\bigr)$, for every w in the RKHS. Up to constants these can be completely characterized through the corresponding centered small ball probabilities and approximation of the center w from the RKHS. Define

$$(5.1) \qquad \phi_w(\varepsilon) = \inf_{h\in\mathbb{H}:\|h-w\|\leq\varepsilon} \tfrac{1}{2}\|h\|_{\mathbb{H}}^2 - \log \mathrm{P}\bigl(\|W\| < \varepsilon\bigr).$$

For $w = 0$ this agrees with the negative exponent $\phi_0(\varepsilon)$ of the small ball probability $\mathrm{P}\bigl(\|W\| < \varepsilon\bigr) = e^{-\phi_0(\varepsilon)}$ defined previously. Up to constants this quantity gives the exponent of the small ball probability at center w.

Lemma 5.3. *For any w in the support of W and every $\varepsilon > 0$,*

$$\phi_w(\varepsilon) \leq -\log \mathrm{P}\bigl(\|W - w\| < \varepsilon\bigr) \leq \phi_w(\varepsilon/2).$$

Proof. For any $h \in \mathbb{H}$ with $\|h - w\| \leq \varepsilon$ we have $\|W - w\| \leq \varepsilon + \|W - h\|$ and hence $\mathrm{P}\bigl(\|W - w\| < 2\varepsilon\bigr) \geq \mathrm{P}\bigl(\|W - h\| < \varepsilon\bigr)$. The latter probability can be bounded below by $\exp\bigl(-\frac{1}{2}\|h\|_{\mathbb{H}}^2\bigr)\mathrm{P}\bigl(\|W\| < \varepsilon\bigr)$, in view of the preceding lemma. We conclude by optimizing over $h \in \mathbb{H}$.

The set $B_\varepsilon = \{h \in \mathbb{H}: \|h - w\| \leq \varepsilon\}$ is convex and closed in $\mathbb{H}$, because the RKHS topology is stronger than the norm topology. Therefore the (convex) map $h \mapsto \|h\|_{\mathbb{H}}^2$ attains a minimum on B_ε at some point h_ε. Because $(1-\lambda)h_\varepsilon + \lambda h \in B_\varepsilon$ for every $h \in B_\varepsilon$ and $0 \leq \lambda \leq 1$, it follows that $\|(1 - \lambda)h_\varepsilon + \lambda h\|_{\mathbb{H}}^2 \geq \|h_\varepsilon\|_{\mathbb{H}}^2$, which implies that $2\lambda\langle h - h_\varepsilon, h_\varepsilon\rangle_{\mathbb{H}} + \lambda^2\|h - h_\varepsilon\|_{\mathbb{H}}^2 \geq 0$. The fact that this is true for every $0 \leq \lambda \leq 1$ can be seen to imply that $\langle h, h_\varepsilon\rangle_{\mathbb{H}} \geq \|h_\varepsilon\|_{\mathbb{H}}^2$ for every $h \in B_\varepsilon$.

By Theorem 4.3 the process W can be written as $W = \sum_{i=1}^\infty (Uh_i)h_i$, for any given complete orthonormal system $h_1, h_2, \ldots$ in $\mathbb{H}$, where the series converges almost surely in norm. The truncated series $W^m = \sum_{i=1}^m (Uh_i)h_i$ takes its values in $\mathbb{H}$. If $\|W - g - w\| < \varepsilon$ and some arbitrary $g \in \mathbb{H}$, then $\|W^m - g - w\| < \varepsilon$ for sufficiently large m, almost surely. Equivalently, $W^m - g \in B_\varepsilon$ and hence the preceding paragraph implies that $\langle W^m - g, h_\varepsilon\rangle_{\mathbb{H}} \geq \|h_\varepsilon\|_{\mathbb{H}}^2$, eventually as $m \to \infty$, almost surely. Here $\langle W^m, h_\varepsilon\rangle_{\mathbb{H}} = \sum_{i=1}^m (Uh_i)\langle h_i, h_\varepsilon\rangle_{\mathbb{H}} = U\sum_{i=1}^m h_i\langle h_i, h_\varepsilon\rangle_{\mathbb{H}}$. By the continuity of U the right side converges in $L_2(\Omega, \mathcal{U}, \mathrm{P})$ to Uh_ε as $m \to \infty$, and hence almost surely along a subsequence. We conclude that $Uh_\varepsilon - \langle g, h_\varepsilon\rangle_{\mathbb{H}} \geq \|h_\varepsilon\|_{\mathbb{H}}^2$ almost surely on the event $\{\|W - g - w\| < \varepsilon\}$. In particular the choice $g = -h_\varepsilon$ yields that $Uh_\varepsilon \geq 0$ almost surely on the event $\{\|W + h_\varepsilon - w\| < \varepsilon\}$.

By Lemma 3.2,

$$\begin{aligned} \mathrm{P}\bigl(W \in w + \varepsilon\mathbb{B}_1\bigr) &= \mathrm{P}\bigl(W - h_\varepsilon \in w - h_\varepsilon + \varepsilon\mathbb{B}_1\bigr) \\ &= \mathrm{E}e^{-Uh_\varepsilon - \frac{1}{2}\|h_\varepsilon\|_{\mathbb{H}}^2} 1_{W\in w-h_\varepsilon+\varepsilon\mathbb{B}_1} \leq e^{-\frac{1}{2}\|h_\varepsilon\|_{\mathbb{H}}^2}\mathrm{E}1_{W\in w-h_\varepsilon+\varepsilon\mathbb{B}_1}, \end{aligned}$$

by the preceding paragraph. The probability on the right side is smaller than $\mathrm{P}(W \in \varepsilon\mathbb{B}_1)$ by Anderson's lemma. ■

6. Small Ball Probability and Entropy

The unit ball of the RKHS not only expresses the shape of the Gaussian measure, but also allows a quantitative estimate of the small ball probability $e^{-\phi_0(\varepsilon)} = \mathrm{P}\bigl(\|W\| < \varepsilon\bigr)$ through its entropy within the Banach space.

Let $N\big(\varepsilon, \mathbb{H}_1, \|\cdot\|\big)$ be the smallest number of balls of radius $\varepsilon > 0$ needed to cover the unit ball $\mathbb{H}_1$ of the RKHS. This is bounded by the maximal number $D(\varepsilon)$ of points h_i in $\mathbb{H}_1$ with $\|h_i - h_j\| \geq \varepsilon$ for $i \neq j$. Because each ball of radius $\varepsilon/2$ around a point h_i has probability at least $e^{-1/2}\mathrm{P}\big(\|W\| < \varepsilon/2\big)$ by Lemma 5.2 and these balls are disjoint, it follows that $1 \geq D(\varepsilon)e^{-1/2}\mathrm{P}\big(\|W\| < \varepsilon/2\big)$, whence $D(\varepsilon)$ is finite for every $\varepsilon > 0$. This shows that the RKHS unit ball $\mathbb{H}_1$ is precompact in $\mathbb{B}$.

The following results, which were proved by [7] and [10], refine this argument, and show roughly that for regularly behaved entropy $\varepsilon \mapsto \log N\big(\varepsilon, \mathbb{H}_1, \|\cdot\|\big)$ and small ball exponent $\varepsilon \mapsto \phi_0(\varepsilon)$, and for small ε,

$$\log N\Big(\frac{\varepsilon}{\sqrt{\phi_0(\varepsilon)}}, \mathbb{H}_1, \|\cdot\|\Big) \asymp \phi_0(\varepsilon).$$

However, the exact statement has several constants in it.

Lemma 6.1. *Let $f\colon (0,\infty) \to (0,\infty)$ be regularly varying at zero. Then*

(i) $\log N\big(\varepsilon/\sqrt{2\phi_0(\varepsilon)}, \mathbb{H}_1, \|\cdot\|\big) \gtrsim \phi_0(2\varepsilon)$.

(ii) If $\phi_0(\varepsilon) \lesssim f(\varepsilon)$, then $\log N\big(\varepsilon/\sqrt{f(e)}, \mathbb{H}_1, \|\cdot\|\big) \lesssim f(\varepsilon)$.

(iii) If $\log N\big(\varepsilon, \mathbb{H}_1, \|\cdot\|\big) \gtrsim f(\varepsilon)$, then $\phi_0(\varepsilon) \gtrsim f\big(\varepsilon/\sqrt{\phi_0(\varepsilon)}\big)$.

(iv) If $\log N\big(\varepsilon, \mathbb{H}_1, \|\cdot\|\big) \lesssim f(\varepsilon)$, then $\phi_0(2\varepsilon) \lesssim f\big(\varepsilon/\sqrt{\phi_0(\varepsilon)}\big)$.

Lemma 6.2. *For $\alpha > 0$ and $\beta \in \mathcal{R}$, as $\varepsilon \downarrow 0$, $\phi_0(\varepsilon) \asymp \varepsilon^{-\alpha}(\log 1/\varepsilon)^{\beta}$ if and only if $\log N\big(\varepsilon, \mathbb{H}_1, \|\cdot\|\big) \asymp \varepsilon^{-2\alpha/(2+\alpha)}(\log 1/\varepsilon)^{2\beta/(2+\alpha)}$.*

7. RKHS under Transformation

If a Gaussian process is transformed into another Gaussian process under a one-to-one, continuous, linear map, then the RKHS is transformed in parallel.

Lemma 7.1. *Let $T\colon \mathbb{B} \to \underline{\mathbb{B}}$ be a one-to-one, continuous, linear map from a separable Banach space $\mathbb{B}$ into a Banach space $\underline{\mathbb{B}}$ and let W be a Borel measurable, zero-mean Gaussian random element in $\mathbb{B}$ with RKHS $\mathbb{H}$. Then the RKHS of the Gaussian random element TW in $\underline{\mathbb{B}}$ is equal to $T\mathbb{H}$ and $T\colon \mathbb{H} \to \underline{\mathbb{H}}$ is an isometry for the RKHS-norms.*

Proof. Let $T^*\colon \underline{\mathbb{B}}^* \to \mathbb{B}^*$ be the adjoint of T. The RKHS $\underline{\mathbb{H}}$ of TW is by definition the completion of the set of Pettis integrals

$$\underline{S}\underline{b}^* = \mathrm{E}(TW)\underline{b}^*(TW) = T(\mathrm{E}W\underline{b}^*(TW)) = TST^*\underline{b}^*,$$

for the inner product

$$\langle \underline{S}\underline{b}_1^*, \underline{S}\underline{b}_2^* \rangle_{\underline{\mathbb{H}}} = \mathrm{E}\underline{b}_1^*(TW)\underline{b}_2^*(TW) = \mathrm{E}(T^*\underline{b}_1^*W)(T^*\underline{b}_2^*W) = \langle ST^*\underline{b}_1^*, ST^*\underline{b}_2^* \rangle_{\mathbb{H}}.$$

It follows that the element $\underline{S}\underline{b}^*$ of $\underline{\mathbb{H}}$ is the image under T of the element $ST^*\underline{b}^*$ of $\mathbb{H}$, and its norm is the same: $\|\underline{S}\underline{b}^*\|_{\underline{\mathbb{H}}} = \|ST^*\underline{b}^*\|_{\mathbb{H}}$. Thus $T\colon ST^*\underline{\mathbb{B}}^* \subset \mathbb{H} \to \underline{\mathbb{H}}$ is an isometry for the RKHS-norms. It extends by continuity to a linear map from the completion $\mathbb{H}_0$ of $ST^*\underline{\mathbb{B}}^*$ in $\mathbb{H}$ to $\underline{\mathbb{H}}$. Because T is continuous for the norm of $\mathbb{B}$, this extension agrees with T. Because $T\colon ST^*\underline{\mathbb{B}}^* \to \underline{S}\underline{\mathbb{B}}^*$ is onto, T is an isometry for the RKHS-norms, and $\mathbb{H}_0$ and $\underline{\mathbb{H}}$ are by definition the completions of $ST^*\underline{\mathbb{B}}^*$ and $\underline{S}\underline{\mathbb{B}}^*$, we have that $T\colon \mathbb{H}_0 \to \underline{\mathbb{H}}$ is an isometry onto $\underline{\mathbb{H}}$. It remains to be shown that $\mathbb{H}_0 = \mathbb{H}$.

Because T is one-to-one, the range $T^*\underline{\mathbb{B}}^*$ of its adjoint is weak-* dense in $\mathbb{B}^*$ ([12], Corollary 4.12). By Lemma 2.2 the map $S\colon \mathbb{B}^* \to \mathbb{H}$ is continuous relative to the weak-* and RKHS topologies. Combined this yields that $S(T^*\underline{\mathbb{B}}^*)$ is dense in $S\mathbb{B}^*$ for the RKHS-norm of $\mathbb{H}$ and hence is dense in $\mathbb{H}$.

Taken together the preceding shows that $T\colon \mathbb{H} \to \underline{\mathbb{H}}$ is an isometry onto $\underline{\mathbb{H}}$. ∎

8. RKHS Relative to Different Norms

A stochastic process W can often be viewed as a map into several Banach spaces. For instance, a process indexed by the unit interval with continuous sample paths is a Borel measurable map in the space $C[0,1]$, but also in the space $L_2[0,1]$; a process with continuously differentiable sample paths is a map in $C[0,1]$, but also in $C^1[0,1]$. The RKHS obtained from using a weaker Banach space is typically the same.

Lemma 8.1. *Let* $(\mathbb{B}, \|\cdot\|)$ *be a separable Banach space on* $\mathbb{B}$ *and let* $\|\cdot\|'$ *be a norm on* $\mathbb{B}$ *with* $\|b\|' \le \|b\|$. *Then the RKHS of a Borel measurable zero-mean Gaussian random element in* $(\mathbb{B}, \|\cdot\|)$ *is the same as the RKHS of this map viewed in the completion of* $\mathbb{B}$ *under* $\|\cdot\|'$.

Proof. Let $\mathbb{B}'$ be the completion of $\mathbb{B}$ relative to $\|\cdot\|'$. The assumptions imply that the identity map $I\colon (\mathbb{B}, \|\cdot\|) \to (\mathbb{B}', \|\cdot\|')$ is continuous, linear and one-to-one. The proposition therefore is a consequence of Lemma 7.1. ∎

Example 8.1. *Let* W *be a mean zero Gaussian process indexed by the unit interval* $[0,1]$ *with covariance function* $K(s,t) = \mathrm{E}W_sW_t$.

If W *has continuous sample paths, then it is a random element in* $C[0,1]$. *The RKHS of* W *viewed as a random element in* $C[0,1]$ *is the completion of the linear span of the functions* $K(t,\cdot)$ *under the inner product (2.2).*

If W *is a measurable process and* $\int_0^1 W_s^2\,ds < \infty$ *surely, then* W *is a random element in* $L_2[0,1]$. *The dual space of* $L_2[0,1]$ *consists of the maps* $g \mapsto \int g(s)f(s)\,ds$ *for* f *ranging over* $L_2[0,1]$, *and* $Sf(t) = \mathrm{E}W_t \int W_s f(s)\,ds = \int K(s,t)f(s)\,ds$. *Therefore, the RKHS of* W *viewed as a random element in* $L_2[0,1]$ *is the completion of the linear span of the functions* $t \mapsto \int K(s,t)\,f(s)\,ds$ *for* f *ranging over* $L_2[0,1]$ *under the inner product* $\langle Sf, Sg\rangle_{\mathbb{H}} = \int\int K(s,t)f(s)g(t)\,dsdt$.

If W *has continuous sample paths, then its covariance kernel is continuous, and it can be shown by direct arguments that the two RKHSs agree. This also follows from the preceding lemma.*

9. RKHS under Independent Sums

If a given Gaussian prior misses certain desirable ‘directions’ in its RKHS, then these can be filled in by adding independent Gaussian components in these directions. A closed linear subspace $\mathbb{B}_0 \subset \mathbb{B}$ of a Banach space $\mathbb{B}$ is *complemented* if there exists a closed linear subspace $\mathbb{B}_1$ with $\mathbb{B} = \mathbb{B}_0 + \mathbb{B}_1$ and $\mathbb{B}_0 \cap \mathbb{B}_1 = \{0\}$.

Lemma 9.1. *Let* V *and* W *be independent Borel measurable, zero-mean, jointly Gaussian maps from a given probability space into a separable Banach space with supports* $\mathbb{B}^V$ *and* $\mathbb{B}^W$ *such that* $\mathbb{B}^V \cap \mathbb{B}^W = \{0\}$ *and the subspace* $\mathbb{B}^V$ *is complemented by a subspace that contains* $\mathbb{B}^W$. *Then the RKHS of* $V + W$ *is the direct sum of the RKHSs of* V *and* W *and the RKHS norms satisfy* $\|h^V + h^W\|^2_{\mathbb{H}^{V+W}} = \|h^V\|^2_{\mathbb{H}^V} + \|h^W\|^2_{\mathbb{H}^W}$.

Proof. By the independence of V and W the Pettis integral $S^{V+W}b^* = \mathrm{E}(W+V)b^*(V+W)$ can be written as $S^{V+W}b^* = S^V b^* + S^W b^*$. The assumptions of trivial intersection $\mathbb{B}^V \cap \mathbb{B}^W = \{0\}$ and of complementation of $\mathbb{B}^V$ entail that there exists a continuous linear map $\Pi\colon \mathbb{B} \to \mathbb{B}^V$ such that $\Pi b = b$ if $b \in \mathbb{B}^V$ and $\Pi b = 0$ if $b \in \mathbb{B}^W$; (cf., [6], 29.2). Then $b^* \circ \Pi \in \mathbb{B}^*$, $\Pi V = V$ and $\Pi W = 0$ almost surely, whence $S^W(b^* \circ \Pi) = 0$ and hence $S^{V+W}(b^* \circ \Pi) = S^V b^*$. It follows that $S^V \mathbb{B}^* \subset S^{V+W}\mathbb{B}^*$ and by symmetry $S^W \mathbb{B}^* \subset S^{V+W}\mathbb{B}^*$. Also, for any $b_1^*, b_2^* \in \mathbb{B}^*$,

$$\begin{aligned} \langle S^V b_1^*, S^W b_2^* \rangle_{\mathbb{H}^{V+W}} &= \langle S^{V+W} b_1^* \circ \Pi, S^{V+W} b_2^* \circ (I - \Pi) \rangle_{\mathbb{H}^{V+W}} \\ &= \mathrm{E}\big(b_1^* \circ \Pi(V+W)\big)\,\big(b_2^* \circ (I-\Pi)(V+W)\big) \\ &= \mathrm{E}(b_1^* V)\,(b_2^* W) = 0. \end{aligned}$$

We conclude that $S^V \mathbb{B}^* \perp S^W \mathbb{B}^*$ in $\mathbb{H}^{V+W}$, so that $\mathbb{H}^{V+W}$ is the direct (orthogonal) sum of $\mathbb{H}^V$ and $\mathbb{H}^W$. Furthermore $\|S^{V+W}b^*\|^2_{\mathbb{H}^{V+W}} = \mathrm{E}(b^*(V+W))^2 = \mathrm{E}(b^*V)^2 + \mathrm{E}(b^*W)^2$. ∎

By the Hahn–Banach theorem the assumption of complementation is certainly satisfied as soon as one of the supports of V and W is finite-dimensional.

The assumption that $\mathbb{B}^V \cap \mathbb{B}^W = \{0\}$ can be interpreted as requiring 'linear independence' rather than some form of orthogonality of the supports of V and W. The stochastic independence of V and W translates the linear independence into orthogonality in the RKHS of $V + W$.

The assumption requires trivial intersection of the supports of the variables V and W, rather than of sets that carry probability one. Because the RKHS is independent of the norm (Lemma 8.1) the closure operation involved in computing the support may be taken for the strongest norm which is defined on the random elements.

The assumption that $\mathbb{B}^V \cap \mathbb{B}^W = \{0\}$ cannot be removed. For instance, if $V = \sum_i \mu_i Z_i \psi_i$ and $W = \sum_i \mu_i' Z_i' \psi_i$ are series expansions with independent standard normal variables $(Z_i), (Z_i')$ on a common basis (ψ_i), then the sum process can be written $V + W = \sum_i \mu_i'' Z_i'' \psi_i$ for $\mu_i'' = \sqrt{\mu_i^2 + (\mu_i')^2}$ and Z_i'' independent standard normal variables. The RKHS of $V + W$ is then the set of series $\sum_i w_i \psi_i$ with coefficients (w_i) satisfying $\sum_i (w_i/\mu_i'')^2 < \infty$ (see Section 4). Thus the RKHS is not an orthogonal sum and, asymptotically as $i \to \infty$, the eigenvalues $(\mu_i'')^2$, which determine the presence of the directions ψ_i in the RKHS, are determined by the slowest of the two sequences μ_i and μ_i'. If $\mu_i/\mu_i' \to 0$, then the RKHS of $V + W$ is essentially the same as the RKHS of W.

10. Examples

The RKHS of standard Brownian motion, viewed as a random element in $C[0,1]$, is well known to be the set

$$\{f\colon [0,1] \to \mathcal{R}, f \in AC, f(0) = 0, \int f'(t)^2\,dt < \infty\}, \tag{10.1}$$

where $f \in AC$ is the assumption that f is absolutely continuous. The RKHS inner product is

$$\langle f, g \rangle_{\mathbb{H}} = \int_0^1 f'(t) g'(t)\,dt.$$

Lemma 10.1. *The RKHS of a standard Brownian motion W on $[0,1]$ is given by (10.1) with the inner product as indicated.*

Proof. We use the definition of the RKHS in Section 2.1 and the fact that the covariance kernel of Brownian motion is given by $s\wedge t$. The RKHS is the completion of the linear span of the functions $t\mapsto s\wedge t$ as s ranges over $[0,1]$, under the inner product determined by

$$\langle s_1\wedge\cdot, s_2\wedge\cdot\rangle_{\mathbb{H}} = s_1\wedge s_2 = \int (s_1\wedge t)'(s_2\wedge t)'\,dt,$$

where the prime denotes differentiation relative to t, in the sense of absolute continuity.

The linear span of the functions $t\mapsto s\wedge t$ contains every function that is 0 at 0, continuous, and piecewise linear on a partition $0=s_0<s_1\cdots<s_N=1$. Indeed to obtain such a function with slopes $\alpha_1,\ldots,\alpha_N$ on the intervals $(s_0,s_1),\ldots,(s_{N-1},s_N)$, first determine the coefficient of $s_N\wedge\cdot$ to have a correct slope on (s_{N-1},s_N), next determine the coefficient of $s_{N-1}\wedge\cdot$ to have a correct slope on (s_{N-2},s_{N-1}), etc. The derivatives of these functions are piecewise constant, and the set of piecewise constant functions is dense in $L_2[0,1]$. ∎

Given the RKHS of Brownian motion it is now easy to derive the RKHS of several processes related to it.

- To release Brownian motion at zero, we may start it at an independent standard normal variable Z, giving the process $t\mapsto Z+W_t$. The RKHS of the constant process $t\mapsto Z$ are the constant functions, which have trivial intersection with the RKHS of Brownian motion. A given function $f\colon[0,1]\to\mathcal{R}$ can be decomposed as $f=f(0)+\bigl(f-f(0)\bigr)$, where the second part is in the RKHS of Brownian motion if it is absolutely continuous with square integrable derivative. By Lemma 9.1, the RKHS of $Z+W$ is the set of all absolutely continuous functions $f\colon[0,1]\to\mathcal{R}$ equipped with the inner product $\langle f,g\rangle_{\mathbb{H}}=f(0)g(0)+\int f'(s)g'(s)\,ds$.
- To smooth Brownian motion we may consider its k-fold integral $I^k_{0+}W$, where $(I^1_{0+}f)(t)=\int_0^t f(s)\,ds$ and $I^k_{0+}=I^{k-1}_{0+}I^1_{0+}$. Taking a primitive is a continuous, linear, one-to-one map from $C[0,1]\to C[0,1]$, and hence by Lemma 7.1 the RKHS of $I^k_{0+}W$ is the set of functions $I^k_{0+}f$ for f in the RKHS of Brownian motion, equipped with the inner product $\langle I^k_{0+}f, I^k_{0+}g\rangle_{\mathbb{H}}=\int_0^1 f'(s)g'(s)\,ds$. This space can be described simply as the set of all functions $f\colon[0,1]\to\mathcal{R}$ that are k-times differentiable with an absolutely continuous kth derivative with square-integrable $f^{(k+1)}$, equipped with the inner product $\langle f,g\rangle_{\mathbb{H}}=\int_0^1 f^{(k+1)}(s)g^{(k+1)}(s)\,ds$.
- The sample paths of k-fold integrated Brownian motion $I^k_{0+}W$ have k vanishing derivatives at zero, which negatively affects its approximation properties to smooth functions. (See Example 10.1 below.) We can release the derivatives by adding a polynomial and considering the process $t\mapsto\sum_{i=0}^k Z_it^i/i!+(I^k_{0+}W)_t$, for $Z_0,\ldots,Z_k$ i.i.d. standard normal variables, independent of W. The supports of the polynomial process $t\mapsto\sum_{i=0}^k Z_it^i/i!$ and $I^k_{0+}W$ in $C[0,1]$ do not have a trivial intersection, and hence we cannot apply Lemma 9.1 in that setting. However, we may consider these processes as Borel measurable random elements in the space $C^{(k)}[0,1]$ of k-times differentiable functions, equipped with the norm $\|f\|_k=\|f\|_\infty+\|f^{(k)}\|_\infty$. According to Lemma 8.1, this does not change the RKHS. The support of the process $I^k_{0+}W$ in $C^{(k)}[0,1]$ contains only functions with k vanishing derivatives at 0, and hence does have trivial intersection with the support of the

polynomial process $t \mapsto \sum_{i=0}^k Z_i t^i/i!$, which is the set of kth degree polynomials. Applied in this setting Lemma 9.1 yields that the RKHS of the process $t \mapsto \sum_{i=0}^k Z_i t^i/i! + (I_{0+}^k W)_t$ is the set of functions $f\colon [0,1] \to \mathcal{R}$ that are k-times differentiable with an absolutely continuous kth derivative with square-integrable $f^{(k+1)}$, equipped with the inner product $\langle f, g\rangle_{\mathbb{H}} = \sum_{i=0}^k f^{(i)}(0)g^{(i)}(0) + \int_0^1 f^{(k+1)}(s)g^{(k+1)}(s)\,ds$. To see the latter, note that any f can be uniquely written as $f = P_k + (f - P_k)$ for $P_k(t) = \sum_{i=0}^k f^{(i)}(0)t^i/i!$, the kth degree Taylor polynomial and $f - P_k$ a function with k vanishing derivatives at zero. The polynomial P_k is contained in the RKHS of the polynomial process $t \mapsto \sum_{i=0}^k Z_i t^i/i!$ with square RKHS-norm $\sum_{i=0}^k P_k^{(i)}(0)^2$ by Example 4.2, and the function $f - P_k$ is contained in the RKHS of $I_{0+}^k W$ by the preceding.

The preceding can be extended to fractional integrals of Brownian motion. Rather than studying the fractional integral operator in detail, we give a direct derivation of the RKHSs. For $\alpha > 0$ and W a standard Brownian motion the *Riemann–Liouville process with Hurst parameter* $\alpha > 0$ is defined as

$$R_t^\alpha = \int_0^t (t-s)^{\alpha - 1/2}\,dW_s, \quad t \geq 0.$$

The process R^α is a centered Gaussian process with continuous sample paths. It can be viewed as a multiple of the $(\alpha + 1/2)$-fractional integral of the "derivative dW of Brownian motion". For $\alpha > 0$ and a (deterministic) measurable function f on $[0,1]$ the (left-sided) *Riemann–Liouville fractional integral of f of order* α (if it exists) is defined as (cf., [13])

$$I_{0+}^\alpha f(t) = \frac{1}{\Gamma(\alpha)} \int_0^t (t-s)^{\alpha-1} f(s)\,ds.$$

For α a natural number, the function $I_{0+}^\alpha f$ is just the α-fold iterated integral of f, and for $\alpha > 1/2$ the Rieman–Liouville process is equal to $\Gamma(\alpha + 1/2) I_{0+}^{\alpha - 1/2} W$ for I_{0+}^α the fractional integral.

Lemma 10.2. *The RKHS of the Riemann–Liouville process with parameter $\alpha > 0$ viewed as a random element in $C[0,1]$ is $\mathbb{H} = I_{0+}^{\alpha+1/2}(L_2[0,1])$ and the RKHS-norm is given by*

$$\|I_{0+}^{\alpha+1/2} f\|_{\mathbb{H}} = \frac{\|f\|_2}{\Gamma(\alpha + 1/2)}.$$

Proof. We use the characterization of the RKHS as the completion of the functions (2.1) under the inner product (2.2). With f_s the function defined by $f_s(u) = (s - u)_+^{\alpha - 1/2}$, we have, for all $s, t \geq 0$ and $\langle \cdot, \cdot \rangle_2$ the inner product of $L_2[0,1]$,

$$E R_t^\alpha R_s^\alpha = \int (t-u)_+^{\alpha-1/2} (s-u)_+^{\alpha-1/2}\,du = \langle f_t, f_s\rangle_2 = \Gamma(\alpha + 1/2) I_{0+}^{\alpha+1/2} f_s(t).$$

Hence every simple element of $\mathbb{H}$ of the form (2.1) is given by $I_{0+}^{\alpha+1/2} f$ for some $f \in L_2[0,1]$. Moreover, the inner product (2.2) of two such elements $I_{0+}^{\alpha+1/2} f$ and $I_{0+}^{\alpha+1/2} g$ is given by

$$\left\langle I_{0+}^{\alpha+1/2} f, I_{0+}^{\alpha+1/2} g \right\rangle_{\mathbb{H}} = \frac{\langle f, g\rangle_2}{\Gamma^2(\alpha + 1/2)}. \tag{10.2}$$

It follows that the RKHS $\mathbb{H}$ is a subspace of the Hilbert space obtained by endowing $I_{0+}^{\alpha+1/2}(L_2[0,1])$ with the inner product (10.2). To prove the converse inclusion, suppose that $g \in I_{0+}^{\alpha+1/2}(L_2[0,1])$ is orthogonal to $\mathbb{H}$. Then $g = I_{0+}^{\alpha+1/2} f$ for some $f \in L_2[0,1]$ and g is, in particular, orthogonal to every element $I_{0+}^{\alpha+1/2} f_t$ of $\mathbb{H}$. Hence, for every $t \in [0,1]$,

$$0 = \Big\langle I_{0+}^{\alpha+1/2} f, I_{0+}^{\alpha+1/2} f_t \Big\rangle_{\mathbb{H}} = \frac{\langle f, f_t \rangle_2}{\Gamma^2(\alpha+1/2)} = \frac{I_{0+}^{\alpha+1/2} f(t)}{\Gamma(\alpha+1/2)}.$$

The injectivity of the operator $I_{0+}^{\alpha+1/2}: L_2[0,1] \to L_2[0,1]$ (see [13], Theorem 13.1) then implies that $f = 0$, whence $g = 0$. We conclude that $\mathbb{H} = I_{0+}^{\alpha+1/2}(L_2[0,1])$, and the inner product on $\mathbb{H}$ is given by (10.2). ■

Example 10.1. *The process $t \mapsto Z + \int_0^t W_s\,ds$, for Z a standard normal variable and W an independent Brownian motion, has sample paths of regularity $3/2$ and can take any value at 0, but the derivative at 0 is 0. We shall show that the latter makes the process inappropriate as a prior model for $3/2$-smooth functions.*

By similar arguments as before the RKHS $\mathbb{H}$ of the process can be seen to be the set of all functions $h: [0,1] \to \mathcal{R}$ with absolutely continuous derivative such that $\int_0^1 h''(s)^2\,ds < \infty$ and $h'(0) = 0$, with square norm

$$\|h\|_{\mathbb{H}}^2 = \|h''\|_2^2 + h(0)^2.$$

We shall show that for the identity function id we have

$$\inf\{\|h\|_{\mathbb{H}}^2 : h \in \mathbb{H}, \|h - id\|_\infty < \varepsilon\} \gtrsim \frac{1}{\varepsilon}.$$

This may be contrasted with the approximation by the RKHS of the process $t \mapsto Z_0 + Z_1 t + \int_0^t W_s\,ds$, which is of order $(1/\varepsilon)^{2/3}$ for every function in $C^{3/2}[0,1]$ (see [15]).

To prove the claim note that $\|h - id\|_\infty < \varepsilon$ implies that $h(3\varepsilon) - h(0) > \varepsilon$. Therefore the quantity in the display is bounded below by

$$\inf\Big\{\int_0^{3\varepsilon} h''(s)^2\,ds : h(3\varepsilon) - h(0) > \varepsilon, h'(0) = 0\Big\}.$$

For a given h as in the display we can define g by

$$g(y) = \frac{h(3\varepsilon y) - h(0)}{\varepsilon}.$$

Then $g'(y) = 3h'(3\varepsilon y)$, $g''(y) = 9h''(3\varepsilon y)\varepsilon$, and

$$g(0) = 0, \qquad g'(0) = 0, \qquad g(1) > 1.$$

$$\int_0^{3\varepsilon} h''(s)^2\,ds = \int_0^{3\varepsilon} g''(s/(3\varepsilon))^2 \frac{1}{(9\varepsilon)^2}\,ds = \int_0^1 g''(u)^2 \frac{1}{27\varepsilon}\,du.$$

Thus the preceding display is bigger than

$$\Big(\frac{1}{27\varepsilon}\Big) \inf\Big\{\int_0^1 g''(u)^2\,du : g(1) > 1, g(0) = g'(0) = 0\Big\}.$$

The infimum is nonzero, because $g'' = 0$ implies that g is a linear function, hence identically 0 because $g(0) = g'(0) = 0$, contradicting $g(1) > 1$.

Acknowledgments

Professor van Zanten's research was partially supported by the Netherlands Organization for Scientific Research NWO.

References

[1] Borell, C. (1975). The Brunn-Minkowski inequality in Gauss space. *Invent. Math.* **30**, 2, 207–216. MR0399402 (53 #3246)

[2] Cirel′son, B. S. (1975). Density of the distribution of the maximum of a Gaussian process. *Teor. Verojatnost. i Primenen.* **20**, 4, 865–873. MR0394834 (52 #15633)

[3] de Acosta, A. (1983). Small deviations in the functional central limit theorem with applications to functional laws of the iterated logarithm. *Ann. Probab.* **11**, 1, 78–101. MR682802 (84m:60038)

[4] Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28**, 2, 500–531. MR1790007 (2001m:62065)

[5] Ghosal, S. and Roy, A. (2006). Posterior consistency in nonparametric regression problem under gaussian process prior. *Ann. Statist. 34*, 2413–2429.

[6] Jameson, G. J. O. (1974). *Topology and normed spaces.* Chapman and Hall, London. MR0463890 (57 #3828)

[7] Kuelbs, J. and Li, W. V. (1993). Metric entropy and the small ball problem for Gaussian measures. *J. Funct. Anal.* **116**, 1, 133–157. MR1237989 (94j:60078)

[8] Kuelbs, J., Li, W. V., and Linde, W. (1994). The Gaussian measure of shifted balls. *Probab. Theory Related Fields* **98**, 2, 143–162. MR1258983 (95c:60004)

[9] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces.* Vol. **23**. Springer-Verlag, Berlin. MR1102015 (93c:60001)

[10] Li, W. V. and Linde, W. (1999). Approximation, metric entropy and small ball estimates for gaussian measures. *The Annals of Probability* **27**, 3, 1556–1578.

[11] Li, W. V. and Shao, Q.-M. (2001). Gaussian processes: inequalities, small ball probabilities and applications. In *Stochastic processes: theory and methods.* Handbook of Statist., Vol. **19**. North-Holland, Amsterdam, 533–597. MR1861734

[12] Rudin, W. (1973). *Functional analysis.* McGraw-Hill Book Co., New York. McGraw-Hill Series in Higher Mathematics. MR0365062 (51 #1315)

[13] Samko, S. G., Kilbas, A. A., and Marichev, O. I. (1993). *Fractional integrals and derivatives.* Gordon and Breach Science Publishers, Yverdon. MR1347689 (96d:26012)

[14] Tokdar, S. and Ghosh, J. (2005). Posterior consistency of gaussian process priors in density estimation. *J. Statist. Plann. Inf. 137*, 34–42.

[15] Van der Vaart, A. and Van Zanten, J. (2006). Rates of contraction of posterior distributions based on gaussian process priors. *Preprint*.

[16] van der Vaart, A. W. (1988). *Statistical estimation in large parameter spaces.* CWI Tract, Vol. **44**. Stichting Mathematisch Centrum Centrum voor Wiskunde en Informatica, Amsterdam. MR927725 (89e:62049)

[17] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics. MR1385671 (97g:60035)

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 224
© Institute of Mathematical Statistics, 0000

A Bayesian Semi-Parametric Model for Small Area Estimation

Donald Malec[1] **and Peter Müller**[2]

U.S. Census Bureau and M.D. Anderson Cancer Center

Abstract: In public health management there is a need to produce subnational estimates of health outcomes. Often, however, funds are not available to collect samples large enough to produce traditional survey sample estimates for each subnational area. Although parametric hierarchical methods have been successfully used to derive estimates from small samples, there is a concern that the geographic diversity of the U.S. population may be oversimplified in these models. In this paper, a semi-parametric model is used to describe the geographic variability component of the model. Specifically, we assume Dirichlet process mixtures of normals for county-specific random effects. Results are compared to a parametric model based on the base measure of the Dirichlet process, using binary health outcomes related to mammogram usage.

Contents

1. Introduction

Large national surveys are generally constructed to provide estimates for a diverse group of data users. Estimates, from a single survey, often cover scores of topics and are usually provided for as many population groups as the sample size will support. Population groups of interest consist of both large and demographic groups and subnational areas. As the data users are thought to hold a variety of prior opinions, estimates for a particular group are usually constructed using only sample data from the group in question along with randomization theory. Often estimates are needed for small population subgroups that do not contain enough data for precise estimation. In this situation, small area estimation methods, based on statistical models and parametric exchangeability assumptions that use data from similar

[1] Statistical Research Division, U.S. Bureau of the Census Washington, DC, US
This report is released to inform interested parties of ongoing research to encourage discussion of work in progress. The views expressed on statistical, methodological, technical or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau

[2] Department of Biostatistics, M.D. Anderson Cancer Center, Houston TX 77030

Keywords and phrases: Dirichlet process, Mixture models, National Health Interview Survey

groups are often used. These exchangeability assumptions are generally applied across small areas if there is no a priori evidence to the contrary. Of course, this assumption may be incorrect and, without alternatives to this assumption, will remain throughout the inference.

We propose a semi-parametric, hierarchical model for small area estimation. The model produces estimates for a small area that selectively use data from only subsets of other small areas. The uncertainty about which subsets are valid is automatically accounted for. This model is useful when, *a priori*, one suspects that all small areas are not equally similar but one does not know where the dissimilarities are. The model is an extension of currently used parametric hierarchical models for small area estimation. Although the parametric hierarchical models are flexible enough to adjust for the amount of 'borrowing', it is required that the subsets where 'borrowing' can take place be prespecified. In practice, subset identity may not be known. In addition, some small areas may be outliers and should not be used in drawing inference about other areas and, vice versa, other areas should not be used in drawing inference about them. The semiparametric model, proposed here, will adapt to such outliers. The desired flexibility is accomplished by using a Dirichlet process (DP) prior to define a probability model on possible partitions of small areas. Specifically, the DP prior defines partitions of the set of all counties. Conditional on an assumed partition the model assumes a homogeneous population within all counties in a partitioning subset. This alleviates the difficult problem of specifying a large number of prior probabilities for possible partitions as in Malec and Sedransk (1992). In addition, the DP model is computationally tractable. The model is applied to the National Health Interview Survey (NHIS) for estimates of mammography utilization.

We begin with a historical overview of methodological improvements for small area estimation in Section 2. This is followed, in Section 3, by a description of the NHIS and how the variables were selected for the model of mammography utilization. In Section 4 we motivate and describe the prior model on the random effects. Section 5 discusses details of the implementation by Markov chain Monte Carlo simulation. In Section 6 we apply the model to the NHIS and compare results with alternative approaches.

2. Early and Current Methods for Small Area Estimation

Early estimates for small areas were based on the assumption that an outcome is primarily a function of demographic class. By knowing the demographic characteristics of a small area and estimating the national-level prevalence of outcomes within the same demographic groups, one can derive small area estimates. Estimates based on this assumption are known as 'synthetic estimates'. One of the first publications using this method pertained to synthetic state estimation of disability [14].

A systematic, empirical study of biases in synthetic estimation was undertaken by Schaible et al. [20]. Little data is usually available to evaluate the bias of actual small area estimates. However, using a wide range of outcomes available in the U.S. decennial census, it was shown that the basic homogeneity assumptions underlying synthetic estimation were frequently unwarranted. To alleviate this problem, the biases in synthetic small area estimates can be modeled as random effects. Related estimates of this type have been proposed in Fay and Herriot [3], Dempster and Tomberlin [2] and Battese and Fuller [1]. The model-based approach using random effects opens up the possibility for a variety of models, resulting in realistic estimators that may combine data in a nonlinear manner.

Much of the recent development in small area estimation has concentrated on obtaining estimates from random effect models without resorting to asymptotic approximations using Bayesian Inference and Markov chain Monte Carlo methods (see, e.g., [10]). In current research, the distributional assumptions of the random effects models are being questioned. For example, alternative models which allow for spatial effects have been employed by Ghosh et al. [5] and others. Maiti [7] uses finite mixture models for random effects.

For a recent more detailed review of small area estimation, see Rao [16].

3. The NHIS and the Selection of Variables for Estimation

The NHIS is typical of many national, representative, personal interview surveys. A personal interview can be relatively costly but response rates are generally higher than for other modes of interview and complicated questions, especially questions involving visual cues, can be asked. A substantial contributor to the total cost of a personal interview survey is travel cost. To minimize travel costs, each interviewer usually conducts interviews near his or her residence. Hence, sampling units are usually small geographic land areas. As a result, there is a relatively large sample size in these selected areas but relatively few areas selected across the country. In addition, the overall sample size is selected to produce traditional, randomization-based survey estimates at the national level. Due to high sampling costs, the sample size is usually inadequate to make precise design-based estimates for most small areas. However, estimates are usually desired at smaller levels, for example, for allocation of funds or for assessing and administering health services. The required precision of these estimates is usually of the same order of magnitude as that required for national estimates, resulting in an extreme shortfall of data for making small area estimates.

The NHIS is redesigned every ten years following the decennial census, taking advantage of up-to-date population data. The NHIS is based on a complex sample design of clusters of housing units, called segments. Segments are selected from within a Primary Sampling Unit (PSU), which is usually a group of contiguous counties. PSUs are sampled from strata, constructed to contain PSUs with similar socio-demographic content. See Massey et al. (1989) for details. The segment is the smallest sampling unit for many questions asked in the NHIS. Subsamples, within segments, are also collected on special health topics from either a subsample of households or from people within households. The questions on mammography usage, studied here, are from a subsample of housing units, within segments sampled during years 1993 and 1994.

The sampled outcomes from the NHIS are not an i.i.d. sample from the population. They should not be considered a simple random sample due to both non-response and to possible selection bias due to the sampling design. Nonresponse is accounted for by poststratifying by county, age and race groups. Sample selection effects are countered by including design variables in the model. An example of an alternative approach, in which the sample selection process is modeled directly, is given by Malec et al. [8].

In this paper we consider inference related to one of the questions in the NHIS which asks whether a mammogram had been obtained within the past two years. Periodic mammograms for screening of breast cancer are highly recommended for women of certain ages. Estimation of the mammography utilization rate by age and race for local areas is important for evaluation purposes. Although NHIS data based on single years of age by race within each county is available, as are census population estimates for counties by race and five-year age groups, we chose to evaluate the $D = 6$ demographic groups defined by the three age groups (30-39, 40-49, 50+) crossed with race (black, non-black).

We consider a model:

$$P(y_{idk} = 1) = p_{id},$$

where $y_{idk} = 1$ if the k-th person in demographic group d and county i did receive a mammogram in the past two years, and $y_{idk} = 0$ otherwise. It is assumed throughout that there may be differences between local effects of mammography utilization due to demographic groups. This is why six different prevalence rates, p_{id}, $d = 1, \ldots, D$, are used for each county i. The first step in the model selection process is to select county covariates to account for small area effects. The county variables used are from the Area Resource File [22]. Twenty-two county covariates were evaluated. These covariates comprised regional status of the county, urban variables, percent of population working in white collar jobs, percent of population working in agriculture and construction, urban status, unemployment rate, percent of households that are renters, percent minority (black, asian, hispanic and mexican), degree of education, population density, and median home value. Let x_i denote a vector of all possible county-level covariates and let μ_d denote demographic effects, $d = 1, \ldots, D$. Using a model

$$P_{id} = \text{logit}(p_{id}) = \mu_d + x_i' b, \tag{1}$$

and Schwarz's criterion we identified two county-level covariates: percent of workforce in white collar jobs (x_{i1}) and percent of persons aged 25+ with no more than a ninth-grade education (x_{i2}). A second model-fitting step was used to determine whether interactions of the two county variables with the D demographic groups were present. Using a logistic regression model like (1) with additional interactions of county variables and demographic groups, we again used Schwarz's criterion and by forward stepwise selection identified the best set of interactions. We found interactions of both covariates with the demographic group of 30-39 year old whites. The estimated interaction term amounted to dropping all covariate effects on this particular demographic group. As a last check, plots of average residuals were used to determine that a linear relationship with the logits appeared reasonable.

Although the county covariates used in the model selection are close to the variables that were used to design sampling strata, Malec et al. [10] found that strata effects may still be present in the data. Sampling strata were defined by grouping the county-based PSUs into 198 strata so that the PSUs in each stratum have similar summary measures of socio-economic status [11]. By assuring that each stratum contains a prespecified sample size, a more systematic coverage of the population is possible. We will include stratum effects in the model to account for this important prior knowledge based on the design. We shall use an additional subindex $_s$ in p_{sid} and x_{sid} to indicate strata.

In the prior specification, instead of assuming that all stratum effects are distributed i.i.d., we allow their variability to be distinct within broad mega strata. We defined two mega-strata by grouping the 198 strata into those representing the very large metropolitan areas (like the New York City area) versus all others.

4. A Semi-Parametric Randomized Block Model

Based on the discussion above, we include random stratum effects ν_s as well as random (county by demographic) group effects β_{sid}. Let n_{sid} denote the number of individuals interviewed in demographic domain d, $d = 1, \ldots, D$, county i, $i = 1, \ldots, I$, and stratum s, $s = 1, \ldots, S$. Counties are nested within strata, and demographic domains are crossed with counties.

Let $y_{sid} \sim Bin(n_{sid}, p_{sid})$ denote the number of positive responses among these n_{sid} individuals. Let $s(i)$ denote the stratum containing county i, and let $\beta_i = (\beta_{sid},\ s = s(i), d = 1, \ldots, D)$ denote the D dimensional random effects vector for county i. Since county indices i are unique across stata, the subindex $_s$ in β_{sid} is redundant and we use it only when it helps to clarify the hierarchical structure of the model. We assume a logistic regression of the success probabilities p_{sid} on some quantitative covariates x_{sid} and county-specific and stratum-specific random effects:

$$P_{sid} = \text{logit}(p_{sid}) = x'_{sid}b + \beta_{sid} + v_{s(i)}, \tag{2}$$

with priors

$$\beta_i \sim h(\beta_i), \quad v_s \sim N(0, \delta^2_{m(s)}), \quad b \sim N(m_b, V_b), \quad \delta_m^{-2} \sim \text{Ga}(a_\delta, b_\delta)$$

$m = 1, \ldots, M$. Here $m(s)$ denotes the mega-stratum containing stratum s, and $\text{Ga}(a, b)$ denotes a gamma distribution with mean a/b. In words, the random effects β_i are generated by some distribution $h(\cdot)$, details of which will be discussed below. The stratum random effects are *a priori* normal with random variance δ_m. The model is completed with conjugate hyperpriors for δ_m and b, with the latter possibly non-informative by choosing $V_b^{-1} = 0$. While a large number of experimental units are available to inform inference about the county-specific random effects β_i, only a moderate (198) and small (2) number of stratum and mega-stratum-specific effects are included in the model. Therefore, we use fully parametric random effects distributions for ν_s and δ_m, but a flexible non-parametric model for β_i.

The choice of the prior model $h(\cdot)$ is guided by the following considerations. First, health related outcomes always vary by age and race, but are correlated within counties. Thus we need a multivariate prior on $(\beta_{i1}, \ldots, \beta_{iD})$ allowing for different effects in different demographic domains, and interactions between these effects. Second, the model includes only few covariates, leaving significant heterogeneity due to other un-recorded covariates. To account for such heterogeneity the model needs to allow for possible clusters of sub-populations not identified by the given covariates and overdispersion. Third, as with any health outcome the model needs to accommodate outliers without unduly influencing inference. Finally, the model should be a natural generalization of more conventional multivariate normal random effects distributions.

These considerations lead us to use a mixture of normals prior model. Let $\varphi_\theta(\cdot)$ denote a multivariate normal probability density function with moments $\theta = (\mu, \Sigma)$. Then

$$\beta_i \sim \sum_{j=1}^{\infty} w_j N(\beta_i; \underbrace{\mu'_j, \Sigma'_j}_{\theta'_j}) = \int \varphi_\theta(\beta_i) dG(\theta),$$

where

$$G = \sum_{j=1}^{\infty} w_j \delta_{\theta'_j}.$$

The mixture of normals model (4) allows for heterogeneity, outliers, skewness etc., as desired. The model includes a simple multivariate normal prior model as a special case. By choosing a hyperprior on G which *a priori* favors a few dominating terms in the mixture we formalize the idea that *a priori* we assume simple structure, but as the data dictates the model allows introduction of more complicated structure *a posteriori*, like a discrete mixture of a few dominating normal kernels. This is achieved using a DP prior on G, $G \sim DP(\alpha G_\nu)$. Here G_ν is the (standardized) base measure and $\alpha > 0$ is the total mass parameter. The base measure can possibly depend on further hyperparameters ν. Below, we will specify a base measure and hyperprior on ν and the total mass parameter α. See Ferguson [4] for a complete description of the DP; a discussion of DP and DP mixture models like (4) can be found in, e.g., [6], [13], and [23].

We summarize some properties which are relevant in the context of our application.

1. The base measure G_ν has an interpretation as the prior mean for the random measure; the total mass parameter α is a precision parameter. For any measurable set A we have
$$E\{G(A)\} = G_\nu(A) \text{ and } Var\{G(A)\} = G_\nu(A)\{1 - G_\nu(A)\}/(1 + \alpha).$$
2. The random measure G is a.s. discrete. Let F_0 and F denote the c.d.f. of G_ν and G, respectively. Then F can be thought of as a random step function approximating F_0. The size w_j of the steps depends on α. The larger the α, the smaller the weights, i.e., steps sizes, w_j.
3. Consider a random sample $\theta_i \sim G$, $i = 1, \ldots, n$. Because of the discreteness of G, with positive probability some of the θ_i will be identical. Specifically, we have the prior probabilities $P(\theta_i = \theta_j | \theta_h) = 1/(\alpha + n - 1)$ for $j < i,\ i \notin \{j, h\}$.

A commonly used device in posterior simulation when the likelihood involves a mixture like in (4) is to break the mixture by introducing latent variables [18]. We implement this by rewriting the DP mixture model (4) as

$$\beta_i \sim N(\beta_i; \mu_i, \Sigma_i), \quad (\mu_i, \Sigma_i) \sim G, \quad G \sim DP(\alpha, G_\nu). \tag{3}$$

The model is completed by specifying a base measure G_ν,

$$G_\nu(\mu, \Sigma) = N(\mu; m, B)\ W[\Sigma^{-1}; s, (sS)^{-1}], \tag{4}$$

and a hyperpior for $\nu = (m, B, S)$ and α,

$$S \sim W[q, (R/q)], \quad m \sim N(a, A), \quad B^{-1} \sim W[c, (cC)^{-1}], \quad \alpha \sim G(a_\alpha, b_\alpha). \tag{5}$$

Here $W(n, A)$ denotes a Wishart distribution with scalar parameter n and matrix parameter A, and $G(a, b)$ denotes a Gamma distribution with shape parameter a and scale parameter b, parameterized such that the expected value of $G(a, b)$ is a/b.

The model is summarized in Figure 1.

We will refer to model (2) together with prior (4) or, equivalently, (3) and hyperprior (4) and (5) as the MDP (Dirichlet process mixture) model. Alternatively to the MDP prior one could choose fully parametric normal mixture models as proposed by, for example, Roeder and Wasserman [19] or Richardson and Green [17]. Based on experience with similar models, we would expect predictive inference,

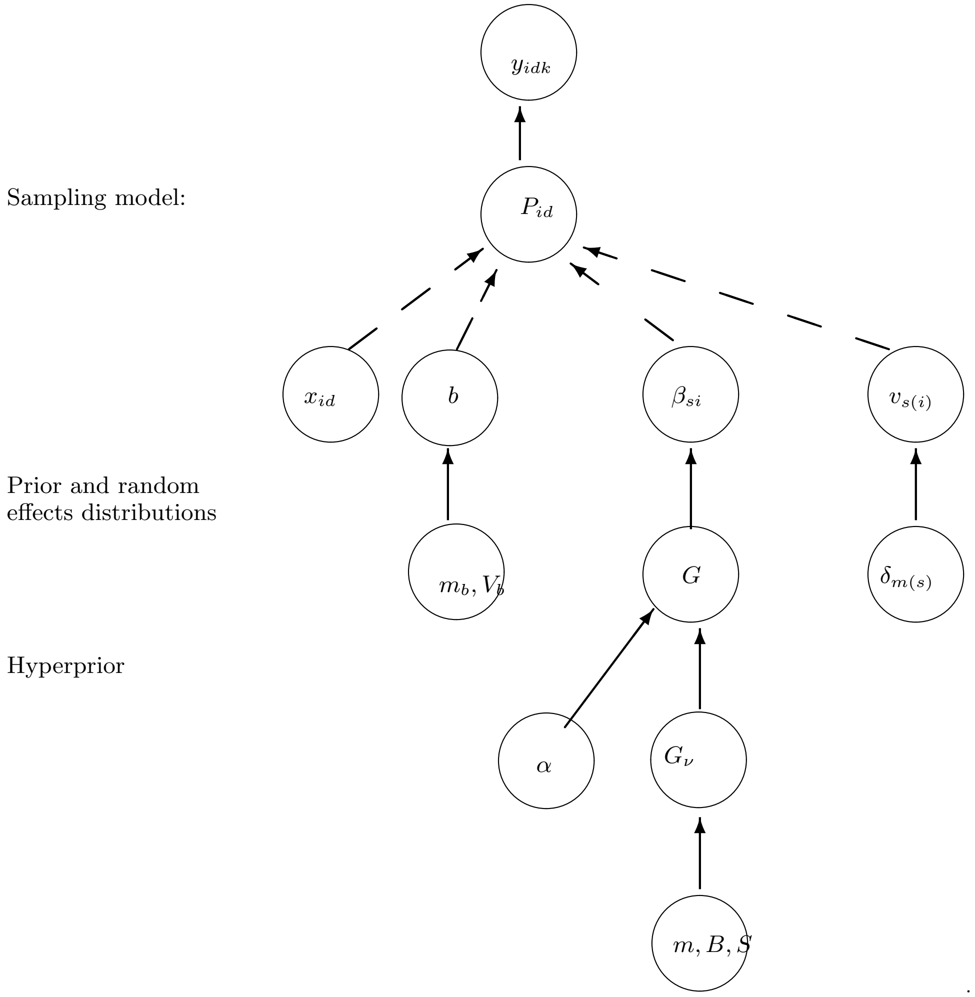


Fig 1. *Graphical model representation. Circles indicate the variables, including the data, parameters, covariates and hyperparameters. Solid arrows indicate that the probability models for the target variable are indexed by the variable at the sources of the arrow. Dashed lines represent the deterministic expression for P_{sid}. Fixed hyperparameters are not indicated.*

like the posterior predictive inference for state totals, including counties without samples, not to be much different under these models than inference using mixture models with DP priors. Still, we prefer the semi-parametric formulation for several reasons. First, the prior parameters G_ν and α in the DP prior model have a straightforward interpretation as the mean and dispersion parameters ('equivalent prior sample size'). It is not immediately clear what implication certain choices for priors on the parameters of the mixture model would have on the marginal distribution of y in the finite mixture model. Second, the computational effort is comparable in all three models.

5. Posterior Simulation

Model (2) with prior (3) – (5) can be estimated by Markov chain Monte Carlo simulation. We shall use β for $\beta = (\beta_i, i = 1, \ldots, I)$, μ for $\mu = (\mu_1, \ldots, \mu_I)$, etc. Let y denote the data vector, $y_i = (y_{sid}, d = 1, \ldots, D)$ denote the data vector for county i (with $s = s(i)$ equal to the stratum containing county i), and $y_s = (y_{sid},\ i : s(i) = s,\ d = 1, \ldots, D)$ denote the data vector for stratum s. An entry of the type $a|b, c, y$ indicates that parameter a is being updated conditional on the currently imputed values of b and c and the data y. Absence of a variable d in the conditioning set indicates conditional independence of a and d or that the model is being marginalized over d. We outline the sequence of the updating scheme, with details discussed below.

Let

$$(i)\ b|\beta, v, y, \quad (ii)\ \beta_i|\mu_i, \Sigma_i, b, v, y_i, \quad (iii)\ v|b, \beta, \delta, y, \quad (iv)\ \delta|v, \quad (v)\ \mu, \Sigma, \nu, \alpha|\beta.$$

Steps (i) - (v) describe the transition probability of a Markov chain in $(b, \beta, v, \mu, \Sigma)$. Step (v) refers to updating the parameters of the mixture model $h(\cdot)$, including $\mu_i, \Sigma_i,\ i = 1, \ldots, I$, the hyperparameters ν of the base measure G_ν, and the total mass parameter α. In a straightforward Gibbs sampler implementation each of the updating steps would draw from the respective complete conditional posterior distribution to generate a new imputed value of the respective parameter. Unfortunately this is only possible for Step (iv). In all other steps the conditional posterior distribution is not in a format allowing efficient random variate generation. The appropriate MCMC implementation for steps (i) - (iii) is explicated in the Appendix. Step (v) is described in [15].

By construction, the stationary distribution of this chain is the desired posterior distribution $p(b, \beta, v, \mu, \Sigma|y)$. Most posterior inferences take the form of integrals with respect to the posterior, such as the posterior mean $\bar{b} = \int b\, dp(b, \beta, v, \mu, \Sigma|y)$, and such posterior integrals can be approximated by ergodic averages. For example $\bar{b} \approx 1/T\ \sum_{t=1}^{T} b^t$, where b^t denotes the value of b after t iterations of the Markov chain.

The aim of the small area estimation model (2) is to provide inference for both the sampled units and for the sampling units for which no data is available, and to summarize such inference by subpopulations of interest, like states, etc. The posterior simulation described earlier in this section allows us to compute such inference with minimal additional computational cost. Index with $i = I + 1, \ldots, J$ counties not included in the available sample. Denote with $N_{sid}, i = 1, \ldots, J,\ s = s(i),\ d = 1, \ldots, D$, the total populations in each cell of county and demographic domain. The population totals N_{sid} and covariates x_{sid} are available for all counties $i = 1, \ldots, J$, from census data. To compute inference for state totals, we proceed by the following steps. Index the states with $a = 1, \ldots, A$. Denote with $a(i)$ the index of the state a containing county i. Let Y_{sid} denote the *total* number of subjects in county i and demographic domain d who had a mammogram within the last two years. Let $Y_a = \sum_{\{i:a(i)=a\}} \sum_d Y_{sid}$ denote the total for state a.

At each iteration of the Markov chain Monte Carlo simulation we have imputed values for b, β_{sid}, $i = 1, \ldots, I$, and v_s, $s = 1, \ldots, S$. To impute random effects for counties absent in the sample, $i = I+1, \ldots, J$, we simulate values for $\theta_i = (\mu_i, \Sigma_i)$, $i = I+1, \ldots, J$, using the following probabilities:

$$P(\theta_i = \theta_j) = 1/(\alpha + i - 1), \quad j = 1, \ldots, i-1,$$
$$P(\theta_i \neq \theta_j,\ j < i) = \alpha/(\alpha + i - 1) \text{ and } P(\theta_i | \theta_i \neq \theta_j, j < i) = G_\nu(\theta_i), \tag{6}$$

and $P(\beta_i|\theta_i) = N(\mu_i, \Sigma_i)$. Given imputed values for the random effects for *all* counties and all strata, and for the logistic regression coefficients b, we can now impute success probabilities p_{sid} for all counties, and simulate total counts $Y_{sid} \sim Bin(N_{sid}, p_{sid})$ for all counties. Adding up the state totals $Y_a = \sum_{i:a(i)=a} \sum_d Y_{sid}$ we get simulated values $Y_a \sim p(Y_a|y)$ from the posterior distribution on the state totals.

6. The National Health Interview Survey

Figure 2 shows the final inference on the state totals (as a percentage of total population N_a for state a). Note how inference from the semi-parametric model corresponds to a compromise between the oversmoothing synthetic estimates, and the overfitting empirical means (based on observed data in each state only). The states are sorted by decreasing imputed probability to facilitate comparison. The posterior standard deviations in the state percentages are between 2 and 6 percentage points. This is the uncertainty due to estimating the state totals based on the partial sample information only. Another source of error is due to evaluating the posterior means numerically by simulation only. The corresponding uncertainties are negligible relative to the inherent posterior standard deviations. Figure 3 shows the same information in a geographical map of the U.S.

Figure 4 shows a summary of the imputed random effects β_{sid}. Clearly, assuming random effects for different domains to be i.i.d. generated from some univariate random effects distribution would be inappropriate. Posterior correlations of β_{sid} across d range from -0.5 to 0.6. Figure 5 shows some features of the estimated mixing distribution G. A fully parametric model would correspond to either a point mass G, or a conjugate multivariate normal (in μ) measure G, corresponding to a hierarchical model. Neither seems to be a good approximation to G. Figure 6 shows the posterior distribution on the number of distinct μ_i in the mixture model (4). The estimated stratum-specific random effects v_s are shown in Figure 7.

7. Discussion

The proposed approach addressed some limitations of currently used models in small area estimation by substituting a non-parametric model for the random effects distribution of the county- and domain-specific random effects. Resulting inference allows more fidelity to the data than oversmoothing synthetic estimates, without completely abandoning the borrowing of strength across counties, domains and strata as formalized in the hierarchical model. Still, several assumptions remain

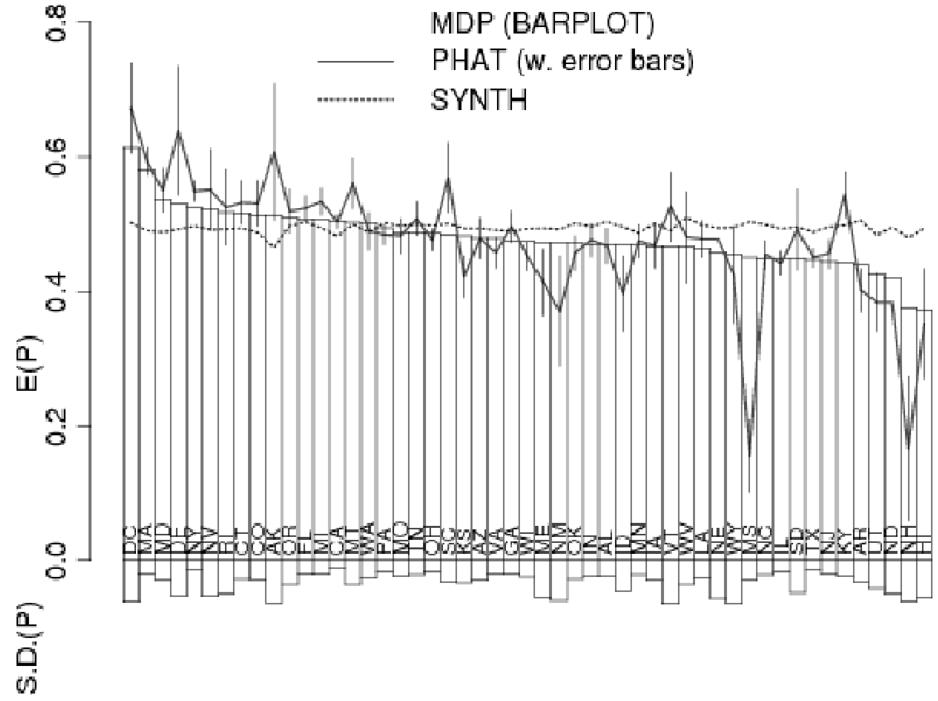


Fig 2. *Estimated state totals $E(Y_a|y)$ under the proposed model (bar plot), the synthetic method (dashed line) and as sample averages over observed samples in each state (dotted line). The short bars below the horizontal axis show one posterior s.d. $SD(Y_a|y)$ for the state totals. Error bars at each sample average estimate indicate corresponding sampling errors (assuming that all observations in a given state were independent). The sample did not include any data for NE and ND. Thus, there is no 'sample estimate' for these states.*

in the model. This includes the linear regression implied by the term $x'_{sid}b$ in (2) and the normal prior on the stratum-specific random effects. While residual plots (not shown here) indicate that the linear regression assumption was not severely contraindicated by the data, a less restrictive approach is desirable. Given the massive data available to fit the model, more general non-parametric regression models are possible. Also, the skewed distribution in Figure 7 indicates the need for less restrictive prior models on the stratum-specific random effects v_s. Both generalizations are possible with methods discussed in this paper and will be pursued in future research.

Other extensions could add additional robustness by, for example, replacing the informal variable selection that we used for the demographic covariates by model averaging, by allowing stratum-specific random effects for the regression on demographic covariates, and by including spatial smoothing. Besides the basic test of fit by residual plots one could carry out specific model validation to investigate the use of such model extensions.

Appendix: Resampling the logistic regression parameters

Steps (i), (ii) and (iii) in Section 5 require the updating of logistic regression parameters. We describe an independence chain implementation for Step (i). Steps

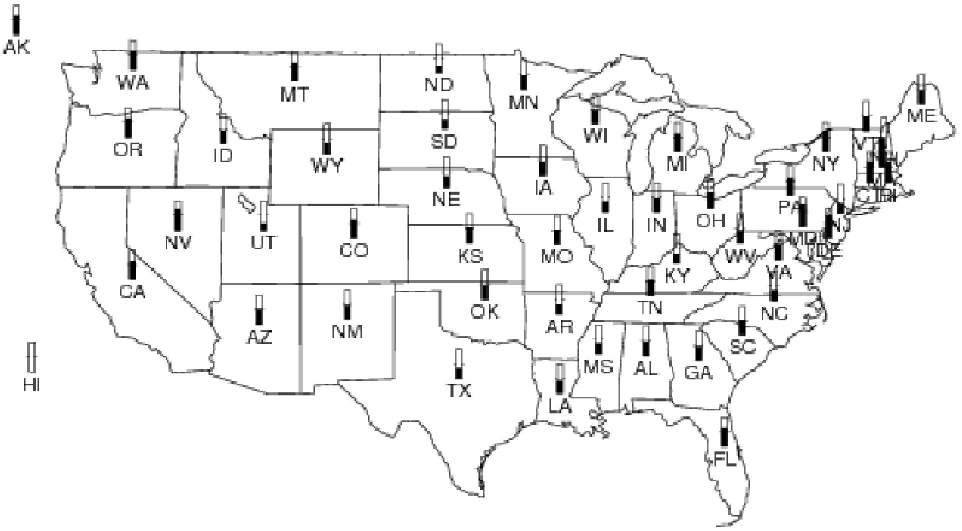


Fig 3. *Posterior predictive means per state for the percentage of women who had a mammogram done in the past two years. The solidly filled fraction of the thermometer in each state corresponds to the estimated mammogram utilization in that state. To highlight the differences between states, the measurements are shown relative to the minimum and maximum estimated percentage usage. A fully filled thermometer corresponds to 58.4%, an entirely empty thermometer corresponds to 37%.*

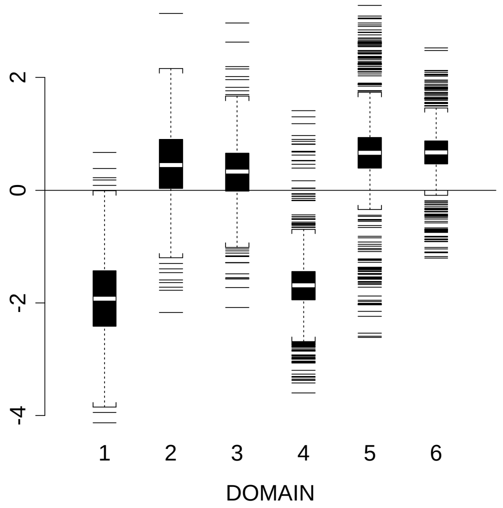


Fig 4. *Boxplots of imputed random effects β_{sid}, split by domain d, based on the random effects as imputed after 3000 iterations of the simulation.*

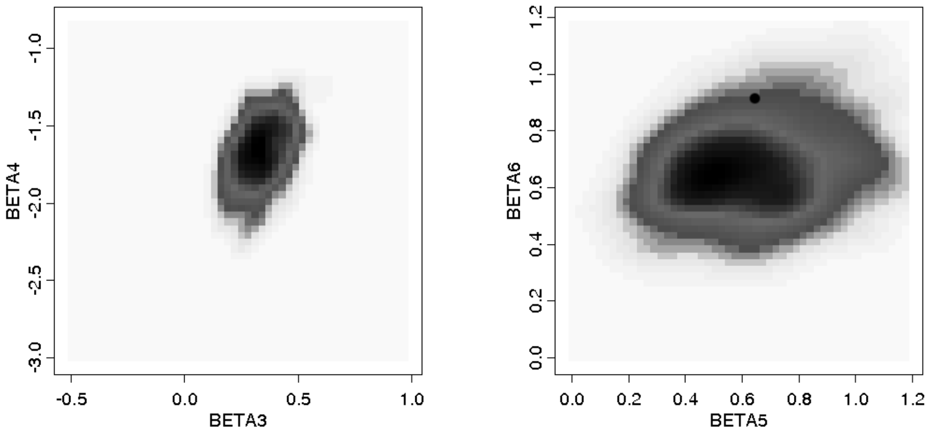


FIG 5. *Estimated mixing distribution $G(\mu)$.*

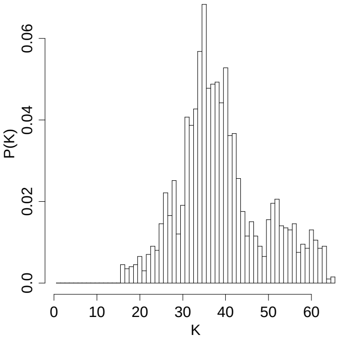


FIG 6. *The posterior distribution $p(k|y)$ on the number k of distinct μ_i in the mixture model (4). There is a total of $I = 3060$ counties in the sample, i.e., $1 \leq k \leq 3060$.*

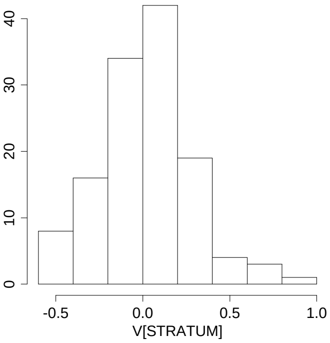


FIG 7. *$E\{v_s|y\}$, $s = 1..S$. The figure shows a histogram of imputed stratum-specific effects v_s.*

(*ii*) and (*iii*) allow analogous implementations.

To update $b = (b_1, \dots, b_p)$, we first compute the moments of the normal approximation to the likelihood based on the linearized logistic link function in a logistic regression model (cf., [12], p41). Denote the moments of the resulting p-dimensional normal approximation by (m_1, V_1).

The $N(m_1, V_1)$ approximation of the likelihood is now combined with the $N(m_b, V_b)$ prior to obtain a bivariate normal approximation of the conditional posterior $p(b|\beta, v, D)$. Let m_3 and V_3 denote the moments of this approximation. Draw a candidate $\tilde{b} \sim N(m_3, V_3)$, and compute the following acceptance probability. Let ω denote the full parameter vector, and $\tilde{\omega}$ the parameter vector with b replaced by $\tilde{b}$.

$$a(b, \tilde{b}) = \min\left\{1, \frac{p(y|\tilde{\omega})\ \varphi(b; m_1, V_1)}{p(y|\omega)\ \varphi(\tilde{b}; m_1, V_1)}\right\}, \tag{7}$$

where $\varphi(x; m, V)$ denotes a normal density with moments m and V, evaluated at x. Replace b by $\tilde{b}$ with probability a.

To derive expression (7) consider the general expression for acceptance probabilities in a Metropolis–Hastings algorithm (see, for example, [21]). Denote with $g(\tilde{\omega}|\omega)$ the proposal distribution. Then

$$a(\omega, \tilde{\omega}) = \min\left\{1, \frac{p(\tilde{\omega}|y)}{p(\omega|y)} \frac{g(\omega|\tilde{\omega})}{g(\tilde{\omega}|\omega)}\right\}.$$

Substitute for $g(\tilde{\omega}|\omega)/g(\omega|\tilde{\omega})$:

$$\frac{\varphi(\tilde{\beta}_i; m_3, V_3)}{\varphi(\beta_i; m_3, V_3)} = \frac{\varphi(\tilde{\beta}_i; m_1, V_1)\varphi(\tilde{\beta}_i; m_2, V_2)}{\varphi(\beta_i; m_1, V_1)\varphi(\beta_i; m_2, V_2)}$$

Also, $p(\tilde{\omega}|y)/p(\omega|y) = [p(y_i|\tilde{\omega}_i)\varphi(\tilde{\beta}_i; m_2, V_2)]/[p(y_i|\omega_i)\varphi(\beta_i; m_2, V_2)]$. The factors $\varphi(\ \cdot\ ; m_2, V_2)$ in $p(\tilde{\omega}|y)/p(\omega|y)$ cancel against the same factors in $g(\tilde{\omega}|\omega)/g(\omega|\tilde{\omega})$ and we arrive at (7).

References

[1] Battese, G. and Fuller, W. (1981). Prediction of county crop areas using survey and satellite data. In *Proceedings of the American Statistical Association, Survey Research Section*, 500–505.

[2] Dempster, A. and Tomberlin, T. (1980). The Analysis of Census Undercount form a Postenumeration Survey. In Proceedings of the Conference on Census Undercount, Arlington, VA, 88–94.

[3] Fay, R. and Herriot, R. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *J. Amer. Statist. Assoc.* **74**, 269–277.

[4] Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–230.

[5] Ghosh, M., Natarajan, K., Stroud, T. W. F. and Carlin, B. (1998). Generalized linear models for small-area estimation *J. Amer. Statist. Assoc.* **93**, 273–282.

[6] MacEachern, S. N. and Müller, P. (2000). Efficient MCMC Schemes for Robust Model Extensions using Encompassing Dirichlet Process Mixture Models. In Ruggeri, F. and Ríos-Insua, D. (eds), *Robust Bayesian Analysis*, Springer-Verlag: New York, NY, 295–316.

[7] Maiti, T. (2001). Robust generalized linear mixed models for small area estimation. *J. Statist. Plan. Inf.* , **98**, 225-238.

[8] Malec, D., Davis W. and Cao, X. (1999). Model Based Small Area Estimates of Overweight Prevalence Using Sample Selection Adjustment, *Statist. Med.* **18**, 3189–3200.

[9] Malec, D. and Sedransk, J. (1992). Bayesian Methodology for Combining Results From Different Experiments When the Specifications for Pooling Are Uncertain. Biometrika **79**, 593–601.

[10] Malec, D., Sedransk, J., Moriarity, C. and Le Clere, F. (1997). State estimates of disability using a hierarchical model: An empirical evaluation. *J. Amer. Statist. Assoc.* **92**, 815–826.

[11] Massey, J.T., Moore, T.F., Parsons, V.L. and Tadros, W. (1989). Design and Estimation for the National Health Interview Survey, 1985-94. National Center for Health Statistics, Vital and Health Statistics, **2**, 110.
(http://www.cdc.gov/nchswww/data/sr2_110.pdf)

[12] McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models, 2 Ed.* Chapman and Hall, New York, NY.

[13] Müller, P. and Quintana, F. (2004). Nonparametric Bayesian Data Analysis, *Statistical Science*, **19**, 95–110.

[14] National Center for Health Statistics (1968). Synthetic State Estimates of Disability. PHS Publication No. 1759. Washington, D.C., U.S. Government Printing Office.

[15] Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models, *Journal of Computational and Graphical Statistics*, **9**, 249–265.

[16] Rao, J. N. K. (2003). *Small Area Estimation.*, Wiley, Hoboken, NJ.

[17] Richardson, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Roy. Statist. Soc. Ser. B* **59**, 731–792.

[18] Robert, C. (1995). Inference in mixture models. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (Eds.), *Markov Chain Monte Carlo in Practice.* Chapman and Hall, London.

[19] Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *J. Amer. Statist. Assoc.* **92**, 894–902.

[20] Schaible, W., Brock, D. and Schnack, G. (1977). An empirical comparison of the simple inflation, synthetic and composite estimators for small area statistics. In *Proceedings of the American Statistical Association, Social Statistics Section*, 1017–1021.

[21] Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Statist.* **4**, 1701–1762.

[22] U.S. Department of Health and Human Services (1989). *The Area Resource File (ARF) System.* Office of Data Analysis and Management (ODAM) Report No. 7–89.

[23] Walker, S., Damien, P., Laud, P., and Smith, A. (1999). Bayesian nonparametric inference for distributions and related functions (with discussion). *Journal of the Royal Statistical Society, Series B* **61**, 485–527.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 238

arXiv: math.PR/0000000

A Hierarchical Bayesian Approach for Estimating the Origin of a Mixed Population

Feng Guo[1], Dipak Dey[2,*] and Kent Holsinger[3]

Abstract: We propose a hierarchical Bayesian model to estimate the proportional contribution of source populations to a newly founded colony. Samples are derived from the first generation offspring in the colony, but mating may occur preferentially among migrants from the same source population. Genotypes of the newly founded colony and source populations are used to estimate the mixture proportions, and the mixture proportions are related to environmental and demographic factors that might affect the colonizing process. We estimate an assortative mating coefficient, mixture proportions, and regression relationships between environmental factors and the mixture proportions in a single hierarchical model. The first-stage likelihood for genotypes in the newly founded colony is a mixture multinomial distribution reflecting the colonizing process. The environmental and demographic data are incorporated into the model through a hierarchical prior structure. A simulation study is conducted to investigate the performance of the model by using different levels of population divergence and number of genetic markers included in the analysis. We use Markov chain Monte Carlo (MCMC) simulation to conduct inference for the posterior distributions of model parameters. We apply the model to a data set derived from grey seals in the Orkney Islands, Scotland. We compare our model with a similar model previously used to analyze these data. The results from both the simulation and application to real data indicate that our model provides better estimates for the covariate effects.

Contents

1. Introduction

Fisheries scientists and marine biologists are often faced with the problem of identifying proportions of individuals in a single catch that come from different stocks.

[*] Corresponding author

[1] Department of Statistics, Virginia Tech, Blacksburg, VA 24060, e-mail: feng.guo@vt.edu

[2] Department of Statistics, University of Connecticut, Storrs, CT 06269, e-mail: dey@stat.uconn.edu

[3] Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, e-mail: kent@darwin.eeb.uconn.edu

AMS 2000 subject classifications: Primary 60K35, 60K35; secondary 60K35

Keywords and phrases: Hierarchical Bayes, MCMC, Multinomial

Estimating these proportions is necessary for evaluating the effect of commercial fisheries on individual fisheries stocks and for understanding the ecological factors that influence the relative contributions of different stocks. Similarly, those who study marine mammals are often interested in identifying the source populations for newly founded colonies as well as environmental or demographic factors that influence the relative contributions of different sources. The increasing ease with which genetic data are collected and the tendency for populations of species to become genetically differentiated over time has led to the increase in using genetic markers to estimate the proportional contribution of source populations to mixed stocks. The rationale is simple: allele frequencies are likely to differ among source populations, and genotype frequencies in the harvest site/new habitat are determined by the proportional contributions of the source populations. Both the differences among source populations and the mixture proportions can be detected by appropriate statistical models.

Several methods have been developed for the inference of the proportional contribution, $\mathbf{m}$, where m_i is the percentage of individuals in the mixed population originating from source i. Conditional Maximum Likelihood Estimates (MLEs) have been widely used [8, 9]. The conditional MLE assumes the sampled source populations are exhaustive lists of all possible sources and the allele frequency of the sources are known without error. Neither assumption is satisfied for real samples. Smouse *et al* [12] extended conditional MLEs to unconditional MLEs in which the source allele frequencies are treated as estimated parameters and unknown sources are allowed to be presented in the model.

In recent years, several authors have applied Bayesian methods to the stock mixture problem [2–4, 9, 11]. One advantage of a Bayesian model is that the influence of the genotype data from the mixture population on the estimation of source allele frequency is fully incorporated through the joint likelihood and is reflected in the posterior distribution of allele frequencies. Another advantage is that we can include non-genetic information in the model through appropriate priors. We can, for example, set the prior distribution of mixture proportions as a function of ecological or demographic parameters.

Until recently interest has largely focused on inference of the proportional stock contributions, but there is increasing interest in understanding the ecological factors that influence those proportions, e.g., source population size and the distance between the source and mixture habitat [3, 9]. While including these relationships is difficult to implement in classical models, a hierarchical Bayesian model can easily incorporate these relationships into the prior for the vector of proportional contributions, $\mathbf{m}$. Existing approaches for inference on $\mathbf{m}$ either use an additive logistic transformation with parameters assumed to have normal priors on the logistic scale [9] or model $\mathbf{m}$ directly on the simplex using a Dirichlet with parameters assumed to be lognormal [3]. In this paper, we propose an alternative formulation of Dirichlet prior structure that is both more efficient in separating mean and variance effects and also directly interpretable. We conduct a simulation study to demonstrate our approach and investigate its performance by varying the level of differentiation among source populations and number of genetic markers. We apply our model to the data derived from grey seals in the Orkney Islands and compare our results with those obtained with a Dirichlet-lognormal prior [3] and with a model using a uniform prior for $\mathbf{m}$.

2. Models

We conduct our analysis in a Bayesian framework. The parameters, such as (relative) allele frequency, $\mathbf{P}$, and proportional contribution, $\mathbf{m}$, are considered as random variables and the statistical inference is based on the posterior distributions of parameters. In this analysis, the genetic data from source and mixture populations are included in the likelihood function and the covariate information is included through a hierarchical prior structure.

The likelihood of the data is derived from genetic theory. The genetic data consist of two parts: the allele counts from source populations and genotype counts from the mixed population. Gaggiotti's model [3] deals with the situation where there is one new colony and several source populations that might contribute to the founding group of the new colony. The data are collected from the first generation descendants of migrants, but the model allows for non-random, assortative mating, i.e., individuals from the same source population are more likely to mate with one another than those from different source populations.

Consider a first generation descent individual k in the new colony whose mother is from population i and father is from population j. Denote $P(\mathbf{y}_k|ij)$ as the probability that this individual has genotype $\mathbf{y}_k$, which includes L loci. Denote (a_{1lk}, a_{2lk}) as the genotype of individual k at locus l. First consider individuals with both parents from the same population, i.e., $i = j$. Assume mating is random among those individuals and Hardy–Weinberg Equilibrium (HWE), which states that the frequency of the heterozygous genotype is twice that of the homozygous genotype, holds. The probability of genotype $\mathbf{y}_k$ is,

$$P(\mathbf{y}_k|ii) = \prod_{l=1}^{L} \delta_{lk} p_{a_{1lk;li}} p_{a_{2lk;li}}, \tag{2.1}$$

where $p_{a_{1lk;li}}$ is the allele frequency of a_{1lk} at locus l in population i, and δ_{lk} is an indicator variable defined as

$$\delta_{lk} = \begin{cases} 1 & \text{if } a_{1lk} = a_{2lk} \\ 2 & \text{if } a_{1lk} \neq a_{2lk}. \end{cases}$$

When the parents are from different populations, i.e., $i \neq j$, $P(\mathbf{y}_k|ij)$ is given by

$$P(\mathbf{y}_k|ij) = \prod_{l=1}^{L} (p_{a_{1lk;li}} p_{a_{2lk;lj}} + \gamma_{lk} p_{a_{2lk;li}} p_{a_{1lk;lj}}), \tag{2.2}$$

where

$$\gamma_{lk} = \begin{cases} 0 & \text{if } a_{1lk} = a_{2lk} \\ 1 & \text{if } a_{1lk} \neq a_{2lk} \end{cases}.$$

Parameter γ_{lk} indicates that when alleles at a locus are different, there are two different ways of assigning them to parents in different source populations.

When mating happens assortatively, i.e., individuals tend to mate with those from the same source population, HWE is not valid. The assortative mating can be modeled by an assortative mating coefficient $\omega \in (0, 1)$. Specifically, a proportion ω of first generation descendants arise from assortative mating among individuals from the same source and a proportion $1 - \omega$ arise from random mating among all migrants. Consequently, the likelihood of finding the genotype $\mathbf{y}_k$ in a sample from the new colony is as follows:

$$\begin{aligned} P(\mathbf{y}_k|\omega, \mathbf{P}, \mathbf{m}) =&\omega \sum_{i=1}^{I} m_i P(\mathbf{y}_k|ii)+ \\ &(1-\omega)\left[\sum_{i=1}^{I} m_i^2 P(\mathbf{y}_k|ii) + \sum_{i=1}^{I}\sum_{j\neq i} m_i m_j P(\mathbf{y}_k|ij)\right], \end{aligned} \tag{2.3}$$

where $P(\mathbf{y}_k|ii)$ and $P(\mathbf{y}_k|ij)$ is as in (2.1) and (2.2).

Since HWE is assumed for source populations, the genotype frequency is determined by the allele frequency. It is easy to show that the likelihood associated with the genotype frequencies is equivalent to a multinomial with parameters corresponding to the allele frequencies and response variables as allele counts from source populations. If we assume independence among the genotype counts across loci and populations, the likelihood function for source allele counts is a product multinomial:

$$P(\mathbf{N}|\mathbf{P}) \sim \prod_{i=1}^{I}\prod_{l=1}^{L}\prod_{j=1}^{A_l} p_{jli}^{N_{jli}},$$

where N_{jli} is the allele count for source population i at locus l for allele a_j and A_l is the number of alleles at locus l.

The prior distributions for ω and $\mathbf{P}$ reflect prior beliefs about plausible values for these parameters. We choose vague/noninformative priors to reflect the fact that we have no reason to expect particular values for these parameters. In particular, we assume an uniform prior on $(0, 1)$ for the assortative mating coefficient ω. For allele frequencies in the source populations, $\mathbf{P}$, we assume a Dirichlet prior

$$\pi(\mathbf{p}|\alpha) \propto \prod_{i=1}^{I}\prod_{l=1}^{L}\prod_{j=1}^{A_l} p_{jli}^{\alpha_{jl}-1}.$$

As there is no previous data or preference for $\mathbf{P}$, it is reasonable to take α's all equal to 1, leading to a symmetric Dirichlet prior with parameter 1.

The key to this analysis is how to incorporate the demographic/environmental factors into the estimation of the proportional contribution $\mathbf{m}$. Information on $\mathbf{m}$ is obtained only indirectly through its influence on genotype frequencies in the mixed population. Thus, there is no simple data likelihood connecting $\mathbf{m}$ and the covariates. In a Bayesian framework, however, we can assign an informative prior for $\mathbf{m}$ containing the information from the covariates. The challenge in doing a regression type analysis is that the sum of the components of $\mathbf{m}$ must be equal to 1. Since covariates have to be considered for every source, an ordinary linear model or logit transformation does not fit here. Okuyama and Bolker [9] overcome this problem by using an additive log ratio transformation based on an additive logistic normal distribution. However, a baseline population has to be selected and the covariates need to be adjusted according to the baseline population, which makes it hard to accommodate multiple covariates and the interpretation of the coefficients is not straightforward. Gaggiotti *et al*[3] use a hierarchical Dirichlet prior to address this problem. In Gaggiotti's setup, the first level prior distribution for $\mathbf{m}$ is a Dirichlet distribution in which the individual parameters follow a lognormal distribution, i.e.,

$$\mathbf{m} \sim \mathcal{D}(\boldsymbol{\psi})$$

$$\log(\psi_i) \sim \mathcal{N}(\mu_i, \sigma^2) \tag{2.4}$$

$$\mu_i = \alpha_0 + \sum_{r=1}^{p} \alpha_r G_{ri},$$

where G_{ri} is the value of the r^{th} factor for source population i and $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_p)$ is the vector of regression coefficients. The value of the covariates are standardized and the prior for α_r is $\mathcal{N}(0, \sigma_p^2)$. The covariates affect the prior through the parameter $\boldsymbol{\psi}$.

We introduce a new hierarchical prior structure for the mixture proportions $\mathbf{m}$. The first level prior for $\mathbf{m}$ is a Dirichlet distribution with parameters $((1-\rho)/\rho)\boldsymbol{\varphi}$, where $\rho \in (0,1)$ and $\boldsymbol{\varphi}$ are the hyperparameters of the prior subject to the constraint $\sum_{i=1}^{I} \varphi_i = 1$. This form of prior is widely used in population genetics for its relationship with the measure of population differentiation, e.g., Wright's F_{ST} [1, 5, 6]. Due to the fact that the covariates are also observed, another hierarchical level is added to incorporate the randomness. A Dirichlet prior with parameter $\boldsymbol{\eta}$ is assigned to $\boldsymbol{\varphi}$. The covariates are included in the model by setting the logarithm of $\boldsymbol{\eta}$ to be a function of a linear combination of the covariates, i.e.,

$$\mathbf{m} \sim \mathcal{D}(\frac{1-\rho}{\rho}\boldsymbol{\varphi}),$$

$$\boldsymbol{\varphi} \sim \mathcal{D}(\boldsymbol{\eta}), \tag{2.5}$$

$$\log(\eta_i) = \alpha_0 + \sum_{r=1}^{p} \alpha_r G_{ri}.$$

Since the covariates are normalized, the regression coefficients α_r's are assumed to be independent of each other. Normal priors with mean zero and a large variance, $\sigma_p^2 = 10$, are assigned to parameter $\boldsymbol{\alpha}$. The full model is:

$$\begin{aligned} &\pi(\mathbf{P}, \omega, \mathbf{m}, \rho, \boldsymbol{\varphi}, \boldsymbol{\alpha} | \mathbf{Y}, \mathbf{N}) \\ &\qquad \propto \; p(\mathbf{Y}|\mathbf{P}, \omega, \mathbf{m})\pi(\mathbf{m}|\rho, \boldsymbol{\varphi})\pi(\boldsymbol{\varphi}|\boldsymbol{\alpha})\pi(\boldsymbol{\alpha})\pi(\omega)p(\mathbf{N}|\mathbf{P})\pi(\mathbf{P}), \end{aligned} \tag{2.6}$$

where $\mathbf{Y}$ is the genotype data of the new colony, $\mathbf{N}$ is the allele count in source populations, and $\mathbf{P}$ is the allele frequency. The Directed Acyclic Graphs (DAGs) of the Dirichlet-Dirichlet and Dirichlet-Lognormal models are presented in Figure 1 and Figure 2.

The prior structure of our model puts the support of ρ between $(0,1)$ and the value of $((1-\rho)/\rho)\varphi_i$ on the entire positive line. This setup brings several advantages. First, a natural vague prior for ρ is simply a uniform distribution between 0 and 1, $\mathcal{U}(0,1)$. Second, when we use $\mathcal{U}(0,1)$ as a prior, the posterior mean of ρ can be used as an indicator of the dispersion of the regression of $\mathbf{m}$ on demograpic/environmental factors. The variance associated with component m_i is $\varphi_i(1-\varphi_i)\rho$, and $1-\rho$ is roughly the proportion of variance in $\mathbf{m}$ explained by the regression (cf.,[6]). Third, an informative prior on ρ can be used to influence the variance of the prior and the relative weight of environmental covariates and genetic data on the posterior of $\mathbf{m}$. To see this, observe that $(1-\rho)/\rho$ increases when ρ decreases and the variance of $\mathcal{D}(((1-\rho)/\rho)\boldsymbol{\varphi})$ is decided by the absolute value of $((1-\rho)/\rho)\boldsymbol{\varphi}$. Thus a small value of ρ corresponds to a small variance. As the variance of the prior usually decides the relative weight of the prior information

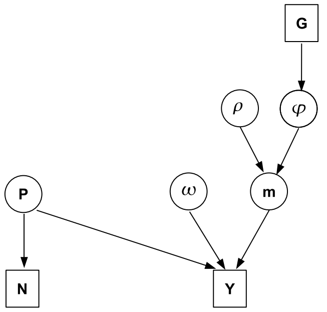


FIG 1. *Dirichlet-Dirichlet Model*

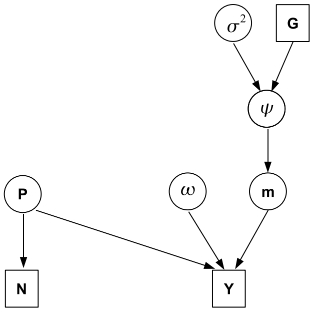


FIG 2. *Dirichlet-Lognormal Model*

on the posterior distribution, ρ indicates the relative weight of the covariates in the posterior distribution of $\mathbf{m}$.

In the models of [3], [9], and the model proposed in this paper, the effect of the covariates is incorporated through the prior for $\mathbf{m}$. The parameter of interest, $\mathbf{m}$, is determined by both genetic information through the likelihood and demographic/environmental information through the prior. An important question is, what is the relative influence of these two sources of information on posterior inference? The influence of the prior is usually directly related to its variance: with a large variance the posterior is dominated by the likelihood while with a small variance the posterior is dominated by the prior. For the Dirichlet-lognormal prior in (2.4), the mean of $\mathbf{m}$ is controlled by the value of $\boldsymbol{\psi}$. At the same time the variance of $\mathbf{m}$ is affected by both the magnitude of $\boldsymbol{\psi}$ and the distribution of σ^2. The interaction among these two parameters tends to increase the uncertainty in the posterior distribution. In contrast, the prior proposed in (2.5) clearly separates the roles of the parameters: the mean of $\mathbf{m}$ is determined by $\boldsymbol{\varphi}$ and the variance of $\mathbf{m}$ is controlled by ρ. This separation is due to the constraint that $\boldsymbol{\varphi}$ is on the simplex. In the simulation study and the application to a real life dataset, we illustrate that the Dirichlet–Dirichlet prior shows less variation in estimating the covariate effects while providing comparable coverage of interval estimates.

3. Simulation Study

We conduct a simulation study to investigate the performance of our proposed model by simulating data from populations with different levels of genetic differentiation as well as different numbers of genetic markers. As discussed above, the estimation of the mixture proportions relies on the divergence among source populations. We are interested in how different level of divergence among source populations would affect the posterior distributions of the parameters of interest. From a practical point of view, population divergence level cannot be controlled by researchers. Instead, researchers can determine how many genetic markers are to be assigned and included in the analysis. Hence, we are also interested in the relationship between the number of loci and the posterior distribution of the parameters.

We consider three simulation scenarios. Under the first scenario, the level of population differentiation is moderate and there are a relatively small number of genetic markers, e.g., 8 loci are available. Under the second scenario, the number of

TABLE 1
Normalized Covariates

Source	Distance	Productivity
1	-0.295	1.298
2	-0.849	1.285
3	-0.822	-0.238
4	-0.562	-1.256
5	-0.326	-0.729
6	1.533	0.286
7	1.320	-0.646

loci is the same as the first scenario but the level of genetic differentiation among source populations is higher. Under the last scenario, the genetic variation is the same as that of the first scenario but the number of genetic markers is doubled, i.e., 16 genetic markers are available. The number of source populations, number of individuals in the mixed population, and allele counts in the source populations are comparable with those in the grey seal data we analyze later.

In the first part of the simulation we generate allele counts in the source populations, which should reflect the level of genetic differentiation among them. This is realized through a hierarchical population structure. We assume that the allele frequencies of the source populations are from a common hyper-population, which has fixed allele frequencies $\boldsymbol{\psi}$, a $L \times A$ matrix with L being the number of loci and A being the number of alleles at each locus. (Without loss of generality, we assume all loci have the same number of alleles.) The allele frequencies $\mathbf{p}_{li}$, a $1 \times A$ vector, for source i and locus l are random samples from a Dirichlet distribution, $\mathcal{D}(((1-\theta)/\theta)\boldsymbol{\psi}_l)$, where $\boldsymbol{\psi}_l$, a $1 \times A$ vector, is the allele frequency of locus l for the hyper-population, and θ is a population divergence measure used widely in population genetic studies, namely Wright's Fst. Note that $E[p_{jli}] = \psi_{jl}$ and $Var[p_{jli}] = \theta\psi_{jl}(1-\psi_{jl})$, where ψ_{jl} is the allele frequency of locus l, allele j for the hyper-population. We choose $\theta = 0.05$ and $\theta = 0.2$ for small and large divergence scenarios, respectively.

The detailed simulation is described as follows. Step 1: generate allele frequencies of the hyper population, $\boldsymbol{\psi}$, by generating a random sample L times from an A dimensional symmetric Dirichlet distribution with parameter 1. Step 2: generate allele frequencies, $\mathbf{p}_{li}$, from the Dirichlet distribution $\mathcal{D}(((1-\theta)/\theta)\boldsymbol{\psi}_l)$ with predefined θ. Step 3: generate allele counts, N_{li}, for source i and locus l, from a multinomial distribution with total allele counts $N = 400$, and probability $\mathbf{p}_{li}$ (from step 2).

In the second part of the simulation we generate genotypes of individuals from the mixed population, which requires the proportional contributions $\mathbf{m}$ and the probability of each genotype. We adopt fixed proportional contributions, which are a function of the two covariates. Note that in both the Dirichlet–Dirichlet (2.5) and the Dirichlet-lognormal (2.4) models, the conditional expectations of the prior for $\mathbf{m}$ are the same, namely,

$$E[m_i|\boldsymbol{\alpha}] = \frac{e^{\boldsymbol{\alpha}\cdot\mathbf{G}_i}}{\sum_{i=1}^{I} e^{\boldsymbol{\alpha}\cdot\mathbf{G}_i}},$$

where $\boldsymbol{\alpha}$ is the vector of regression coefficients and $\mathbf{G}_i$ is the vector of covariates for source i. We use two covariates with the values shown in Table 1 and the coefficients are set to $\alpha_1 = -0.5$ and $\alpha_2 = 0.5$.

The genotype of an individual k is generated by the following steps. First, we decide whether its parents are from the same source by comparing a uniform random number on [0,1] with a preset assortative coefficient $w = 0.05$. The second step is to generate the genotype frequency at each locus. If the parents are from the same source population, the probability of genotype $\mathbf{y}_k$ is $\sum_{i=1}^{I} m_i P(\mathbf{y}_k|ii)$, where $P(\mathbf{y}_k|ii)$ is as in (2.1). If the parents come randomly from the source populations then the probability of genotype $\mathbf{y}_k$ is $\sum_{i=1}^{I} m_i^2 P(\mathbf{y}_k|ii)+\sum_{i=1}^{I}\sum_{j\neq i} m_i m_j P(\mathbf{y}_k|ij)$, where $P(\mathbf{y}_k|ij)$ is the probability of parents from different source populations as in (2.2). Once we have the probability of each genotype for individual k at locus l, we can easily generate the genotype from this probability. Step 2 is repeated for each locus of the individual to get the complete genotype of individual k. The above steps are repeated 160 times to get the genotypes of 160 individuals in the mixed population.

We generate 50 data sets for each of the three scenarios, and we fit both the Dirichlet–Dirichlet prior (2.5) proposed in this paper and the Dirichlet-lognormal prior (2.4) to each data set using a MCMC method. Since most of the parameters are vectors on a simplex, we use a multi-dimensional logit transformation to put the support of the transformed parameters on the real line and remove the simplex constraint. A normal proposal density is then used to conduct a Metropolis–Hastings update nested in the Gibbs sampling. Details of the MCMC update procedure are presented in the Appendix. For most of the data set, we conduct 30,000 iterations in the simulation with 5,000 burn-in and thin the MCMC output by 5. For chains showing suspicious convergence behavior, longer iterations and fine tuning are used to ensure convergence.

Table 2 presents a summary of the posterior analysis, including the average of the posterior means, posterior standard deviations, root mean square error (RMSEs), and the lengths of the 95% highest probability density (HPD) intervals. In general, the posterior means of $\mathbf{m}$ are reasonably close to the true values in all scenarios. The effects of population divergence and number of loci are reflected mainly in the posterior dispersion of $\mathbf{m}$. As shown in the Table, the lengths of the 95% HPD intervals, the posterior standard deviations, and the RMSEs, all indicate that the posterior dispersion of $\mathbf{m}$ decreases with the increase of population differentiation. Given the same level of population differentiation, increasing the number of genetic markers also significantly improves the precision of posterior estimation for $\mathbf{m}$. These results suggest that although in practice the population divergence is always fixed, collecting and including more genetic markers in the analysis can significantly improve the estimation of the proportional contribution parameters $\mathbf{m}$.

For the regression coefficient $\boldsymbol{\alpha}$, both models provide reasonable estimates for the posterior means. However, the posterior variation is large and the 95% HPD intervals all contain zero. Results from the simulation study indicate that neither level of population differentiation nor number of loci has significant effects on the precision of $\boldsymbol{\alpha}$ estimates. We consider this as a reasonable result since the covariate coefficients are essentially a regression over 7 data points, i.e., the 7 source populations. The level of divergence and number of loci improve the precisions of the posterior variance sfor $\mathbf{m}$, which only affect $\boldsymbol{\alpha}$ indirectly. With only 7 data points, few simulated data sets will be able to provide strong support for a regression relationship. Thus, increasing the number of loci or studying highly differentiated populations will do little to improve posterior estimates of $\boldsymbol{\alpha}$. A larger number of populations would be required to provide statistically supportable evidence of the effects.

The advantages of the Dirichlet–Dirichlet prior proposed in this paper are seen primarily in the reduced posterior variation of the regression coefficients, α_1 and α_2. Under the Dirichlet–Dirichlet prior, the posterior standard deviations and RMSEs

Table 2
Posterior summary of simulation study

		Dirichlet–Dirichlet			Dirichlet-lognormal		
	TRUE	$\theta = 0.05$	$\theta = 0.20$	$\theta = 0.05$	$\theta = 0.05$	$\theta = 0.2$	$\theta = 0.05$
		L=8	L=8	L=16	L=8	L=8	L=16
m_1	0.249	* 0.256	0.254	0.239	0.258	0.255	0.239
		** 0.044	0.033	0.036	0.045	0.032	0.036
		*** 0.063	0.052	0.053	0.065	0.052	0.054
		****0.170	0.127	0.138	0.173	0.126	0.139
m_2	0.327	0.326	0.377	0.350	0.328	0.379	0.351
		0.045	0.036	0.039	0.046	0.036	0.039
		0.068	0.076	0.058	0.069	0.079	0.059
		0.175	0.140	0.150	0.178	0.141	0.151
m_3	0.151	0.134	0.125	0.137	0.135	0.125	0.138
		0.038	0.026	0.031	0.038	0.026	0.031
		0.052	0.053	0.048	0.052	0.053	0.049
		0.145	0.099	0.121	0.149	0.100	0.121
m_4	0.079	0.073	0.065	0.077	0.072	0.064	0.075
		0.033	0.020	0.027	0.033	0.020	0.026
		0.042	0.038	0.039	0.044	0.038	0.039
		0.119	0.075	0.103	0.121	0.074	0.100
m_5	0.092	0.094	0.078	0.090	0.094	0.077	0.090
		0.033	0.022	0.027	0.034	0.021	0.028
		0.051	0.036	0.044	0.051	0.037	0.044
		0.1259	0.082	0.103	0.126	0.082	0.105
m_6	0.060	0.069	0.060	0.060	0.066	0.059	0.059
		0.033	0.020	0.025	0.033	0.020	0.024
		0.045	0.032	0.038	0.046	0.033	0.038
		0.115	0.074	0.090	0.114	0.074	0.089
m_7	0.042	0.047	0.041	0.048	0.047	0.040	0.048
		0.025	0.015	0.022	0.026	0.016	0.0232
		0.038	0.029	0.037	0.039	0.029	0.036
		0.088	0.057	0.078	0.089	0.056	0.079
α_1	-0.500	-0.398	-0.429	-0.430	-0.485	-0.493	-0.493
		0.477	0.469	0.479	0.613	0.578	0.587
		0.516	0.496	0.503	0.659	0.615	0.625
		1.849	1.815	1.849	2.383	2.268	2.305
α_2	0.500	0.449	0.520	0.433	0.538	0.618	0.515
		0.416	0.407	0.408	0.535	0.519	0.523
		0.437	0.425	0.422	0.555	0.551	0.535
		1.637	1.600	1.602	2.097	2.041	2.056
ω	0.050	0.037	0.014	0.014	0.037	0.014	0.014
		0.035	0.013	0.014	0.035	0.013	0.014
		0.040	0.039	0.038	0.040	0.039	0.038
		0.106	0.040	0.042	0.106	0.040	0.042

*: average of posterior means;
**: average of posterior standard deviations;
***: average of RMSEs;
****: average length of 95% HPD intervals.

are uniformly smaller than that of the Dirichlet-lognormal model even though the prior variances for $\boldsymbol{\alpha}$ are all set to the same value, i.e., $\sigma_p^2 = 10$. We consider this as mainly due to the confounding of the effects that both $\boldsymbol{\psi}$ and σ^2 have on the variances of $\mathbf{m}$. Another possible reason is the effects of the prior for $\tau = 1/\sigma^2$. In any case, the Dirichlet–Dirichlet prior has the advantage of leading to more precise estimation of regression coefficients and ease in picking a non-informative prior without sacrificing nominal coverage of credible intervals.

4. Application to the Grey Seal Data Set

To illustrate the usefulness of our approach, we apply it to data from grey seal, *Helicoerus grypus*, populations in the Orkney Islands, which were also analyzed by [3] and [4]. The data consist of allele frequencies of 8 loci for seven source colonies and the genotype frequencies for a newly established colony on Stronsay island.

TABLE 3
Model Evaluation

Models	Dbar	pD	DIC	LPML
Dirichlet-Dirichlet	8044	336	8380	-3023
Dirichlet-lognormal	8042	337	8379	-3023
Uniform	8045	336	8381	-3023

There are two explanatory variables associated with each source population: distance between the source island and Stronsay island (α_1), and the 'productivity' index, which is related to the population density and size of the source population, (α_2). The genetic data were collected from the first generation descendants of migrants to Stronsay. We use the likelihood in equation (2.3) to allow for the possibility that migrants are more likely to mate with other individuals from the island from which they migrated. Since there is no closed form for the posterior distribution, we use MCMC methods for posterior inference.

We compare results from three models with different priors: the Dirichlet–Dirichlet prior, the Dirichlet-Lognormal prior, and a model with the symmetric Dirichlet prior with parameter 1 for $\mathbf{m}$, which corresponds to a model in which covariate effects are not incorporated. Our results reveal that differences among the models rarely lead to substantial differences in the mean posterior likelihood, which is intuitively reasonable. The part of the likelihood function concerned with source population allele frequencies is identical across all models, and the part of the likelihood concerned with colony allele frequencies is tightly tied to the observed genotypes. The three models differ only through the prior for the proportional contributions $\mathbf{m}$, which has limited impact on the likelihood unless the source population differs substantially. A direct consequence of these properties is the similarity among metrics for model evaluation measures that use only the likelihood, e.g., DIC [13] and the logarithm of the pseudomarginal likelihood (LPML) [7]. As shown in Table 3, neither DIC nor LPML provides strong support for any of the models relative to the others.

The posterior densities of the model parameters are given in Figure 3. Table 4 gives the posterior means and 95% HPD intervals of the parameters of interest: $\mathbf{m}$,τ,$\boldsymbol{\alpha}$ and ρ. It can be seen that the posterior means of $\mathbf{m}$ are quite different for the model with symmetric Dirichlet prior (with no covariates) and the models using covariate information. Specifically, models using covariate information suggest a larger proportion from sources 2 and 3 than the uniform model, which is a reasonable result since sources 2 and 3 are the closest source islands to the new colony and posterior analysis indicates distance has a moderate effect on the proportion contribution.

The 95% HPD intervals of all regression coefficients in all models include zero, which is not surprising from the analysis of the simulation results. The large variation in $\boldsymbol{\alpha}$ is presumably due to the small number of source populations. Nonetheless, there is some support for the notion that the coefficient associated with distance, α_1, is negative. The posterior probability that α_1 is negative is more than 0.785 for the Dirichlet–Dirichlet model and 0.850 for the Dirichlet-lognormal model. The posterior probability that α_2 is positive in our model is 0.593 and 0.596 in the Dirichlet-lognormal model. The Dirichlet–Dirichlet model shows a shorter HPD interval compared to the Dirichlet-lognormal model, which is also consistent with the simulation results. In short, distance has a negative effect on the proportional contributions and population sizes have minor positive effects on the proportional contributions.

As discussed above, the parameter ρ in our hierarchical model is analogous to

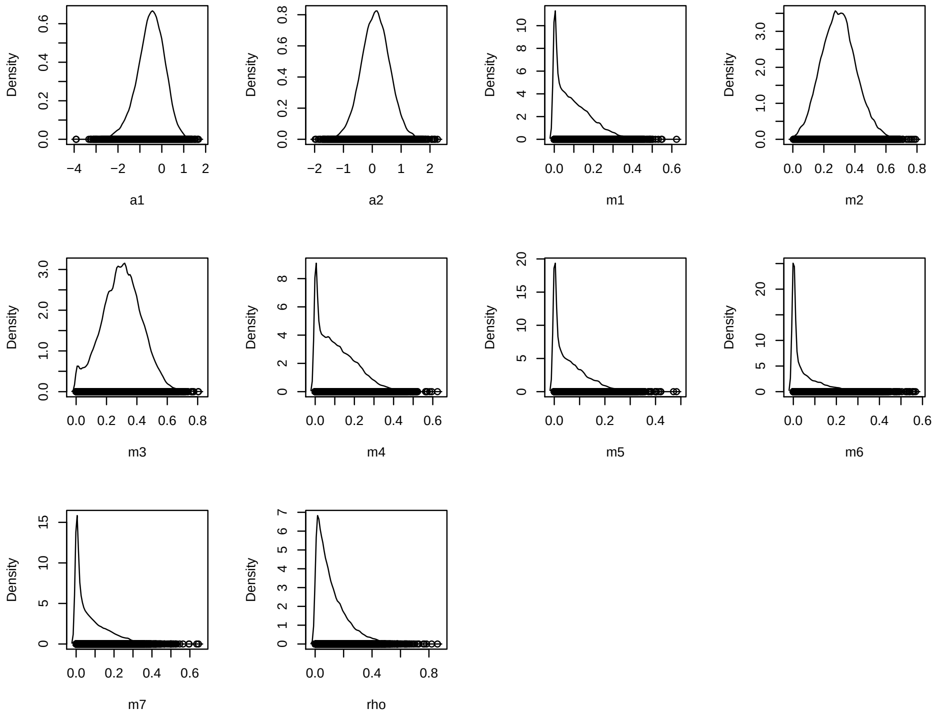


Fig 3. *The posterior densities of parameters in the Dirichlet–Dirichlet model*

Table 4
Posterior means and 95% HPD intervals

	Dirichlet–Dirichlet		Dirichlet-lognormal		Uniform	
	Mean	95% HPD	Mean	95%HPD	Mean	95% HPD
m_1	0.097	(0,0.280)	0.101	(0,0.282)	0.099	(0,0.250)
m_2	0.297	(0.086,0.526)	0.305	(0.085,0.542)	0.243	(0.046,0.436)
m_3	0.3	(0,0.514)	0.324	(0.045,0.586)	0.258	(0.035,0.475)
m_4	0.113	(0,0.300)	0.11	(0,0.305)	0.104	(0,0.256)
m_5	0.061	(0,0.196)	0.07	(0,0.195)	0.081	(0,0.198)
m_6	0.052	(0,0.214)	0.037	(0,0.179)	0.092	(0,0.234)
m_7	0.079	(0,0.270)	0.053	(0,0.230)	0.123	(0,0.289)
α_1	-0.494	(-1.808,0.668)	-1.03	(-3.005,0.961)	0	(0,0)
α_2	0.113	(-0.864,1.084)	0.182	(-1.191,1.680)	0	(0,0)
ω	0.609	(0.105,1.000)	0.613	(0.110,1.000)	0.616	(0.056,0.986)
ρ	0.118	(0,0.343)				
τ			1.453	(0.183,4.377)		

a 'goodness of fit' measure for the relationship between the covariates and $\mathbf{m}$. Specifically, $1 - \rho$ is roughly the proportion of variance in $\mathbf{m}$ explained by the regression. As the results in Table 4 show, the posterior mean for $1 - \rho$ is near 0.9, which indicates a fairly tight regression in spite of the uncertainties associated with α. In summary, we conclude that there is moderate support for the hypothesis that increasing distances between the source and colony populations decrease the proportional contributions of the sources to the colony.

5. Conclusions

The primary goal of this analysis is to incorporate environmental/demographic information into the estimation of the proportional contributions of source populations to a new colony through appropriate informative priors. Two other models are available which satisfy the constraint that the sum of the proportional contributions must equal one, i.e., additive logistic transformation [9] and Dirichlet-lognormal model [3]. We introduce a parametrization for the Dirichlet prior derived from population genetics in which we specify the mean, φ_i, and variance, $\rho(1-\varphi_i)\varphi_i$, of the mixture parameters and a linear model for the parameters of a second Dirichlet that determines φ_i. The Dirichlet–Dirichlet prior has several advantages over the alternatives. First, the parameter ρ has a natural vague prior distribution, a uniform distribution [0,1]. Second, ρ controls the variance of the Dirichlet prior and $1 - \rho$ has a natural interpretation as the proportion of variance explained by regression. Finally, the mean of the proportional contributions is not affected by the parameter ρ and the regression coefficients have a direct interpretation as regression effects on proportional representation. The separation of mean effect and variance effect is a major advantage of the proposed formulation compared to alternative models where the proportional contribution for any given population depends on the relative magnitude of coefficients associated with other regression components and their random effects.

The simulation study indicates that larger population divergence would lead to more precise estimation of the proportional contributions $\mathbf{m}$. Given a particular level of population divergence, better estimates of $\mathbf{m}$ can also be achieved by including more loci in the analysis. The simulations show that the Dirichlet–Dirichlet prior has better performance in estimating the regression coefficients in term of posterior variation than a Dirichlet-lognormal prior.

When we apply our model to the grey seal data we find that the distance between a source island and the new colony play a moderate role in its proportional contribution but that the effect of source population productivity is weak. These results are consistent with those presented in [3], but the posterior variability of the regression coefficients is smaller, as in the simulation study. The advantages of the formulation presented here seem likely to be generally available in the analysis of compositional data. In particular, a formulation similar to the one used here may be generally useful in modeling situations where additive logistic transformations have been the norm, both because of direct interpretability of regression coefficients and the natural interpretation of $1-\rho$ as a goodness of fit measure.

Appendix: A General Approach for Updating a Proportional Vector

The conditional distributions of model parameters are non-standard distributions; hence, we use a Metropolis–Hastings algorithm nested within Gibbs sampling to conduct each MCMC update. Several vector parameters, $\mathbf{P}$, $\mathbf{m}$,and φ, are subject to the constraint that the support of their components is on [0,1] and the summation equals to one. We use a multidimensional logit transformation to 'de-constrain' the parameters and perform Metropolis–Hastings updating using a Normal proposal density. Let $\boldsymbol{\theta}$ be a vector of dimension $p{+}1$ with constraints $\theta_i > 0$ and $\sum_{i=1}^{p+1}\theta_i = 1$. Let

$$\theta_i = \frac{\exp(\xi_i)}{1+\sum_{j=1}^{p}\exp(\xi_j)}.$$

The Jacobian matrix $\partial f(\boldsymbol{\theta})/\partial\boldsymbol{\xi}$ is the matrix with entries

$$x_{ij} = \begin{cases} \frac{e^{\xi_i}+e^{\xi_i}(\sum_{j=1}^{p} e^{\xi_j})-e^{2\xi_i}}{(1+\sum_{j=1}^{p} e^{\xi_j})^2} & \text{for } i=j \\ \frac{-e^{\xi_i+\xi_j}}{(1+\sum_{j=1}^{p} e^{\xi_j})^2} & \text{for } i \neq j. \end{cases}$$

It can be shown that the determinant of the Jacobian matrix is

$$\frac{e^{\sum_{j=1}^{p}\xi_j}}{(1+\sum_{j=1}^{p} e^{\xi_j})^{p+1}}.$$

The full conditional distribution of ξ is

$$f(\xi|D) = f(\boldsymbol{\theta}|D)\frac{e^{\sum_{j=1}^{p}\xi_j}}{(1+\sum_{j=1}^{p} e^{\xi_j})^{p+1}}.$$

Instead of sampling $\boldsymbol{\theta}$, we conduct a Metropolis-Hastings update for $\boldsymbol{\xi}$ using a normal proposal density $N(\hat{\xi}, \hat{\sigma}^2_{\hat{\xi}})$, where $\hat{\xi}$ is the maximizer of $\pi(\xi|D)$ and $\hat{\sigma}^2_{\hat{\xi}}$ is the estimated variance, which could be a fixed value based on a pilot run or the inverse of the score matrix. Alternatively, we can use a Normal proposal density centered at the current value. The algorithm operates as follows:
Step 1. Let $\boldsymbol{\xi}$ be the current value. Find the maximum likelihood estimate of $\boldsymbol{\xi}$, $\hat{\boldsymbol{\xi}}$.
Step 2. Generate a proposal value $\boldsymbol{\xi}^*$ from $N(\hat{\boldsymbol{\xi}}, \hat{\sigma}^2_{\hat{\boldsymbol{\xi}}})$.
Step 3. A move from $\boldsymbol{\xi}$ to $\boldsymbol{\xi}^*$ is made with probability

$$\min\left\{\frac{f(\boldsymbol{\xi}^*|\mathbf{D})\Phi(\frac{\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}}{\hat{\sigma}\boldsymbol{\xi}})}{f(\boldsymbol{\xi}|\mathbf{D})\Phi(\frac{\boldsymbol{\xi}^*-\hat{\boldsymbol{\xi}}}{\hat{\sigma}\boldsymbol{\xi}})}, 1\right\}$$

where Φ is the standard normal probability density function. The ξ is then converted back to its expression in terms of θ.

Acknowledgments

We are indebted to William Amos for sharing the data on grey seals that motivated this study. Collection of these data was supported by NERC grant GR3/11662 to W. Amos. This research is supported by NIH grant 1R01-GM068449-01A1.

References

[1] Balding, D. J. and Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* **96** 3–12.

[2] Bolker, B., Okuyama, T. Bjorndal, K. and Bolten, A. (2003). Stock estimation for sea turtle populations using genetic markers: accounting for sampling error of rare genotypes. Ecological Applications *Ecological Applications* **13** 763–775.

[3] Gaggiotti, O. E., Brooks, S. P., Amos, W. and Harwood, J. (2004). Combining demographic, environmental and genetic data to test hypotheses about colonization events in metapopulations. *Molecular Ecology* **13** 811–825.

[4] Gaggiotti, O. E., Jones, F., Lee, W. M., Amos, W., Harwood, J. and Nichols, R. A. (2002). Patterns of colonization in a metapopulation of grey seals. *Nature* **13** 424–427.

[5] Holsinger, K. E. (1999). Analysis of genetic diversity in hierarchically structured populations: a bayesian perspective. *Hereditas* **130** 245–255.

[6] Holsinger, K. E. and Wallace, L. E. (2004). Bayesian approaches for the analysis of population structure: an example from Platanthera leucophaea (Orchidaceae). *Molecular Ecology* **13** 887–894.

[7] Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2001). *Bayesian Survival Analysis.* Springer-Verlag.

[8] Millar, R. B. (1987). Maximum likelihood estimation of mixed stock fishery composition. *Can. J. Fish. Aquat. Sci.* **44** 583–590.

[9] Milner, G. B., Teel, D. J., Utter, F. M., and Winans, G. A. (2005). A genetic method of stock identification in mixed populations of pacific salmon, oncorhynchus spp. *Mar. Fish. Rev.* **47** 1–8.

[10] Okuyama, T. and Bolker, B. M. (2005). Combining genetic and ecological data to estimate sea turtle origins. *Ecological Applications* **15** 315–325.

[11] Pella, J. and Masuda, M. (2001). Bayesian methods for analysis of stock mixtures from genetic characters. *Fish. Bull.* **99** 151–167.

[12] Smouse, P. E., Waples, R. S. and Towrek, J. A. (1990). A genetic mixture analysis for use with incomplete source population data. *Can. J. Fish. Aquat. Sci.* **47** 620–634.

[13] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B* **64 NO.4** 583–639.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 252
© Institute of Mathematical Statistics, 0000
arXiv: math.PR/0000000

Kendall's Tau in High-Dimensional Genomic Parsimony

Pranab K. Sen

University of North Carolina, Chapel Hill
e-mail: pksen@bios.unc.edu

Abstract: High-dimensional data models, often with low sample size, abound in many interdisciplinary studies, genomics and large biological systems being most noteworthy. The conventional assumption of multinormality or linearity of regression may not be plausible for such models which are likely to be statistically complex due to a large number of parameters as well as various underlying restraints. As such, parametric approaches may not be very effective. Anything beyond parametrics, albeit, having increased scope and robustness perspectives, may generally be baffled by the low sample size and hence unable to give reasonable margins of errors. Kendall's tau statistic is exploited in this context with emphasis on dimensional rather than sample size asymptotics. The Chen-Stein theorem has been thoroughly appraised in this study. Applications of these findings in some microarray data models are illustrated.

Contents

1. Introduction

The past three decades have witnessed a phenomenal growth of research literature on statistical methods for large dimensional data models. Such models abound in various interdisciplinary fields, especially in the evolving field of genomics and bioinformatics. *Knowledge discovery and data mining* (KDDM) or statistical learning tools are usually advocated for such high dimensional data models, often on primarily computational or heuristic justifications. The curse of dimensionality is so overwhelming that classical likelihood (principle) based statistical inference tools, baffled with an excessive number of parameters, may not be robust or efficient. Conventional assumptions of multinormality of errors and linearity of regression models may not be generally tenable in such contexts. Moreover, having a large number

Departments of Biostatistics and Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599-7420, U.S.A.

AMS 2000 subject classifications: Primary 62G10, 62G99; secondary 62P99

Keywords and phrases: Bioinformatics, Chen-Stein theorem, dimensional asymptotics, FDR, multiple hypotheses testing, nonparametrics, permutational invariance, U-statistics

of coordinate variables, the assumption of their stochastic independence may not be realistic in a majority of cases. On top of that, at least a part of the response variables may be discrete or even purely qualitative in nature; often, the categorical responses may not reveal any (partial) ordering. In that sense, discrete multivariate analysis may appear to be more appropriate than conventional multinormal model based analysis. Even for multinormal models, the high-dimensionality may demand a far larger sample size in order to implement a full likelihood based asymptotic analysis. That is, we need the conventional $n \gg K$ environment for drawing appropriate statistical conclusions with reasonable precision.

Typically, in such high-dimensional models, one encounters a $K \gg n$ environment, where K is the dimension of the data and n is the sample size. In such *high-diensional low sample size*, HDLSS, models, effective dimension reduction may be a challenging statistical task, usually beyond the scope of KDDM. For example, in neuronal spike train models, there are literally tens of thousands of neurons (nerve cells), and in the presence of external stimuli, the spike trains for any observable subset of neurons exhibit a high-degree of nonstationarity. Further, recording of such spike trains in a large number of nerve cells may be invasive to the brain functioning due to the destructive nature of recording ([15], Ch. 3). Essentially, we have a very high dimensional counting process. Doubly stochastic Poisson processes have been considered in the literature, albeit without much claim of optimal resolutions. In magnetic resonance imaging, MRI, there could be tens of thousands of microscopic units producing an enormously high dimensional spatial data model. More complexities may arise in case of (functional) fMRI models. For such HDLSS models, parametric asymptotics may not have adequate scope or good statistical interpretation.

The transition from conventional normal theory to nonparametric linear models has been well fortified along with the development of nonparametric or robust statistical methods based on R- statistics (ranks), M-statistics (maximization) and linear combinations of order statistics or L-statistics; see, for example, [8] where other pertinent references have been extensively cited. In a more general setup, nonparametric regression functionals have been formulated wherein the linearity of regression or a specific nonlinear form are not assumed to hold.

In the context of testing monotonicity of nonparametric regression, without assuming a linear or any specific nonlinear form, Ghosal et al. [5] considered suitable U-processes based on a locally smoothed Kendall's tau statistic. They provided general asymptotics for such locally smoothed Kendall's tau processes when both the independent and dependent variates are stochastic, and illustrated their effective use in the postulated hypothesis testing problem. Such local versions of Kendall's tau statistics have simple statistical interpretation, albeit, in view of possibly slower rate of convergence, the impact of large sample size is apparent in their analysis. In the contemplated bioinformatics area, as we shall see, the HDLSS scenario calls for alternative approaches, and some of these will be explored in this study.

In a simple regression setup, the Theil-Sen (point as well as interval) estimates of the regression slope based on the Kendall tau statistic [14], have simple forms, and are computationally tractable and statistically robust. Another advantage of the Kendall tau statistic is its adaptability for count data as well as latent-effect models. Further, a test for the null hypothesis of no regression based on the Kendall tau statistic (being distribution-free under the null hypothesis of invariance) remains valid and efficient for such complex models. Our contemplated models, unlike [5], entail a high dimensional data with relatively (and often inadequately) smaller sample size, i.e., the HDLSS ($K \gg n$) environment. As we shall see in the next section, there may not be a genuine temporal pattern. In addition, there may be other complications arising from lack of spatial-compactness, spatial homogeneity and other spatial dependence patterns.

For better motivation, in Section 2, an illustration is made with a microarray data model where HDLSS models typically arise. Section 3 deals with the appropriateness of statistical modeling and analysis based on a pseudo-marginal approach incorporating coordinatewise construction of the Kendall tau statistic, in such $K \gg n$ environments. Section 4 is devoted to the dimensional asymptotics for the Kendall tau process in such HDLSS models where there are two basic problems : (i) group divergence, and (ii) classification of genes into disease and nondisease types. For the first problem, a pseudo-marginal approach based on the Hamming distance has been explored in [17] while in the latter context, multiple hypotheses testing (MHT) problems in HDLSS setups arise in a different perspective and call for some alternative novel tools for valid and efficient statistical appraisals. Motivated by these perspectives in such HDLSS models, some applications of the Chen-Stein [3] theorem in such $K \gg n$ environments are presented in the last section. These generalizations cover both the MHT and the gene-environment interaction testing problems.

2. An Illustrative Data Model

We consider a genomic model arising in microarray data analysis as an illustration. The microarray technology allows simultaneous studies of thousands of genes, K, possibly differentially expressed under diverse biological / experimental setups, with only a few, n, arrays. We may refer to Lobenhofer et al. (2002) where for a set of 1900 genes, arranged in rows, the gene expressions were recorded at 6 time points, with 8 observations at each time point. Thus $1900 = K \gg n = 48$. The gene-expression levels are measured by their color intensity (or luminosity) as a quantitative (nonnegative) variable, either on the (0, 1) or 0 - 100 per cent scale, or (based on the log-scale) on the real line $\Re$. A gene associated (causally or statistically) with a target disease is known as a *disease gene*, DG, while the others as *nondisease genes*, NDG. Gene expression levels under different environments cast light on plausible *gene-environment interactions* (or associations) so that if the arrays are properly designed, *mapping disease genes* may be facilitated with such microarray studies. One of the main issues is identifying differentially expressed genes among thousands of genes, tested simultaneously, across experimental conditions. Typically, for a target disease, there are only a few DG while the NDG comprise the vast majority. A NDG is expected to have a low gene expression level while a DG is expected to have generally higher expression levels. Thus, a natural *stochastic ordering* of gene expression levels of the DG with varying disease severity is plausible while the NDG expression levels are expected to be stochastically unaffected by such disease level differentials.

Microarray data go thorough a lot of standardization and normalization so that conventional simple models, such as the classical MANOVA models, may rarely be totally adaptable. If the arrays are indexed by an explanatory or design variate (t) that possesses an ordering (not necessarily linear), then the stochastic ordering could be exploited through suitable nonparametric techniques. The main difficulty in modeling and statistically analyzing microarray data stems from the high dimensionality of the genes compared to the number of arrays. While the different arrays may sometimes be taken to be at least statistically independent, the genes may not. Moreover, not much is known about the spatial topology of the genes or their genetic distances. There is another factor that merits our attention. The gene expression levels for the different genes in an array are neither expected to be stochastically independent nor (marginally) identically distributed. Sans such an i.i.d. clause, standard parametrics typically adaptable for fMRI models (albeit mostly done in a Bayesian coating) may encounter roadblocks for fruitful adaptation in microarray data models. Thus, structurally, such data models are different from those usually encountered in nonparametric functional regression models. For this reason, a pseudo-marginal approach is highlighted here. This approach exploits the marginal nonparametrics fully and renders some useful modeling and analysis convenience.

3. Some HDLSS Formulations

Motivated by microarray data models introduced in Section 2, we consider here a set of n arrays (sample observations) where there is a design variate t_i associated with the ith array, for $i = 1, \dots, n$. Without loss of generality, we assume the t_i are ordered, i.e.,

$$t_1 \le t_2 \le \cdots \le t_n, \tag{3.1}$$

with at least one strict inequality. We do not, however, impose any linear or specific parametric ordering of these design variates. The multisample (ordered alternative) model is a particular case where n can be partitioned into I subsets of sizes $n_1, \dots, n_I$ such that within each subgroup, the t_i are the same while they are ordered over the I different subsets. For the ith array, corresponding to the K genes (positions), we have a gene expression level denoted by $X_{ik},\ k = 1, \dots, K$, so that we have K-vectors $\mathbf{X}_i = (X_{i1}, \dots, X_{iK})'$, for $i = 1, \dots, n$. The joint distribution function of $\mathbf{X}_i$ is denoted by $F_i(\mathbf{x}),\ \mathbf{x} \in \Re^K$. Further, for the kth gene in the ith array, i.e., X_{ik}, the marginal distribution is denoted by $F_{ik}(x),\ x \in \Re$, for $k = 1, \dots, K;\ i = 1, \dots, n$. For a given i, the $F_{ik},\ k = 1, \dots, K$ may not be generally the same, and moreover, the $X_{ik},\ k = 1, \dots, K$ may not be all stochastically independent.

If a gene k is NDG and the t_i reflect the variability of the disease level, then the $F_{ik}, i = 1, \dots, n$ should be the same. On the other hand, for a DG k, for $i < i'$, X_{ik} should be stochastically smaller than $X_{i'k}$ in the sense that the $F_{ik}, i = 1, \dots, n$ should have the ordering

$$F_{1k}(x) \ge F_{2k}(x) \ge \cdots \ge F_{nk}(x), \forall\ x \in \Re. \tag{3.2}$$

Therefore, we could force a characteriation of DG and NDG based on the following stochastic ordering: For a NDG k, the $F_{ik},\ i = 1, \dots, n$ are all the same, this being denoted by the null hypothesis H_{0k}, while for a DG k, the stochastic ordering in (3.2) holds which we denote by H_{1k}, for $k = 1, \dots, K$. In this marginal formulation, we have a set of K hypotheses corresponding to the K genes, and whatever appropriate test statistic (say T_{nk}) we use for testing H_{0k} vs. H_{1k}, these statistics may not be, generally, stochastically independent. The basic problem is therefore to test simultaneously for

$$H_0 = \cap_{k=1}^K H_{0k} \text{ vs } H_1 = \cup_{k=1}^K H_{1k}, \tag{3.3}$$

without ignoring possible dependence of the test statistics for the component hypotheses testing H_{0k} vs H_{1k}, for $k = 1, \dots, K$. This makes it appealing to follow the general guidelines of the Roy [12] *union-intersection principle* (UIP), albeit in a marginalization (i.e., adapting a finite union and finite intersection scheme), and thus permitting a more general framework so as to allow simultaneous testing and classification into DG / NDG groups. In a very parametric setup, some order restricted inference problems have been considered by [11]. However, in our setup, such normality based parametric models may not be very appropriate.

Our approach is based on the classical Kendall tau statistics for each of the K genes and the incorporation of these (possibly dependent) marginal statistics in a composite scheme for classification. For the kth gene, based on the n observations $X_{ik},\ i = 1, \ldots, n$, and the tagging variables $t_1, \ldots, t_n$, we define the Kendall tau statistic as

$$T_{nk} = \binom{n}{2}^{-1} \sum_{1 \le i < i' \le n} \operatorname{sign}(X_{i'k} - X_{ik}) \operatorname{sign}(t_{i'} - t_i), \tag{3.4}$$

for $k = 1, \ldots, K$. Conventionally, we take $\operatorname{sign}(0) = 0$. Note that T_{nk} is a (generalised) U-statistic of degree 2 [7]. Further, note that by (3.1), we may set $\mathcal{S} = \{(i, i') : t_i < t_{i'}; 1 \le i < i' \le n\}$ and let N be the cardinality of the set $\mathcal{S}$. Then by (3.1), $n - 1 \le N \le \binom{n}{2}$. Moreover, we may rewrite T_{nk} as

$$T_{nk} = \binom{n}{2}^{-1} \sum_{\mathcal{S}} \operatorname{sign}(X_{i'k} - X_{ik}),\ k = 1, \ldots, K, \tag{3.5}$$

where $\mathcal{S}$ depends on the ordering of the t_j and therefore remains the same for every $k = 1, \ldots, K$. Note further that whenever $N < \binom{n}{2}$, the range of variation of T_{nk} is $(\ -N/\binom{n}{2}, N/\binom{n}{2}\ )$ which is contained in the interval $(-1, 1)$. That is why we shall find it convenient to take the modified or rescaled Kendall tau as

$$T^o_{nk} = N^{-1} \sum_{\mathcal{S}} \operatorname{sign}(X_{i'k} - X_{ik}), \tag{3.6}$$

whose range is exactly $(-1, 1)$, albeit the distribution being still discrete.

Note that for any $k = 1, \ldots, K$, under H_{0k}, for every $i \ne i'$, the difference $X_{i'k} - X_{ik}$ is symmetrically distributed around 0, and hence, $E_{0k}\{\operatorname{sign}(X_{i'k} - X_{ik})\} = 0$ so that

$$E_{0k}\{T_{nk}\} = E_{0k}\{T^o_{nk}\} = 0,\ \forall\ k = 1, \ldots, K. \tag{3.7}$$

Further, the marginal distribution of T_{nk} under H_{0k} is generated by the $n!$ equally likely permutations of the X_{ik} among themselves. Therefore when all the F_{ik} are continuous, ties among the observations being negligible with probability 1, T_{nk} (or T^o_{nk}) is distribution-free under H_{0k}. This distribution may depend on the set $\mathcal{S}$ but that being the same for all k, we conclude that under H_0 in (3.3), marginally each T^o_{nk} is distribution-free and these K statistics all have the same marginal distribution. If all the t_i were stochastic and continuous then $N = \binom{n}{2}$ and we will have

$$\operatorname{Var}_0\{T_{nk}\} = 2(2n + 5)/\{9n(n - 1)\}. \tag{3.8}$$

On the other hand, in general for $N \le \binom{n}{2}$, the t_i are not distinct and may be even nonstochastic, and hence, the variance is equal to

$$\operatorname{Var}_0\{T^o_{nk}\} = N^{-2}\{(2/3)(N_1 - N_2) + N\} = \nu_n^2,\ \text{say}, \tag{3.9}$$

where N_1 is the cardinality of the set $\{(i,i'),(i,i'') : t_i < t_{i'}, t_i < t_{i''}, t_{i'} \neq t_{i''}\}$ and N_2 is the cardinality of the set $\{(i,i'),(i'',i) : t_i > t_{i''} \neq t_{i'}\}$. For small values of n and given (3.1), one can enumerate $\mathcal{S}$ and obtain the exact distribution of T^o_{nk} under H_{0k}. If n is large, the standardized form of the statistic, i.e., T^o_{nk}/ν_n has closely a standard normal distribution. In our setup, perhaps the exact permutation distribution plays a greater role and this will be illustrated later on.

The behavior of T^o_{nk} under alternatives would naturally depend on the stochastic ordering in (3.2) and these statistics will not be exact distribution-free nor possibly have identical marginal laws. Nevertheless, under (3.2), for every $i < i',\ X_{i'k} - X_{ik}$ has a distribution tilted to the right, so that

$$E\{T^o_{nk} \mid H_{1k}\ \} \geq 0,\ \forall\ k = 1, \ldots, K. \tag{3.10}$$

This motivates us to use tests based on the marginal statistics T^o_{nk} using the right hand side critical region, or equivalently the right-hand sided p-values. Recall that the distribution of each T^o_{nk}, at least for n not too large, is discrete, but that is not going to be of any particular concern. A greater concern is to incorporate possible stochastic dependence among the K statistics $T^o_{nk},\ k = 1, \ldots, K$ (even under the null hypothesis) and their possible heterogeneity when some of the H_{1k} are true. A basic problem is to formulate suitable multiple hypothesis testing procedures to assess which hypotheses are to be rejected subject to a suitably defined Type I error rate. This is elaborated in the next section.

4. Dimensional Asymptotics and the Union Intersection Test

Although independence across microarrays may be assumed, their i.d. structure may be vitiated if the arrays relate to different biological or experimental setups. Moreover, for different genes, the gene expression (marginal) distributions are likely to be different when there is gene-environment interaction. Taking into account such plausible inter-gene stochastic dependence and heterogeneity, we need to prescribe statistical modeling and analysis tools. This will be accomplished through dimensional asymptotics where K is made to increase indefinitely while n, being small compared to K, may or may not be adequately large.

In view of (3.3), it is tempting to appeal to the union-intersection principle [12], or UIP, to construct suitable test statistics which will cover the genome-wise picture in a reasonable way. Towards this, we may note that as under H_0 (i.e., $H_{0k}, \forall k$), marginally each T^o_{nk} has the same distribution (which does not depend on the underlying F_{ik}). Thus, corresponding to any $c : -1 \leq c \leq 1$, the tail probability $P_0\{T^o_{nk} > c\}$ is the same for all k and this can be evaluated by using the exact permutation distribution generated by the $n!$ permutations of the $X_{ik}, 1 \leq i \leq n$. The UIP then leads to the following union-intersection test, UIT, statistic:

$$T^{*o}_n = \max\{T^o_{nk} : 1 \leq k \leq K\}, \tag{4.1}$$

where the test function is given by $\phi(T^{*0}_n) = 1, \gamma$, or 0, accordingly as T^{*o}_n is $>, =$ or $< c$ and $\gamma : (0 \leq \gamma \leq 1)$ is so chosen that $E_0\{\phi(T^{*o}_n)\} = \alpha$, the preassigned *level of significance*. Note that for n not adequately large, the null distribution of T^o_{nk} is essentially discrete and hence this usual randomization test function is aimed to take care of this problem.

The crux of the problem is therefore to determine such a critical level c_α. The joint distribution of the $T^o_{nk}, 1 \leq k \leq K$, even under the null hypothesis H_0, depends on the underlying K-dimensional distribution F_i, and hence, in general will not be distribution-free. Thus, the usual technique of finding out the critical level of T_n^{*o} from this joint distribution may be intractable.

One possibility is to incorporate the fact that under H_0, the K-vectors $\mathbf{X}_i, i = 1, \ldots, n$, are i.i.d. and hence their joint distribution remains invariant under any permutation of these vectors among themselves. Thereby we can evaluate such critical values by an to appeal to the permutation distribution generated by the $n!$ equally likely permutations of the K-vectors $\{\mathbf{X}_i\}$ among themselves. This permutation law generates the (unconditional null) marginal laws of the T_{nk}^o, and provides some conditional versions of their joint distributions of various orders. Since this permutation law is a conditional law (given the collection of all these K-vectors), the critical values obtained in this manner are themselves stochastic, thus introducing another layer of variation. Nevertheless, it provides a conditionally distribution-free test. One discouraging feature of this permutation approach is that the permutation invariance does not hold under the alternative hypothesis, and hence critical levels computed from the permutation law involving an observed set of $\{\mathbf{X}_i\}$ may be sensitive to the data conformity to the null situation.

If we assume that all the T_{nk}^o are stochastically independent, then we have for any c, $-1 \le c \le 1$, under H_0,

$$P_0\{T_n^{*o} \le c\} = [P_0\{T_{n1}^o \le c\}]^K, \tag{4.2}$$

so that the distribution-free nature of the T_{nk} under the null hypothesis provides the access to the computation of the test function and the critical level. If n is at least moderately large, in view of the asymptotic normality of T_{nk}^o/ν_n, the randomization test function may be replaced by a conventional normal theory test function, where for the individual tests, a significance level α^* is so chosen that

$$\alpha = 1 - (1 - \alpha^*)^K. \tag{4.3}$$

Generally, if we let $\alpha^* = (\alpha/K)$, then the size of the UIT is $\le \alpha$ no matter whether the T_{nk}^o are stochastically independent or not. There is, therefore, a certain amount of conservativeness in this specification.

In passing, we may remark that by the classical asymptotics on Hoeffding's U-statistics, any pair (T_{nk}^o, T_{nq}^o), with $k \ne q$, is a bivariate U-statistic, for α^* sufficiently small, so using the bivariate extreme statistics results (viz., [18]), we can claim that the events $\{T_{nk}^o > c_{\alpha^*}\}$ and $\{T_{nq}^o > c_{\alpha^*}\}$ will be asymptotically (as $K \to \infty$) independent so that $P_0\{T_{nk}^o > c_{\alpha^*}, T_{nq}^o > c_{\alpha^*}\}$ can be well approximated by $[P_0\{T_{nk}^o > c_{\alpha^*}\}]^2$. In a similar manner, the third order probability terms can be handled, and the Bonferroni bound retaining the second and third order probabilities provide a good approximation : $\alpha = K\alpha^* - \binom{K}{2}\alpha^{*2} + \binom{K}{3}\alpha^{*3} + o(\alpha^{*3})$. As a result, $\alpha^* = (\alpha/K)$ provides a good approximation to the level of significance. Therefore, for the UIT, when K is large, even when the genes are not stochastically independent, letting $\alpha^* = (\alpha/K)$ we may consider the following multiple hypothesis testing scheme:

For a chosen $\alpha^* = K^{-1}\alpha$, obtain the marginal distributional critical level c_{α^*}, and reject those H_{0k}; $k \in \{1, \ldots, K\}$ for which the corresponding T_{nk}^o exceeds c_{α^*}

A randomization test function can be prescribed when n is not adequately large. Thus, the UIT provides a bound on the *family wise error rate*, FWER. If we take $\alpha* \sim \alpha/K$ and K is large, we need to make sure that n is so large that $\nu_n^{-1} c_{\alpha^*} < 1$; this will imply that if we are to use the permutation null distribution of any T_{nk}^o , being attracted by the permutational central limit theorem, it has a nonzero mass point beyond c_{α^*}/ν_n. If $\nu_n^2 = O(n^{-1})$, as is typically the case, then $c_{\alpha^*} = O(n^{-1/2}\sqrt{-2\log\alpha^*})$ so that $\log K = O(n)$ and this does not appear to be a serious concern in real life applications. For example, if we have three groups of arrays, say within each group there are 5 arrays, the total number of partitioning 15 units into 3 subsets of 5 each is equal to $(15)!/(5!)^3$ and this is so large (756,756) that even if K is as large as 30,000, it would not be a problem. However, for large K, the UIT, like the classical likelihood ratio test, will have little power, and hence alternative test procedures need to be explored. This illustrates the important role of the design of the study and the number of arrays required in trying to include a very large K.

Roy's UIT can be adapted by exploring the information contained in the ordered p-values. If the T^o_{nk} are all stochastically independent (and as they are identically distributed under the null hypothesis H_0) then one can adapt Simes' [19] theorem (which is a restatement of the classical Ballot theorem (viz., [9]) introduced some twenty years earlier). If $P_1, \ldots, P_K$ are the p-values for the K marginal tests and $P_{K:1} \le \cdots \le P_{K:K}$ are the corresponding order statistics, then assuming that under H_0 the P_k have a uniform (0, 1) distribution (i.e., tacitly assuming that the T^o_{nk}/ν_n have a continuous distribution under H_0), Simes' theorem asserts that for every $\alpha : 0 < \alpha < 1$,

$$P\{P_{K:k} \; > \; k\alpha/K, \; \forall\, k = 1, \ldots, K \;|H_0\} = 1 - \alpha. \tag{4.4}$$

Suppose now we define the *anti-ranks* $S_1, \ldots, S_K$ by letting

$$P_{K:k} = P_{S_k}, \; k = 1, \ldots, K, \tag{4.5}$$

where again ties among the ranks are neglected under the assumption of continuity of the distribution of the P_k. Whereas Simes' theorem provides a test of the overall hypothesis, Hochberg [6] derived a step-up procedure for multiple hypotheses testing based on the following : For every $\alpha \in (0, 1)$,

$$P\{P_{K:k} \ge \alpha/(K - k + 1), \; \forall\, k = 1, \ldots, K \;|H_0\} = 1 - \alpha. \tag{4.6}$$

Benjamini and Hochberg [2] considered a step-up procedure based on the Simes theorem. Their multiple hypothesis testing procedure is the following:

Reject those null hypotheses $\{H_{0S_k}\}$ *for which* $P_{S_k} \le k\alpha/K, \; k = 1, \ldots, K,$ *and accept those null hypotheses in the complementary set.*

For some related developments in a parametric setup, we refer to [2], [4], [10], [20], and [13], among others.

These developments paved the way for other measures of error rates which are more adaptable in the $K \gg n$ environment. Some of these will be discussed later on. There are two basic concerns that can be voiced in this respect. The whole setup is based on the assumed uniform distribution of the P_k under the null hypothesis. However, if we look into the statistics T^o_{nk} in our setup, we may note that though they have a specified distribution, the latter is a discrete one defined over the interval $(-1, 1)$. Noting that there are a set of discrete mass points, ties among the T^o_{nk}/ν_n (and hence P_k) can not be neglected with probability one, and moreover, the P_k will have a set of probability mass points on $[0, 1]$ with non-zero masses. Thus, technically the above probability results are not strictly usable (unless n is indefinitely large, contradicting the $K \gg n$ environment). Secondly, as was stressed earlier, the T^o_{nk} across the set of genes are generally not stochastically independent. Controlling the FWER when K is very large may generally entail undue conservativeness of multiple hypotheses testing schemes. On the other hand, using a level of significance for each marginal hypothesis testing problem may lead to a large FWER.

In the context of microarrays suppose that there are K_1 disease genes (DG) and $K_0 = K - K_1$ NDG; thus, we have a set of K_0 null hypotheses which are true and a complementary set of K_1 hypotheses which are not true. Suppose that based on our multiple hypotheses testing procedure, we accept m_0 out of K_0 true null hypothesis so that the remaining $K_0 - m_0 = m_1$ true null hypotheses are rejected. Similarly, among the K_1 not true null hypotheses, l_0 are accepted as true and l_1 accepted in favor of the alternative. Thus, a totality of $R = m_1 + l_1$ hypotheses are rejected while $K - R$ are accepted. Mind that though we observe R, through our chosen multiple hypotheses testing procedure, individually m_1, l_1 are not observable; all these (R, l_1, m_1) are stochastic in nature. A natural modification of the FWER, to suit such $K \gg n$ environments, is the *per-comparison error rate* (PCER) defined as

$$\text{PCER} \ = \ E(m_1)/K, \tag{4.7}$$

which is the expected proportion of Type I errors among the K hypotheses. A related measure is the *per-family error rate* (PFER), defined as

$$\text{PFER} \ = \ E(m_1), \tag{4.8}$$

which is the expected total number of Type I errors among the K hypotheses. Obviously, PFER $= K$.PCER ,and is generally large when K is large (unless the PCER is very small). Moreover,

$$\text{PFER} = E(m_1) = \sum_{r \geq 1} rP\{m_1 = r\} \geq P\{m_1 > 0\}, \tag{4.9}$$

so that PFER $\geq$ FWER.

If our observed $R = 0$ then no true null hypothesis is rejected and hence there is no false discovery. For $R \geq 1$, the proportion of false discovery is given by $Q = m_1/R$; conventionally, it is taken $Q = 0$ when $R = 0$, so that Q is properly defined for every nonnegative R and m_1. However, Q is not observable. Hence, the *false discovery rate* (FDR) is defined as

$$\text{FDR } = E\{Q\} = \sum_{r \geq 1} P\{R = r\} E\{m_1/R | R = r\}. \tag{4.10}$$

Since, conventionally, we have forced $Q = 0$ for $R = 0$, this definition of FDR may produce a negative bias. An alternative definition, known as the ^{p}FDR, is defined as

$$^{p}\text{FDR} = E\{Q | R > 0\} = \text{FDR}/P\{R \geq 1\}. \tag{4.11}$$

Naturally, ^{p}FDR $\geq$ FDR.

In the formulation of FDR and ^{p}FDR it is not necessary to assume that all of the test statistics have continuous distributions under the null hypothesis. If these distributions are all continuous then of course the p-values have a uniform (0,1) distribution under the null hypothesis, and hence, the multiple hypotheses testing schemes discussed earlier can be conveniently adapted. In our setup, each test statistic has marginally the same null distribution, albeit that is discrete. So, it might be necessary, especially when n is not large, to make use of this otherwise

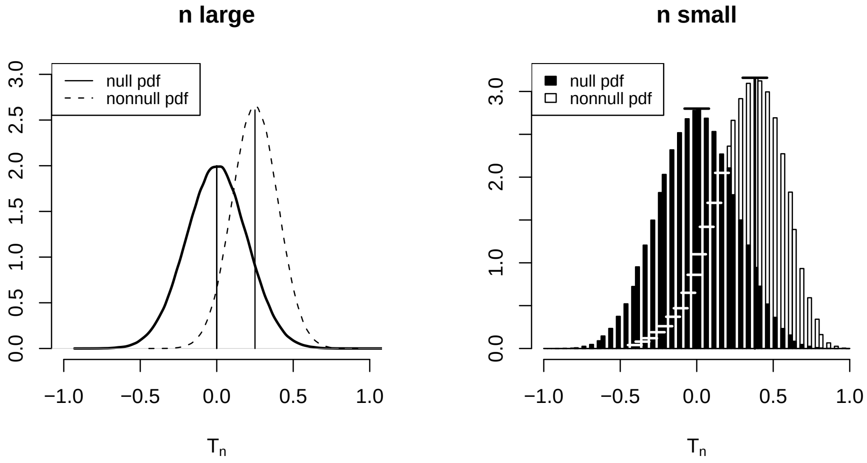


FIG 1. Comparison of the null distribution with the alternative distribution.

completely specified, discrete distribution without assuming a uniform distribution for the associated p-values under the null hypothesis.

We may simulate the permutation distribution of any marginal test statistics and thereby take into account possible dependence among the gene expressions without assuming any specific pattern. Of course, marginally, each test statistic has the same null distribution. So, if we consider the set $\{T^o_{nk}: \ k = 1, \ldots, K\}$ and define the empirical distribution

$$G_K(t) = K^{-1} \sum_{k=1}^{K} I(T^o_{nk} \leq t), \ t \in (-1, 1), \tag{4.12}$$

then $E_0\{G_K(t)\} = G(t), \forall t \in (-1, 1)$ where $G(t)$ is the common marginal distribution of the T^o_{nk} under the null hypothesis. The summands in $G_K(t)$ are all bounded variables, nondecreasing in $t \in (-1, 1)$ and $G(t)$ is also nondecreasing and assumes values on $(0, 1)$. Thus, whenever $G_K(t)$ stochastically converges pointwise to $G(t)$, it does so uniformly in $t \in (-1, 1)$. Further $G_K(t) - G(t)$ is a bounded r.v., and hence, if it converges in probability, it converges in the rth mean for every $r > 0$. Therefore it might suffice to assume that the dependence pattern satisfies the condition:

$$\text{Var}(G_K(t)) \to 0, \text{as } K \to \infty. \tag{4.13}$$

Then we conclude that $\|G_K(.) - G(.)\| = \sup\{|G_K(t) - G(t)| : \ t \in (-1, 1)\}$ stochastically converges to 0. Further, (4.13) holds under quite general dependence patterns.

It is naturally tempting to explore weak convergence (invariance principles) results for $\sqrt{K}(G_K(.) - G(.))$ wherein K is taken indefinitely large but not n. Since $G(t), t \in (-1, 1)$ is a discrete distribution function with mass points over $(-1, 1)$, the jump-discontinuities of $G(.)$ may vitiate the usual compactness (or tightness) properties possessed in the continuous case, albeit by strengthening (4.13) to

$$\limsup_K \ K\text{Var}(G_K(t)) < \infty, \forall \ t \ \in \ (-1, 1), \tag{4.14}$$

pointwise, the asymptotic normality (as $K \to \infty$) follows under quite general dependency conditions. If we have some linear functional of $G_K(.)$ as a test statistic, this weak convergence would have been quite useful in deriving the asymptotic (in K) normality of the test statistic under the null hypothesis; (4.14) would have been sufficient in that context. However, in our case, we have some functional of $G_K(.)$, of extremal order statistic type, namely, the extreme quantiles of a set of dependent r.v.s, and hence we may need somewhat different regularity conditions. This perspective is appraised more elaborately in the next section.

5. Dimensional Asymptotics and Chen–Stein Theorem

In the previous section we have briefly discussed the plausibility of some K_o NDG and K_1 DG with $K_o + K_1 = K$, the total number of genes. Neither K_1 nor the DG positions are known and hence we have a dual problem of estimating K_1 as well as identifying the positions of these K_1 DG's. It is conceivable that the NDG having stochastically smaller expression levels (than the DG) and the stochastic dependence among the DG may not be insignificant. We intend to incorporate this stochastic dependence structure among the gene expressions in a suitable model. Unfortunately, sans any positional ordering of the K genes, it might be difficult to assume suitable mixing conditions under which central limit theorems may apply. As for considering alternative limit theorems for dependent sequences, we intend to incorporate the Chen–Stein theorem [3] and its ramifications wherein Poisson approximations for more general dependent sequences are advocated. For our convenience, let us state the Chen-Stein Theorem in a slightly updated version [1].

Theorem 1. *(Chen–Stein): Let $\mathcal{I}$ be an index set with elements $i \in \mathcal{I}$ and let K be the cardinality of the set $\mathcal{I}$. For each $i \in \mathcal{I}$ let Y_i be an indicator random variable and let*

$$P\{Y_i = 1\} = 1 - P\{Y_i = 0\} = p_{Ki},\ i \in \mathcal{I}. \tag{5.1}$$

Let $W = \sum_{i\in\mathcal{I}} Y_i$ the total number of occurrence of the events $\{Y_i = 1\}$, $i \in \mathcal{I}$, and let $\lambda_K = \sum_{i\in\mathcal{I}} p_{Ki} = E(W)$. For each $i \in \mathcal{I}$, we define a set $\mathcal{J}_i \in \mathcal{I}$ and its complement $\mathcal{J}_i^c$ as the set of dependence of i and its complement, set of independence of i. Thus, it is tacitly assumed that Y_i is independent of $\{Y_j,\ j \in \mathcal{J}_i^c\}$, for every $i \in \mathcal{I}$. Further, let

$$\begin{aligned} b_1 &= \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}_i} E(Y_i)E(Y_j); \\ &= \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}_i} p_{Ki}p_{Kj}, \end{aligned} \tag{5.2}$$

$$b_2 = \sum_{i\in\mathcal{I}} \sum_{j(\neq i)\in\mathcal{J}_i} E(Y_iY_j), \tag{5.3}$$

and

$$b_3 = \sum_{i\in I} E|\{E(Y_i - E(Y_i)|\{Y_j, \forall j \in \mathcal{J}_i^c\})|. \tag{5.4}$$

Finally, let Z be a random variable having Poisson distribution with parameter $E(Z) = \lambda_K$. Then

$$\begin{aligned} \|\mathcal{L}(W) - \mathcal{L}(Z)\| &\leq 2(b_1 + b_2 + b_3)\frac{1 - e^{-\lambda_K}}{\lambda_K} \\ &\leq 2(b_1 + b_2 + b_3)\min\{1, \lambda_K^{-1}\}. \end{aligned} \tag{5.5}$$

A direct corollary to Theorem 1 is the following:

$$|P\{W=0\}-e^{-\lambda_K}| \leq 2(b_1+b_2+b_3)\min\{1,\lambda_K^{-1}\}. \tag{5.6}$$

An interesting feature of this Theorem is the dual control of λ_K, the expectation and b_1, b_2, and b_3, the dependence functions. In line with our intended application we consider a natural extension of this result. With the same notation as in Theorem 1, we replace the $Y_i, i \in \mathcal{I}$, by a sequence of processes $Y_i(t), i \in \mathcal{I}, t \in T$, where $T=(0,a)$, for some $a>0$, and assume that for each i, $Y_i(t)$ is nondecreasing in t and yet a zero-one valued random variable. Further assume that the sets $\mathcal{J}_i$ do not depend on $t \in T$. For every $i \in \mathcal{I},\ t \in T$, we denote by $p_{Ki}(t)=E(Y_i(t))$, and the corresponding parameters by $\lambda_K(t),\ b_1(t), b_2(t)$ and $b_3(t)$. Let $\mathbf{W}_K=\{W_K(t), t \in T\}$ be the sum process and corresponding to Z, we introduce a Poisson process $\mathbf{Z}_K=\{Z_K(t), t\in T\}$ whose expectation process is $\{\boldsymbol{\lambda}_K=\{\lambda_K(t), t \in T\}$. Then

$$\|\mathcal{L}(\mathbf{W}_K)-\mathcal{L}(\mathbf{Z}_K)\| \ \leq 2\ \sup\{(b_1(t)+b_2(t)+b_3(t))\frac{1-e^{-\lambda_K(t)}}{\lambda_K(t)} : t\in T\}.$$

The proof of this extension is along the lines of Theorem 1 and hence we omit the details.

In our study, unless n is large, we may not have a continuous time parameter ($t \in T$). Thus, we consider an intermediate result that remains applicable for small n as well.

Theorem 2. *Consider a set of M discrete time points $-1 \leq \tau_1 < \ldots < \tau_M \leq 1$ with respective probability masses $\eta_{n1}, \ldots, \eta_{nM}$ where M may depend on n. Also, let $\nu_{nj}=\sum_{i\leq j}\eta_{ni},\ j=1,\ldots,M$. Further, let $Y_i(\tau_j), i=1,\ldots K,\ j=1,\ldots,M$ be an array of zero-one valued random variables where $Y_i(\tau_j)$ is nondecreasing in τ_j and $E(Y_i(\tau_j))=\nu_{nj},\ j=1,\ldots,M$. Define $\mathbf{W}_K=\{W_K(\tau_j), j=1,\ldots,M\}$ where $W_K(\tau_j)=\sum_{i=1}^K Y_i(\tau_j)$ for $j=1,\ldots,M$. Similarly, let $\mathbf{Z}_K=\{Z_K(\tau_j), j=1,\ldots,M\}$ be a discrete time parameter Poisson process with the drift function $\boldsymbol{\nu}_K=\{\nu_{nj}, j=1,\ldots,M\}$. Define the parameters $b_{K1}(\tau_j), b_{K2}(\tau_j), b_{K3}(\tau_j),\ j=1,\ldots,M$ as in (5.2), (5.3), and (5.4); assume that as $K\to\infty$,*

$$\max\{(b_{K1}(\tau_j)+b_{K2}(\tau_j)+b_{K3}(\tau_j))\frac{1-e^{-\nu_{nj}}}{\nu_{nj}} : j \leq M\} \to 0. \tag{5.7}$$

Then, as K increases indefinitely,

$$\|\mathcal{L}(\mathbf{W}_K)-\mathcal{L}(\mathbf{Z}_K)\| \to 0. \tag{5.8}$$

Again, being a finite-dimensional version of Theorem 1, this does not need an elaborate proof.

In the present context, under the null hypothesis, all the T^o_{nk} have a common distribution over $(-1,1)$; this is discrete but symmetric about 0, and is completely known (though could be computationally intensive if n is not too small). Let us denote the distinct mass points for T^o_{nk} by $-1=a_1<a_2<\cdots, a_L=1$ and let

$$\tau_j = P_0\{T^o_{nk}\geq a_{L-j+1}\},\ j=1,\ldots,L. \tag{5.9}$$

Then $0\leq \tau_1<\tau_2<\cdots<\tau_L\leq 1$. Also, let us write

$$Y_k(\tau_j)=I(T^o_{nk}\geq a_{L-j+1}),\ j=1,\ldots,L,\ k=1,\ldots,K. \tag{5.10}$$

Further, let

$$W_K(\tau_j) = \sum_{k=1}^{K} Y_k(\tau_j),\ j = 1, \dots, L. \tag{5.11}$$

Also, let $J = \max\{j : 1 \le j \le L;\ \tau_j \le \eta\}$ for some pre-assigned $\eta > 0$. Basically, we would like to pursue the distributional features of the partial sequence $\{W_K(\tau_j),\ j \le J\}$, and incorporate Theorem 2. Note that in this way, we avoid the conventional assumption of a continuous null distribution of the coordinate-wise test statistics. Of course, if n is adequately large, the assumption of a uniform distribution of the p-values (under the null hypothesis) would be reasonable. For example, if we have a three sample situation with $n_1 = n_2 = n_3 = 4$ then $L = (12)!/(4!)^3 = 34,650$ so that we could choose $J = 1$ and use the Poisson approximation. It is also possible to choose $J = 2$ with an appropriate cut-off point and still stick to a FWER around 0.05. In any case, under alternatives (of stochastic ordering) the distribution of the T^o_{nk} will be tilted towards the right, still confined to the interval (-1, 1), and hence, their centering would be shifted to the right of the origin with a negatively skewed distribution.

Corresponding to the known points $\tau_1 < \cdots < \tau_J$, let us consider the partial process $W_K(\tau_j), j = 1, \dots, J$, as defined above. Also, let us choose a set of nonnegative integers $r_1 \le \cdots \le r_J$ in such a way that

$$P_0\{W_K(\tau_j) > r_j, \text{for some } j \le J\} = \alpha, \tag{5.12}$$

where α may not be exactly equal to a specified level (such as 0.05) but can be approximated very well through the above Poisson process result. If we let

$$A_j = [\ W_K(\tau_j) > r_j\ ],\ j = 1, \dots, J, \tag{5.13}$$

then (5.12) can be written as $P\{\ \cup_{j \le J} A_j\ \}$, so that by the Bonferroni inequality,

$$\begin{aligned} P\{\ \cup_{j\le J} A_j\} &= \sum_{j \le J} P\{A_j\} - \sum_{1 \le j < j' \le J} P\{A_j A_{j'}\} \\ &+ \sum_{1 \le j<k<l \le J} P\{A_j A_k A_l\} + \cdots + (-1)^K P\{A_1 \cdots A_K\}. \end{aligned} \tag{5.14}$$

Next, we use the Poisson approximation to each $P\{A_j\}$ wherein we use the following:

$$P\{A_j\} \sim e^{-\nu_{nj}}\{\sum_{r > r_j} \nu^r_{nj}/r!\},\ j \ge 1. \tag{5.15}$$

Further, note that $W_K(\tau_j)$ is a nondecreasing (step) function in j so that using the Markov property and Theorem 2 we may evaluate $P\{A_j A_{j'}\}$. Actually, we write $P\{A_j A_k\} = P\{A_j\} \cdot P\{A_k | A_j\}$, for $k > j$, and use Theorem 2 to approximate the conditional probability by $P\{Z_k > r_k | Z_j > r_j\}$ where $r_k \ge r_j, \forall j < k$. Also, typically terms involving more than 2 events (A_j) will be small and can usually be neglected. Nevertheless, even if they are not small, the Markov property embedded in Theorem 2 can be used to provide a good approximation. Alternatively, we may write $P\{\cup_{j \le J} A_j\} = 1 - P\{\cap_{j \le J} A^c_j\}$ and using Theorem 2, write $P\{\cap_{j \le J} A^c_j\}$ as a J-tuple sum over Poisson distributional probabilities. For small $r_j, j \le J$, as is typically the case, this computation does not appear to be a formidable task.

Led by these findings, let us now consider the following testing procedure:

Compute the $W_K(\tau_j),\ j \leq J$ *as above. If* $W_K(\tau_j) \leq m_j,\ \forall j \leq J$*, accept the null hypothesis that there is no DG. On the other hand, if* $W_K(\tau_j)$ *is greater than* m_j *for at least one* $j \leq J$*, then reject the null hypothesis that all the genes are NDG, and proceed to detect those genes* $k \in \mathcal{K}$ *as DG where*

$$\mathcal{K} = \{k \in \{1, \dots, K\} : Y_k(\tau_j) = 1,\ \text{for some } j \leq J\}. \tag{5.16}$$

Note that if for some k, $Y_k(\tau_j) = 1$ for some $j \leq J$, then $Y_k(\tau_{j'}) = 1,\ \forall\ j' \geq j$. Further, note that $\mathcal{K}$ is a stochastic subset of $\{1, \dots, K\}$, and $R =$ cardinality of $\mathcal{K}$ is a (nonnegative) integer valued random variable. The overall significance level of this testing procedure is well approximated by the preassigned level α.

Let us denote the following exclusive events by

$$B_1 = A_1;\quad B_j = A_1^c \cdots A_{j-1}^c A_j,\ j \leq J. \tag{5.17}$$

Then, by definition, $A_j = \cap_{j \leq J} B_j$. With the same notation as in (4.7)—(4.11), we study the other measures (viz., PCER, PFER, FDR and ^{p}FDR). Towards this, we consider the nonnull situation where K_0 are NDG and $K_1 = K - K_0$ are DG. To handle the distribution of R, the total number of rejections, we let

$$\tau_j^* = (K_0\tau_j + K_1\beta_j)/K = \tau_j + (K_1/K)(\beta_j - \tau_j),\ j \geq 1, \tag{5.18}$$

where

$$\beta_j = K_1^{-1} \sum_{k \in DG} P\{T_{nk}^o \geq a_{L-j+1} | k \in\ \{1, \dots, K\} - \mathcal{K}_0\}, \tag{5.19}$$

for $j = 1, \dots, J$. Note by arguments similar to those in Sections 3 and 4, $\beta_j \gg \tau_j, \forall j \leq J$. We may write

$$E(m_1) = \sum_{j \leq J} E(m_1 I(B_j)). \tag{5.20}$$

Next note that the events $B_j, j \leq J$, depend on the partial process $W_K(\tau_j), j \leq J$ and are thereby governed by Theorem 2 with $\nu_{nj} = Kt_j^*, j \leq J$. On the other hand, the distribution of m_1 is governed by the process $W_K^o(\tau_j), j \leq J$, where the drift function for $W_K^o(\tau_j))$ is $\nu_{nj}^o = K_0\tau_j, j \leq J$. Using Theorem 2 and the reproductive property of the Poisson distribution, we may well approximate the (conditional) distribution of m_1, given R, by a binomial law with parameters $(R, K_0\tau_j/(K_0\tau_j + K_1\beta_j))$ whenever B_j holds. Thus, we are able to provide a good approximation to the PFER by writing

$$E(m_1) = \sum_{j \leq J} E\{E(m_1/R \mid R, B_j) R I(B_j)\}. \tag{5.21}$$

If $J = 1$, the conditional binomial law directly applies and we have the approximation

$$\frac{K_0 t_1}{K_0 t_1 + K_1 t_1^*}.E(RI(R > r_1)) = K_0 t_1 P\{R \geq r_1\}, \tag{5.22}$$

where the last step follows from the fact that for a Poisson variable X with parameter np, $E(XI(X > r)) = npP\{X \geq r\}$. For $J \geq 2$, we have to apply the conditional binomial law under the sets B_j, followed by the distribution of R over the sets B_j, and this can be done by repeated quadrature procedures. Numerical studies have thereby good scope.

By construction, rejection of the null hypothesis H_0 entails that $R > r_1$ and may even be greater than r_1 if B_j pertains for some $j \geq 1$. As such, we do not have any problem in applying the original definition of FDR (in (4.10)). We write

$$\text{FDR} = E(Q) = \sum_{j \leq J} E(QI(B_j)) = \sum_{j \leq J} E\{E(Q|R \in B_j)I(B_j)\} \tag{5.23}$$

and use the conditional binomial law for each term in the right hand side. Detailed numerical study is planned for a future communication.

We conclude this section with some pertinent remarks and observations. First, the use of the Chen–Stein theorem in a multi-state context can be done under fairly mild regularity conditions regarding the dependence of the genes. Secondly, by our choice of the $r_j, j \leq J$ and allowing possibly $J \geq 1$, we are not only in a position to allow more flexibility in the choice of statistical inference procedures but also to enforce the rejection of null hypothesis under a more structured setup. This allows us to study the FDR, etc., under more diverse setups. Further, using Kendall's tau statistic for each gene separately, we are in a position to allow heterogeneity of the gene expressions across the K genes in a completely arbitrary manner, while under the null hypothesis, the distribution of the $T^o_{nk}, k = 1, \ldots, K$ being completely known provides an easy access to the incorporation of the Chen–Stein theorem. Finally, instead of using Kendall's tau statistic (coordinate-wise), it might be attractive to use more general rank statistics [16]. Though the distribution-free aspect holds under the null hypothesis, such distributions are more complex to evaluate and the associated Poisson processes have more complex drift functions. Further, such linear rank statistics involve some design variables which assume more structure on the $F_{ik}, k = 1, \ldots, K$, not necessary with the use of Kendall's tau.

Acknowledgments.

Research partially supported by the C. C. Boshamer Research Foundation at the University of North Carolina, Chapel Hill. The author is grateful to the reviewers for their critical reading of the manuscript and most helpful comments. Thanks are also due to Dr. Moonsu Kang and Sunil Suchandran for providing the Figure in the text.

References

[1] Arratia, R., Goldstein, L. and Gordon, L. (1990). Poisson approximation and the Chen-Stein method : Rejoinder. *Statistical Science* **5**, 432–434.

[2] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B* **57**, 289–300.

[3] Chen, L. H. Y. (1975). Poisson approximation for dependent trials. *Ann. Probab.* **3**, 534–545.

[4] Dudoit, S., Shaffer, J. and Boldrick, J. (2003). Multiple hypothesis testing in microarray experiments. *Statist. Sci.* **18**, 71–103.

[5] Ghosal, S., Sen, A. and van der Vaart, A. W. (2000). Testing monotonicity of regression. *Ann. Statist.* **28**, 1054–1081.

[6] Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800–802.

[7] Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19**, 293–325.

[8] Jurečková, J. and Sen, P. K. (1996). *Robust Statistical Procedures: Asymptotics and Interrelations.* John Wiley, New York.

[9] Karlin, S. (1969). *A First Course in Stochastic Processes.* Academic Press, New York.

[10] Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. *Ann. Statist.* **33**, 1138–1154.

[11] Peddada, S., Harris. S., Zajd, J. and Harvey, E. (2005). ORIGEN: Order restricted inference ordered gene expression data. *Bioinformatics* **21**, 3933–3934.

[12] Roy, S. N. (1953). A heuristic method of test construction and its use in multivariate analysis. *Ann. Math. Statist.* **24**, 220–238.

[13] Sarkar, S. K. (2006). False discovery and false nondiscovery rates in singlestep multiple testing procedures. *Ann. Statist.* **34**, 394–415.

[14] Sen, P. K. (1968). Estimates of regression coefficients based on Kendall's tau. *J. Amer. Statist. Assoc.* **63**, 1379–1389.

[15] Sen, P. K. (2004). *Excursions in Biostochastics: Biometry to Biostatistics to Bioinformatics* Invited Lecture Ser. No. 5, Institute of Statistical Studies, Academia Sinica, Taipei.

[16] Sen, P. K. (2006). Robust statistical inference for high-dimensional data models with applications to genomics. *Austrian J. Statist.* **35**, 197–214.

[17] Sen, P. K, Tsai, M.-T. and Jou, Y.-S. (2007). High-dimension low sample size perspectives in constrained statistical inference : The SARSCoV RNA genome in illustration. *J. Amer. Statist. Assoc.* **102**, 686–694.

[18] Sibuya, M. (1959). Bivariate extreme statistics. *Ann. Inst. Statist. Math.* **11**, 195–210.

[19] Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754.

[20] Storey, J. (2007). The optimal discovery procedure: A new approach to simultaneous significance testing. *J. Roy. Statist. Soc. B* **69**, 1 – 22.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 268
© Institute of Mathematical Statistics, 0000
arXiv: math.PR/0000000

Orthogonalized Smoothing for Rescaled Spike and Slab Models

Hemant Ishwaran[1] **and Ariadni Papana**[2]

Cleveland Clinic and Case Western Reserve University

Abstract: Rescaled spike and slab models are a new Bayesian variable selection method for linear regression models. In high dimensional orthogonal settings such models have been shown to possess optimal model selection properties. We review background theory and discuss applications of rescaled spike and slab models to prediction problems involving orthogonal polynomials. We first consider global smoothing and discuss potential weaknesses. Some of these deficiencies are remedied by using local regression. The local regression approach relies on an intimate connection between local weighted regression and weighted generalized ridge regression. An important implication is that one can trace the effective degrees of freedom of a curve as a way to visualize and classify curvature. Several motivating examples are presented.

Contents

1. Introduction

Rescaled spike and slab models were introduced in [1] as a Bayesian variable selection method in linear regression models. Such models were shown to possess a *selective shrinkage* property in orthogonal models. This property allows the posterior mean for the coefficients to shrink to zero for truly zero coefficients while for the

[1] Department of Quantitative Health Sciences Wb4, Cleveland Clinic, 9500 Euclid Avenue, Cleveland, Ohio 44195, e-mail: hemant.ishwarant@gmail.com

[2] Department of Statistics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106, e-mail: ariadni.papana@case.edu

AMS 2000 subject classifications: Primary 62J07; secondary 62J05
Keywords and phrases: selective shrinkage, penalization, effective degrees of freedom

non-zero coefficients posterior estimates are similar to the ordinary least squares (OLS) estimates. In [2], rescaled spike and slab models were used to analyze multi-group microarray data (an extension of previous work [3]). Selective shrinkage was shown to be a sufficient condition for oracle-like total misclassification. A finite sample adaptive method for selecting variables using this principle was given.

In this manuscript we extend the application of rescaled spike and slab models to smoothing problems. Given an outcome value Y related to a variable x through an unknown function $f(x)$, we would like accurate recovery of $f(x)$. Smoothing is a prediction problem, and an important contribution of the paper is advancing applications of rescaled spike and slab models to prediction settings. However this does not mean selective shrinkage, a core ingredient to model selection, is not at play in a prediction paradigm. Indeed, as shown, selective shrinkage plays a crucial role in adaptive selection of over-parameterized basis functions in response to curvature of $f(x)$.

We consider global smoothing via orthogonal polynomial regression as well as local regression using orthogonal polynomials. Orthogonality is a key ingredient in our approach. Not only does it allow us to exploit the selective shrinkage property of rescaled spike and slab models, which follow as a consequence of orthogonality, but it also greatly improves the computationally efficiency of our algorithms. While much work has been done in the area of smoothing, we note there are novel features in our approach potentially useful in applied settings. One important feature being that selective shrinkage allows for greater adaptivity to curvature and greater robustness to misspecification of dimension of basis functions. Secondly, in local regression settings, adaptivity via selective shrinkage can be interpreted in terms of dimensionality and curvature. From this we provide an effective degrees of freedom plot for graphing estimated dimensionality of $f(x)$ as a function of x. Such plots provide a simple and powerful way to register curves. Several applications are provided as illustration.

2. Rescaled Spike and Slab Models

We begin by first reviewing background theory for rescaled spike and slab models. The underlying setting is the linear regression model where $Y_1,\ldots,Y_n$ are independent responses such that

$$Y_i = \beta_1 x_{i,1} + \cdots + \beta_d x_{i,d} + \varepsilon_i = \mathbf{x}_i^t \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1,\ldots,n. \tag{2.1}$$

Here $\mathbf{x}_1,\ldots,\mathbf{x}_n$ are non-random (fixed design) d-dimensional covariates and $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_d)$ is the unknown coefficient vector. The ε_i are independent random variables (but not necessarily identically distributed) such that $E(\varepsilon_i) = 0$, $E(\varepsilon_i^2) = \sigma_0^2$ and $E(\varepsilon_i^4) \leq M$ for some $M < \infty$ (the last condition is needed to invoke a triangular central limit theorem later, but is not crucial and can certainly be relaxed). The variance $\sigma_0^2 > 0$ is assumed to be unknown. Throughout we assume $\mathbf{x}_i$ are standardized so that $\sum_{i=1}^n x_{i,k} = 0$ and $\sum_{i=1}^n x_{i,k}^2 = n$ for $k = 1,\ldots,d$ (without loss of generality we assume that there is no intercept term in (2.1)). We shall also assume throughout that $\mathbf{X}$, the $n \times d$ design matrix, is orthogonal, i.e., $\mathbf{X}^t\mathbf{X} = n\mathbf{I}$. As mentioned in the Introduction, this will allow us to exploit certain elegant theories for rescaled spike and slab methods, although, of course, the spike and slab method works for general design matrices.

Spike and slab methods first appeared in the works of [4] and [5] for subset selection in linear regression models. The expression "spike and slab", coined by

Mitchell and Beauchamp in [5], referred to the prior for the regression coefficients used in their hierarchical formulation. This was chosen so that the coefficients were mutually independent with a two-point mixture distribution made up of a uniform flat distribution (the slab) and a degenerate distribution at zero (the spike). In [6] a different type of prior was used. This involved a scale mixture of two normal distributions. In particular, the use of a normal prior was highly advantageous and led to a Gibbs sampling method that highly popularized the spike and slab approach; see [7–11].

As pointed out in [1], priors involving a normal scale mixture distribution, of which [6] is a special example, constitute a wide class of models termed "spike and slab models". A modified class of spike and slab models called "rescaled spike and slab models" was introduced [1]. These new models differed in that the original Y_i values were replaced by new values scaled by the square root of the sample size and divided by the square root of an estimate for σ_0^2. Rescaling was shown to induce a non-vanishing penalization effect for the posterior mean, and when used in tandem with a continuous bimodal prior, the resulting posterior mean was shown to possess a selective shrinkage property in orthogonal models [1].

A *rescaled spike and slab model* was defined in [1] to denote a Bayesian hierarchical model specified as follows:

$$\begin{array}{rcll} (Y_i^*|\mathbf{x}_i,\boldsymbol{\beta}) & \overset{\text{ind}}{\sim} & \mathrm{N}(\mathbf{x}_i^t\boldsymbol{\beta}, n), & i=1,\ldots,n, \\ (\boldsymbol{\beta}|\gamma) & \sim & \mathrm{N}(\mathbf{0},\boldsymbol{\Gamma}) & \\ \boldsymbol{\gamma} & \sim & \pi(d\boldsymbol{\gamma}). & \end{array} \tag{2.2}$$

Here Y_i^* are the rescaled Y_i values defined by $Y_i^* = \hat{\sigma}^{-1}n^{1/2}Y_i$, where $\hat{\sigma}^2 = ||\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_0||^2/(n-d)$ is the unbiased estimator for σ_0^2 based on the full model, and $\hat{\boldsymbol{\beta}}_0$ is the OLS estimate for $\boldsymbol{\beta}$ (other estimators for σ_0^2 are also possible; these details, however, play a minor role). The value of n used in the first level of the hierarchy in (2.2) is a variance inflation factor introduced to compensate for the rescaling. Moreover, inclusion of n in the hierarchy was shown in [1] to be necessary for selective shrinkage to take place. Without rescaling, shrinkage for the posterior mean vanishes in the limit due to the prior becoming swamped by the likelihood [1].

In (2.2), $\mathbf{0}$ denotes a d-dimensional zero vector, $\boldsymbol{\Gamma} = \mathrm{diag}(\gamma_1,\ldots,\gamma_d)$ is a $d \times d$ diagonal matrix and π is the prior measure for $\boldsymbol{\gamma} = (\gamma_1,\ldots,\gamma_d)^t$. A Bayesian parameter σ^2 can also be introduced in (2.2) at the first level of the hierarchy. However, we avoid this approach here and opt for the simpler set up ((2.2)). The rationale for this is the following: (i) we have already removed the effect of σ_0^2 when rescaling Y_i, and (ii) the simpler setup enforces a sparse solution for the posterior mean in ill-determined settings when d is of the same size, or larger, than n. Point (ii) is especially relevant as this is the setting we are interested in here.

2.1. Rescaling, the Choice of $\boldsymbol{\pi}$, and Implications for Shrinkage

In addition to rescaling the response, the prior for γ_k must satisfy certain requirements in order for selective shrinkage to occur. A sufficient condition requires the prior to have a bimodal property such that the right tail of the distribution is continuous and such that there is a spike in the distribution near zero (see Theorem 6 of [1] for precise details). One such example is the continuous bimodal prior used in [1–3]. This prior is induced by a parameterization involving a binary variable and a positive variable with an inverse-gamma distribution. More precisely, define

γ_k by $\gamma_k = I_k \tau_k^2$, where I_k and τ_k^2 are parameters with priors specified according to

$$\begin{aligned}
(I_k|v_0, w) &\overset{\text{iid}}{\sim} (1-w)\,\delta_{v_0}(\cdot) + w\,\delta_1(\cdot), \qquad k = 1, \ldots, d, \\
(\tau_k^{-2}|a_1, a_2) &\overset{\text{iid}}{\sim} \text{Gamma}(a_1, a_2) \\
w &\sim \text{Uniform}[0,1]. \qquad (2.3)
\end{aligned}$$

The choice for v_0 (a small positive value) and a_1 and a_2 (the shape and scale parameters for a gamma density) are selected so that γ_k has a continuous bimodal distribution with a spike at v_0 and a right continuous tail (see Figure 1). Such a prior allows the posterior to shrink a coefficient to zero depending upon the value for γ_k. Small values heavily shrink a coefficient towards zero.

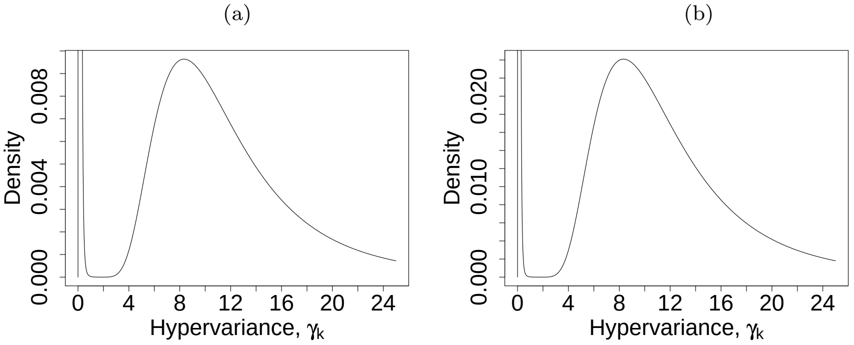


FIG 1. *Conditional density for γ_k given w: (a) $w = 0.1$, (b) $w = 0.25$. Observe that only the densities height changes as w is varied. One can think of w as a complexity parameter controlling model dimension. Prior based on hyperparameters $a_1 = 5, a_2 = 50$ and $v_0 = 0.005$ as in [1–3].*

2.2. Selective Shrinkage Recast in Terms of Penalization

One can view the posterior mean as a solution to a constrained least squares optimization problem in which the hypervariances are related to penalty terms. This provides us with another way to think about the effects of selective shrinkage. As before, we consider the orthogonal setting where $\mathbf{X}^t\mathbf{X} = n\mathbf{I}$. Let $V_k = E(\nu_k|\mathbf{Y}^*)$ where $\nu_k = \gamma_k/(1+\gamma_k)$. For our argument it will be easier to think of penalization in terms of $\hat{\boldsymbol{\beta}} = \hat{\sigma} n^{-1/2} \hat{\boldsymbol{\beta}}^*$, where $\hat{\boldsymbol{\beta}}^* = E(\boldsymbol{\beta}|\mathbf{Y}^*)$ denotes the posterior mean for $\boldsymbol{\beta}$ under our rescaled spike and slab model. It can be shown that

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ ||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 + n \sum_{k=1}^{d} \frac{1-V_k}{V_k} \beta_k^2 \right\}. \qquad (2.4)$$

Observe how each β_k coefficient in (2.4) is penalized by a unique value $(1-V_k)/V_k$. The closer V_k is to 1, the smaller the penalty and the less the shrinkage for β_k, while the closer V_k is to zero, the larger the penalty, and the more β_k is shrunk to zero. It is clear the more adaptive V_k is to the true coefficient value, the more accurate variable selection becomes.

This argument can be formalized by studying the asymptotic behavior of V_k. Using a similar argument as in [2], one can show that under the spike and slab model (2.2) with continuous bimodal prior (2.3) (specified in Figure 1), the following holds:

Theorem 2.1. *Assume that* $\max_{1\le i\le n}||\mathbf{x}_i||/\sqrt{n} \to 0$. *If (2.1) represents the true data model, and the coefficient* β_k *for variable* k *is truly non-zero, then*

$$V_k \overset{\mathrm{d}}{\rightsquigarrow} 1. \tag{2.5}$$

Moreover, if β_k *is truly zero, then*

$$E(\nu_k|w,\mathbf{Y}^*) \overset{\mathrm{d}}{\rightsquigarrow} \frac{\int_0^1 \nu\exp\big(\nu Z_k^2/2\big)(1-\nu)^{-3/2}f\big(\nu/(1-\nu)|w\big)\,d\nu}{\int_0^1 \exp\big(\nu Z_k^2/2\big)(1-\nu)^{-3/2}f\big(\nu/(1-\nu)|w\big)\,d\nu}, \tag{2.6}$$

where $f(\cdot|w) = (1-w)g_0(\cdot) + wg_1(\cdot)$ *is the prior density for* γ_k *given* w, *where* $g_0(u) = v_0u^{-2}g(v_0u^{-1})$, $g_1(u) = u^{-2}g(u^{-1})$ *and*

$$g(u) = \frac{a_2^{a_1}}{(a_1-1)!}u^{a_1-1}\exp(-a_2u)$$

and Z_k *has a* $N(0,1)$ *distribution.*

Result (2.5) of Theorem 2.1 shows that V_k approaches the value 1 in the case where there is true signal, and hence the penalty for β_k in (2.4) vanishes as the sample size increases, and β_k will not be shrunk, just as we'd expect and hope for.

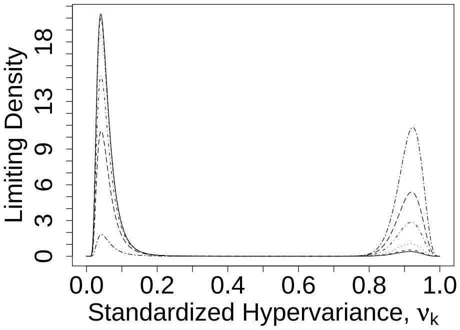


FIG 2. *Limiting density for* ν_k *conditioned on* $w = 0.1$ *and* Z_k^2 *under the null that* β_k *is truly zero. Values for* Z_k^2 *selected from the 25th, 50th, 75th, 90th, 95th and 99th percentiles of a* χ_1^2*-distribution. Mode on the left increases as* Z_k^2 *decreases, whereas mode on the right decreases.*

Result (2.6) applies to the case when β_k is really zero. The term Z_k appearing in (2.6) is the limit of the posterior mean $\hat\beta_k^*$ under the null, and thus Z_k reflects the effect of the data on the posterior under the null. In particular, unless $\hat\beta_k^*$ is unduly large, the posterior mean for ν_k should be relatively close to the value under the prior. This has implications for sparse settings. In such cases, the posterior value for w will be small and the posterior for ν_k conditioned on w (which will look like the prior given w) will be concentrated near zero. Thus, the left-hand side of (2.6) should be small and the posterior mean penalized and shrunk towards zero. On the other hand, if $\hat\beta_k^*$ is large, then the left-hand side of (2.6) will be large, and there will be less penalization and less shrinkage for β_k. A large value for $\hat\beta_k^*$ is unlikely under the null and in fact is expected only when β_k is really non-zero, which is another way to see why (2.5) holds. Figure 2 illustrates how ν_k might depend upon $\hat\beta_k^*$ in a sparse setting under the null that β_k is truly zero.

3. Orthogonal Polynomials: First Illustration

For our first illustration we consider a dataset related to spinal bone mineral density (BMD) (see [12] for a more complete description of the data). The response is the relative change in spinal BMD as a function of age in male and female adolescents. Figure 3 plots the results of our analysis. Predicted values for Y based on the posterior of the rescaled spike and slab model (2.2) are superimposed on the figure as solid dark and dashed dark lines for men and women, respectively. The analysis on the left side of the plot is based on an orthogonal polynomial design matrix with $d = 10$ basis functions. Also superimposed are OLS estimates (gray lines).

While the left side of Figure 3 shows some difference between the methods, discrepancies become more apparent if d is allowed to increase. We re-ran the same analysis but using an overly parameterized design involving $d = 25$ basis functions. The right side of Figure 3 records the result. Notice how badly OLS overfits, whereas rescaled spike and slab predictors remain relatively unaffected.

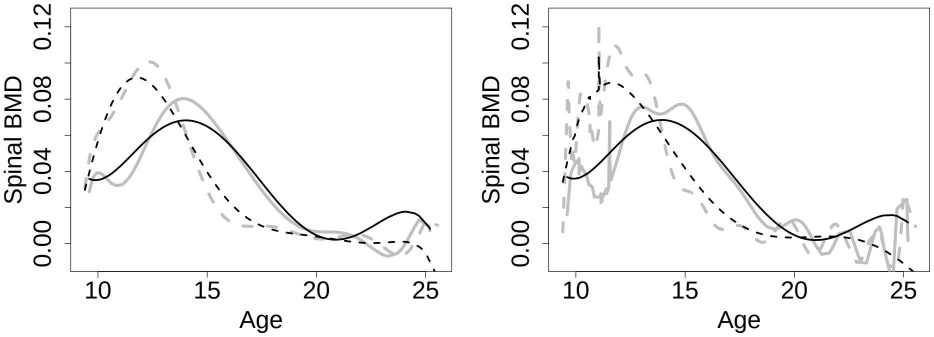


FIG 3. *Left plot: Relative change in spinal BMD as a function of age. Solid dark and dashed dark lines are spike and slab predicted curves for men and women using an orthogonal polynomial design matrix with $d = 10$ basis functions. Gray solid and dashed lines are OLS estimates for men and women. Right plot: Analysis similar as before, but now using an over parameterized basis function, $d = 25$. Note the spiky behavior of the OLS.*

3.1. Comparative Analysis Using Effective Kernels

A more formal comparison between the two approaches can be based on an effective kernel analysis. Effective kernels were introduced in ([13], Chapter 2.8), as a way to evaluate the differences between kernel smoothers. Suppose we have data (x_i, Y_i), $i = 1, \dots, n$, where Y_i are the response values. It is assumed that

$$Y_i = \mathrm{f}_i + \varepsilon_i, \qquad i = 1, \dots, n, \tag{3.1}$$

where $\mathrm{f}_i = \mathbf{x}_i^t \boldsymbol{\beta}$ are unknown mean values and $\mathbf{x}_i \in \mathbb{R}^d$ are the values of the pre-chosen underlying d basis functions evaluated at x_i. Call a smoother $s(x)$ linear in Y if for each x_0

$$s(x_0) = \sum_{j=1}^{n} S_j(x_0) Y_j,$$

where $S_j(x_0)$ depends only upon the x-values and not the responses. More generally, let $\mathbf{f} = (\mathrm{f}_1, \dots, \mathrm{f}_n)^t$. If $s(x_i)$ is a linear smoother for f_i, then

$$\hat{\mathbf{f}} = \mathbf{S}\mathbf{Y}$$

is a linear smoothed estimate of $\mathbf{f}$, where $\mathbf{S}$ is the $n\times n$ *smoother matrix*, $\mathbf{S}=\{s_{i,j}\}$ for $s_{i,j}=S_j(x_i)$. The value $S_i(x_i)=s_{i,i}$ is often referred to as the *effective kernel at* x_i [13, 14]. The effective kernel measures the influence of x_i on Y_i. The set of values $\{s_{i,j}:j=1,\dots,n\}$, which is the ith row of $\mathbf{S}$, is called the effective kernel for Y_i. Plotting the effective kernel is a way to compare different smoothers [13].

This idea can be adapted to our setting as follows. First we derive the effective kernel for the OLS estimate. Consider the orthogonal regression setting in which $\mathbf{X}^t\mathbf{X}=n\mathbf{I}$. Let $\mathbf{x}_{(k)}$ denote the kth column of the design matrix $\mathbf{X}$. It follows that $\hat{\mathbf{f}}=\mathbf{S}\mathbf{Y}$, where

$$\mathbf{S}=\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t=n^{-1}\sum_{k=1}^{d}\mathbf{x}_{(k)}\mathbf{x}_{(k)}^t. \tag{3.2}$$

The effective kernel for Y_i is $n^{-1}\sum_{k=1}^{d}x_{i,k}\mathbf{x}_{(k)}^t$ and the effective kernel at x_i is $s_{i,i}=n^{-1}\mathbf{x}_i^t\mathbf{x}_i$.

The notion of an effective kernel needs to be slightly modified to handle adaptive penalization. We adopt the notion of an adaptive smoother matrix that allows the effective kernel to depend upon both x_i and Y_i. Define the spike and slab predictor as $\hat{\mathbf{f}}^*=\mathbf{X}\hat{\boldsymbol{\beta}}$.

Theorem 3.1. *Under othogonality, the spike and slab predictor for Y_i in (3.1) can be written as $\hat{\mathbf{f}}^*=\mathbf{S}^*\mathbf{Y}$, where*

$$\mathbf{S}^*=n^{-1}\sum_{k=1}^{d}V_k\mathbf{x}_{(k)}\mathbf{x}_{(k)}^t. \tag{3.3}$$

One can conceptualize $\mathbf{S}^*$ as an adaptive linear smoother matrix. Consequently, the effective kernel for Y_i is defined as $n^{-1}\sum_{k=1}^{d}V_kx_{i,k}\mathbf{x}_{(k)}^t$, and the effective kernel at x_i is $s_{i,i}^*=n^{-1}\sum_{k=1}^{d}V_kx_{i,k}^2$.

Figure 4 shows the effective kernels at x_i for the OLS and rescaled spike and slab predictors, where x_i is age. The plots are based on the over-parameterized orthogonal polynomial design involving $d=25$ basis functions. The large number of predictors helps to emphasize the non-robustness of the OLS estimate. Note especially how the OLS estimate is affected by the points near the edges of the plots. In contrast, note the robustness of the spike and slab approach.

3.2. Effective Degrees of Freedom

The smoother matrix provides information about the nature of a predicted curve. Ideally, however, we would like a rigorous and systematic manner in which to summarize this information as a way to *register* (classify) a curve. One way to compare rigorously curves is to use the notion of the *effective degrees of freedom* [13]. For any smoother matrix $\mathbf{S}$, the effective degrees of freedom, $\mathscr{D}_f$, is defined as

$$\mathscr{D}_f(\mathbf{S})=\mathrm{tr}(\mathbf{S})=\sum_{i=1}^{n}s_{i,i}.$$

For the OLS smoother matrix (3.2), $\mathscr{D}_f(\mathbf{S})=n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{d}x_{i,k}^2=d$ (the last identity on the right is due to orthogonality). Meanwhile, for the spike and slab smoother matrix (3.3), we have the following corollary to Theorem 3.1.

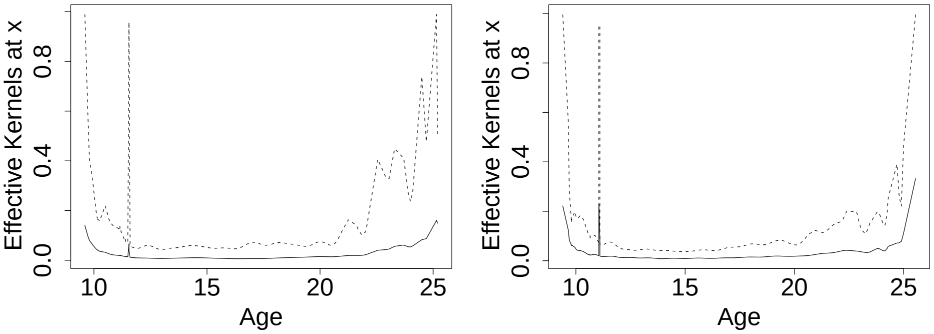


FIG 4. *Left plot: Effective kernels at x_i, $i = 1, \ldots, n$, for men using over-parameterized design, $d = 25$ (rescaled spike and slab values depicted by thick line; OLS by dashed line). Right plot: Effective kernels for women.*

Corollary 3.1. *Under the conditions of Theorem 3.1,*

$$\mathscr{D}_f(\mathbf{S}^*) = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{d} V_k x_{i,k}^2 = \sum_{k=1}^{d} V_k \le d.$$

Hence, the degrees of freedom for the spike and slab smoother is bounded by the dimension of the underlying polynomial basis.

Observe how $\{V_k\}$, the shrinkage parameters, dictate the degrees of freedom. The larger the value, the more degrees of freedom used up, and the less shrinkage there is. In the analysis presented earlier using a saturated design ($d = 25$), the effective degrees of freedom are 4.2 and 5.8 for men and women, respectively, indicating more overall shrinkage for men and evidence of differences in the two curves.

Effective degrees of freedom are useful for assessing overall differences between curves. However the method is limited in its ability to register a curve, as it reduces the overall properties of a curve to a single summary value. In the next section we illustrate a much more effective way to register curves.

4. Local Regression

In this section we illustrate how rescaled spike and slab models can be used for local regression, an alternative method of smoothing [15, 16]. By exploiting orthogonality, and by drawing connections to generalized ridge regression, we show that rescaled spike and slab predictors can be viewed as local regression smoothers with a local smoother matrix whose effective degrees of freedom can be traced over x as a way to characterize curvature of the underlying function. Another nice feature of using rescaled spike and slab models, just like in global smoothing, is that we end up being fairly robust to the choice of the dimension of the underlying basis functions.

First let's review some background on local regression. In local regresssion, for a given x_i, rather than performing a global regression to estimate f_i, one instead fits a weighted regression model using weighted least-squares, with weights for an observation x chosen by how close they are to x_i. This results in a local estimator

$$\hat{\mathbf{f}}(x_i) = (\hat{\mathrm{f}}_{i,1}, \ldots, \hat{\mathrm{f}}_{i,n})^t$$

in which the ith coordinate, $\hat{\mathrm{f}}_{i,i}$, is used as an estimator for $\mathrm{f}_i = E(Y_i)$. Unlike (3.1), however, the relationship between f_i and x_i can vary with i.

As well known, a local regression predictor is nothing more than a weighted least squares predictor. That is, for a given x_i, let $\mathbf{b}_{i,j} = (b_{i,j,1}, \cdots, b_{i,j,d})^t$ be the values of the d basis functions chosen for x_i evaluated at x_j. The local regression predictor is defined as $\hat{\mathbf{f}}(x_i) = \mathbf{B}_i \hat{\boldsymbol{\beta}}_W$, where

$$\hat{\boldsymbol{\beta}}_W = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{j=1}^n \Big(Y_j - \sum_{k=1}^d \beta_k b_{i,j,k} \Big)^2 K\left(\frac{x_j - x_i}{h}\right) \right\}, \tag{4.1}$$

and $K(\cdot)$ is a positive kernel function with unknown bandwidth parameter $h > 0$. Solving, it can be shown that $\hat{\mathbf{f}}(x_i)$ is the weighted least squares predictor,

$$\hat{\mathbf{f}}(x_i) = \mathbf{B}_i (\mathbf{B}_i^t \mathbf{W}(x_i) \mathbf{B}_i)^{-1} \mathbf{B}_i^t \mathbf{W}(x_i) \mathbf{Y} \tag{4.2}$$

where $\mathbf{B}_i$ is the $n \times d$ design matrix with jth row $\mathbf{b}_{i,j}$, and $\mathbf{W}(x_i) = \mathrm{diag}\{W_{i,j}\}$ is the $n \times n$ diagonal weight matrix, where

$$W_{i,j} = K\left(\frac{x_j - x_i}{h}\right), \qquad j = 1, \ldots, n.$$

See [14] for details.

Example 4.1. A popular basis function expansion for local regression is in terms of polynomials [17, 18]. In this case, the design matrix is

$$\mathbf{B}_i = \begin{pmatrix} 1 & x_1 - x_i & \cdots & (x_1 - x_i)^d \\ 1 & x_2 - x_i & \cdots & (x_2 - x_i)^d \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n - x_i & \cdots & (x_n - x_i)^d \end{pmatrix}_{n \times (d+1)}. \tag{4.3}$$

Note that $\mathbf{B}_i$ has rank $d+1$ because we always include an intercept term. The rationale for using a polynomial expansion follows by considering an expansion of of $E(Y_j) = f(x_j)$ around x_i. Since,

$$f(x_j) = f(x_i) + \sum_{k=1}^p \frac{(x_j - x_i)^k}{k!} f^{(k)}(x_i) + R_j,$$

where R_j is a small remainder term, the local weighted regression (4.1) is (approximately) the value for $\boldsymbol{\beta}$ minimizing

$$\sum_{j=1}^n \left(f(x_i) + \sum_{k=1}^p \frac{(x_j - x_i)^k}{k!} f^{(k)}(x_i) - \Big(\beta_0 + \sum_{k=1}^d \beta_k (x_j - x_i)^k \Big) \right)^2 K\left(\frac{x_j - x_i}{h}\right).$$

If $d = p$, then β_0 estimates $f(x_i)$ while β_k estimates $f^{(k)}(x_i)/k!$ for $k = 1, \ldots, d$. The local regression predictor is

$$\hat{\mathrm{f}}_{i,j} = \hat{\beta}_{0,W} + \sum_{k=1}^d \hat{\beta}_{k,W} (x_j - x_i)^k, \qquad j = 1, \ldots, n,$$

which should be a good approximation to $f(x_j)$ when x_j is near x_i.

4.1. Rescaled Spike and Slab Weighted Regression

The representation (4.2) presents an immediate tie-in to the spike and slab methodology. The rescaled posterior mean, $\hat{\boldsymbol{\beta}}$, from (2.2) is a model averaged generalized ridge regression (GRR) estimator, expressible as

$$\hat{\boldsymbol{\beta}} = E\left\{\left(\mathbf{X}^t\mathbf{X} + n\boldsymbol{\Gamma}^{-1}\right)^{-1}\mathbf{X}^t\mathbf{Y}\middle|\mathbf{Y}^*\right\}.$$

It is not hard to see that by appropriately introducing a weighting matrix into the hierarchy, that one can arrive at a model averaged weighted GRR estimator, and a smoother of the form (4.2). The advantage of this type of approach is that the resulting smoother will be based on an estimator that uses adaptive penalization.

In this modification, similar to (4.3), we work with a polynomial basis that depends upon i. However, our polynomial basis will be strictly orthogonal. Let

$$\mathbb{I}_{i,h} = \left\{j : K\left(\frac{x_j - x_i}{h}\right) > 0\right\}.$$

For an orthogonal basis we define $\mathbf{B}_i$ to be the design matrix for x_i obtained using a d-degree orthogonal basis using only those x_j values where $j \in \mathbb{I}_{i,h}$.

For each $j \in \mathbb{I}_{i,h}$, define $Y_j^* = \hat{\sigma}_i^{-1} n_i^{1/2} Y_j$, where n_i is the cardinality of $\mathbb{I}_{i,h}$ and $\hat{\sigma}_i^2$ is an estimator for σ_0^2 for the set of responses, $\{Y_j : j \in \mathbb{I}_{i,h}\}$. We use the estimator due to [19],

$$\hat{\sigma}_i^2 = \frac{1}{2(n_i - 1)}\sum_{j=1}^{n_i - 1}(Y_{(j+1)} - Y_{(j)})^2,$$

where $Y_{(j)}$ is the Y-value corresponding to the jth ordered x-value in $\mathbb{I}_{i,h}$ (the estimator is most easily computed by sorting the x values).

Let $\mathbf{Y}_i^*$ be the vector of the rescaled values $Y_j^* = \hat{\sigma}_i^{-1} n_i^{1/2} Y_j$ for $j \in \mathbb{I}_{i,h}$. Let $\mathbf{W}_i$ be the subset of $\mathbf{W}(x_i)$ corresponding to those $j \in \mathbb{I}_{i,h}$. For a given x_i, the modified rescaled spike and slab model is

$$\begin{aligned}
(\mathbf{Y}_i^* | \mathbf{B}_i, \mathbf{W}_i, \boldsymbol{\beta}) &\sim \mathrm{N}(\mathbf{B}_i\boldsymbol{\beta}, n_i\mathbf{W}_i^{-1}) \\
(\boldsymbol{\beta}|\boldsymbol{\gamma}) &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Gamma}) \\
\boldsymbol{\gamma} &\sim \pi(d\boldsymbol{\gamma}).
\end{aligned} \tag{4.4}$$

Consider the following theorem which characterizes the spike and slab predictor $\hat{\mathbf{f}}^*(x_i) = \mathbf{B}_i\hat{\boldsymbol{\beta}}_{i,W}$, where $\hat{\boldsymbol{\beta}}_{i,W}$ is the rescaled posterior mean from (4.4). We use this result later to explicitly characterize the smoother matrix and its effective degrees of freedom under orthogonality.

Theorem 4.1. *Under the Bayesian hierarchy (4.4), the spike and slab local predictor can be expressed as*

$$\hat{\mathbf{f}}^*(x_i) = (\hat{\mathrm{f}}_{i,1}^*, \ldots, \hat{\mathrm{f}}_{i,n}^*)^t = \mathbf{S}_{i,h}^*\mathbf{Y}_i,$$

where $\mathbf{S}_{i,h}^*$ is the model averaged smoothing matrix defined by

$$\mathbf{S}_{i,h}^* = \mathbb{E}\left\{\mathbf{B}_i\left(\mathbf{B}_i^t\mathbf{W}_i\mathbf{B}_i + n_i\boldsymbol{\Gamma}^{-1}\right)^{-1}\mathbf{B}_i^t\mathbf{W}_i\middle|\mathbf{Y}_i^*\right\}.$$

Note that the smoother matrix $\mathbf{S}_{i,h}^*$, unlike (4.2), takes advantage of adaptive penalization.

4.2. Orthogonality

Our construction for the basis ensures that $\mathbf{B}_i^t\mathbf{B}_i = n_i\mathbf{I}$. However, in order to fully exploit orthogonality, we additionally require that

$$\mathbf{B}_i^t\mathbf{W}_i\mathbf{B}_i = n_i\mathbf{I}. \tag{4.5}$$

For (4.5) to hold we must have $\mathbf{W}_i = \mathbf{I}$. The simplest way to satisfy this condition is to use a nearest neighbour kernel. For a fixed bandwidth value h, let

$$K\left(\frac{x}{h}\right) = 1\{|x| < h\}.$$

The nearest neighbour kernel puts a weight of 1 on all values of x within a distance of h to zero. Using such a kernel implies that $\mathbf{W}_i = \mathbf{I}$ and $\mathbb{I}_{i,h} = \{j : |x_j - x_i| < h\}$.

Shrinkage, just as in the global orthogonal regression setting, is intimately related to the degrees of freedom of the smoother matrix. Consider the following corollary to Theorem (4.1) characterizing effective degrees of freedom under orthogonality.

Corollary 4.1. *Under the orthogonality assumption (4.5), the local smoother matrix for the rescaled spike and slab predictor is* $\mathbf{S}^*_{i,h} = n_i^{-1}\mathbf{B}_i\mathbf{V}_i\mathbf{B}_i^t$*. The effective degrees of freedom of* $\mathbf{S}^*_{i,h}$ *equals*

$$\mathscr{D}_f(\mathbf{S}^*_{i,h}) = n_i^{-1} tr(\mathbf{B}_i\mathbf{V}_i\mathbf{B}_i^t) = n_i^{-1} tr(\mathbf{B}_i^t\mathbf{B}_i\mathbf{V}_i) = \sum_{k=1}^d V_{i,k} \le d,$$

where $\mathbf{V}_i = diag\{V_{i,k}\}$ *and*

$$V_{i,k} = \mathbb{E}\left(\frac{\gamma_k}{1+\gamma_k}\middle|\mathbf{Y}_i^*\right), \qquad k = 1,\ldots,d.$$

Hence, the degrees of freedom of the spike and slab local smoother is bounded by the dimension of the local polynomial basis.

The effective degrees of freedom can be used to provide insight into the geometry of a curve f. If the effective degrees of freedom is large, f will possess higher order local curvature, whereas if the degrees of freedom are small, f is likely to be flat. Plotting $\mathscr{D}_f(\mathbf{S}^*_{i,h})$ is therefore a way to register a curve and to identify key differences between curves. We illustrate this concept by way of three different examples.

4.3. Spinal BMD Data Revisited

For our first example, we applied the local rescaled spike and slab model (4.4) to the previously analyzed BMD data. For the analysis, we used a nearest neighbour kernel with a bandwidth set at $h = 1$, corresponding to one year of age. For the basis function we used cubic orthogonal polynomials. Figure 5 plots the spike and slab predictor for men and women (line types as in Figure 3). Also superimposed on the figure are predicted curves using Friedman's supersmoother (implemented in the programming language R by the call 'supsmu(x,y)'). The two methods agree closely, although Friedman's smoother appears to over-smooth the data for men. The right-hand side of Figure 5 plots the effective degrees of freedom for men and women. One can immediately see a phase shift in the figure, signifying distinct modes for the two curves. Also, overall, there is significantly less shrinkage for women with degrees of freedom being positive over a much wider region than men. In both cases, curves eventually flatten out at around age 20.

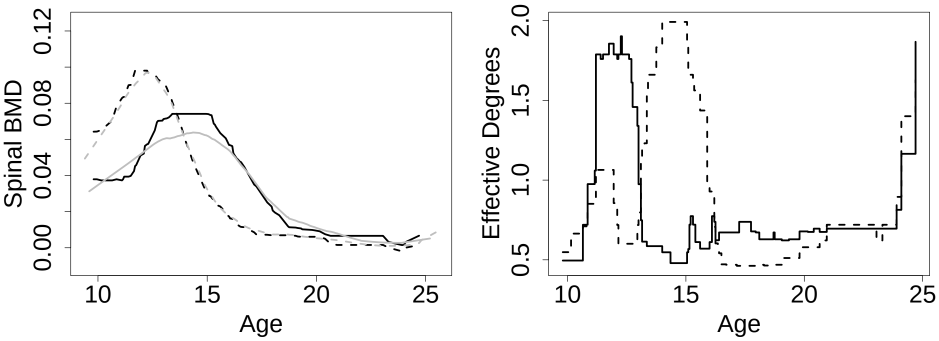


Fig 5. *Left plot: Local kernel regression for BMD data via rescaled spike and slab models with orthogonal polynomial cubic basis functions. Solid dark and dashed dark lines are spike and slab predictors for men and women, respectively. Gray solid and dashed lines are Friedman's supersmoother for men and women. Right plot: Effective degrees of freedom of spike and slab local smoother (solid lines are men, dashed lines are women).*

4.4. Cosmic Microwave Background Radiation

As another illustration we look at data related to cosmic microwave background (CMB) radiation [20]. Here, the value for x is the multipole moment and Y is the estimated power spectrum of the temperature fluctuations. The outcome is sound waves in the cosmic microwave background radiation, which is the heat left over from the big bang.

We used the same strategy and settings as before. For the bandwidth we used $h = 25$ which was estimated prior to fitting using generalized cross-validation. Results from the analysis are depicted in Figure 6 with plots zoomed in on different regions of x in order to help visualize the varying curvature. The bottom right plot of Figure 6 shows the effective degrees of freedom. The plot suggests the presence of at least 4 distinct inflection points. In particular, note that initially for $x < 200$ there is a steep increase in the curve signified by the effective degrees of freedom being roughly constant at 3.0. At around $x = 200$ there is a significant drop in the effective degrees of freedom, followed by an increase and a flattening out until around $x = 400$. The drop at $x = 200$ indicates the first inflection point. At $x = 400$ there is another drop in the effective degrees of freedom. Similarly, there is a drop near $x = 600$ and $x = 800$. All told, this suggests at least 4 distinct inflections, all appearing in multiples of 200 starting at $x = 200$.

4.5. Mass Spectometry Protein Data

The study of proteins is critical to understanding living organisms at the molecular level as proteins are the main components of physiological pathways of cells. Proteomics, the study of proteins on a large scale, is often considered the natural step after genomics in the study of biological systems. Greatly complicating any system-wide analysis of proteins, however, is the dynamic nature of the proteome, which constantly changes through its biochemical interactions with the genome and the environment. While the challenges faced by proteomics are great, the benefits at the same time are potentially huge. For example, by studying protein differences for diseased individuals, one might be able to discover pathways responsible for these differences, which in turn could lead to novel biomarkers for identifying disease.

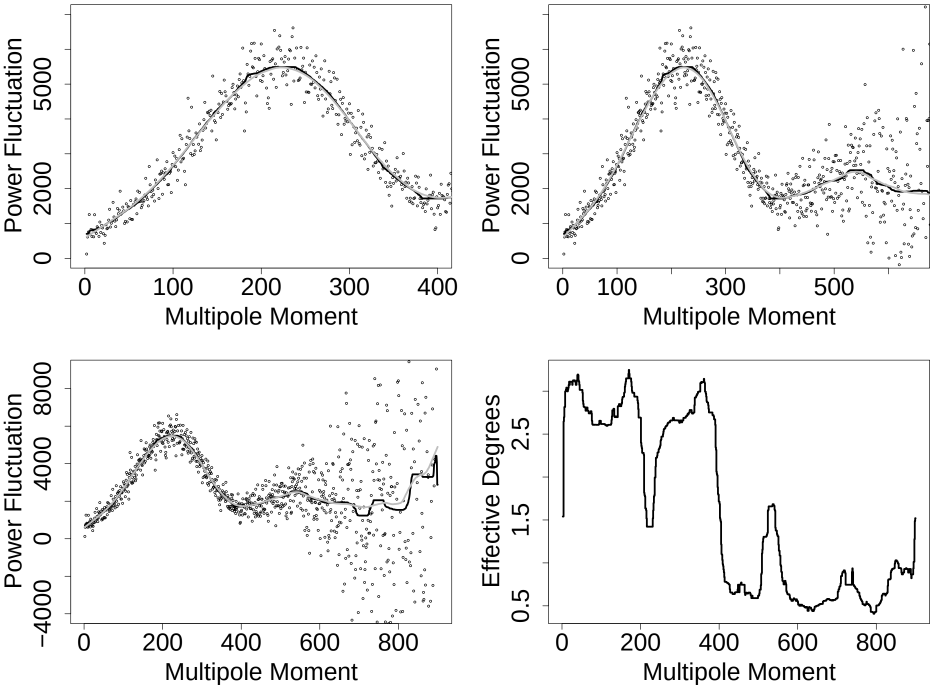


Fig 6. *First 3 plots (top to bottom left to right) are rescaled spike and slab local predictors (thick dark lines) for CMB data. Gray lines are Friedman's smoother. Bottom right plot: Effective degrees of freedom for rescaled spike and slab predictor. Note the presence of 4 modes suggested by this last plot.*

One promising technology for profiling protein behavior is SELDI-TOF-MS (surface enhanced laser desorption/ionization time-of-flight mass spectrometry). In this technology, homogeneous biological samples are placed on the active surface of an array. The protein samples are washed and an energy absorbing molecule solution is placed on the surface of the array and allowed to crystalize. The array is then queried by a laser which ionizes the proteins in the sample. Charged gaseous peptides are emitted and their intensity is detected downstream. The mass over charge ratio (m/z) of a peptide-ion is determined from the recorded TOF (time-of-flight). The data collected from a SELDI-TOF-MS experiment consists of the intensity (abundance) of proteins in the sample for a given m/z ratio. One can think of the set of these two values as constituting a spectra. Each biological sample produces one spectra and it is of interest to study differences in spectra as a function of phenotype. See [21] for more details and further references.

Identification of unique peaks in the spectra, a method commonly referred to as peak identification, is a crucial part of analyzing mass spectrometry data. From a statistical perspective, peak identification can be recast as a smoothing problem where the goal is to identify modes in the data after appropriate smoothing. The outcomes are the spectometry intensity measurements, whereas x is the specific m/z ratio. To illustrate how our spike and slab method can be used for peak detection, we analyzed a set of 8 calibration spectra available as part of the "PROcess" library [21] in the Bioconductor R-suite. The data is unique because it is known a priori that the same 5 proteins are present in each of the 8 samples.

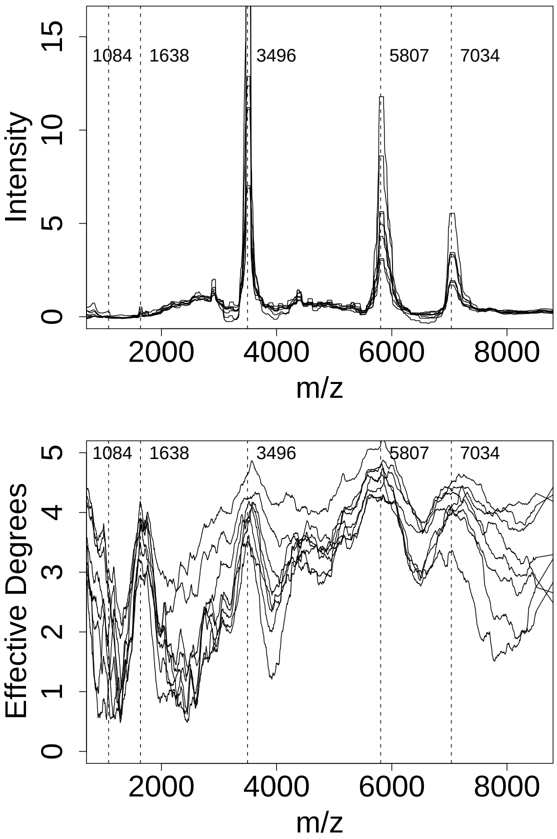


FIG 7. *Mass spectometry calibration data: 8 spectra, each comprising 13,468 distinct m/z ratios (horizontal axis constrained to a subset of observed m/z ratios to help zoom in figure). Top plot: Solid lines are rescaled spike and slab predictors and dashed vertical lines indicate known unique proteins. Bottom plot: Effective degrees of freedom.*

The results from the analysis are plotted in Figure 7 (we note that the data was first baseline normalized prior to analysis). The black lines in the top plot are the rescaled spike and slab smoothed predictors for each spectra. We used orthogonal polynomials of degree 5 (the high degrees of freedom used due to the spiky nature of the data). The bandwidth was set at $h = 50$. Superimposed are 5 dashed vertical lines indicating the 5 distinct proteins. Interestingly, we find that 3 of the 5 proteins are clearly identified in all 8 spectra. However, the two smallest proteins $m/z = 1084$ and $m/z = 1638$ are less visible, the protein at $m/z = 1084$ especially so. There is also evidence of at least 2 additional peaks at approximately $m/z = 3500$ and $m/z = 4500$. The effective degrees of freedom plot, also given in Figure 7, confirms these findings. The plot also indicates that overlap of spectra is sub-par suggesting further normalization of the data is needed.

Acknowledgements

Hemant Ishwaran was supported by the National Science Foundation grant DMS-0405675. The authors are grateful to Bertrand Clarke and the referee who reviewed an earlier version of the paper. We also thank Liang Li for his comments.

References

[1] Ishwaran, H. and Rao, J.S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Ann. Statist.* **33** 730–773.

[2] Ishwaran, H. and Rao, J.S. (2005). Spike and slab gene selection for multigroup microarray data. *J. Amer. Stat. Assoc.*, **100** 764–780.

[3] Ishwaran, H. and Rao, J.S. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *J. Amer. Stat. Assoc.*, **98** 438–455.

[4] Lempers, F. B. (1971). *Posterior Probabilities of Alternative Linear Models.* Rotterdam University Press, Rotterdam.

[5] Mitchell, T.J. and Beauchamp, J.J. (1988). Bayesian variable selection in linear regression. *J. Amer. Stat. Assoc.*, **83** 1023–1036.

[6] George, E.I. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *J. Amer. Stat. Assoc.*, **88** 881–889.

[7] George, E.I. and McCulloch, R.E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica.*, **7** 339–373.

[8] Chipman H. (1996). Bayesian variable selection with related predictors. *Canad. J. Statist.*, **24** 17–36.

[9] Clyde, M., DeSimone, H. and Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *J. Amer. Stat. Assoc.*, **91** 1197–1208.

[10] Geweke, J. and Meese, R. (1981). Estimating regression models of finite but unknown order. *International Econ. Rev.*, **22** 55–70.

[11] Chipman, H.A. and George, E.I. and McCulloch R.E. (2001). The practical implementation of Bayesian model selection. In P. Lahiri, editor, *IMS Monograph 38, Model Selection.* IMS, Hayward, California.

[12] Bachrach, L.K., Hastie, T., Wang, M-C. and Narasimhan, B. (1999). Bone mineral acquisition in healthy asian, hispanic, black and caucasian youth. A longitudinal study. *J. Clin. Endocrinol. Metab.*, **84** 4702–4712.

[13] Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

[14] Hastie, T. and Loader, C. (1993). Local regression: automatic kernel carpentry. *Stat. Science*, **8** 120–143.

[15] Stone C.J. (1977). Consistent nonparametric regression. *Ann. Statist.*, **5** 595–620.

[16] Cleveland W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Stat. Assoc.*, **74**, 829–836.

[17] Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Stat. Assoc.*, **87** 998–1004.

[18] Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaptation. *J. Royal Statist. Soc. B.*, **57** 371–394.

[19] Rice J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.*, **12** 1215–1230.

[20] Genovese, C.R., Miller, C.J., Nichol, R.C., Arjunwadkar, M. and Wasserman L. (2004). Nonparametric inference for the cosmic microwave background. *Statist. Sci.*, **19** 308–321.

[21] Li, X., Gentleman, R., Lu, X., Shi, Q., Iglehart, J.D., Harris, L., and Miron, A. (2005). SELDI-TOF mass spectometry protein data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Gentleman et al. (eds.). Springer, New York, 91–108.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 283

Nonparametric Statistics on Manifolds with Applications to Shape Spaces

Abhishek Bhattacharya and Rabi Bhattacharya [1]

University of Arizona

Abstract: This article presents certain recent methodologies and some new results for the statistical analysis of probability distributions on manifolds. An important example considered in some detail here is the 2-D shape space of k-ads, comprising all configurations of k planar landmarks ($k > 2$) -modulo translation, scaling and rotation.

Contents

1. Introduction

The statistical analysis of shape distributions based on random samples is important in many areas such as morphometrics (discrimination and classification of biological shapes), medical diagnostics (detection of change or deformation of shapes in some organs due to some disease, for example) and machine vision (e.g., digital recording and analysis based on planar views of 3-D objects). Among the pioneers on foundational studies leading to such applications, we mention Kendall [20] (also see Kendall et al. [21]) and Bookstein [9]. The geometries of the spaces are those of differentiable manifolds often with appropriate Riemannian structures.

Our goal in this article is to establish some general principles for nonparametric statistical analysis on such manifolds and apply those to some shape spaces, especially Kendall's two-dimensional shape space Σ_2^k of the so-called k-ads, i.e., the

[1] Department of Mathematics, University of Arizona, Tucson, AZ 85721. emails: abhishek@math.arizona.edu; rabi@math.arizona.edu

AMS 2000 subject classifications: Primary 62G20; secondary 62E20, 62H35

Keywords and phrases: Extrinsic and intrinsic means and variations, Kendall's shape spaces, two-sample nonparametric tests

space of configurations of k points on the plane (not all identical), identified modulo size and under Euclidean motions of translation and rotation. Two sample tests for the comparison of both extrinsic and intrinsic Fréchet mean shapes and mean variations of two distributions on Σ_2^k are provided. As far as we know the explicit computations of these tests are new. In the case of the intrinsic mean and variation, the usual support criterion (see e.g., Le [24] and Bhattacharya and Patrangenaru [6–8]) is significantly relaxed, thereby substantially enhancing the applicability of the tests.

For recent results on statistical analysis of 3-D shapes, which we do not consider here, we refer to Dryden et al.[11] and Bandulasiri et al.[2].

Sometimes the sample sizes in shape analysis are only moderately large. Under such circumstances, one may more effectively use Effron's bootstrap methods (Effron [14]), whose superiority over the classical CLT-based confidence regions and tests may be established via higher order asymptotics (see e.g., Babu and Singh [1], Bhattacharya and Qumsiyeh[5], Bhattacharya and Ghosh [4], Ghosh [16], Hall [17]).

We next turn to the specific example of main interest to us, namely, Σ_2^k. For purposes of medical diagnostics, classification of biological species, etc., one may use expert help to choose a suitable ordered set of k points or landmarks in the plane, or a *k-ad*,

$$\mathbf{z} = \{(x_j, y_j), 1 \leq j \leq k\},$$

on a two-dimensional image of an object under consideration. One assumes that not all k points are the same, and $k > 2$. Kendall's shape space Σ_2^k comprises the equivalence classes of all such k-ads under translation, rotation and scaling. For a given k-ad $\mathbf{z}$, the effect of translation is removed by considering $\mathbf{z} - \langle \mathbf{z} \rangle$ where $\langle \mathbf{z} \rangle$ is the vector whose elements are all equal to the mean location of the k-ad, namely, $(1/k)\sum_{j=1}^k (x_j, y_j)$. The translated k-ads then lie in the $(2k-2)$-dimensional hyperplane H of $(\Re^2)^k \approx \Re^{2k}$, given by

$$H = \{(x_j, y_j)_{1 \leq j \leq k} : \sum x_j = 0, \ \sum y_j = 0\},$$

and they comprise all of H except the origin. The effect of scale, or length, is removed by dividing $\mathbf{z} - \langle \mathbf{z} \rangle$ by $\|\mathbf{z} - \langle \mathbf{z} \rangle\|$ where $\|.\|$ is the usual Euclidean norm in $(\Re^2)^k$,

$$\|(u_j, v_j)_{1 \leq j \leq k}\| = [\sum (u_j^2 + v_j^2)]^{1/2}.$$

The resulting transformed k-ad $\mathbf{w} = (\mathbf{z} - \langle \mathbf{z} \rangle)/\|\mathbf{z} - \langle \mathbf{z} \rangle\|$ is called the *preshape* of the k-ad $\mathbf{z}$. The set of preshapes is then naturally identified with the unit sphere in H, which is basically the same as the unit sphere S^{2k-3} in $\Re^{2k-2}$. Finally, the *shape* $[\mathbf{z}]$ of a k-ad $\mathbf{z}$ is given by the *orbit* of $\mathbf{w} = (u_j, v_j)'_{1 \leq j \leq k}$ under rotation, namely,

$$[\mathbf{z}] = \left[\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} u_j \\ v_j \end{pmatrix}_{1 \leq j \leq k} \right], \ -\pi < \theta \leq \pi. \tag{1.1}$$

Thus Σ_2^k is a *quotient space* of S^{2k-3}, namely, S^{2k-3}/S^1, and it has dimension $2k - 4$.

We will use a mathematically more convenient way of describing Σ_2^k as achieved by viewing a k-ad as an element of $\mathbb{C}^k$, namely, $\mathbf{z} = (x_j + iy_j)_{1 \leq j \leq k}$. Then $\langle \mathbf{z} \rangle$ is the complex k-vector whose elements are all equal to $(1/k)\sum_{j=1}^k (x_j + iy_j)$. The translated k-ad then lies in the complex $(k-1)$-dimensional hyperplane

$$H^{k-1} = \{(a_j)_{1\le j\le k} \in \mathbb{C}^k : \sum_{j=1}^{k} a_j = 0\}$$

The norm $\|\mathbf{z} - \langle\mathbf{z}\rangle\|$ has the same value as before. But the rotation by an angle θ of $\mathbf{w} = (\mathbf{z} - \langle\mathbf{z}\rangle)/\|\mathbf{z} - \langle\mathbf{z}\rangle\|$ may now be expressed as $e^{i\theta}\mathbf{w}$. For a system of coordinate neighborhoods, or parametrization, of this *spherical representation* of Σ_2^k as a quotient space of S^{2k-3}, see Gallot et al. ([15], pp. 32, 34).

Another parametrization of Σ_2^k, compatible with the above, is obtained by viewing the shape of a k-ad $\mathbf{z} \equiv (x_j + iy_j)_{1\le j\le k}$ as the orbit

$$\{z_0(\mathbf{z} - \langle\mathbf{z}\rangle) : \ z_0 \in \mathbb{C} \setminus \{0\}\}.$$

Note that $z_0 = \lambda e^{i\theta}$ for $\lambda = |z_0|$ and some $\theta \in (-\pi, \pi]$, so that the orbit, namely, a complex line through the origin in H^{k-1}, is independent of both scale and rotation and, therefore, a representation of the shape of $\mathbf{z}$. Thus Σ_2^k is (isomorphic to) the space of all complex lines through the origin in $\mathbb{C}^{k-1}$, the *complex projective space* $\mathbb{C}P^{k-2}$, a familiar and important example in differential geometry. For a system of coordinate neighborhoods for Σ_2^k viewed as $\mathbb{C}P^{k-2}$, see Gallot et al. ([15], pp 9, 10, 64, 65).

We next consider an *extrinsic distance* on Σ_2^k corresponding to a special embedding, namely, the *Veronese-Whitney embedding* ϕ_E of Σ_2^k into the space $S(k, \mathbb{C})$ of $k \times k$ complex Hermitian matrices:

$$\phi_E([\mathbf{z}]) = \mathbf{w}\mathbf{w}^* \tag{1.2}$$

where $\mathbf{w} = (\mathbf{z} - \langle\mathbf{z}\rangle)/\|\mathbf{z} - \langle\mathbf{z}\rangle\|$ is the preshape of $\mathbf{z}$. Here $\mathbf{w}$ is regarded as a column vector of k complex numbers, $\mathbf{w} = (w_1, w_2, \ldots, w_k)'$, and $\mathbf{w}^*$ is the transpose of its complex conjugate. Observe that the right side of (1.2) is constant on the orbit $\{e^{i\theta}w : \ -\pi < \theta \le \pi\}$ of the preshape w and is, therefore, a function of the shape $[z]$ of the k-ad. Also, this function is one-to-one on Σ_2^k into $S(k, \mathbb{C})$. The vector space $S(k, \mathbb{C})$, with the real scaler field $\Re$, has dimension k^2. This is because a $k \times k$ Hermitian matrix is specified by k real numbers on the diagonal and $\binom{k}{2}$ complex numbers (i.e., $2\binom{k}{2}$ real numbers) as lower-right off-diagonal elements. On $S(k, \mathbb{C})$ define the norm $\|.\|$ and distance d by

$$\begin{aligned}\|A\|^2 &= \text{Trace}AA^* = \text{Trace}A^2,\\ d^2(A,B) &= \|A - B\|^2 = \text{Trace}(A - B)^2.\end{aligned} \tag{1.3}$$

Note that this is the same as the Euclidean norm and distance in $\Re^{2k^2}$. The induced distance ρ_E on Σ_2^k is then given by

$$\begin{aligned}\rho_E^2([\mathbf{z}], [\mathbf{w}]) &= d^2(\phi_E([\mathbf{z}]), \phi_E([\mathbf{w}])) = \text{Trace}(\mathbf{u}\mathbf{u}^* - \mathbf{v}\mathbf{v}^*)^2\\ &= \sum_{j=1}^{k} |u_j|^2 + \sum_{j=1}^{k} |v_j|^2 - \sum_{j=1}^{k}\sum_{j'=1}^{k} (u_j\bar{u}_{j'}v_{j'}\bar{v}_j + v_j\bar{v}_{j'}u_{j'}\bar{u}_j)\\ &= 2 - 2|\mathbf{u}^*\mathbf{v}|^2\end{aligned} \tag{1.4}$$

where $\mathbf{u}$ and $\mathbf{v}$ are the preshapes of $[\mathbf{z}]$ and $[\mathbf{w}]$ respectively. The distance ρ_E is known as the full *Procrustes distance* (Kendall [22] , Kent [23] , Dryden and Mardia [12]).

Let X_j, $1 \leq j \leq n$, be i.i.d. k-ads such that their shapes $[X_j]$, $1 \leq j \leq n$, have the common distribution Q. Let $\tilde{\mu}$ denote the Euclidean mean of $\tilde{Q} = Q \circ \phi_E^{-1}$ viewed as a probability measure on $S(k, \mathbb{C})$. Let $\tilde{M} = \phi_E(\Sigma_2^k)$, and denote the Euclidean projection of $\tilde{\mu}$ on $\tilde{M}$ by $P\tilde{\mu} \equiv P_{\tilde{M}}\tilde{\mu}$. The *extrinsic mean* of Q is then $\mu_E = \phi^{-1}(P\tilde{\mu})$. It minimizes the *Fréchet function* (2.1) with respect to the distance ρ_E. Similarly, for the sample extrinsic mean, calculate $P\overline{\tilde{X}}$ where $\overline{\tilde{X}} = (1/n)\sum_{j=1}^n \phi_E([X_j])$ is a coordinate-wise average of the matrix elements $W_j W_j^*$ and W_j is the preshape of X_j $(1 \leq j \leq n)$. The asymptotic distribution of $\sqrt{n}(P\overline{\tilde{X}} - P\tilde{\mu})$ is given by that of its projection on the tangent space $T_{P\tilde{\mu}}\tilde{M}$ at $P\tilde{\mu}$, since its projection on the complement of $T_{P\tilde{\mu}}\tilde{M}$ is negligible. For computation of this projection, one chooses a suitable orthonormal basis of $S(k, \mathbb{C})$ (considered as a single orthonormal frame for its constant tangent spaces), and calculates the differential of the projection map $P = P_{\tilde{M}} : S(k, \mathbb{C}) \to \tilde{M}$ in terms of these coordinates. One thus arrives at a non-singular $(2k-4)$-dimensional Normal distribution in the limit (see Sections 3.1-3.4 for details).

Turning to the *intrinsic mean* on a Riemannian manifold M, with geodesic distance d_g, the first problem to resolve is its existence as the unique minimizer of the Fréchet function $\int d_g^2(p,m)Q(dm)$. Here a result of Karchar [19] on the existence of a unique minimizer is greatly improved by a result of Kendall [22], which allows the radius r of a geodesic ball $B(p,r)$ containing the support of Q to be twice as large as required by Karchar [19] (Proposition 4.1). On such a ball, the map $\phi = \exp_p^{-1}$ (the inverse of the *exponential map* at p), is a diffeomorphism onto its image in the tangent space T_pM at p. Using the coordinates of the vector space T_pM, called *normal coordinates*, one arrives at a central limit theorem for the sample intrinsic mean μ_{nI} (Theorem 4.2), following Bhattacharya and Patrangenaru [8]. Note that, with the (non-Euclidean) distance on T_pM induced by ϕ from the geodesic distance d_g on M, the image $\mu_n = \phi(\mu_{nI})$ of μ_{nI} is the minimizer of the Fréchet function

$$F_n(x) \equiv \int d_g^2(\phi^{-1}x, \phi^{-1}y)\tilde{Q}_n(dy)$$

where $\tilde{Q}_n = Q_n \circ \phi^{-1}$, $Q_n = (1/n)\sum_{j=1}^n \delta_{[X_j]}$. Thus μ_n is a *M-estimator* in the Euclidean space T_qM. The assumptions in Theorem 4.2 guarantee that this M-estimator is asymptotically Gaussian around $\mu = \phi(\mu_I)$. The asymptotic distribution of the test statistic (4.5) follows from this.

The computation of the test statistic (4.5) is generally more involved than that used for comparing extrinsic means (see, e.g., (3.17) for the case $M = \Sigma_2^k$). This involves, in particular, the metric tensor of M to compute geodesics and normal coordinates. We refer to [3] for the asymptotic theory for intrinsic means, with explicit computations of parameters especially for the planer shape space of k-ads. However in Section 5 of the present article, we display numerical values of the intrinsic two-sample test statistics, along with the corresponding p-values, in two examples. It may be noted that for highly concentrated data in each of these examples, the extrinsic and intrinsic distances are close and hence the extrinsic and intrinsic test statistics have virtually the same values.

The minimum value attained by the Fréchet function is called the *Fréchet variation* of Q and it is a measure of spread of the distribution Q. The sample Fréchet variation is a consistent estimator of the Fréchet variation of Q as proved in Proposition 2.4. If the Fréchet mean exists, we derive the asymptotic distribution of the sample Fréchet variation in Theorem 2.5. This can be used to construct a nonparametric test statistic to compare the spread of two populations on M. We compute numerical values of the test statistic, along with the p-values for $M = \Sigma_2^k$ in Section 5. For highly concentrated data as in the examples considered in Section 5, the Fréchet variations of the distributions are very small. Then the mean comparison is usually sufficient to discriminate between the populations and the variations show no significant difference.

We conclude this section with two brief remarks. First, the main objective of inference in the two-sample problem on Σ_2^k is to discriminate between two different distributions on it. It turns out, in most practical problems that arise, that the means and variations (extrinsic or intrinsic) are generally adequate for this discrimination. More elaborate procedures such as nonparametric density estimation suffer from the 'curse of dimensionality' on this commonly high-dimensional space. One can, however, do such density estimation on a tangent space (e.g., on $T_{\mu_I}M$, via the inverse exponential map $\exp_{\mu_I}^{-1}$), as in the Euclidean case. Excepting for the computation in normal coordinates, this presents no novelty. Secondly, in examples with real data sets that we have studied (e.g., those in Section 5), the p-values of the nonparametric two-sample tests for comparing means, developed in this article, are always much smaller (often by an order of magnitude or more) than those based on existing, mostly parametric, tests in the literature (see Dryden and Mardia [12]). This seems to indicate that the tests proposed here may be more powerful than those that have been used in the past, for many data sets that arise in practice. This perhaps also points to the inadequacy of parametric models of shapes popularly used in the literature in capturing certain important shape features.

2. Fréchet Mean and Variation on Metric Spaces

Let (M, ρ) be a metric space, ρ being the distance on M. For a given probability measure Q on (the Borel sigma-field of) M, define the *Fréchet function* of Q as

$$F(p) = \int_M \rho^2(p, x) Q(dx), \quad p \in M. \tag{2.1}$$

2.1. Fréchet Mean

Definition 2.1. Suppose $F(p) < \infty$ for some $p \in M$. Then the set of all p for which $F(p)$ is the minimum value of F on M is called the *Fréchet mean set* of Q, denoted by C_Q. If this set is a singleton, say $\{\mu_F\}$, then μ_F is called the *Fréchet mean* of Q. If $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) with common distribution Q, and $Q_n \doteq (1/n)\sum_{j=1}^n \delta_{X_j}$ is the corresponding empirical distribution, then the Fréchet mean set of Q_n is called the *sample Fréchet mean set*, denoted by C_{Q_n}. If this set is a singleton, say $\{\mu_{F_n}\}$, then μ_{F_n} is called the *sample Fréchet mean.*

The following result has been proved in Theorem 2.1, Bhattacharya and Patrangenaru [7].

Proposition 2.1. *Suppose every closed and bounded subset of M is compact. If the Fréchet function $F(p)$ of Q is finite for some p, then C_Q is nonempty and compact.*

The next result establishes the strong consistency of the sample Fréchet mean. For a proof, see Theorem 2.3, Bhattacharya and Patrangenaru [7].

Proposition 2.2. *Assume* (i) *that every closed bounded subset of M is compact, and* (ii) *F is finite on M. Then given any $\epsilon > 0$, there exists an integer valued random variable $N = N(\omega, \epsilon)$ and a P-null set $A(\omega, \epsilon)$ such that*

$$C_{Q_n} \subset C_Q^\epsilon \equiv \{p \in M : \rho(p, C_Q) < \epsilon\}, \ \forall n \geq N \tag{2.2}$$

outside of $A(\omega, \epsilon)$. In particular, if $C_Q = \{\mu_F\}$, then every measurable selection μ_{F_n} from C_{Q_n} is a strongly consistent estimator of μ_F.

Remark 2.1. It is known that a connected Riemannian manifold M which is complete (in its geodesic distance) satisfies the topological hypothesis of Propositions 2.1 and 2.2: every closed bounded subset of M is compact (see Theorem 2.8, Do Carmo [10], pp 146-147). We will investigate conditions for the existence of the Fréchet mean of Q (as a unique minimizer of the Fréchet function F of Q) in the subsequent sections.

Remark 2.2. One can show that the reverse of (2.2), that is, '$C_Q \subset C_{Q_n}^\epsilon \ \forall \ n \geq N(\omega, \epsilon)$' does not hold in general. See, for example, Bhattacharya and Patrangenaru ([7], Remark 2.6.

Next we consider the asymptotic distribution of μ_{F_n}. For Theorem 2.3, we assume M to be a differentiable manifold of dimension d. Let ρ be a distance metrizing the topology of M. For a proof of the following result, see Theorem 2.1, Bhattacharya and Patrangenaru [8].

Theorem 2.3. *Suppose the following assumptions hold:*
(i) *Q has support in a single coordinate patch, (U, ϕ), $\phi : U \longrightarrow \Re^d$ smooth. Let $Y_j = \phi(X_j)$, $j = 1, \ldots, n$.*
(ii) *The Fréchet mean μ_F of Q is unique.*
(iii) *$\forall x$, $y \mapsto h(x, y) = \rho^2(\phi^{-1}x, \phi^{-1}y)$ is twice continuously differentiable in a neighborhood of $\phi(\mu_F) = \mu$.*
(iv) $E(\mathrm{D}_r h(Y_1, \mu))^2 < \infty \ \forall r$.
(v) $E(\sup_{|u-v| \leq \epsilon} |\mathrm{D}_s\mathrm{D}_r h(Y_1, v) - \mathrm{D}_s\mathrm{D}_r h(Y_1, u)|) \to 0$ *as* $\epsilon \to 0 \ \forall \ r, s$.
(vi) $\Lambda = E\left(\mathrm{D}_s\mathrm{D}_r h(Y_1, \mu)\right)$ *is nonsingular.*
(vii) $\Sigma = \mathrm{Cov}\left(\mathrm{D}h(Y_1, \mu)\right)$ *is nonsingular.*
Let μ_{F_n} be a measurable selection from the Fréchet sample mean set, and write $\mu_n = \phi(\mu_{F_n})$. Then under the assumptions (i)-(vii),

$$\sqrt{n}(\mu_n - \mu) \xrightarrow{\mathcal{L}} N(0, \Lambda^{-1}\Sigma(\Lambda^{'})^{-1}). \tag{2.3}$$

2.2. Fréchet Variation

Definition 2.2. The *Fréchet variation* V of Q is the minimum value attained by the Fréchet function F defined by (2.1) on M. Similarly the minimum value attained by the *sample Fréchet function*,

$$F_n(p) = \frac{1}{n}\sum_{j=1}^{n} \rho^2(X_j, p) \tag{2.4}$$

is called the *sample Fréchet variation* and denoted by V_n.

From Proposition 2.1 it follows that if the Fréchet function $F(p)$ is finite for some p, then V is finite and equals $F(p)$ for all p in the Fréchet mean set C_Q. Similarly the sample variation V_n is the value of F_n on the sample Fréchet mean set C_{Q_n}. The following result establishes the strong consistency of V_n as an estimator of V.

Proposition 2.4. *Suppose every closed and bounded subset of M is compact, and F is finite on M. Then V_n is a strongly consistent estimator of V.*

Proof. In view of Proposition 2.2, for any $\epsilon > 0$, there exists $N = N(\omega, \epsilon)$ such that

$$|V_n - V| = |\inf_{p\in M} F_n(p) - \inf_{p\in M} F(p)| \leq \sup_{p\in \overline{C_Q^\epsilon}} |F_n(p) - F(p)| \tag{2.5}$$

for all $n \geq N$ almost surely. From the proof of Theorem 2.3 in Bhattacharya and Patrangenaru [7], it follows that for any compact set $K \subset M$,

$$\sup_{p\in K} |F_n(p) - F(p)| \longrightarrow 0 \text{ a.s. as } n \to \infty.$$

Since $\overline{C_Q^\epsilon}$ is compact, it follows from (2.5) that

$$|V_n - V| \longrightarrow 0 \text{ a.s. as } n \to \infty.$$

□

Remark 2.3. The sample variation is a consistent estimator of the population variation even when the Fréchet function F of Q does not have a unique minimizer.

Next we derive the asymptotic distribution of V_n when there is a unique population Fréchet mean.

Theorem 2.5. *Let M be a differentiable manifold. Using the notation of Theorem 2.3, under assumptions* (i)-(vii) *and assuming $E(\rho^4(X_1, \mu_F)) < \infty$, one has*

$$\sqrt{n}(V_n - V) \xrightarrow{\mathcal{L}} N\left(0, \mathrm{var}(\rho^2(X_1, \mu_F))\right). \tag{2.6}$$

Proof. Let

$$F(x) = \int \rho^2(\phi^{-1}(x), m) Q(dm),\ F_n(x) = \frac{1}{n}\sum_{j=1}^n \rho^2(\phi^{-1}(x), X_j).$$

Let μ_{F_n} be a measurable selection from the sample mean set and $\mu_n = \phi(\mu_{Fn})$. Then

$$\begin{aligned}\sqrt{n}(V_n - V) &= \sqrt{n}(F_n(\mu_n) - F(\mu)) \\ &= \sqrt{n}(F_n(\mu_n) - F_n(\mu)) + \sqrt{n}(F_n(\mu) - F(\mu)),\end{aligned} \tag{2.7}$$

$$\begin{aligned}\sqrt{n}(F_n(\mu_n) - F_n(\mu)) &= \frac{1}{\sqrt{n}}\sum_{j=1}^n\sum_{r=1}^d (\mu_n - \mu)_r \mathrm{D}_r h(Y_j, \mu) \\ &+ \frac{1}{2\sqrt{n}}\sum_{j=1}^n\sum_{r=1}^d\sum_{s=1}^d (\mu_n - \mu)_r(\mu_n - \mu)_s \mathrm{D}_s\mathrm{D}_r h(Y_j, \mu_n^*)\end{aligned} \tag{2.8}$$

for some μ_n^* in the line segment joining μ and μ_n. By assumption (v) of Theorem 2.3 and because $\sqrt{n}(\mu_n - \mu)$ is asymptotically normal, the second term on the right of (2.8) converges to 0 in probability. Also $(1/n)\sum_{j=1}^n \mathrm{D}h(Y_j, \mu) \to \mathrm{E}\left(\mathrm{D}h(Y_1, \mu)\right) = 0$, so that the first term on the right of (2.8) converges to 0 in probability. Hence (2.7) becomes

$$\begin{aligned}\sqrt{n}(V_n - V) &= \sqrt{n}(F_n(\mu) - F(\mu)) + o_P(1) \\ &= \frac{1}{\sqrt{n}}\sum_{j=1}^n \left(\rho^2(X_j, \mu_F) - \mathrm{E}\rho^2(X_1, \mu_F)\right) + o_P(1).\end{aligned} \tag{2.9}$$

By the CLT for the i.i.d. sequence $\{\rho^2(X_j, \mu_F)\}$, (2.9) converges in law to $N(0, \mathrm{var}(\rho^2(X_1, \mu_F)))$. □

Remark 2.4. Theorem 2.5 requires the population mean to exist for the sample variation to be asymptotically Normal. It may be shown by examples that it fails to give the correct distribution if there is not a unique mean.

Theorem 2.5 can be used to construct a nonparametric test for testing whether two populations have the same spread. Suppose Q_1 and Q_2 are two probability distributions with unique Fréchet means μ_{1F} and μ_{2F} and Fréchet variations V_1 and V_2, respectively. We have i.i.d. samples $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ from Q_1 and Q_2, respectively. Let μ_{F_n} and μ_{F_m} denote the sample means, V_n and V_m denote the sample variations. Then the null hypothesis is

$$H_0 : V_1 = V_2 = V.$$

Under H_0, from (2.6),

$$\sqrt{n}(V_n - V) \xrightarrow{\mathcal{L}} N(0, \sigma_1^2) \tag{2.10}$$

$$\sqrt{m}(V_m - V) \xrightarrow{\mathcal{L}} N(0, \sigma_2^2) \tag{2.11}$$

$$\text{where } \sigma_1^2 = \mathrm{var}(\rho^2(X_1, \mu_{1F})),\ \sigma_2^2 = \mathrm{var}(\rho^2(Y_1, \mu_{2F})).$$

Suppose $n/(m+n) \to p$, $m/(m+n) \to q$, for some $p, q > 0$; $p + q = 1$. Then from (2.10) and (2.11),

$$\sqrt{n+m}(V_n - V_m) \xrightarrow{\mathcal{L}} N(0, \left(\frac{\sigma_1^2}{p} + \frac{\sigma_2^2}{q}\right)), \tag{2.12}$$

$$\frac{V_n - V_m}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}} \xrightarrow{\mathcal{L}} N(0, 1), \tag{2.13}$$

where $s_1^2 = (1/n)\sum_{j=1}^n (\rho^2(X_j, \mu_{F_n}) - V_n)^2$ and $s_2^2 = (1/m)\sum_{j=1}^m (\rho^2(Y_j, \mu_{F_m}) - V_m)^2$ are the sample estimates of σ_1^2 and σ_2^2, respectively. Hence the test statistic used is

$$T_{nm} = \frac{V_n - V_m}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}. \tag{2.14}$$

For a test of size α, we reject H_0 if $|T_{nm}| > Z_{1-(\alpha/2)}$ where $Z_{1-(\alpha/2)}$ is the $(1 - (\alpha/2))^{\text{th}}$ quantile of $N(0, 1)$.

From now on, unless otherwise stated, we assume that $(\mathbf{M}, \mathbf{g})$ is a d-dimensional connected complete Riemannian manifold, g being the Riemannian metric tensor on M. We shall come across different notions of means and variations depending on the distance chosen on M. We begin with the *extrinsic distance* in the next section.

3. Extrinsic Mean and Variation

Let $\phi : M \to \Re^k$ be an embedding of M into $\Re^k$, and let $\tilde{M} = \phi(M) \subset \Re^k$. Define the distance on M as: $\rho(x, y) = \|\phi(x) - \phi(y)\|$, where $\|.\|$ denotes Euclidean norm ($\|u\|^2 = \sum_{i=1}^k {u_i}^2$, $u = (u_1, u_2, .., u_k)'$). This is called the *extrinsic distance* on M.

Assume that $\tilde{M}$ is a closed subset of $\Re^k$. Then for every $u \in \Re^k$ there exists a compact set of points in $\tilde{M}$ whose distance from u is the smallest among all points in $\tilde{M}$. We will denote this set by

$$Pu \equiv P_{\tilde{M}}u = \{x \in \tilde{M} : \|x-u\| \leq \|y-u\| \; \forall y \in \tilde{M}\}.$$

If this set is a singleton, u is said to be a *nonfocal point* of $\Re^k$ (with respect to $\tilde{M}$); otherwise it is said to be a *focal point* of $\Re^k$.

Definition 3.1. Let (M, ρ), ϕ be as above. Let Q be a probability measure on M with finite Fréchet function. The Fréchet mean (set) of Q is called the *extrinsic mean* (set) of Q, and the Fréchet variation of Q is called its *extrinsic variation.* If X_j ($j = 1, \ldots, n$) are iid observations from Q, and $Q_n = \frac{1}{n}\sum_{j=1}^n \delta_{X_j}$ is the empirical distribution, then the Fréchet mean(set) of Q_n is called the *extrinsic sample mean*(set) and the Fréchet variation of Q_n is called the *extrinsic sample variation.*

Let $\tilde{Q}$ and $\tilde{Q}_n$ be the images of Q and Q_n, respectively, on $\Re^k$ under ϕ: $\tilde{Q} = Q \circ \phi^{-1}$, $\tilde{Q}_n = Q_n \circ \phi^{-1}$. The next result gives us a way to calculate the extrinsic mean and establishes the consistency of the sample mean as an estimator of the population mean if that exists. For a proof see Proposition 3.1 in Bhattacharya and Patrangenaru [7].

Proposition 3.1. *(a) If $\tilde{\mu} = \int_{R^k} u\tilde{Q}(du)$ is the mean of $\tilde{Q}$, then the extrinsic mean set of Q is given by $\phi^{-1}(P\tilde{\mu})$. (b) If $\tilde{\mu}$ is a nonfocal point of $\Re^k$ (relative to $\tilde{M}$), then the extrinsic sample mean μ_{nE} (any measurable selection from the extrinsic mean set of Q_n) is a strongly consistent estimator of the extrinsic mean $\mu_E = \phi^{-1}(P\tilde{\mu})$.*

3.1. Asymptotic Distribution of the Sample Extrinsic Mean

We can use Theorem 2.3 to get the asymptotic distribution of the sample extrinsic mean. However, expressions for the parameters Λ and Σ are not easy to get. Here we devise another way to derive the asymptotic distribution. We assume that the mean $\tilde{\mu}$ of $\tilde{Q}$ is a nonfocal point, so that the projection $P\tilde{\mu}$ of $\tilde{\mu}$ on $\phi(M)$ is unique, and the extrinsic mean of Q is $\mu_E = \phi^{-1}(P\tilde{\mu})$. Let $\overline{\tilde{X}} = (1/n)\sum_{j=1}^n \tilde{X}_j$ denote the sample mean of $\tilde{X}_j = \phi(X_j)$. The extrinsic sample mean set is $C_{Q_n} = \phi^{-1}(P\overline{\tilde{X}})$, where $P\overline{\tilde{X}}$ is the set of projection of $\overline{\tilde{X}}$ on $\phi(M)$. In a neighborhood of a nonfocal point such as $\tilde{\mu}$, $P(.)$ is smooth. So we can write

$$\sqrt{n}[P(\overline{\tilde{X}}) - P(\tilde{\mu})] = \sqrt{n}(\mathrm{d}_{\tilde{\mu}}P)(\overline{\tilde{X}} - \tilde{\mu}) + o_P(1) = (\mathrm{d}_{\tilde{\mu}}P)(\sqrt{n}(\overline{\tilde{X}} - \tilde{\mu})) + o_P(1) \tag{3.1}$$

where $\mathrm{d}_{\tilde{\mu}}P$ is the differential (map) of $P(.)$, which takes vectors in the tangent space of $\Re^k$ at $\tilde{\mu}$ to tangent vectors of $\phi(M)$ at $P(\tilde{\mu})$. Hence the left side is asymptotically normal.

For the case of regular submanifolds embedded in an Euclidean space by the inclusion map, a similar asymptotic distribution and a two-sample test were constructed independently by Hendricks and Landsman [18] and, for more general manifolds, by Patrangenaru [26] and Bhattacharya and Patrangenaru [8].

3.2. Application to the Planar Shape Space of k-ads

Consider a set of k points on the plane, e.g., k locations on a skull projected on a plane, not all points being the same. We will assume $k > 2$ and refer to such a set as a *k-ad* (or a set of k *landmarks*). For convenience we will denote a k-ad by k complex numbers ($z_j = x_j + iy_j, 1 \leq j \leq k$), i.e., we will represent k-ads on a complex plane. By the *shape* of a k-ad $\mathbf{z} = (z_1, z_2, \ldots, z_k)$, we mean the equivalence class, or orbit of $\mathbf{z}$ under translation, rotation and scaling. To remove translation, one may substract $\langle \mathbf{z} \rangle \equiv (\langle z \rangle, \langle z \rangle, \ldots, \langle z \rangle)$ ($\langle z \rangle = (1/k) \sum_{j=1}^{k} z_j$) from $\mathbf{z}$ to get $\mathbf{z} - \langle \mathbf{z} \rangle$. Rotation of the k-ad by an angle θ and scaling (by a factor $r > 0$) are achieved by multiplying $\mathbf{z} - \langle \mathbf{z} \rangle$ by the complex number $\lambda = r \exp i\theta$. Hence one may represent the shape of the k-ad as the complex line passing through $\mathbf{z} - \langle \mathbf{z} \rangle$, namely, $\{\lambda(\mathbf{z} - \langle \mathbf{z} \rangle) \colon \lambda \in \mathbb{C} \setminus \{0\}\}$. Thus the space of k-ads is the set of all complex lines on the (complex $(k-1)$-dimensional) hyperplane, $H^{k-1} = \{w \in C^k \setminus \{0\} \colon \sum_1^k w_j = 0\}$. Therefore the shape space Σ_2^k of planer k-ads has the structure of the *complex projective space* $\mathbb{C}P^{k-2}$: the space of all complex lines through the origin in $\mathbb{C}^{k-1}$. As in the case of $\mathbb{C}P^{k-2}$, it is convenient to represent the element of Σ_2^k corresponding to a k-ad $\mathbf{z}$ by the curve $\gamma(z) = [z] = \{e^{i\theta}((z - \langle \mathbf{z} \rangle)/\|z - \langle \mathbf{z} \rangle\|) \colon 0 \leq \theta < 2\pi\}$ on the unit sphere in $H^{k-1} \approx \mathbb{C}^{k-1}$.

If we denote by u the quantity $(\mathbf{z} - \langle \mathbf{z} \rangle)/\|\mathbf{z} - \langle \mathbf{z} \rangle\|$, called the *preshape* of the shape of $\mathbf{z}$, then another representation of Σ_2^k is via the *Veronese-Whitney embedding* ϕ into the space $S(k, \mathbb{C})$ of all $k \times k$ complex Hermitian matrices. $S(k, \mathbb{C})$ is viewed as a (real) vector space with respect to the scaler field $\Re$. The embedding ϕ is given by

$$\begin{aligned} &\phi \colon \Sigma_2^k \to S(k, \mathbb{C}), \\ &\phi([z]) = uu^* \ (u = (u_1, \ldots, u_k)' \in H^{k-1}, \|u\| = 1) \\ &\qquad\quad = ((u_i \bar{u}_j))_{1 \leq i,j \leq k}. \end{aligned} \tag{3.2}$$

The shape of $\mathbf{z}$, $[z] = \{e^{i\theta}u \colon 0 \leq \theta < 2\pi\}$ is the orbit of the vector u under rotation. Note that if $v_1, v_2 \in [z]$, then $\phi([v_1]) = \phi([v_2]) = \phi((z - \langle \mathbf{z} \rangle)/\|z - \langle \mathbf{z} \rangle\|)$. Define the *extrinsic distance* ρ on Σ_2^k by that induced from this embedding, namely,

$$\rho^2([z], [w]) = \|uu^* - vv^*\|^2 \ \ , u \doteq \frac{z - \langle \mathbf{z} \rangle}{\|z - \langle \mathbf{z} \rangle\|} \ , v \doteq \frac{w - \langle \mathbf{w} \rangle}{\|w - \langle \mathbf{w} \rangle\|} \tag{3.3}$$

where for arbitrary $k \times k$ complex matrices A, B,

$$\|A - B\|^2 = \sum_{j,j'} |a_{jj'} - b_{jj'}\|^2 = \text{Trace}(A - B)(A - B)^* \tag{3.4}$$

is just the squared euclidean distance between A and B regarded as elements of $\mathbb{C}^{k^2}$ (or, $\Re^{2k^2}$). Since the matrices uu^*, vv^* in (3.2) are Hermitian, one notes that the image $\phi(\Sigma_2^k)$ of Σ_2^k is a closed subset of $\mathbb{C}^{k^2}$ and the "conjugate-transpose" symbol $*$ may be dropped from (3.4) in computing distances in $\phi(\Sigma_2^k)$.

Let Q be a probability measure on the shape space Σ_2^k, let $[X_1], [X_2], \ldots, [X_n]$ be an i.i.d. sample from Q and let $\tilde{\mu}$ denote the mean vector of $\tilde{Q} \doteq Q \circ \phi^{-1}$, regarded as a probability measure on $\mathbb{C}^{k^2}$ (or, $\Re^{2k^2}$). Note that $\tilde{\mu}$ belongs to the convex hull of $\tilde{M} = \phi(\Sigma_2^k)$ and in particular, is an element of H^{k-1}. Let T be a (complex) orthogonal $k \times k$ matrix such that $T\tilde{\mu}T^* = D = \text{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)$, where $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_k$ are the eigenvalues of $\tilde{\mu}$. Then, writing $v = Tu$ with u

as in (3.3),

$$\begin{aligned}
\|uu^* - \tilde{\mu}\|^2 = \|vv^* - D\|^2 &= \sum_{j=1}^{k} (|v_j|^2 - \lambda_j)^2 + \sum_{j \neq j'} |v_j \overline{v}_{j'}|^2 \\
&= \sum {\lambda_j}^2 + \sum_{j=1}^{k} |v_j|^4 - 2\sum_{j=1}^{k} \lambda_j |v_j|^2 + \sum_{j=1}^{k} |v_j|^2 . \sum_{j'=1}^{k} |v_{j'}|^2 - \sum_{j=1}^{k} |v_j|^4 \\
&= \sum {\lambda_j}^2 + 1 - 2\sum_{j=1}^{k} \lambda_j |v_j|^2 \qquad (3.5)
\end{aligned}$$

which is minimized (on $\phi(\Sigma_2^k)$) by taking $v = e_k = (0, \ldots, 0, 1)'$, i.e., $u = T^* e_k$, a unit eigenvector having the largest eigenvalue λ_k of $\tilde{\mu}$. It follows that the extrinsic mean μ_E, say, of Q is unique if and only if the eigenspace for the largest eigenvalue of $\tilde{\mu}$ is (complex) one-dimensional, and then $\mu_E = [\mu]$, $\mu(\neq 0) \in$ the eigenspace of the largest eigenvalue of $\tilde{\mu}$.

From (3.5), the extrinsic variation of Q has the expression

$$\begin{aligned}
V &= \mathrm{E}\|X_1 X_1^* - \mu\mu^*\|^2 \\
&= \mathrm{E}\|X_1 X_1^* - \tilde{\mu}\|^2 + \|\tilde{\mu} - \mu\mu^*\|^2 \\
&= 2(1 - \lambda_k) \qquad (3.6)
\end{aligned}$$

Therefore, we have the following consequence of Proposition 2.4 and Proposition 3.1.

Corollary 3.2. *Let μ_n denote an eigenvector of $(1/n)\sum_{j=1}^{n} X_j {X_j}^*$ having the largest eigenvalue λ_{kn}. (a) If the largest eigenvalue λ_k of $\tilde{\mu}$ is simple, then the extrinsic sample mean $[\mu_n]$ is a strongly consistent estimator of the extrinsic mean $[\mu]$. (b) The sample extrinsic variation, $V_n = 2(1 - \lambda_{kn})$ is a strongly consistent estimator of the extrinsic variation, $V = 2(1 - \lambda_k)$.*

The distance ρ on Σ_2^k in (3.3) can be expressed as

$$\rho^2([z],[w]) \equiv \|uu^* - vv^*\|^2 = 2(1 - |u^* v|^2). \qquad (3.7)$$

This is the so-called *full Procrustes distance* for Σ_2^k. See Kent [23], Dryden and Mardia [12], and Kendall et al. [21].

3.3. Asymptotic Distribution of Mean Shape

To get the asymptotic distribution of the sample extrinsic mean shape using (3.1), we embed $M = \Sigma_2^k$ into $S(k, \mathbb{C})$, the space of all $k \times k$ complex self-adjoint matrices, via the map ϕ in (3.2). We consider $S(k, \mathbb{C})$ as a linear subspace of $\mathbb{C}^{k^2}$ (over $\Re$) and as such a regular submanifold of $\mathbb{C}^{k^2}$ embedded by the inclusion map, and inheriting the metric tensor:

$$\langle A, B \rangle = \mathrm{Re}\left(\mathrm{Trace}(A\bar{B}')\right).$$

The (real) dimension of $S(k, \mathbb{C})$ is k^2. An orthonormal basis for $S(k, \mathbb{C})$ is given by $\{v_b^a : 1 \leq a \leq b \leq k\}$ and $\{w_b^a : 1 \leq a < b \leq k\}$, defined as

$$v_b^a = \begin{cases} \frac{1}{\sqrt{2}}(e_a e_b^t + e_b e_a^t), & a < b \\ e_a e_a^t, & a = b \end{cases}$$

$$w_b^a = +\frac{i}{\sqrt{2}}(e_a e_b^t - e_b e_a^t), \ a < b.$$

where $\{e_a : 1 \leq a \leq k\}$ is the standard canonical basis for $\Re^k$.

We also take $\{v_b^a : 1 \leq a \leq b \leq k\}$ and $\{w_b^a : 1 \leq a < b \leq k\}$ as the (constant) orthogonal frame for $S(k, \mathbb{C})$. For any $U \in O(k)$ $(UU^* = U^*U = I)$, $\{Uv_b^aU^* : 1 \leq a \leq b \leq k\}$, $\{Uw_b^aU^* : 1 \leq a < b \leq k\}$ is also an orthogonal frame for $S(k, \mathbb{C})$. Assume that the mean $\tilde{\mu}$ of $\tilde{Q}$ has its largest eigenvalue simple. To apply (3.1), we view $\mathrm{d}_{\tilde{\mu}}P : S(k, \mathbb{C}) \to T_{P(\tilde{\mu})}\phi(\Sigma_2^k)$. Choose $U \in O(k)$ such that $U^*\tilde{\mu}U = D \equiv \mathrm{Diag}(\lambda_1, \ldots, \lambda_k)$, $\lambda_1 \leq \ldots \leq \lambda_{k-1} < \lambda_k$ being the eigenvalues of $\tilde{\mu}$.

Choose the basis frame $\{Uv_b^aU^*, Uw_b^aU^*\}$ for $S(k, \mathbb{C})$. Then one can show that

$$\mathrm{d}_{\tilde{\mu}}P(Uv_b^aU^*) = \begin{cases} 0 & \text{if } 1 \leq a \leq b < k,\ \mathrm{a = b = k}, \\ (\lambda_k - \lambda_a)^{-1}Uv_k^aU^* & \text{if } 1 \leq a < k, b = k. \end{cases}$$

$$\mathrm{d}_{\tilde{\mu}}P(Uw_b^aU^*) = \begin{cases} 0 & \text{if } 1 \leq a < b < k \\ (\lambda_k - \lambda_a)^{-1}Uw_k^aU^* & \text{if } 1 \leq a < k, b = k. \end{cases} \tag{3.8}$$

Write

$$\begin{aligned} \sqrt{n}(\bar{\bar{\tilde{X}}} - \tilde{\mu}) = & \sum\sum_{1 \leq a \leq b \leq k} \langle \sqrt{n}(\bar{\bar{\tilde{X}}} - \tilde{\mu}), Uv_b^aU^* \rangle Uv_b^aU^* \\ & + \sum\sum_{1 \leq a < b \leq k} \langle \sqrt{n}(\bar{\bar{\tilde{X}}} - \tilde{\mu}), Uw_b^aU^* \rangle Uw_b^aU^*. \end{aligned} \tag{3.9}$$

Since $\bar{\bar{\tilde{X}}}\mathbf{1}_k = \tilde{\mu}\mathbf{1}_k = 0$, $\lambda_1 = 0$ and $U_{.1} = \alpha\mathbf{1}_k$, $|\alpha| = 1/\sqrt{k}$. Thus

$$\langle \sqrt{n}(\bar{\bar{\tilde{X}}} - \tilde{\mu}), Uv_b^1U^* \rangle = \langle \sqrt{n}(\bar{\bar{\tilde{X}}} - \tilde{\mu}), Uw_b^1U^* \rangle = 0.$$

Therefore,

$$\begin{aligned} & \mathrm{d}_{\tilde{\mu}}P(\sqrt{n}(\bar{\bar{\tilde{X}}} - \tilde{\mu})) \\ & = \sum_{a=2}^{k-1} \langle \sqrt{n}(\bar{\bar{\tilde{X}}} - \tilde{\mu}), Uv_k^aU^* \rangle (\lambda_k - \lambda_a)^{-1} Uv_k^aU^* \\ & + \sum_{a=2}^{k-1} \langle \sqrt{n}(\bar{\bar{\tilde{X}}} - \tilde{\mu}), Uw_k^aU^* \rangle (\lambda_k - \lambda_a)^{-1} Uw_k^aU^*. \end{aligned} \tag{3.10}$$

From (3.10), we see that $\sqrt{n}(P(\bar{\bar{\tilde{X}}}) - P(\tilde{\mu}))$ has an asymptotic Gaussian distribution on a subspace of $S(k, \mathbb{C})$ with asymptotic coordinates

$$T_n(\tilde{\mu}) = \Big(\langle \sqrt{n}(\bar{\bar{\tilde{X}}} - \tilde{\mu}), Uv_k^aU^* \rangle_{a=2}^{k-1},\ \langle \sqrt{n}(\bar{\bar{\tilde{X}}} - \tilde{\mu}), Uw_k^aU^* \rangle_{a=2}^{k-1} \Big)$$

with respect to the basis vector $\{(\lambda_k - \lambda_a)^{-1}Uv_k^aU^*, (\lambda_k - \lambda_a)^{-1}Uw_k^aU^*\}_{a=2}^{k-1}$. Writing $\Sigma(\tilde{\mu})$ for the covariance matrix of $T_n(\tilde{\mu})$, and assuming that it is nonsingular,

$$T_n(\tilde{\mu})'\Sigma(\tilde{\mu})^{-1}T_n(\tilde{\mu}) \longrightarrow \mathcal{X}_{2k-4}^2. \tag{3.11}$$

3.4. Two Sample Testing Problems on Σ_2^k

Let Q_1 and Q_2 be two probability measures on the shape space Σ_2^k, and let μ_1 and μ_2 denote the means of $Q_1 \circ \phi^{-1}$ and $Q_2 \circ \phi^{-1}$, respectively. Suppose $[x_1], \ldots, [x_n]$

and $[y_1], \dots, [y_m]$ are i.i.d. random samples from Q_1 and Q_2 respectively. Let $X_i = \phi([x_i])$, $Y_i = \phi([y_i])$ be their images onto $\phi(\Sigma_2^k)$ which are random samples from $Q_1 \circ \phi^{-1}$ and $Q_2 \circ \phi^{-1}$, respectively. Suppose we are to test if the extrinsic means of Q_1 and Q_2 are equal, i.e.

$$H_0 : P\mu_1 = P\mu_2$$

We assume that both μ_1 and μ_2 have simple largest eigenvalues. Then under H_0, the corresponding eigenvectors differ by a rotation.

Choose $\mu \in S(k, \mathbb{C})$ with the same projection as μ_1 and μ_2. Suppose $\mu = U\Lambda U^*$, where $\Lambda = \text{Diag}(\lambda_1 \leq \lambda_2 \leq \ldots < \lambda_k)$ are its eigenvalues and $U = [U_1, U_2, \dots, U_k]$ are the corresponding eigenvectors. Under H_0, $P\mu_1 = P\mu_2 = U_k U_k^*$. From (3.10),

$$\begin{aligned} &\mathrm{d}_\mu P(\bar{X} - \mu) \\ &= \sum_{a=2}^{k-1} \sqrt{2}\text{Re}(U_a^* \bar{X} U_k)(\lambda_k - \lambda_a)^{-1} U v_k^a U^* + \sum_{a=2}^{k-1} \sqrt{2}\text{Im}(U_a^* \bar{X} U_k)(\lambda_k - \lambda_a)^{-1} U w_k^a U^* \\ &= \sum_{a=2}^{k-1} (\lambda_k - \lambda_a)^{-1} (U_a^* \bar{X} U_k) U_a U_k^* + \sum_{a=2}^{k-1} (\lambda_k - \lambda_a)^{-1} (U_k^* \bar{X} U_a) U_k U_a^*, \end{aligned} \tag{3.12}$$

$$\begin{aligned} &\mathrm{d}_\mu P(\bar{Y} - \mu) \\ &= \sum_{a=2}^{k-1} \sqrt{2}\text{Re}(U_a^* \bar{Y} U_k)(\lambda_k - \lambda_a)^{-1} U v_k^a U^* + \sum_{a=2}^{k-1} \sqrt{2}\text{Im}(U_a^* \bar{Y} U_k)(\lambda_k - \lambda_a)^{-1} U w_k^a U^* \\ &= \sum_{a=2}^{k-1} (\lambda_k - \lambda_a)^{-1} (U_a^* \bar{Y} U_k) U_a U_k^* + \sum_{a=2}^{k-1} (\lambda_k - \lambda_a)^{-1} (U_k^* \bar{Y} U_a) U_k U_a^*. \end{aligned} \tag{3.13}$$

Define

$$T(\mu)_{ij} = \begin{cases} \text{Re}(U_{i+1}^* X_j U_k) & \text{if } 1 \leq i \leq k-2,\ 1 \leq j \leq n \\ \text{Im}(U_{i-k+3}^* X_j U_k) & \text{if } k-1 \leq i \leq 2k-4,\ 1 \leq j \leq n \end{cases}$$

$$S(\mu)_{ij} = \begin{cases} \text{Re}(U_{i+1}^* Y_j U_k) & \text{if } 1 \leq i \leq k-2,\ 1 \leq j \leq m \\ \text{Im}(U_{i-k+3}^* Y_j U_k) & \text{if } k-1 \leq i \leq 2k-4,\ 1 \leq j \leq m \end{cases}$$

$$\bar{T}(\mu) = \frac{1}{n} \sum_{j=1}^{n} T(\mu)_{.j},\ \bar{S}(\mu) = \frac{1}{m} \sum_{j=1}^{m} S(\mu)_{.j}. \tag{3.14}$$

Under H_0, $\bar{T}(\mu)$ and $\bar{S}(\mu)$ have mean zero, and as $n, m \to \infty$,

$$\sqrt{n}\bar{T}(\mu) \xrightarrow{\mathcal{L}} N(0, \Sigma_1(\mu)),\ \sqrt{m}\bar{S}(\mu) \xrightarrow{\mathcal{L}} N(0, \Sigma_2(\mu)) \tag{3.15}$$

where $\Sigma_1(\mu)$ and $\Sigma_2(\mu)$ are the covariances of $T(\mu)_{.1}$ and $S(\mu)_{.1}$, respectively. Suppose $(n/(m+n)) \to p$, $(m/(m+n)) \to q$, for some $p, q > 0$; $p + q = 1$. Then

$$\sqrt{n+m}(\bar{T}(\mu) - \bar{S}(\mu)) \xrightarrow{\mathcal{L}} N_{2k-4}(0, \frac{1}{p}\Sigma_1(\mu) + \frac{1}{q}\Sigma_2(\mu)).$$

Thus assuming $\Sigma_1(\mu)$, $\Sigma_2(\mu)$ and hence $\frac{1}{p}\Sigma_1(\mu) + \frac{1}{q}\Sigma_2(\mu)$ to be nonsingular,

$$(n+m)(\bar{T}(\mu) - \bar{S}(\mu))'(\frac{1}{p}\Sigma_1(\mu) + \frac{1}{q}\Sigma_2(\mu))^{-1}(\bar{T}(\mu) - \bar{S}(\mu)) \xrightarrow{\mathcal{L}} \mathcal{X}_{2k-4}^2. \tag{3.16}$$

Note that the nonsingularity assumption for $\Sigma_1(\mu)$ and $\Sigma_2(\mu)$ are satisfied if, for example, Q_1 and Q_2 have nonzero absolutely continuous components with respect to the volume measure on Σ_2^k (identified with the Riemannian manifold $\mathbb{C}P^{k-2}$). We can choose μ to be any positive linear combination of μ_1 and μ_2. Then under H_0, μ will have the same projection on $\phi(\Sigma_2^k)$ as μ_1 and μ_2. We may take $\mu = p\mu_1 + q\mu_2$. In practice, since μ_1 and μ_2 are unknown, so is μ. Then we may estimate μ by the pooled sample mean $\hat{\mu} = (n\bar{X} + m\bar{Y})/(m+n)$, $\Sigma_1(\mu)$ and $\Sigma_2(\mu)$ by their sample estimates $\hat{\Sigma}_1(\hat{\mu})$ and $\hat{\Sigma}_2(\hat{\mu})$, where

$$\hat{\Sigma}_1(\mu) = \frac{1}{n}T(\mu)T(\mu)^{'} - \bar{T}(\mu)\bar{T}(\mu)^{'},\ \hat{\Sigma}_2(\mu) = \frac{1}{m}S(\mu)S(\mu)^{'} - \bar{S}(\mu)\bar{S}(\mu)^{'}$$

Then the two-sample test statistic in (3.16) can be estimated by

$$T_{nm} = (\bar{T}(\hat{\mu}) - \bar{S}(\hat{\mu}))^{'}(\frac{1}{n}\hat{\Sigma}_1(\hat{\mu}) + \frac{1}{m}\hat{\Sigma}_2(\hat{\mu}))^{-1}(\bar{T}(\hat{\mu}) - \bar{S}(\hat{\mu})). \tag{3.17}$$

Given level α, we reject H_0 if

$$T_{nm} > \mathcal{X}^2_{2k-4}(1-\alpha). \tag{3.18}$$

The expression for T_{nm} depends on the spectrum of $\hat{\mu}$ through the orbit $[U_k(\hat{\mu})]$ and the subspace spanned by $\{U_2(\hat{\mu}), \ldots, U_{k-1}(\hat{\mu})\}$. If the population mean exists, $[U_k(\hat{\mu})]$ is a consistent estimator of $[U_k(\mu)]$ and by perturbation theory (see Dunford and Schwartz [13], p. 598), the projection on $\mathrm{Span}\{U_2(\hat{\mu}), \ldots, U_{k-1}(\hat{\mu})\}$ converges to that on $\mathrm{Span}\{U_2(\mu), \ldots, U_{k-1}(\mu)\}$. Thus from (3.16) and (3.17), T_{nm} has an asymptotic $\mathcal{X}^2_{2k-4}$ distribution. Hence the test in (3.18) has asymptotic level α.

To test if the populations have the same spread around their respective means, we use the test statistic in (2.14), which is

$$T_{nm} = 2\frac{\lambda_{km} - \lambda_{kn}}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}, \tag{3.19}$$

where λ_{kn} and λ_{km} are the largest eigenvalues of $\bar{X}$ and $\bar{Y}$, respectively. Under H_0, T_{nm} has asymptotic Normal distribution.

4. Intrinsic Mean and Variation

Let (M, g) be a d-dimensional connected complete Riemannian manifold, g being the Riemannian metric on M. Let the distance $\rho = d_g$ be the geodesic distance under g. Let Q be a probability distribution on M with finite Fréchet function,

$$F(p) = \int_M d_g^2(p, m)Q(dm),\ p \in M. \tag{4.1}$$

Definition 4.1. The Fréchet mean (set) of Q under the distance d_g is called its *intrinsic mean* (set). The Fréchet variation of Q under d_g is called its *intrinsic variation.* Let $X_1, X_2, \ldots, X_n$ be i.i.d. observations on M with common distribution Q. The sample Fréchet mean (set) is called the *sample intrinsic mean* (set) and the sample Fréchet variation is called the *sample intrinsic variation.*

Let us define a few technical terms related to Riemannian manifolds which we will use extensively in the subsequent sections. For details on Riemannian Manifolds, see DoCarmo [10], Gallot et al. [15] or Lee [25].

1. *Geodesic*: These are curves γ on the manifold with zero acceleration. They are locally length minimizing curves. For example, consider great circles on the sphere or straight lines in $\Re^d$.
2. *Exponential map*: For $p \in M$, $v \in T_pM$, we define $\exp_p v = \gamma(1)$, where γ is a geodesic with $\gamma(0) = p$ and $\dot{\gamma}(0) = v$.
3. *Cut locus*: Let γ be a unit speed geodesic starting at p, $\gamma(0) = p$. Let t_0 be the supremum of all t for which γ is length minimizing on $[0, t]$. Then $\gamma(t_0)$ is called the cut point of p along γ. The *cut locus* of p, $C(p)$, is the set of all cut points of p along all geodesics. For example, $C(p) = \{-p\}$ on S^d.
4. *Convex ball*: A ball B is called convex if, for any $p, q \in B$, a unique geodesic from p to q lies in B, which is also the shortest geodesic from p to q. For example, any ball of radius $\pi/2$ or less in S^d is convex.
5. *Sectional Curvature*: Recall the notion of Gaussian curvature of two dimensional surfaces. On a Riemannian manifold M, choose a pair of linearly independent vectors $u, v \in T_pM$. A two dimensional submanifold of M is swept out by the set of all geodesics starting at p and with initial velocities lying in the two-dimensional section π spanned be u, v. The curvature of this submanifold is called the sectional curvature at p of the section π.

In all subsequent sections, we assume that M has all sectional curvatures bounded above by some $C \geq 0$.

The next result, due to Kendall [22], gives a sufficient condition for the existence of a unique local minimum of F in a geodesic ball of reasonably wide radius.

Proposition 4.1. *If the support of Q is contained in $B(p, r)$ with $r < \pi/(2\sqrt{C})$ and $\overline{B(p, r)} \cap C(p) = \phi$, then the Fréchet function F of Q has a unique local minimum in $B(p, r)$.*

Recall that (Karchar [19]; see also Theorem 2.1 in Bhattacharya and Patrangenaru [7]) if $Q(C(p)) = 0\ \forall p \in M$, then every local minimum μ of F satisfies

$$\int_{T_\mu M} v\tilde{Q}(dv) = 0 \tag{4.2}$$

where $\tilde{Q}$ is the image of Q under the map $\exp_\mu^{-1}$ on $M \setminus C(\mu)$.

4.1. Asymptotic Distribution of the Sample Intrinsic Mean

One can use Theorem 2.3 to get the asymptotic distribution of the sample intrinsic mean. For that we need to verify assumptions (i) to (vii). The next result gives sufficient conditions for those assumptions to hold.

Theorem 4.2. *Suppose the support of Q is contained in a geodesic ball $B(p, r)$ with center p and radius r as in Proposition 4.1. Let $\phi = \exp_p^{-1} : B(p, r) \longrightarrow T_pM (\approx \Re^d)$. Define $h(x, y) = d_g^2(\phi^{-1}x, \phi^{-1}y)$; $x, y \in \Re^d$. Let $((\mathrm{D}_r h))_{r=1}^d$ and $((\mathrm{D}_r \mathrm{D}_s h))_{r,s=1}^d$ be the matrices of first and second order derivatives of $y \mapsto h(x, y)$. Let $\tilde{X}_j = \phi(X_j) (j = 1, \ldots, n)$, $X_1, \ldots, X_n$ being i.i.d. observations from Q. Let $\mu = \phi(\mu_I)$, μ_I being the point of local minimum of F in $B(p, r)$. Let $\mu_n = \phi(\mu_{nI})$, μ_{nI} being the point of local minimum of F_n in $B(p, r)$. Define $\Lambda = \mathrm{E}((\mathrm{D}_r \mathrm{D}_s h(\tilde{X}_1, \mu)))_{r,s=1}^d$, $\Sigma = \mathrm{Cov}((\mathrm{D}_r h(\tilde{X}_1, \mu)))_{r=1}^d$. If Λ and Σ are nonsingular, then*

$$\sqrt{n}(\mu_n - \mu) \xrightarrow{\mathcal{L}} N(0, \Lambda^{-1}\Sigma\Lambda^{-1}). \tag{4.3}$$

Proof. When Q is considered as a probability measure on the compact ball $\overline{B(p,r)}$ (as the underlying metric space), μ_n is a consistent estimator of μ, by Proposition 2.2. In view of Proposition 4.1, as in the proof of Theorem 2.3 in Bhattacharya and Patrengenaru [8], Assumptions (i)-(vii) of Theorem 2.3 are verified. □

Remark 4.1: The nonsingularity of Σ in Theorem 4.2 is a mild condition which holds in particular if Q has a density (component) with respect to the volume measure. The nonsingularity of Λ is a more delicate matter in general, involving a detailed analysis involving curvature and Jacobi fields. These matters are considered in detail in Bhattacharya and Bhattacharya [3].

Remark 4.2: Under the hypothesis of Proposition 4.1 (and Theorem 4.2), the point of local minimum μ_I of F in $B(p,r)$ may not be the global minimizer of F on M. However, if one restricts attention to the closed ball $\overline{B(p,r)}$ as the underlying metric space of interest, this point of local minimum is the intrinsic mean (on $\overline{B(p,r)}$). The advantage of Theorem 4.2 over the earlier result Theorem 2.3 in Bhattacharya and Patrengenaru [8] is that here one allows a much wider support of Q, namely, the radius r here is twice as large as that allowed in the earlier result. This is particularly important in two-sample problems as well as in problems of classification involving several populations. Also from a statistical point of view, the mean shape is perhaps better represented if defined as the Fréchet mean over $\overline{B(p,r)}$ than over the whole of M, since $Q(M \setminus \overline{B(p,r)}) = 0$ and since $B(p,r)$ is a connected Riemannian manifold inheriting the metric of M.

Theorem 4.2 can be used to construct an asymptotic $1-\alpha$ confidence set for μ_I which is given by

$$\{\mu_I : n(\mu_n - \mu)^t(\hat{\Lambda}^{-1}\hat{\Sigma}\hat{\Lambda}^{-1})^{-1}(\mu_n - \mu) \leq \mathcal{X}_d^2(1-\alpha)\} \tag{4.4}$$

where $(\hat{\Sigma}, \hat{\Lambda})$ are consistent sample estimates of (Σ, Λ) and $\mathcal{X}_d^2(1-\alpha)$ is the upper $(1-\alpha)^{\text{th}}$ quantile of the chi-squared distribution with d degrees of freedom.

Also we can perform a nonparametric test to test if two distributions Q_1 and Q_2 have the same intrinsic mean μ_I. Let $\mu = \phi(\mu_I)$. Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be i.i.d. observations from Q_1 and Q_2, respectively. Let Q_n and Q_m be the empirical distributions and μ_{n1} and μ_{m2} be the corresponding sample mean coordinates. We want to test $H_0 : \mu_{1I} = \mu_{2I} = \mu_I$, say, against $H_1 : \mu_{1I} \neq \mu_{2I}$, where μ_{1I} and μ_{2I} are the true intrinsic means of Q_1 and Q_2, respectively. Then the test statistic used is

$$T_{nm} = (n+m)(\mu_{n1} - \mu_{m2})'\hat{\Sigma}^{-1}(\mu_{n1} - \mu_{m2}), \tag{4.5}$$

$$\hat{\Sigma} = (m+n)\left(\frac{1}{n}\hat{\Lambda}_1^{-1}\hat{\Sigma}_1\hat{\Lambda}_1^{-1} + \frac{1}{m}\hat{\Lambda}_2^{-1}\hat{\Sigma}_2\hat{\Lambda}_2^{-1}\right), \tag{4.6}$$

(Λ_1, Σ_1) and (Λ_2, Σ_2) being the parameters in the asymptotic distribution of $\sqrt{n}(\mu_{n1} - \mu)$ and $\sqrt{m}(\mu_{m2} - \mu)$, respectively, as defined in Theorem 4.2. $(\hat{\Lambda}_1, \hat{\Sigma}_1)$ and $(\hat{\Lambda}_2, \hat{\Sigma}_2)$ are consistent sample estimates. In case $n, m \to \infty$ such that $n/(m+n) \to \theta$, $0 < \theta < 1$, then under the hypothesis of Theorem 4.2, assuming H_0 to be true,

$$\sqrt{n+m}(\mu_{n1} - \mu_{m2}) \xrightarrow{\mathcal{L}} N_d(0, \frac{1}{\theta}\Lambda_1^{-1}\Sigma_1\Lambda_1^{-1} + \frac{1}{1-\theta}\Lambda_2^{-1}\Sigma_2\Lambda_2^{-1}). \tag{4.8}$$

So $T_{nm} \xrightarrow{\mathcal{L}} \mathcal{X}_d^2$. We reject H_0 at asymptotic level $1-\alpha$ if $T_{nm} > \mathcal{X}_d^2(1-\alpha)$.

We conclude with the test for the equality of intrinsic variations V_1, V_2 of Q_1 and Q_2. Under the hypothesis of Theorem 4.2, the test for $H_0 : V_1 = V_2$, against $H_1 : V_1 \neq V_2$, is provided by the asymptotically Normal statistic T_{nm} in (2.14), as described at the end of Section 2.

5. Examples

In this section, we record the results of two-sample tests in two examples.

Example 1 (Schizophrenic Children). In this example from Bookstein [9], 13 landmarks are recorded on a midsagittal two-dimensional slice from a Magnetic Resonance brain scan of each of 14 schizophrenic children and 14 normal children. Figures 1a,b show the preshapes of the landmarks for the patient and normal samples along with the respective sample extrinsic mean preshapes. The sample preshapes are rotated appropriately as to minimize their Euclidean distance from the mean preshape. Figure 2 shows the preshapes of the normal and the patient sample extrinsic means along with the pooled sample mean.

The values of the two-sample test statistics (3.17), (4.5) for testing equality of the mean shapes, along with the p-values are as follows.

$$\text{Extrinsic: } T_{nm} = 95.5476, \text{ p-value} = P(\mathcal{X}^2_{22} > 95.5476) = 3.8 \times 10^{-11}.$$
$$\text{Intrinsic: } T_{nm} = 95.4587, \text{ p-value} = P(\mathcal{X}^2_{22} > 95.4587) = 3.97 \times 10^{-11}.$$

The extrinsic sample variations for patient and normal samples are 0.0107 and 0.0093, respectively. The value of the two-sample test statistic (3.19) for testing equality of extrinsic variations is 0.9461, and the p-value is 0.3441. The value of the likelihood ratio test statistic, using the so-called *offset normal shape distribution* (Dryden and Mardia [12], pp. 145-146) is $-2\log\Lambda = 43.124$, p-value $= P(\mathcal{X}^2_{22} > 43.124) = 0.005$. The corresponding values of Goodall's F-statistic and Bookstein's Monte Carlo test (Dryden and Mardia [12], pp. 145-146) are $F_{22,572} = 1.89$, p-value $= P(F_{22,572} > 1.89) = 0.01$. The p-value for Bookstein's test = 0.04.

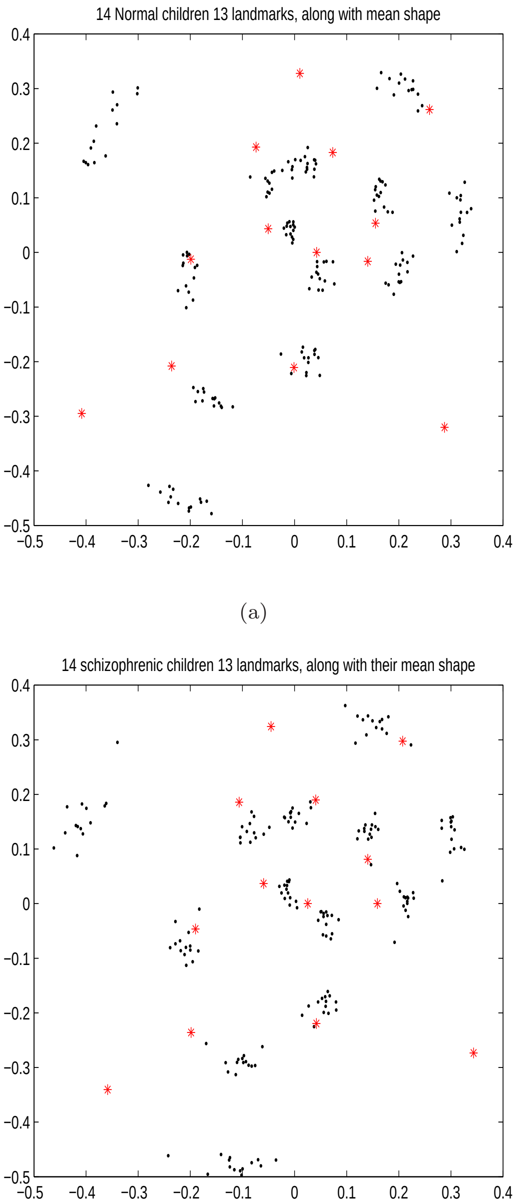
Example 2 (Gorilla Skulls). To test the difference in the shapes of skulls of male and female gorillas, eight landmarks are chosen on the midline plane of the skulls of 29 male and 30 female gorillas. We use the data of O'Higgins and Dryden reproduced in Dryden and Mardia ([12], pp. 317-318). The statistics (3.17) and (4.5) yield the following values:

$$\text{Extrinsic: } T_{nm} = 392.6, \text{ p-value} = P(\mathcal{X}^2_{12} > 392.6) < 10^{-16}.$$
$$\text{Intrinsic: } T_{nm} = 391.63, \text{ p-value} = P(\mathcal{X}^2_{12} > 391.63) < 10^{-16}.$$

The extrinsic sample variations for male and female samples are 0.005 and 0.0038, respectively. The value of the two-sample test statistic (3.19) for testing equality of extrinsic variations is 0.923, and the p-value is 0.356. A parametric F-test (Dryden and Mardia [12], pp. 154) yields $F = 26.47$, p-value $= P(F_{12,46} > 26.47) = 0.0001$. A parametric (Normal) model for Bookstein coordinates leads to the Hotelling's T^2 test (Dryden and Mardia [12], pp. 170-172) yields the p-value 0.0001.

Acknowledgements

This research was supported by NSF Grant DMS 04-06143. The authors greatly appreciate the kind and helpful suggestions by the editor and an anonymous referee which led to a substantial improvement in exposition.



(a)

(b)

Fig 1. *1a and 1b show 13 landmarks for 14 normal and 14 schizophrenic patients, respectively, along with the mean shapes, * correspond to the mean landmarks; 1c shows the sample extrinsic means for the 2 groups along with the pooled sample mean.*

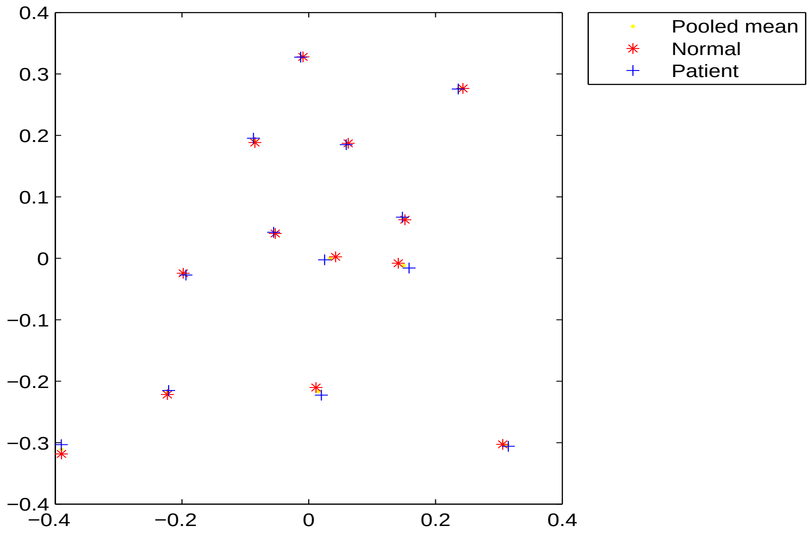


FIG 2. *The sample extrinsic means for the 2 groups along with the pooled sample mean, corresponding to Figure 1.*

References

[1] BABU, G.J. AND SINGH, K. (1984). On one term Edgeworth correction by Efron's bootstrap. *Sankhya Ser. A.* **46** 219-232.

[2] BANDULASIRI, A., BHATTACHARYA, R. N. AND PATRANGENARU, V. (2007). Algorithms for nonparametric inference on shape manifolds with applications in medical imaging. *To appear.*

[3] BHATTACHARYA, A. AND BHATTACHARYA, R. (2007). Statistics on Riemannian Manifolds: Asymptotic Distribution and Curvature. *Proc. Amer. Math. Soc.* In Press.

[4] BHATTACHARYA, R. N. AND GHOSH, J. K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.* **6** 434-451.

[5] BHATTACHARYA, R. N. AND QUMSIYEH, M. (1989). Second order and L^p-comparisons between the bootstrap and empirical Edgeworth expansions. *Ann. Statist.* **17** 160-169.

[6] BHATTACHARYA, R. N. AND PATRANGENARU, V. (2002). Nonparametric estimation of location and dispersion on Riemannian manifolds. *J. Statist. Plann. Inference* **108** 23-35.

[7] BHATTACHARYA, R. N. AND PATRANGENARU, V. (2003). Large sample theory of intrinsic and extrinsic sample means on manifolds-I. *Ann. Statist.* **31** 1-29.

[8] BHATTACHARYA, R. AND PATRANGENARU, V. (2005). Large sample theory of intrinsic and extrinsic sample means on manifolds-II. *Ann. Statist.* **33** 1225-1259.

[9] BOOKSTEIN, F.L. (1991). *Morphometric Tools for Landmark data: Geometry and Biology.* Cambridge Univ. Press.

[10] DO CARMO, M. P. (1992). *Riemannian Geometry.* Birkhauser, Boston. English translation by F. Flaherty.

[11] Dryden, I. L., Le, H. and Wood, A. (2007). The MDS model for shape. *To appear.*

[12] Dryden, I. L. and Mardia, K. V. (1998). *Statistical Shape Analysis.* Wiley N.Y.

[13] Dunford, N. and Schwartz, J. (1958). *Linear Operators-I.* Wiley, New York.

[14] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1-26.

[15] Gallot, S., Hulin, D. and Lafontaine, J. (1990). *Riemannian Geometry, 2nd ed.* Springer.

[16] Ghosh, J. K. (1994). *Higher Order Asymptotics.* Institute of Mathematical Statistics, Hayward.

[17] Hall, P. (1992). *The Bootstrap and Edgeworth Expansion.* Springer. New York.

[18] Hendricks, H. and Landsman, Z. (1998). Mean location and sample mean location on manifolds: Asymptotics, tests, confidence regions. *J. Multivariate Anal.* **67** 227-243.

[19] Karchar, H. (1977). Riemannian center of mass & mollifier smoothing. *Comm. on Pure & Applied Math.* **30** 509-541.

[20] Kendall, D. G. (1984). Shape manifolds, Procrustean metrics, and complex projective spaces. *Bull. London Math. Soc.* **16** 81-121.

[21] Kendall, D. G., Barden, D., Carne, T. K. and Le, H. (1999). *Shape & Shape Theory.* Wiley N.Y.

[22] Kendall, W. S. (1990). Probability, convexity, and harmonic maps with small image-I. Uniqueness and the fine existence. Proc. *London Math. Soc.* **61** 371-406.

[23] Kent, J. T. (1992). New directions in shape analysis. *In The Art of Statistical Science: A Tribute to G. S. Watson* (K. V. Mardia, ed.) 115-128. Wiley, New York.

[24] Le, H. (2001). Locating frechet means with application to shape spaces. *Adv. Appl. Prob.* **33** 324-338.

[25] Lee, J.M. (1997). *Riemannian Manifolds: An Introduction to Curvature.* Springer, New York.

[26] Patrangenaru, V. (1998). *Asymptotic Statistics on Manifolds.* Ph.D. dissertation, Indiana University.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 303

An Ensemble Approach to Improved Prediction from Multitype Data

Jennifer Clarke[1] **and David Seo**[2]

University of Miami School of Medicine

Abstract: We have developed a strategy for the analysis of newly available binary data to improve outcome predictions based on existing data (binary or non-binary). Our strategy involves two modeling approaches for the newly available data, one combining binary covariate selection via LASSO with logistic regression and one based on logic trees. The results of these models are then compared to the results of a model based on existing data with the objective of combining model results to achieve the most accurate predictions. The combination of model predictions is aided by the use of support vector machines to identify subspaces of the covariate space in which specific models lead to successful predictions. We demonstrate our approach in the analysis of single nucleotide polymorphism (SNP) data and traditional clinical risk factors for the prediction of coronary heart disease.

Contents

1. Introduction

In applied research contexts the statistician is often faced with newly available data which may provide information relevant to a recently completed analysis. This scenario is occurring more and more frequently in medical research as genomic data becomes available which may provide information relevant to the determination of

[1] Department of Epidemiology and Public Health, University of Miami Leonard M. Miller School of Medicine, Miami, FL, USA e-mail: `jennifer.clarke@duke.edu`

[2] Department of Medicine, Division of Cardiology, University of Miami Leonard M. Miller School of Medicine, Miami, FL, USA e-mail: `DSeo@med.miami.edu`

AMS 2000 subject classifications: Primary 62M20, 62H30; Secondary 62P10

Keywords and phrases: Prediction, Variable Selection, Model Ensembles, Single Nucleotide Polymorphism (SNP), Support Vector Machines

disease risk, a determination that has been traditionally based on existing clinical data. There is a need for statistical approaches to variable selection and modeling which attempt to provide improved outcome predictions in such contexts by combining information from new and existing data which may be of multiple types.

We have developed one such strategy for utilizing newly available binary data to improve binary outcome predictions from an existing model based on both continuous and binary data. There are many approaches to regression and classification in the machine learning and statistical literature that would be appropriate for modeling binary data, including CART [4], MARS [15], treed models [6], and logic regression [30], to name only a few. Since our interest is specifically in single nucleotide polymorphism (SNP) data, we have chosen to model binary data with logistic regression as well as logic regression models. The logic regression models recognize the often complex interactions that exist among SNPs and attempt to model such interactions in analyzing the relationships between SNPs and outcome status. Logic regression models also perform variable selection and model construction when the number of observations, n, is less than the number of covariates, p, which is a context of particular interest to us.

Our goal is to combine all available information in generating the best outcome predictions possible. In doing so we consider several approaches which borrow ideas from the multimodel ensemble modeling literature [11]. One approach is to take a weighted average of the predictions from the existing model and the binary data model, in the spirit of Bayesian model averaging [7]. A second approach is to build a model from all available covariates and not utilize the existing model, which was built before the newer binary covariates were available. Our final approach is a two-stage approach: determine subspaces of the covariate space on which the predictions from the existing model are accurate, and utilize the predictions from a model of the newer binary covariates on the remaining subspaces. This would yield a more accurate set of predictions overall in situations where neither data type is globally informative, for example, where the data have been collected from a heterogeneous population. To avoid a subspace definition which requires knowledge of the outcome of interest for observations to be predicted, we differentiate these various subspaces via support vector machines (SVM) [3, 8, 39]. As a result our technique yields 'honest' predictions for new observations.

Initially we discuss the model classes and variable selection for binary data. We then discuss how the predictions from such models can be used to improve the predictions from models based on existing data via support vector machines. Our approach is demonstrated in the context of prediction of coronary heart disease from traditional clinical risk factors and genetic (SNP) data.

2. Model Types

We assume a continuous response variable Y and a p-dimensional vector of binary covariates $\mathbf{X}$ (the 'newly available' data). In the case of SNP data each covariate $X_j, j = 1, \ldots, p$, is binary. Since the relationship between the covariates and the response is unknown we consider two model types, logistic regression and logic regression [30]. As logistic regression is a well known modeling technique we will not discuss it in detail. However we will discuss variable selection prior to logistic regression modeling when $n < p$ in Section 2.2. Logic regression is discussed in more detail below.

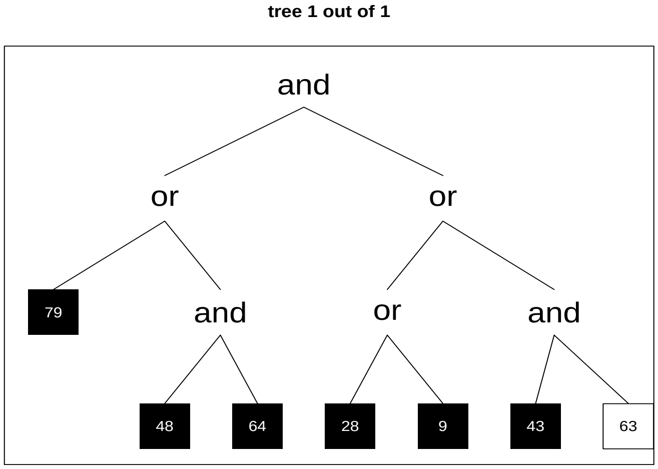


Fig 1. *A logic tree representing the Boolean expression in Equation* (2.2). *White text on a black background denotes the conjugate of a variable.*

2.1. Logic Regression

Logic regression [20, 30] is an adaptive regression methodology for finding Boolean combinations of binary covariates that are associated with an outcome variable. This methodology was developed to address situations where the interaction of many predictors is responsible for differences in the response, which is often the case when all predictors are binary. As described in [30] logic regression models take the form

$$g(E[Y]) = \beta_0 + \sum_{i=1}^{t} \beta_i L_i \tag{2.1}$$

where L_i is a Boolean expression of the covariates X_j. A score function relates fitted values to the response. This framework includes linear regression ($g(E[Y]) = E[Y]$ with score function RSS), logistic regression ($g(E[Y]) = \log(E[Y]/(1-E[Y])$ with score function binomial deviance), as well as classification ($\hat{Y} = I(L = 1)$ where $I(\cdot)$ is the indicator function and the score function is $\sum(Y \neq \hat{Y})$). Logic regression models can be conveniently represented in tree form. For example, the tree in Figure 1 represents the logic expression

$$(((X_{79}^c) \vee ((X_{48}^c) \wedge (X_{64}^c))) \wedge (((X_{28}^c) \vee (X_9^c)) \vee ((X_{43}^c) \wedge X_{63}))). \tag{2.2}$$

where X_j indicates $X_j = 1$ and X_j^c indicates the conjugate ($X_j = 0$).

Note that each L_i in Equation (2.1) may be represented as a tree, and hence logic regression allows for multiple tree models.

The space of possible logic trees is enormous, especially in situations where $n < p$. To search this space efficiently without sacrificing the desire for optimality, either a greedy search or a search via simulated annealing can be employed. These search techniques estimate the L_i and β_i simultaneously (see Equation (2.1)) and use simple 'moves' to search for 'good' logic models (ie., models which minimize the scoring function). Using terminology similar to that of CART [4] these 'moves' include growing, pruning, splitting, and deleting. As greedy searches often lead to models which overfit the data or are suboptimal (as when the search gets 'stuck' in a local minimum) [35] we prefer the use of simulated annealing to search for logic trees. Note that each 'move' mentioned above has a matching 'countermove' (e.g., growing as opposed to pruning) which is important in the Markov chain theory which underlies simulated annealing [38].

We use randomization to both test the null model of no signal in the data and determine the optimal model size (if the test is rejected). For testing the null model we randomly permute the response values and find the best fitting model. If there is no signal, the score of this model should be comparable to the score of the best model fit to the original data. By repeating the above procedure multiple times we can consider the number of runs with model scores better than the score of the best model fit to the original data as a p-value for our test.

The method for finding the optimal model size is based on a series of randomization tests. The null hypothesis for each test is that the optimal model size is k and larger models with better scores are due to noise. Assume the null hypothesis and the best model of size k has score s_k. The fitted values from this model fall into two classes; we now permute the response values within each class and find the best model of any size on the permuted data. If this model has score s_k^* then under the null hypothesis s_k comes from the same distribution as s_k^*. This distribution can be approximated by repeated permutations. We perform the above process for $k \in \{0, \ldots, K\}$ yielding a series of histograms of randomization scores s_k^* for each value of k. The optimal model size is determined by comparing these histograms, for example, one may choose the model size for which only a small proportion of scores s_k^* are better than s_k.

To further avoid overfitting the data set of n observations on p covariates $X_j, j = 1, \ldots, p$, is split into a training set of size n_1 and a test set of size n_2 $(n = n_1 + n_2)$. The logic regression models are fit to the training set and the accuracy of their predictions are evaluated on the test set. The fitting and evaluation of models can be performed in the R package `LogicReg` as described in [29].

2.2. Variable Selection

Unlike logic trees, logistic regression models require that $n < p$. In cases where $n \geq p$ we perform a variable selection via least absolute shrinkage and selection operator (LASSO) [36] prior to regression modeling. LASSO retains the beneficial features of both subset selection and ridge regression by minimizing the residual sum of squares subject to the constraint that the sum of the absolute values of the coefficients on the covariates is less than a constant (ie, a constraint on the L_1 norm of the coefficient vector). This tends to shrink some coefficients and set others to zero, leading to models with improved interpretability and stability. LASSO can be applied to generalized regression models such as logistic regression models; see [36] for details.

Osborne et al. [25] developed an efficient algorithm for computing LASSO estimates which is applicable in the $n < p$ case. We use this algorithm as implemented in the R package `lasso2` [21] in an iterative fashion to perform variable selection, removing those covariates whose coefficients has been set to zero at each iteration.

If the iterative LASSO technique yields $p^* \geq n$ variables with non-zero coefficients we remove variables one at a time between LASSO iterations, starting with those variables with the smallest coefficients, until $p^* < n$. The remaining variables are used in developing a logistic regression model of our response variable via stepwise selection.

It is important to mention that variable selection techniques exist specifically for SNP data. For example, Genomic Control (GC) [9, 10] is an analytic method for SNP selection which controls the false positive rate by separating causal from confounding factors. There are also methods for selecting which SNPs to genotype when presented with a large number of arbitrary SNPs (see, for example, [37, 41]). However, we deemed such methods inappropriate for our context of interest in which we were presented with only the partial results of such methods, i.e., a modest number of SNPs not in linkage disequilibrium (LD) and without haplotype information which had been selected based upon the application of methods similar to those mentioned above (see Section 4 for more details on our context of interest).

3. Comparing and Combining Model Predictions

Our goal is to determine whether the information from new binary covariates $X_j, j = 1, \ldots, p$, can be used to improve predictions of a response Y from a model built on existing covariates $Z_l, l = 1, \ldots, p'$. Let M_1 and M_2 represent the logic regression and logistic regression models fit to $X_j, j = 1, \ldots, p$, respectively, and let M_e represent the existing model fit to $Z_l, l = 1, \ldots, p'$. Suppose we are given a data set of size n' consisting of covariates $\mathbf{X}$ and $\mathbf{Z}$ for which we would like to generate predicted values of the outcome Y. Let $\hat{Y}_1$ be the predictions for this data set from M_1, $\hat{Y}_2$ be the predictions from M_2, and $\hat{Y}_e$ be the predictions from M_e. Possible strategies for generating optimal predictions include the following:

- *Weighted Average of Predictions* $\bar{\hat{Y}}$ Determine whether a weighted average of the predictions from either $\hat{Y}_1$ or $\hat{Y}_2$ and $\hat{Y}_e$ yields better results than $\hat{Y}_e$ alone. A weighed average prediction $\bar{\hat{Y}}$ is defined as

$$\bar{\hat{Y}} = \alpha \hat{Y}_e + (1 - \alpha)\hat{Y_m} \quad m = 1, 2 \tag{3.1}$$

  where $0 \leq \alpha \leq 1$. α is determined by repeated training/test set evaluation.
- *Predictions from Composite Model* $\hat{Y}_c$ We consider whether building a model directly to $\{\mathbf{X}, \mathbf{Z}\}$ will lead to improved predictions. The modeling procedures described in Section 2 are repeated with $\mathbf{Z}$ as well as $\mathbf{X}$ considered as possible covariates. This leads to models M_{c1} (logic regression) and M_{c2} (logistic regression) whose predictions $\hat{Y_{c1}}$ and $\hat{Y_{c2}}$ can be compared to $\hat{Y}_e$.
- *Two-Stage Predictions* $\hat{Y}_s$ Assume a two-class classification problem, i.e., $Y \in \{-1, 1\}$. In Stage 1 we determine for which observations the predictions $\hat{Y}_e$ are correct ($n_c \in \{1, \ldots, n'\}$) or incorrect ($n_{\bar{c}} = \{1, \ldots, n'\}/n_c$). In Stage 2 for observations in $n_{\bar{c}}$ we replace the predictions from $\hat{Y}_e$ with the predictions from either M_1 or M_2. In other words, we create a two-stage model M_s for $Y_i, i = 1, \ldots, n'$, with predictions defined as

$$\hat{Y_{si}} = \begin{cases} \hat{Y_{ei}} & \text{if } Y_{si} = Y_i \\ \hat{Y_{mi}} & \text{if } Y_{si} \neq Y_i \end{cases} \tag{3.2}$$

  where $m = 1, 2$. This predictive scheme may be particularly useful for data from a heterogenous population, where it is possible that the accuracy of the

> predictions from a given model may vary across different subgroups of the population.

Unfortunately M_s, and hence $\hat{Y_s}$, depends on the true response Y. As an alternative we propose the use of a support vector machine (SVM) to discriminate those subspaces of the covariate space on which the results of M_e are correct from those on which the results are incorrect, based on the training data.

3.1. Support Vector Machines

Support vector machines (SVMs) [3, 8, 39] are a group of related supervised learning methods for classification or regression. In the case of two-class classification we consider a set of data points $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where each $y_i \in \{-1, 1\}$ denotes the class to which x_i belongs. The objective of an SVM is to produce a hyperplane which can separate the two classes using only $x_i, i = 1, \ldots, n$ in a way which minimizes the empirical classification error and maximizes the geometric margin between the classes.

More specifically, the (soft margin) support vector machine is the solution to the following optimization problem:

$$\begin{array}{rl} \min_{w,b,\xi} & \frac{1}{2}w^t w + C\sum_{i=1}^{l}\xi_i, \quad C > 0 \\ \text{subject to} & y_i(w^t\phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \end{array}$$

Note that the vectors $x_i, i = 1, \ldots, n$ are mapped to a higher dimensional space by the function ϕ, and the SVM finds a linear separating hyperplane in this higher dimensional space. This SVM has a 'soft margin' in the sense that is allows for misclassified samples; if no hyperplane exists which can separate the two classes, this method will chose the hyperplane which splits the classes as cleanly as possible while still maximizing the geometric margin. The slack variables ξ_i measure the degree of misclassification of the datum x_i. $K(x_i, x_j) = \phi(x_i)^t\phi(x_j)$ is called the kernel function of the SVM. The kernel function typically falls into one of four classes: linear, polynomial, radial basis function, and sigmoid. For more information on SVMs and their implementation we refer the reader to [5, 31].

As stated previously, our use of SVMs is to discriminate those subspaces of the covariate space on which the results of the existing model M_e are correct from those on which the results are incorrect, based on the training data. Let $\mathcal{X}$ be the covariate space and consider a SVM which divides $\mathcal{X}$ into subspaces $\mathcal{X}_c$ and $\mathcal{X}_{\bar{c}}$ where the model results are correct and incorrect, respectively. We now redefine the two-stage model M_s (and $\hat{Y_{si}}$), originally defined in Equation (3.2), independently of Y using the results of the SVMs:

$$\hat{Y_{si}} = \begin{cases} \hat{Y_{ei}} & \text{if } X_i \in \mathcal{X}_c \\ \hat{Y_{mi}} & \text{otherwise} \quad m = 1, 2 \end{cases} \tag{3.3}$$

3.2. Analysis Strategy Recap

Before we move to Section 4 we briefly summarize our analysis strategy. We want a statistical model which can accurately predict the outcome status of an observation given a set of existing predictors (both continuous and binary) and a set of new binary predictors. Our key modeling approach is a two-stage model. In the first stage we build a model from the existing predictors only (a logistic regression model).

Given the predictions from this model we design an SVM which can identify the subspaces in which observations are correctly or incorrectly predicted. In the second stage, in those subspaces where observations are incorrectly predicted, we use a model based only on the new binary predictors (a logistic regression or logic tree model) to generate accurate predictions. In this approach information from the new binary predictors is only utilized where needed, i.e., in subspaces where the existing predictors do not provide enough information to generate accurate outcome predictions.

4. Example: The CATHGEN Study

We demonstrate the use of our method in the analysis of data from a cardiology study. A substantial problem in clinical cardiology is the gap in the ability to detect asymptomatic individuals at high risk for coronary heart disease (CHD) for preventive and therapeutic interventions [17, 26]. Up to 75% of such individuals are designated as low to intermediate risk by standard CHD risk assessment models; however, a substantial number of such individuals who are actually at increased risk may not be identified. One analysis from the Framingham Heart Study found that for individuals that manifested a new CHD event, the initial presentation in over 50% of the cases was myocardial infarction, silent myocardial infarction or sudden cardiac death [1]. Over 50% of individuals with sudden cardiac death have no prior symptoms of CHD [40]. Therefore, it is likely that the traditional risk factors do not account fully for CHD risk [16, 22, 24, 28]. Furthermore, current CHD risk assessment models do not provide one's individual risk. Rather, the calculated assessment is for a population of individuals who share the same demographics and panel of risk factors.

A group of researchers at Duke University Medical Center (DUMC) has pursued an avenue of study evaluating the role of genes and gene variants in the development of atherosclerosis (the AGENDA study) [18, 19, 33]. As a result of their efforts they have compiled a list of candidate genes with a strong statistical correlation with vascular atherosclerosis. Through subsequent analysis for SNPs in these candidate genes, they analyzed 1300+ SNPs for association with significant CHD (stenosis $\geq$75% in at least one coronary artery) in a cohort of 1500 subjects who had undergone cardiac catheterization (CATHGEN). These SNPs were then ranked by their marginal association with the presence of CHD in a cardiac catheterization population.

We conduct an analysis of a subset of the CATHGEN data to test the hypothesis that genetic information in the form of SNPs will improve the ability of risk assessment models that use only traditional risk factors to classify individuals as having high risk for CHD. We developed prediction models for likelihood of significant CHD based on traditional risk factors such as cholesterol, blood pressure, diabetes and smoking, using a group of CATHGEN subjects who underwent cardiac catheterization. A separate set of CATHGEN subject data was used in selecting from the candidate SNP pool those SNPs with the highest marginal association with significant CHD; 81 such SNPs were available for analysis. We then assessed whether including genetic information improved our ability to classify individuals as having significant CHD.

The research was performed under an approved protocol from the Institutional Review Board of DUMC.

Table 1
CATHGEN Data

		Build Set		Evaluation Set	
		Training	Validation	Training	Validation
Male	Young Cases	44	80	69	103
	Controls	34	14	32	18
	Older Cases	79	13	47	11
Female	Young Cases	11	21	11	18
	Controls	59	12	42	18
	Older Cases	15	3	15	4

4.1. Data

Two data sets were constructed from the CATHGEN data, one for SNP selection and model building (build set) and one for the evaluation of model predictions (evaluation or eval set). The evaluation set consisted of white individuals (self-reported race) with complete data for all 81 SNPs and all clinical variables (see Section 4.2). The build set consisted of white individuals with complete data for all 81 SNPs but incomplete clinical data (clinical data was assumed to be missing at random). Within each set individuals were separated into three cohorts: a) controls, ≥ 65 years of age without significant CHD, b) older cases (OC), ≥ 65 years of age with significant CHD, and c) younger cases (YC), ≤ 50 years of age with significant CHD. For each cohort a group of samples was selected for model validation only, those $50 - 55$ years of age with either minimal or significant CHD as defined by coronary angiography (for validation of models of cohorts a) and c)) and those $56 - 65$ years of age with either minimal or significant CHD as defined by coronary angiography (for validation of models of cohorts a) and b)). Each cohort was further split by gender. A table of the study cohorts and number of subjects is shown in Table 1.

We used the 81 SNPs from the AGENDA study with the strongest statistical association with the presence of significant CHD. The strength of association was determined by 1) the p-value of SNP status in a logistic regression model of CHD, including both age and gender as covariates, and 2) the p-value of SNP status from a Cochran-Armitage Test for Trend [2]. We should note that the designation of the top SNPs was performed using a large group of subjects (1500) that included the data used for this study.

Typically with any given single base-pair difference, or single nucleotide polymorphism (SNP), only two out of the four possible nucleotides occur. Since each cell contains a pair of every autosome, we can think of a SNP as a three-level variable X taking the values 0, 1, or 2 (e.g., for nucleotide pairs A/A, A/G, and G/G, respectively). Each SNP can be recoded as a binary variable using either dominant coding ($X_d = 1$ if $X \geq 1$ and $X_d = 0$ otherwise) or recessive coding ($X_r = 1$ if $X = 2$ and $X_r = 0$ otherwise). With the CATHGEN data we chose dominant coding for the SNPs, where $X = 0$ indicates no copy of the minor (less frequently occurring) allele.

4.2. Model Building

Models were constructed on either male or female subjects. Within gender, these models compared either controls with young cases or controls with older cases. We describe the modeling approaches used for each gender/comparison combination. All computations were performed in R [27].

Predictive model using clinical variables(M_e)

For the clinical variables we used standard CHD risk factors as denoted by assessment tools such as the Framingham heart study risk algorithm [40]. We included presence of diabetes, current smoking status, total cholesterol level, HDL cholesterol level, systolic blood pressure and diastolic blood pressure. Clinical variables were collected at the time of cardiac catheterization. We used these variables to train both weighted and unweighted logistic regression models in the evaluation set, as the build set has incomplete clinical data. The weights were chosen to balance the importance of case and control samples. The trained model was then used to classify the validation subjects in the evaluation set as having minimal or significant CHD.

Predictive model using genetic variables(M_1,M_2)

Our SNP data consisted of the 81 SNPs from the AGENDA study typed in our CATHGEN samples, as described in Section 4.1. We constructed two models using only the CATHGEN samples from the build set: 1) LASSO for SNP selection followed by logistic regression (weighted and unweighted) using backwards selection, and 2) logic regression based on all 81 SNPs. Logic models were fit for both classification and logistic regression. These models were then used to classify the subjects in the evaluation set as having minimal or significant CHD. LASSO and logic regression were performed using the R packages `lasso2` and `LogicReg`, respectively.

Predictive model using combined clinical and genetic variables(M_c)

First, logistic regression models (weighted and unweighted) were built using genetic variables, as described above. The SNPs which appear in each model and the clinical variables were combined to train logistic regression models in the evaluation set. These models were then used to classify the validation subjects in the evaluation set as having minimal or significant CHD. A similar procedure was performed for each logic regression model.

Two-Stage Predictive model using the clinical and genetic models(M_s)

First, the trained clinical model was used to classify the subjects in the evaluation set as having minimal or significant CHD. Next an SVM was constructed which could discriminate the subspace of correctly classified samples from the subspace of incorrectly classified samples. For those samples in the subspace of incorrectly classified samples, the trained genetic models were applied to reclassify the subjects into the minimal and significant CHD groups. This resulted in a set of two-stage predictions, as described in Section 3.1 and Equation (3.3). SVMs were based on a radial basis function kernel and computed using the R package `e1071` [12, 23].

4.3. Results

Our interest is in classifying individuals as having non-significant CHD ($Y = 0$) or significant CHD ($Y = 1$). A fitted or predicted probability of significant CHD $\hat{Y}_i$ from a logistic regression model for a given individual was considered an ‘accurate’ classification if $Y_i = 1$ and $\hat{Y}_i \geq 0.5$, or $Y_i = 0$ and $\hat{Y}_i < 0.5$, and considered ‘inaccurate’ otherwise. A fitted or predicted outcome $\hat{Y}_i$ from a logic regression model for

Table 2

Results of Clinical Model and Logic Regression Classification Models on Evaluation Data.

	Training/Test Samples				Validation Samples			
Female, Controls vs Older Cases								
	$\hat{Y}_e$	$\hat{Y}_1$	$\hat{Y}_c$	$\hat{Y}_s$	$\hat{Y}_e$	$\hat{Y}_1$	$\hat{Y}_c$	$\hat{Y}_s$
auROC	71.38	48.81	75.38	**81.31**	50.48	57.62	53.33	**75.24**
acc	71.93	50.88	66.67	**82.46**	50.00	68.18	54.55	**81.82**
fn	36.00	68.00	40.00	**28.00**	57.14	71.43	**42.86**	**42.86**
fp	21.88	34.38	28.13	**9.38**	46.67	13.33	46.67	**6.67**
Male, Controls vs Older Cases								
	$\hat{Y}_e$	$\hat{Y}_1$	$\hat{Y}_c$	$\hat{Y}_s$	$\hat{Y}_e$	$\hat{Y}_1$	$\hat{Y}_c$	$\hat{Y}_s$
auROC	81.97	51.46	**82.97**	74.18	59.52	42.86	59.94	**60.00**
acc	78.48	73.42	78.48	**87.34**	51.72	41.38	83.47	**58.62**
fn	11.48	8.20	11.48	**1.64**	14.29	14.29	5.66	**0.00**
fp	55.56	88.89	55.56	**50.00**	**80.00**	100.0	93.33	**80.00**
Female, Controls vs Younger Cases								
	$\hat{Y}_e$	$\hat{Y}_1$	$\hat{Y}_c$	$\hat{Y}_s$	$\hat{Y}_e$	$\hat{Y}_1$	$\hat{Y}_c$	$\hat{Y}_s$
auROC	80.80	50.15	80.80	**86.53**	63.81	50.00	63.81	**85.71**
acc	79.25	56.60	79.25	**88.68**	61.11	47.22	61.11	**83.33**
fn	33.33	80.95	33.33	**23.81**	42.86	66.67	42.59	**28.57**
fp	12.50	18.75	12.50	**3.13**	33.33	33.33	33.33	**0.00**
Male, Controls vs Younger Cases								
	$\hat{Y}_e$	$\hat{Y}_1$	$\hat{Y}_c$	$\hat{Y}_s$	$\hat{Y}_e$	$\hat{Y}_1$	$\hat{Y}_c$	$\hat{Y}_s$
auROC	84.00	37.01	**86.81**	73.19	60.19	58.33	59.94	**88.11**
acc	83.17	46.53	83.17	**88.12**	81.82	52.07	83.47	**94.21**
fn	4.82	48.19	4.82	**3.61**	8.49	50.00	5.66	**3.77**
fp	72.22	77.78	72.22	**50.00**	86.67	33.33	93.33	**20.00**

a given individual was considered ‘accurate’ if $Y_i = \hat{Y}_i$ and considered ‘inaccurate’ otherwise. We considered overall model accuracy as well as the rate of false positive ($P(\hat{Y}_i \geq 0.5|Y_i = 0)$) and false negative ($P(\hat{Y}_i < 0.5|Y_i = 1)$) model results. In an attempt to balance specificity and sensitivity we also calculated the area under the receiver-operating characteristic (ROC) curve (auROC) for the results of each model. The auROC is equal to the value of the Wilcoxon-Mann-Whitney statistic and can be interpreted as the probability that the model will assign a higher probability of significant CHD to a randomly selected positive sample than to a randomly selected negative sample. The auROC calculations were performed with the R package `ROCR` [34].

The results of the weighted logistic regression models and the logic regression classification and logistic models for each gender and comparison (control vs. young cases or control vs. older cases) on the evaluation set are presented in the Tables 4, 2, and 3. The results of the unweighted logistic regression models are not discussed here due to their similarity to the results of the weighted models. In these Tables $\hat{Y}_e$ = clinical only model, $\hat{Y}_1$ or $\hat{Y}_2$ = SNP only model, $\hat{Y}_c$ = Clinical+SNP model, $\hat{Y}_s$ = Two-Stage Predictions using SVM, acc = accuracy, fn = false negative rate, and fp = false positive rate.

These tables show clearly that the two-stage predictions yield the best results. In some cases the combined clinical+SNP models perform better on the training set in comparison to the two-stage predictions, but their performance deteriorates on the validation samples. This could be due to the fact that the clinical model and the clinical+SNP model were trained on the training/test samples and tested on the validation samples, while the SNP models were tested on both sets of samples (having already been trained on the samples in the build set). This would also explain the consistency of the SNP model results across both the training/test and validation samples.

Interestingly the combined clinical+SNP models did not perform better than

TABLE 3

Results of Clinical Model and Logic Regression Logistic Models on Evaluation Data.

	Training/Test Samples				Validation Samples			
Female, Controls vs Older Cases								
	$\hat{Y}_e$	$\hat{Y}_1$	$\hat{Y}_c$	$\hat{Y}_s$	$\hat{Y}_e$	$\hat{Y}_1$	$\hat{Y}_c$	$\hat{Y}_s$
auROC	71.38	52.50	72.50	**84.63**	50.48	59.05	48.57	**79.53**
acc	71.93	49.12	63.16	**84.21**	50.00	54.55	31.82	**72.73**
fn	36.00	20.00	44.00	**12.00**	57.14	28.57	42.86	**14.29**
fp	21.88	75.00	31.25	**18.75**	46.67	53.33	80.00	**26.67**
Male, Controls vs Older Cases								
	$\hat{Y}_e$	$\hat{Y}_1$	$\hat{Y}_c$	$\hat{Y}_s$	$\hat{Y}_e$	$\hat{Y}_1$	$\hat{Y}_c$	$\hat{Y}_s$
auROC	81.97	59.38	82.97	**88.39**	59.52	43.81	63.33	**82.86**
acc	78.48	49.36	82.28	**91.14**	51.72	44.83	51.72	**82.76**
fn	11.48	59.02	**4.92**	6.56	**14.29**	85.71	**14.29**	**14.29**
fp	55.56	22.22	61.11	**16.67**	80.00	26.67	80.00	**20.00**
Female, Controls vs Younger Cases								
	$\hat{Y}_e$	$\hat{Y}_1$	$\hat{Y}_c$	$\hat{Y}_s$	$\hat{Y}_e$	$\hat{Y}_1$	$\hat{Y}_c$	$\hat{Y}_s$
auROC	80.80	52.60	**81.40**	80.21	63.81	50.00	64.13	**80.95**
acc	79.25	56.60	70.70	**83.02**	61.11	47.22	55.56	**77.78**
fn	**33.33**	66.67	47.62	**33.33**	42.86	66.67	47.62	**38.10**
fp	12.50	28.13	15.63	**6.25**	33.33	33.33	40.00	**0.00**
Male, Controls vs Younger Cases								
	$\hat{Y}_e$	$\hat{Y}_1$	$\hat{Y}_c$	$\hat{Y}_s$	$\hat{Y}_e$	$\hat{Y}_1$	$\hat{Y}_c$	$\hat{Y}_s$
auROC	84.00	47.52	**91.43**	77.78	60.19	57.83	61.32	**80.97**
acc	83.17	49.50	81.13	**92.08**	81.82	57.20	78.51	**91.74**
fn	4.82	49.40	4.82	**0.00**	8.49	44.34	12.26	**4.72**
fp	72.22	55.56	50.00	**44.44**	86.67	40.00	86.67	**33.33**

TABLE 4

Results of Clinical Model and Weighted Logistic Regression Models on Evaluation Data.

	Training/Test Samples				Validation Samples			
Female, Controls vs Older Cases								
	$\hat{Y}_e$	$\hat{Y}_2$	$\hat{Y}_c$	$\hat{Y}_S$	$\hat{Y}_e$	$\hat{Y}_2$	$\hat{Y}_c$	$\hat{Y}_s$
auROC	71.38	54.56	77.38	**84.00**	50.48	60.48	49.52	**75.24**
acc	71.93	56.14	70.18	**85.96**	50.00	72.73	50.00	**81.82**
fn	36.00	64.00	**32.00**	**32.00**	57.14	71.43	**42.86**	**42.86**
fp	21.88	28.13	28.13	**0.00**	46.67	**6.67**	53.33	**6.67**
Male, Controls vs Older Cases								
	$\hat{Y}_e$	$\hat{Y}_2$	$\hat{Y}_c$	$\hat{Y}_s$	$\hat{Y}_e$	$\hat{Y}_2$	$\hat{Y}_c$	$\hat{Y}_s$
auROC	81.97	47.27	**93.62**	77.78	59.52	45.00	66.19	**73.33**
acc	78.48	65.82	88.61	**89.87**	51.72	51.72	48.28	**72.41**
fn	11.48	22.95	6.56	**0.00**	14.29	35.71	21.43	**0.00**
fp	55.56	72.22	**27.78**	44.44	80.00	60.00	80.00	**53.33**
Female, Controls vs Younger Cases								
	$\hat{Y}_e$	$\hat{Y}_2$	$\hat{Y}_c$	$\hat{Y}_s$	$\hat{Y}_e$	$\hat{Y}_2$	$\hat{Y}_c$	$\hat{Y}_s$
auROC	80.80	45.47	**83.33**	81.77	63.81	52.38	49.52	**78.57**
acc	79.25	56.60	75.47	**84.91**	61.11	38.89	50.00	**75.00**
fn	**33.33**	100.0	**33.33**	**33.33**	**42.86**	95.24	**42.86**	**42.86**
fp	12.50	6.25	18.75	**3.13**	33.33	13.33	53.33	**0.00**
Male, Controls vs Younger Cases								
	$\hat{Y}_e$	$\hat{Y}_2$	$\hat{Y}_c$	$\hat{Y}_s$	$\hat{Y}_e$	$\hat{Y}_2$	$\hat{Y}_c$	$\hat{Y}_s$
auROC	84.00	51.94	**95.79**	82.13	60.19	62.52	66.19	**91.45**
acc	83.17	47.62	**93.07**	92.08	81.82	50.41	48.28	**95.04**
fn	4.82	54.22	**2.41**	**2.41**	8.49	52.83	21.43	**3.77**
fp	72.22	44.44	**27.78**	33.33	86.67	26.67	80.00	**13.33**

the clinical only or SNP only models on the validation samples. In many comparisons the clinical and clinical+SNP models performed comparably, with the SNP models performing quite poorly. We surmise that the population under study is quite heterogeneous, and that no one data type provides information predictive of CHD in all subpopulations. The clinical data is predictive for some samples while the genetic data is predictive for others. The results of a weighted average of the predictions from the clinical only and SNP only models ($\bar{\hat{Y}}$; results not shown) were not successful because both data types are not relevant for all samples; often the data types provide conflicting information. The two-stage predictions are an attempt to use an SVM to define subpopulations for which the clinical data or the genetic data are predictive. Using the SVM results we can identify which data type is predictive for a given sample, leading to more accurate predictions overall.

In Figure 2 we display the quality of the two-stage predictions on the evaluation set (train/test subset or validation subset) for each model and each comparison. No single model performs consistently best in all comparisons. By averaging the performance measures (auROC or accuracy) on the validation samples across comparisons we find in terms of auROC the logic regression logistic models perform 1.43% better than the weighted logistic regression models, which in turn perform 2.38% better than the logic regression classification models. In terms of accuracy the logic regression logistic models perform 0.19% better than the weighted logistic regression models, which in turn perform 1.57% better than the logic regression classification models. Hence we conclude that overall the logic regression logistic models perform best, followed by the weighted logistic regression models and finally the logic regression classification models. However, the difference in average performance between any two model types is quite small.

The only comparison in which model performances are clearly distinguished is the male controls vs. older cases comparison. The relatively poor performance of the two-stage predictions from the logic regression classification models is striking; the results in Table 2 reveal that the logic regression classification model had false positive rates of 89% for the training/test patients and 100% for the validation patients. Unfortunately the clinical model also had a high false positive rate of 80%. Both data types (and consequently the two-stage predictions) failed to predict those with minimal CHD. This is possibly a result of SNP selection; of the 9, 8, and 8 SNPs selected by the weighted logistic, logic regression classification and logistic models, respectively, only 2 appear in all three models and no other SNPs are shared by any two models. No definitive conclusions can be drawn without an independent data set on which to validate our results.

5. Discussion

We have presented a two-stage approach to generating combined predictions from models built from different data sources. One model is built on existing data of multiple types (e.g., traditional clinical risk factors), while a second set of models are built on newly available binary predictors only (in our case genetic SNP data). This two-stage approach uses an SVM to distinguish the covariate subspaces on which the existing data model generates accurate or inaccurate predictions. The existing model is used to generate predictions for samples in the 'accurate' subspace while a model built on the newly available data is used to generate predictions for samples in the 'inaccurate' subspace. This approach appears to perform well in generating predictions for a heterogeneous population for which no single data type provides

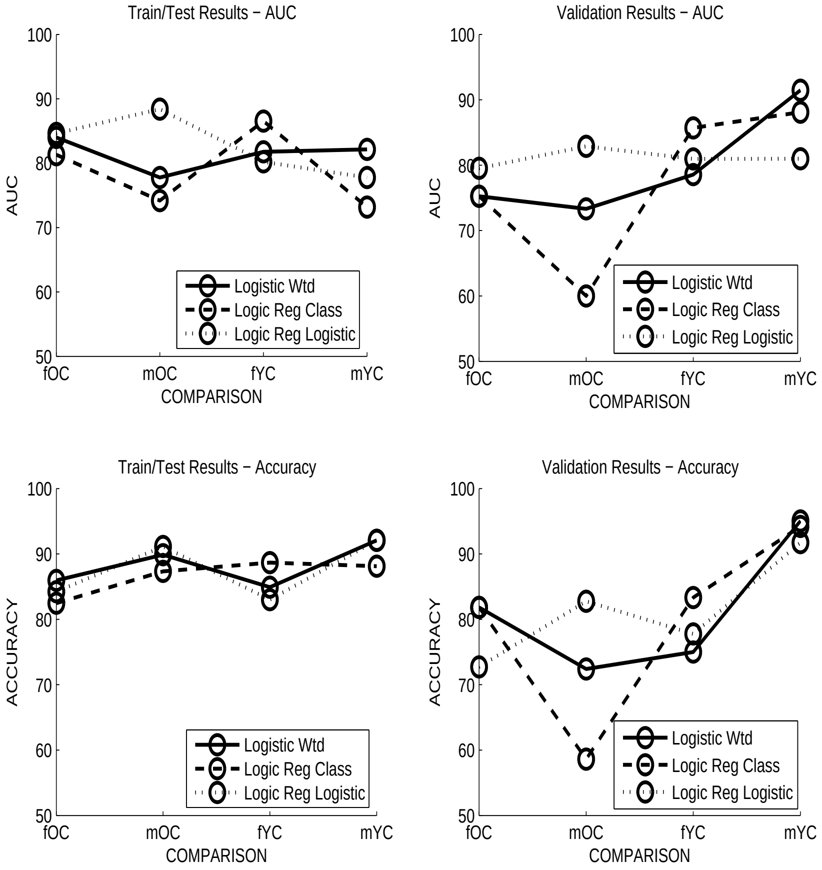


FIG 2. *Quality of Two-Stage Predictions on Evaluation Data*

predictive information for all samples.

As discussed briefly in Section 1 there exist modeling approaches other than logistic and logic regression models which could have been employed here. We chose logic trees because of their ability to capture higher order interactions, an issue of great importance in regression and a key to variable selection. However, similar models could be constructed by Bayesian model averaging with lower-dimensional logistic regression models that allow for interactions among covariates. We also could have employed neural networks or projection pursuit models. These alternative approaches would require careful prior variable selection in any context where $n < p$, but would be worth considering in future work.

Our modeling approach is similar in spirit to ensemble methods [11], learning algorithms which construct a set of classifiers and then generate predictions by taking a (weighted) average or vote of their predictions. One such approach is boosting [13, 14, 32], a method for converting a weak learning algorithm into one with high accuracy. This is done by training classifiers on weighted versions of the training data, giving higher weight to misclassified samples, and forming the final classifier as a linear combination of the training classifiers. This approach does not apply different models to different covariate subspaces, but does attempt to improve model performance in subspaces where the model performs poorly. Our approach is a type of ensemble method in which each classifier gets either a single, fully weighted vote or no vote depending upon the subspace in which the sample of interest is located. It would be of interest to compare our two-stage predictive approach to an approach aimed at building a boosted classifier from all available covariates. The results of such a comparison would help in determining the necessity of building a subspace-dependent classifier.

Several different model types were used in generating predictions from the newly available binary data, including logistic regression and logic regression models. No single model type performed significantly better than the others, although a slight performance advantage was observed when using the two-stage predictions from the logic regression logistic models. Across comparisons within gender and case the best models generated two-stage predictions with validation accuracies between 81.82% and 94.21%. It should be noted, however, that the sizes of the validation sets for some comparisons are quite small and all comparisons were conducted within a single population (CATHGEN). Also, our inferences are done conditional on a fixed chosen model; the variability of the models is not considered in the inference procedure. This is a weakness in our approach as model uncertainty can be substantial in high dimensional data contexts. Hence we regard our results as a ‘proof-of-concept’ for our analysis approach. We are planning an analysis of a second, independent population and await the results of such an analysis before making any definitive conclusions regarding the predictive power of our method.

Acknowledgements

This research was supported by NIH Grant 5K25CA111636-03. The authors wish to thank the following for their assistance: Bertrand Clarke, Department of Statistics, University of British Columbia; Ed Iversen, Department of Statistical Science, Duke University; Pascal Goldschmidt, Dean, Leonard M. Miller School of Medicine, University of Miami.

References

[1] American Heart Association (2006). *Heart Disease and Stroke Statistics - 2006 Update.* 2–10.

[2] Armitrage, P. (1955). Tests for Linear Trends in Proportions and Frequencies. *Biometrics* **11**, 375–386.

[3] Boser, B., Guyon, I. and Vapnik, V. (1992). A Training Algorithm for Optimal Margin Classifiers. In Haussler, D. (Ed.) *5th Annual ACM Workshop on COLT*. ACM Press, 141–152.

[4] Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees.* Wadsworth Press, Belmont, CA.

[5] Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM - A Library for Support Vector Machines.* Software available at URL: http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[6] Chipman, H., George, E, and McCullough, R. (2002). Bayesian Treed Models. *Machine Learning* **48**, 299–320.

[7] Clyde, M. (1999). Bayesian Model Averaging and Model Search Strategies. In Bernardo, J., Berger, J., Dawid, A. and Smith, A. (Eds.) *Bayesian Statistics 6.* Oxford University Press, Oxford, UK. 157–185.

[8] Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. *Machine Learning* **20**, 273–297.

[9] Devlin, B., Bacanu, S.-A. and Roeder, K. (2004). Genomic Control to the extreme. *Nature Genetics* **36**, 1129–1130.

[10] Devlin, B. and Roeder, K. (1999). Genomic Control for Association Studies. *Biometrics* **55**, 997–1004.

[11] Dietterich, T. (2000). Ensemble Methods in Machine Learning. *Lecture Notes in Computer Science* **1857**, 1–15. URL: citeseer.ist.psu.edu/dietterich00ensemble.html.

[12] Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. and Weingessel, A. (2006). The e1071 Package: Miscellaneous Functions of the Department of Statistics (e1071), Technische Universität Wien, Austria. Version 1.5-16.

[13] Freund, Y. (1995). Boosting a Weak Learning Algorithm by Majority. *Information and Computation* **121**, 256–285.

[14] Freund, Y. and Schapire, R. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences* **55**, 119–139.

[15] Friedman, J (1991). Multivariate Adaptive Regression Splines (with Discussion). *Annals of Statistics* **19**, 1–141.

[16] Greenland, P., Knoll, M., Stamler, J., Neaton, J., Dyer, A., Garside, D. and Wilson, P. (2003). Major Risk Factors as Antecedents of Fatal and Nonfatal Coronary Heart Disease Events. *Journal of the American Medical Association* **290**, 891–897.

[17] Greenland, P., Smith, S. and Grundy, S. (2001). Improving Coronary Heart Disease Risk Assessment in Asymptomatic People: Role of Traditional Risk Factors and Noninvasive Cardiovascular Tests. *Circulation* **104**, 1863–1867.

[18] Hauser, E., Crossman, D., Granger, C., Haines, J., Jones, C., Mooser, V., McAdam, B., Winkelmann, B., Wiseman, A., Muhlstein, J., Bartel, A., Dennis, C., Dowdy, E., Estabrooks, S., Eggleston, K., Francis, S., Roche, K., Clevenger, P., Huang, L., Pedersen, B., Shah, S., Schmidt, S., Haynes, C., West, S., Asper, D., Booze, M., Sharma, S., Sundseth, S., Middleton, L., Roses, A., Hauser, M., Vance, J.,

PERICAK-VANCE, M. AND KRAUS, W. (2004). A Genomewide Scan for Early-Onset Coronary Artery Disease in 438 Families: The GENECARD Study. *American Journal of Human Genetics* **75**, 436–447.

[19] KARRA, R., VERMULLAPALLI, S., DONG, C., HERDERICK, E., SONG, X., SLOSEK, K., NEVINS, J., WEST, M., GOLDSCHMIDT-CLERMONT, P. AND SEO, D. (2005). Molecular Evidence for Arterial Repair in Atherosclerosis. *Proceedings of the National Academy of Sciences U.S.A.* **102**, 16789–16794.

[20] KOOPERBERG, C., RUCZINSKI, I., LEBLANC, M. AND HSU, L (2001). Sequence Analysis using Logic Regression. *Genetic Epidemiology* **21**, S626–S631.

[21] LOKHORST, J., VENABLES, B., TURLACH, B. AND MAECHLER, M. (2006). The lasso2 Package: L1 Constrained Estimation aka 'lasso'. University of Western Australia School of Mathematics and Statistics. Version 1.2-5. URL: http://www.maths.uwa.edu.au/ berwin/software/lasso.html.

[22] MAGNUS, P. AND BEAGLEHOLE, R. (2001). The Real Contribution of the Major Risk Factors to the Coronary Epidemics: Time to End the "Only-50% " Myth. *Archives of Internal Medicine* **161**, 2657–2660.

[23] MEYER, D. (2006). Support Vector Machines: The Interface to libsvm in package e1071, Technische Universität Wien, Austria.

[24] MOSCA, L. (2002). C-Reactive Protein: To Screen or Not to Screen? *New England Journal of Medicine* **347**, 1615–1617.

[25] OSBORNE, M., PRESNELL, B. AND TURLACH, B. (2000). On the LASSO and its dual. *Journal of Computational and Graphical Statistics* **9**, 319–337.

[26] PASTERNAK, R., ABRAMS, J., GREENLAND, P., SMAHA, L., WILSON, P. AND HOUSTON-MILLER, N. (2003). Task Force #1 - Identification of Coronary Heart Disease Risk - Is There a Detection Gap? *Journal of the American College of Cardiology* **41**, 1863–1874.

[27] R DEVELOPMENT CORE TEAM (2006). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: http://www.R-project.org.

[28] RIDKER, P., RIFAI, N., ROSE, L., BURING, J. AND COOK, N. (2002). Comparison of C-Reactive Protein and Low-Density Lipoprotein Cholesterol Levels in the Prediction of First Cardiovascular Events. *New England Journal of Medicine* **347**, 1557–1565.

[29] RUCZINSKI, I., KOOPERBERG, C. AND LEBLANC, M. (2002). Logic Regression - methods and software. In DENISON, D., HANSEN, M., HOLMES, B., MALLICK, B. AND YU, B. (Eds.), *Proceedings of the MSRI workshop on Nonlinear Estimation and Classification.* Springer Verlag, New York. 333–344.

[30] RUCZINSKI, I., KOOPERBERG, C. AND LEBLANC, M. (2003). Logic Regression. *Journal of Computational and Graphical Statistics* **12**, 475–511.

[31] SCHÖLKOPF, B. AND SMOLA, A. (2002). *Learning with Kernels.* MIT Press, Cambridge, MA.

[32] SCHAPIRE, R. (1990). The Strength of Weak Learnability. *Machine Learning* **5**, 197–227.

[33] SEO, D., WANG, T., DRESSMAN, H., HERGERICK, E., IVERSEN, E., DONG, C., VATA, K., MILANO, C., RIGAT, F., PITTMAN, J., NEVINS, J., WEST, M. AND GOLDSCHMIDT-CLERMONT, P. (2004). Gene Expression Phenotypes of Atherosclerosis. *Atherosclerosis, Thrombosis, and Vascular Biology* **24**, 1922–1927.

[34] SING, R., SANDER, O., BEERENWINKEL, N. AND LENGAUER, T. (2005). ROCR: Visualizing Classifier Performance in R. *Bioinformatics* **21**, 3940–3941. URL: http://rocr.bioinf.mpi-sb.mpg.de/.

[35] SUTTON, C. (1991). Improving Classification Trees with Simulated Annealing. In Kazimadas, E. (Ed.), *Proceedings of the 23rd Symposium on the Interface.* Interface Foundation of North America, 333–344.

[36] TIBSHIRANI, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

[37] TZENG, J-Y., BYERLEY, W., DEVLIN, B., ROEDER, K. AND WASSERMAN, L. (2003). Outlier Detection and False Discovery Rates for Whole-Genome DNA Matching. *Journal of the American Statistical Association* **98**, 236–246.

[38] VAN LAARHOVEN, P. AND AARTS, E. (1987). *Simulated Annealing: Theory and Applications.* Kluwer Academic Publishers, Norwell, MA.

[39] VAPNIK, V. (1995). *The Nature of Statistical Learning Theory.* Springer-Verlag, New York.

[40] WILSON, P., D'AGOSTINO, R., LEVY, D., BELANGER, A., SILBERSHATZ, H. AND KANNEL, W. (1998). Prediction of Coronary Heart Disease using Risk Factor Categories. *Circulation* **97**, 1837–1847.

[41] XU, H., GREGORY, S., HAUSER, E., STENGER, J., PERICAK-VANCE, M., VANCE, J., ZUCHNER, S. AND HAUSER, M. (2005). SNPselector: a Web Tool for Selecting SNPs for Genetic Association Studies. *Bioinformatics* **21**, 4181–4186. URL: http://primer.duhs.duke.edu/.

IMS Lecture Notes–Monograph Series
Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh
Vol. 0 (0000) 320
© Institute of Mathematical Statistics, 0000
arXiv: math.PR/0000000

Sharp Failure Rates for the Bootstrap Particle Filter in High Dimensions

Peter Bickel[1], Bo Li[2] and Thomas Bengtsson[3]

University of California-Berkeley, Tsinghua University, and Bell Labs

Abstract: We prove that the maximum of the sample importance weights in a high-dimensional Gaussian particle filter converges to unity unless the ensemble size grows exponentially in the system dimension. Our work is motivated by and parallels the derivations of Bengtsson, Bickel, and Li (2007); however, we weaken their assumptions on the eigenvalues of the covariance matrix of the prior distribution and establish rigorously their strong conjecture on when weight collapse occurs. Specifically, we remove the assumption that the nonzero eigenvalues are bounded away from zero, which, although the dimension of the involved vectors grow to infinity, essentially permits the effective system dimension to be bounded. Moreover, with some restrictions on the rate of growth of the maximum eigenvalue, we relax their assumption that the eigenvalues are bounded from above, allowing the system to be dominated by a single mode.

Contents

1. Introduction

Bayesian filtering methods are a commonly employed tool for combing physical models and data. The filters treat the unknown system state as a random variable and resolve its probability density conditional on the data (and the system dynamics) through Monte Carlo sampling techniques. When applied sequentially in time, these methods are commonly referred to as particle filters ([8], [10]). For a diverse collection of applications and an excellent introduction to the field in general, see

[1] Peter Bickel, 367 Evans Hall, Department of Statistics, University of California, Berkeley, 94710-3860, CA USA. E-mail: bickel@stat.berkeley.edu

[2] Bo Li, 440 Weilun Hall, School of Economics and Management, Tsinghua University, Beijing, 100084, China. E-mail: libo@sem.tsinghua.edu.cn

[3] Thomas Bengtsson, Bell Labs, 600 Mountain Avenue, Murray Hill, 07904 NJ USA, E-mail: tocke@research.bell-labs.com

Keywords and phrases: curse of dimensionality, Monte Carlo, importance sampling, ensemble forecast, numerical weather prediction, Bayesian filter, large deviations, ensemble methods, sample size requirements, state-space model

the edited volume by Doucet [6]. The particle filter method relies heavily on a likelihood based reweighting mechanism of the involved sample draws. This reweighting scheme produces the so called importance weights, and these weights are the primary focus of our work. Specifically, in a Gaussian filter context, we examine the behavior of the importance weights as a function of the system dimension and of sample size.

The popularity of the particle filter is no doubt due to the flexibility of the model framework to handle both non-linear and non-gaussian structures. However, in spite of its generality, the method is not without flaws: the particle filter is known to require large Monte Carlo ensembles and frequent resampling to estimate the desired densities (c.f., [9]). This drawback is particularly prevalent in higher dimensional systems where the filter becomes unstable and quickly collapses onto a single point mass. In recent work, for a single Bayes update step in a Gaussian setting, Bengtsson, Bickel, and Li [3] give a derivation of the weight collapse as a function of the system dimension and of sample size. To shed further light on the weight collapse, this paper establishes conjectures (given in [3]) which make their arguments fully rigorous. Just as significantly, we exhibit that collapse is a function of the *effective dimension* (defined in Section 3), rather than the absolute dimension. As in [3], our analysis is given in the context of a stylized Gaussian example, but we conjecture (and simulations show) that our results are informative for situations that depend on similarly defined reweighting schemes. The results imply that to avoid collapse, the sample size must grow super-exponentially in the effective dimension. We do not investigate refinements of particle filters methods, such as simulated tempering [4], although our discussion in Section 2.1 suggests that their approach is not a solution to avoid collapse in truly high-dimensional settings.

Our work is outlined as follows. The next section describes the particle filter, provides notation, and describes the use of the ensemble method for approximating posterior densities. The main developments are then presented in Section 3, where we give several results establishing the conditions under which the maximum sample weight in a Gaussian particle filter converges to unity. All technical results are proved in the Appendix. (We note that some material in Section 2.1 and Section 3 is given in [3], but is reproduced here for completeness.)

2. Model setting

2.1. The Particle filter

Let X_t represent the unknown system state at time t, Y_t be a noisy data measurement of X_t, and let $\mathbf{Y}^t$ represent all data up to and including time t. Based on the data $\mathbf{Y}^t$ and (some) knowledge of the time-evolution of the system state from X_{t-1} to X_t, we seek the posterior distribution $p(X_t|\mathbf{Y}^t)$. We assume we have available a random sample $\{X_{t,i}^f\}$ of size n from the prior distribution $p(X_t|\mathbf{Y}^{t-1})$. Associated with the prior sample is a set of weights $\{w_i^f\}$. We assume further that the likelihood density $p(Y_t|X_t)$ is computable for arbitrary X_t.

The particle filter seeks to recursively in time estimate the probability distribution of the unknown state X_t. At each time t, the probability distribution is represented by the sample ensemble $\{X_{t,i}^f, w_i^f\}$, and the distribution can be propagated forward one time-step by evolving each $X_{t,i}^f$ using the system dynamics. Once new data Y_t is available, Bayes theorem is used to adjust the weights based on how 'close' the associated sample points are to the data. The following schematic describes the particle filter:

$$p(X_t|\mathbf{Y}^{t-1}), Y_t \overset{\text{Bayes}}{\longrightarrow} p(X_t|\mathbf{Y}^t) \overset{G(\cdot)}{\longrightarrow} p(X_{t+1}|\mathbf{Y}^t), Y_{t+1} \overset{\text{Bayes}}{\longrightarrow} p(X_{t+1}|\mathbf{Y}^{t+1})$$

Here, at time t (on the left), Bayes theorem combines $p(X_t|\mathbf{Y}^{t-1})$ and Y_t to produce $p(X_t|\mathbf{Y}^t)$. The system dynamics, in the above represented by $G(\cdot)$ (middle), is used to propagate $p(X_t|\mathbf{Y}^t)$ one time step and this yields $p(X_{t+1}|\mathbf{Y}^t)$. Bayes theorem is then again employed to find the posterior $p(X_{t+1}|\mathbf{Y}^{t+1})$ (right).

In a particle filter, the above schematic is straightforwardly implemented (at least conceptually) using a random sample. We note first that the change-of-variables problem represented by the propagation of $p(X_t|\mathbf{Y}^t)$ can be solved by evaluating $G(\cdot)$ at each sample point. We will not discuss the implementation of the forecast step here; instead, our focus is on the Bayes update step. As mentioned, the particle filter implements the Bayes step by reweighting the prior sample according to the likelihood. We note in passing that the particle filter may be derived as a (sequential) importance sampler (e.g., [2]) where the proposal distribution is given by the prior and the target distribution is given by the posterior. In the schematic below, which describes a bootstrap-likelihood filter, the prior sample is 'converted' to a posterior sample by resampling (with replacement) each member $X^f_{t,i}$ with probability proportional to $w^f_i \times p(Y_t|X^f_{t,i})$, i.e.,

$$\overbrace{\{X^f_{t,1},\ldots,X^f_{t,n}\}}^{prior\ ensemble},\ Y_t \quad \overset{resample}{\longrightarrow} \quad \overbrace{\{X^u_{t,1},\ldots,X^u_{t,n}\}}^{posterior\ ensemble}.$$

Although the particle filter has been successfully applied to a variety settings, it often produces highly varying importance weights. Remedies to stabilize the filter include resampling (renormalizing) the involved empirical measure at regular time intervals [8, 9] and marginalizing or restricting the sample space by conditioning on a larger information set [10, 11]. Another approach is given by simulated tempering [4], which makes use of the regularized likelihood $p(Y_t|X^f_{t,i})^\alpha$, where $0 < \alpha < 1$. However, as can be seen from our derivations, e.g. Proposition 3.1, a fixed α does not alter the conclusion of collapse. Moreover, for each time point, to obtain samples from the target density, simulated tempering generates a sequence of ensembles from kernels $K_i(\cdot)$, $(i = 1,\ldots,I)$ such that $K_I(\cdot)$ approaches the desired kernel $K(\cdot)$ associated with the posterior density. Unfortunately, for truly high dimensional systems, we conjecture that the number of intermediate sampling steps I would be prohibitively large and render it practically unfeasible. Thus, such remedies do not fundamentally address performance when the filter is applied to *very large scale* systems. For example, as noted by ([1], [13]), when applied in high dimensions, the filter collapses to a point mass after a few (or even one!) observation cycles. In particular, as will be shown in Section 3, it is the normalized quantity $w_i = p(Y_t|X^f_{t,i})/\sum_j p(Y_t|X^f_{t,j})$ that behaves singularly.

The next section sets up the necessary notation and formalizes our problem.

2.2. Monte Carlo Scheme

We formalize our problem as follows. Consider a set of n sample points $\mathbf{X} = \{X_1,\ldots,X_n\}$, where $X_i \in \Re^d$ and both the sample size n and system dimension d are 'large'. (To lighten notation, we have dropped the time subscript and the forecast superscript.) We assume that the sample $\mathbf{X}$ is drawn randomly from the prior (or proposal) distribution $p(X)$. New data Y is related to the state X by the conditional density $p(Y|X)$. For concreteness, a functional relationship $Y = f(X) + \varepsilon$ is assumed, and ε is taken to be independent of the state X. The goal is to estimate posterior expectations using the importance ratio, i.e., for some function $h(\cdot)$, we want to estimate

$$E(h(X)|Y) = \int h(X) \frac{p(Y|X)p(X)}{\int p(Y|X)p(X)\mathrm{d}X} \mathrm{d}X,$$

and use

$$\hat{E}(h(X)|Y) = \sum_{i=1}^{n} h(X_i) \frac{p(Y|X_i)}{\sum_{j=1}^{n} p(Y|X_j)}$$

as an estimator. Based on this formulation, the weights attached to each ensemble member

$$w_i = \frac{p(Y|X_i)}{\sum_{j=1}^{n} p(Y|X_j)} \tag{1}$$

are the primary objects of our study. As mentioned, in large scale analyzes, the weights in (1) are highly variable and often produce estimates $\hat{E}(\cdot)$ which are collapsed onto a point mass with $max(w_i) \approx 1$. As illuminated in [3], this degeneracy is pervasive for high-dimensional systems, and appears to hold for a variety of prior and likelihood distributions.

We next consider the case when both the prior and the likelihood distributions are Gaussian.

3. Gaussian Case

We assume a data model given by $Y = HX + \varepsilon$, where Y is a $d \times 1$ vector, H is a known $d \times q$ matrix, and X is a $q \times 1$ vector. Both the proposal distribution and the error distribution are Gaussian with $p(X) = N(\mu_X, \Sigma_X)$ and $p(\varepsilon) = N(0, \Sigma_\varepsilon)$, and the noise ε is taken independent of the state X. Since the data model can be pre-rotated by $\Sigma_\varepsilon^{-1/2}$, we set $\Sigma_\varepsilon = I_d$ without loss of generality (wlog). Moreover, since $EY = EHX$, we can replace X_i by $(X_i - EX_i)$ and Y by $(Y - EY)$ and leave $p(Y|X)$ unchanged. Hence, wlog we also set $\mu_X = 0$. Further, define, for conformable A and B, the inner product $\langle A, B\rangle = A^T B$ (where the superscript ^{T} denotes matrix transpose), and let $\|A\|^2 = \langle A, A\rangle$.

With $p(Y|X) \sim N(HX, I_d)$, the weights in (1) can be expressed as

$$w_i = \frac{\exp\big(-\|Y - HX_i\|^2/2\big)}{\sum_{j=1}^{n} \exp\big(-\|Y - HX_j\|^2/2\big)}. \tag{2}$$

To establish weight collapse for high-dimensional Gaussian $p(Y|X)$ and $p(X)$, we first write the exponent in (2) in terms of the singular values of $cov(HX)$.

Let $d' = rank(H)$. With $\lambda_1^2, \ldots, \lambda_{d'}^2$ the singular values of $cov(HX)$, define the $d' \times d'$ matrix $D = diag(\lambda_1, \ldots, \lambda_{d'})$. Then, with Q an orthogonal matrix obtained by the singular value decomposition of $cov(HX)$, define the $d' \times 1$ vector V such that

$$Q^T HX = DV.$$

Note that V_i corresponding to X_i is $N(0, I_{d'})$. Since Q is orthogonal, we can write

$$\|Y - HX_i\|^2 = \|Q^T Y - DV_i\|^2 = \sum_{j=1}^{d'} \lambda_j^2 W_{ij}^2 + \sum_{j=d'+1}^{d} \epsilon_{0j}^2, \tag{3}$$

where, conditional on Y, $[W_{i1}, \ldots, W_{id'}]^T$ is $N(\xi, I_{d'})$, and where ϵ_{0j} is the j:th component of the observation noise vector ε. The mean vector $\xi = [\mu_1, \ldots, \mu_{d'}]^T$ is given by

$$\xi = D^{-1} Q^T Y = V + D^{-1}\varepsilon', \tag{4}$$

where V and ε' are independent $N(0, I_{d'})$.

Now, for $i = 1, \ldots, n$, define

$$S_i = \frac{\sum_{j=1}^{d'} \lambda_j^2 (W_{ij}^2 - (1 + \mu_j^2))}{\left(2 \sum_{j=1}^{d'} \lambda_j^4 (1 + 2\mu_j^2)\right)^{1/2}}. \tag{5}$$

Note that the second term in (3) is constant for every X_i, and will not influence the weight w_i.

By (2), we can express the maximum weight as

$$w_{(n)} = \frac{1}{1 + T_{n,d'}}, \tag{6}$$

where $T_{n,d'} = \sum_{\ell=2}^{n} e^{-\sigma_{d'}\sqrt{d'}(S_{(\ell)} - S_{(1)})}$ with $\sigma_{d'}^2 = \frac{2}{d'} \sum_{j=1}^{d'} \lambda_j^4 (1 + 2\mu_j^2)$. Thus, to prove weight collapse, we need to show convergence of the denominator in (6) to unity. We now state the following.

Proposition 3.1. *Let $S_i, i = 1, \ldots, n$, be independent random variables with cumulative distribution function (cdf) $G_d(\cdot)$ satisfying the conditions specified in Lemma 3.4 and Lemma 3.5 stated in the Appendix. Let $S_{(1)} \le \cdots \le S_{(n)}$ be the ordered sequence of $S_1, \ldots, S_n$, and define, for some $\sigma > 0$, $T_{n,d} = \sum_{\ell=2}^{n} e^{-\sigma\sqrt{d}(S_{(\ell)} - S_{(1)})}$. Then, as, $n, d \to \infty$, if $\frac{\log n \log d}{d} \to 0$, we have*

$$\sqrt{\frac{\sigma^2 d}{2 \log n}} E(T_{n,d}) \to 1.$$

A proof of the result is provided in the Appendix. For the Gaussian case considered here, an immediate implication of Proposition 3.1 is weight collapse. Specifically, with two additional assumptions, we may assert the following.

Proposition 3.2. *We assume, for the Gaussian case considered here,*

A1: *There is a positive constant δ such that $\frac{1}{\delta} \ge \lambda_1, \cdots, \lambda_{d'} > \delta$; and*

A2: $\tau_{d'}^2 = \frac{2}{d'} \sum_{j=1}^{d'} (3\lambda_j^4 + 2\lambda_j^2) \to \sigma^2 > 0$.

Then, if $\frac{\log n \log d'}{d'} \to 0$, we have $w_{(n)} \xrightarrow{P} 1$.

Proposition 3.2 follows by Lemma 3.6 (Appendix) and Proposition 3.1.

The above result implies that, unless n grows super-exponentially in d', we have weight collapse. We note that Proposition 3.2 is a sharpening of the convergence rate as compared to that implied by Section 3.1 of [3]. The $\log d'$ term appears only because $\max|\mu_j| = O_p(\sqrt{\log d'})$, and we need to make our analysis conditional on the $\{\mu_j\}$.

The results in Proposition 3.2 suggest that large d' leads to collapse. However, we argue now that what really matters is the *effective dimension* of X, defined as the sum of the singular values of $cov(HX)$. We shall assume that

B : $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{d'} \geq \cdots$ are part of an infinite sequence.

Our arguments can be modified to the case where $\{\lambda_j : 1 \leq j \leq d'\}$ is a double array, but we eschew this complication.

There are two possibilities,

$$\text{(i)}\ \sum_{j=1}^{\infty} \lambda_j^2 < \infty, \quad \text{or} \quad \text{(ii)}\ \sum_{j=1}^{\infty} \lambda_j^2 = \infty.$$

We claim that if (i) holds, there is no weight collapse. That is, if, say, $g : \mathcal{R} \mapsto \mathcal{R}$ is bounded and continuous,

$$\sum_{i=1}^{n} w_i g(X_i^*) \xrightarrow{P} Eg(X|Y). \tag{7}$$

In the above, X_i^* is drawn from the empirical measure $\sum_{j=1}^{n} w_i \delta(X_i)$, where $\delta(\cdot)$ represents the delta function, and where, as before, the w_i represents the likelihood-defined weights.

To verify the convergence in (7), note that

$$w_i = U_i / \sum_{j=1}^{n} U_j,$$

where

$$U_i = c_{d'}^{-1} \exp\{-\frac{1}{2} \sum_{j=1}^{d'} \big[\lambda_j^2(Z_{ij}^2 - 1) + 2\lambda_j^2 \mu_j Z_{ij}\big]\}. \tag{8}$$

In (8), the Z_{ij}'s are i.i.d. $N(0,1)$ and

$$c_{d'} = E\big[\exp\{-\frac{1}{2} \sum_{j=1}^{d'} \big[\lambda_j^2(Z_{ij}^2 - 1) + 2\lambda_j^2 \mu_j Z_{ij}\big]\}\big] = \Pi_{j=1}^{d'} \big[(1+\lambda_j^2)^{-1/2} e^{\lambda_j^2/2} e^{\frac{\lambda_j^4 \mu_j^2}{2(1+\lambda_j^2)}}\big].$$

Now, since (i) implies that $\Pi_{j=1}^{d'}(1+\lambda_j^2)^{-1/2} e^{\lambda_j^2/2}$ converges and

$$E\big[\sum_{j=1}^{\infty} \frac{\lambda_j^4 \mu_j^2}{1+\lambda_j^2}\big] = \sum_{j=1}^{\infty} \frac{\lambda_j^4 E(\mu_j^2)}{1+\lambda_j^2} = \sum_{j=1}^{\infty} \lambda_j^2,$$

we have $E(U_1) = 1$ and $c_{d'} \to c$ (with c a constant).

Arguing as in Proposition 4.1 in [3], we can show that

$$Var\big[\frac{1}{n} \sum_{i=1}^{n} U_i g(X_i)\big] \leq \frac{1}{n} E\big[U_1^2 g^2(X_1)\big] \to 0,$$

since a straightforward computation shows that $E(U_1^2) \leq M < \infty$ for all d'. Thus, under (i), the importance weights have the correct expectation and vanishing variance.

On the other hand, if (ii) holds, we can state the following proposition.

Proposition 3.3. *Under* B, *if* $\sum_{j=1}^{\infty} \lambda_j^2 = \infty$ *and* $(\log n \log d')/\tau_{d'}^2 \to 0$, *we have*

$$\frac{\tau_{d'}}{\sqrt{2 \log n}} E(T_{n,d'}) \to 1.$$

We note that our conditions imply that

$$\frac{\max_{1 \le j \le d'} \lambda_j^2 (1 + y_{0j}^2)}{\tau_{d'}^2} \to 0$$

so that asymptotic normality holds. The proof requires Lemma 3.4 and Lemma 3.6.

The form reveals that it is possible to have much slower collapse than what Proposition 3.2 suggests. For instance, if $\lambda_j^2 = 1/j$, B holds but $\tau_{d'}^2 = \log d'(1+o(1))$. In fact, the requirement that the λ_j form an infinite sequence as above can be weakened to requiring simply that the λ_j be bounded above uniformly, and this can be verified using a subsequence argument.

In conclusion, on the basis of Proposition 3.3, provided that the nonzero λ_j's are commensurate, it seems reasonable to define $\sum_{j=1}^{d'} \lambda_j^2$ as the *effective dimension*. We note that the form of the effective dimension also plays a crucial role in the work of [7], who study Monte Carlo sample size requirements in the ensemble Kalman filter framework.

Appendix

We first introduce two lemmas that pertain to Edgeworth expansion type uniform normal approximations of the distribution (the cdf and the density respectively) of independent sums of random variables. The two lemmas lay the groundwork for the proof of Proposition 3.1. Valid for moderately large deviations, the first result (Lemma 3.4) is a special case of Theorem 2.5 in [12], and is stated here without proof.

Lemma 3.4. *Let* $\xi_1, \dots, \xi_d$ *be independent random variables with* $E\xi_j = 0$ *and* $\sigma_j^2 = Var(\xi_j^2) < \infty$. *Set*

$$S_d = \frac{1}{B_d}(\xi_1 + \cdots + \xi_d),$$

where $B_d^2 = \sum_{j=1}^{d} \sigma_j^2$, *and define the Lyapunov quotients*

$$L_{k,d} = \frac{1}{B_d^k} \sum_{j=1}^{d} E\left|\xi_j\right|^k, \quad k = 1, 2, \dots.$$

We also suppose $|E(Z_j^k)| \le k! \gamma_j^{k-2} \sigma_j^2, k \ge 3$, *where* $\gamma_1, \dots, \gamma_d$ *are constant terms.*

With these conditions, as $d \to \infty$, *there exist analytic functions* $P_d(x) = \sum_{k=3}^{\infty} \lambda_{k,d} x^k$ *with* $|\lambda_{k,d}| \le Ac^k d^{-\frac{k-2}{2}}$ *for some* A, c *and all* d, *such that the cdf of* S_d, *denoted* $G_d(\cdot)$, *satisfies,*

$$1 - G_d(x) = (1 - \Phi(x)) exp(P_d(x))(1 + o(1)),$$

$$G_d(-x) = \Phi(-x) \exp(P_d(-x))(1 + o(1))$$

uniformly for all $x \geq 0$ and $x = o(B_d/K_d)$, where $K_d = \max_{1\leq j\leq d}\{\gamma_j, \sigma_j\}$. Furthermore, P_d satisfies

$$|P_d(x)| \leq cx^3/B_d \tag{9}$$

for some constant $c > 0$. We use c generically as a constant independent of d.

Lemma 3.4 gives a normal approximation for the cdf of independent sums, and serves as the basis for the normality conditions of Proposition 3.1. Next we give a lemma for a normal approximation of the density of independent sums, which can be directly derived from Proposition 2 and Theorem 3 of [5].

Lemma 3.5. *With the same notation and conditions as in Lemma 3.4, we assume $\xi_{j,d}$ has density $g_{j,d}$ such that $sup_x\{|g_{j,d}(x)| : 1 \leq j \leq d\} \leq M < \infty$. Then, as $d \to \infty$, the density of S_d, $g_d(\cdot) = G_d(\cdot)$, satisfies*

$$g_d(x) = \phi(x)exp(P_d(x))(1 + o(1))$$

$$g_d(-x) = \phi(-x)exp(P_d(-x))(1 + o(1))$$

uniformly for all $x \geq 0$ and $x = o(B_d/K_d)$, where $K_d = \max_{1\leq j\leq d}\{\gamma_j, \sigma_j\}$.

We note in passing that the condition of uniform boundedness of the $g_{j,d}$ does not hold for Z_j, the Gaussian–Gaussian case. However, the sum of $\lambda_1^2 Z_1^2 + \lambda_2^2 Z_2^2$, where $\lambda_1, \lambda_2 > 0$ and Z_1, Z_2 are independent Gaussian, does indeed satisfy the condition. This may be verified by a direct calculation of the density of the convolution.

The next Lemma is given for the purpose of verifying the Lyapunov quotients conditions appearing in Lemma 3.4 and Lemma 3.5.

Lemma 3.6. *Let $Z_j, V_j, \epsilon_j, j = 1, \ldots, d$, be iid N(0,1). Let $\lambda_1 \geq \lambda_2 \geq \cdots$ where $\sum_{j=1}^{\infty}\lambda_j^2 = \infty$. Then, given $\mu_j \equiv V_j + \frac{\epsilon_j}{\lambda_j}$, for all j, we have*

$$\lambda_j^{2k}E\big(|(Z_j+\mu_j)^2 - (1+\mu_j^2)|^k\big|\mu_j\big) \leq \frac{O_p(\sqrt{\log d})^k}{k!}\rho^k\lambda_j^4 E\big((Z_j+\mu_j)^2 - (1+\mu_j^2)\big|\mu_j\big), \tag{10}$$

for $k \geq 3$.

Thus, given the mean vector $\xi = [\mu_1, \mu_2, \cdots, \mu_d]$ defined in (4), Lemma 3.6 states that the Lyapuanov conditions required by Lemma 3.4 hold, with probability tending to 1. We note that our argument also implies Lemma 3.4.

Proof of Lemma 3.6. Since $(Z_j + \mu_j)^2 - (1 + \mu_j^2) = (Z_j^2 - 1) + 2\mu_j Z_j$, it is enough to bound

$$\lambda_j^{2k}E\big(|(Z_j^2-1) + 2\mu_jZ_j|^k\big|\mu_j\big) \leq 2^k\big(\lambda_j^{2k}E|Z_j^2-1|^k + 2^k(|\mu_j|\lambda_j^2)^k E|Z_j|^k\big).$$

By standard properties of the Gaussian moments, for some positive constant C,

$$E|Z_j^2 - 1|^k \leq C^k k!, \quad \text{and} \quad E|Z_j|^k \leq C^k k!\ .$$

Since $E(Z_j^2 - 1 + 2\mu_jZ_j)^2 = 2 + 4\mu_j^2$ we see that (10) follows from the bound

$$\lambda_j^{2k}|\mu_j|^k \leq \lambda_j^2\mu_j^2 \max\{|\lambda_\ell^2\mu_\ell|^{k-2} : 1 \leq \ell \leq d\} = \big(O_p(\sqrt{\log d})\big)^{k-2},$$

since the $\lambda_j^2\mu_j$ are independent $N(0, \lambda_j^2 + \lambda_j^4)$ so that

$$\max\{|\lambda_\ell^2\mu_\ell| : 1 \leq \ell \leq d\} \leq (\lambda_j^2 + \lambda_j^4)^{1/2}\max\{|V_\ell| : 1 \leq \ell \leq d\}$$

where the V_ℓ are i.i.d. $N(0,1)$. The lemma follows. □

The remainder of the Appendix is devoted to the proof of the main result given in Proposition 3.1.

Proof of Proposition 3.1. Let S_j $(j=1,\dots,n)$ be as defined in the Proposition and let $S_{(1)}$ be the minimum. Note that

$$E(T_{n,d}|S_{(1)}) = \frac{(n-1)\int_{S_{(1)}}^{\infty}\exp\big(-\tau_d(z-S_{(1)})\big)\mathrm{d}G_d(z)}{\bar{G}_d(S_{(1)})}, \tag{11}$$

since, given $S_{(1)}$, the remaining $(n-1)$ observations are i.i.d. with cdf equal to $G_d(z)/\bar{G}_d(S_{(1)})$, $z \geq S_{(1)}$.

Let ε_d be a sequence of constants such that $\varepsilon_d \to 0$ and $\varepsilon_d\tau_d/\sqrt{2\log n} \to \infty$ as $n, d \to \infty$. We first define, for $x < \varepsilon_d\tau_d$,

$$h_{n,d}(x) := \int_x^{\infty} \exp\big(-\tau_d(z-x)\big)\mathrm{d}G_d(z). \tag{12}$$

To evaluate $h_{n,d}(x)$, we break the integral into two parts: the first part yields the integral from x to $x+\varepsilon_d\tau_d$, and the second part yields the tail integral from $x+\varepsilon_d\tau_d$ to ∞. By using the normal approximations of Lemma 3.4 and 3.5, under the assumption that $(\log n)/\tau_d^2 \to 0$, one can show that the second part is $o\big(\sqrt{2\log n}/n\tau_d\big)$.

To deal with the first part, we shall show that as $x \to -\infty$ and $x > -\varepsilon_d\tau_d$,

$$\int_x^{x+\varepsilon_d\tau_d} \exp\big(-\tau_d(z-x)\big)\mathrm{d}G_d(z) = \frac{1}{\tau_d}\phi(x)\exp\big(P_d(x)\big)(1+o(1)) \tag{13}$$

To this end, applying Lemma 3.5 with $\ell = 3$, we obtain,

$$\begin{aligned}
R_d(x) &:= \int_x^{x+\varepsilon_d\tau_d} \exp\big[-\tau_d(z-x) - \frac{1}{2}(z^2-x^2) + P_d(z) - P_d(x)\big]\mathrm{d}z(1+o(1)) \\
&= \int_0^{\Delta_{n,d}} \exp\big[-\tau_d v - \frac{1}{2}((x+v)^2 - x^2) \\
&\qquad\qquad + \sum_{k=3}^{\infty}\lambda_{k,d}((x+v)^k - x^k)\big]\mathrm{d}v(1+o(1)) \\
&= \frac{1}{|x|}\int_0^{|x|\varepsilon_d\tau_d} \exp\big[-(-1+\frac{\tau_d}{|x|})w - \frac{w^2}{2|x|^2} \\
&\qquad\qquad + \sum_{k=3}^{\infty}\lambda_{k,d}\sum_{j=1}^{k}(-1)^{k-j}C_{k,j}|x|^{k-2j}w^j\big]\mathrm{d}w(1+o(1))
\end{aligned}$$

$$= \frac{1}{|x|}\int_0^{|x|\varepsilon_d\tau_d} \exp\big[-b_1 w - b_2 w + \sum_{j=3}^{\infty} b_j w^j\big]\mathrm{d}w(1+o(1)), \tag{14}$$

where

$$\begin{aligned}
b_1 &= \frac{\sigma\sqrt{d}}{|x|} - b_1^* = -1 + \frac{\tau_d}{|x|} - \sum_{k=3}^{\infty}(-1)^{k-1}C_{k,1}\lambda_{k,d}|x|^{k-2}, \\
b_2 &= \frac{1}{2|x|^2} - b_2^* = \frac{1}{2|x|^2} - \sum_{k=3}^{\infty}(-1)^{k-2}C_{k,2}\lambda_{k,d}|x|^{k-4}, \quad \text{and} \\
b_j &= \sum_{k=j}^{\infty}(-1)^{k-j}C_{k,j}\lambda_{k,d}|x|^{k-2j}.
\end{aligned}$$

Note $|\lambda_{k,d}| \leq Ac_0^k \tau_d^{-(k-2)}$ and $C_{k,j} < c^k$, for some constants A, c_0, c where $\mu_j = V_j + \epsilon_j/\lambda_j$. Hereafter, we use c as a generic positive constant that does not depend on x and d. Under the assumptions that $x \to -\infty$, $|x| < \varepsilon_d \tau_d$ (hence $|x|/\tau_d \to 0$), and $|x|\Delta_{n,d} \to \infty$, we have, firstly,

$$\begin{aligned} b_1^* &\leq \sum_{k=3}^{\infty} Akc_0^k \tau_d^{-(k-2)} |x|^{k-2} \\ &= Ac_0^2 \big[3(c_0|x|/\tau_d)/(1-(c_0|x|/\tau_d)) + (c_0|x|/\tau_d)^2/(1-(c_0|x|/\tau_d))^2\big] \\ &= o(1), \end{aligned} \tag{15}$$

secondly,

$$b_2^* \leq x^{-2} \sum_{k=3}^{\infty} c(c|x|/\tau_d)^{k-2} = |x|^{-2}(c|x|/\tau_d)/(1-c|x|/\tau_d) = o(|x|^{-2}), \tag{16}$$

and thirdly,

$$\begin{aligned} b_j &\leq \big(\sum_{k=j}^{2j-1} + \sum_{k=2j}^{\infty}\big) A(cc_0)^k \tau_d^{-(k-2)} |x|^{k-2j} \\ &= \sum_{k=j}^{2j-1} A(cc_0)^k (|x|/\tau_d)^{k-j} |x|^{-j} \tau_d^{-(j-2)} + \sum_{k=2j}^{\infty} A(cc_0)^k (|x|/\tau_d)^{k-2j} \tau_d^{2-2j} \\ &\leq |x|^{-2}(c|x|\tau_d)^{-(j-2)} + c\tau_d^{2-2j} \\ &\leq 2|x|^{-2}(c|x|\tau_d)^{-(j-2)}. \end{aligned} \tag{17}$$

Since $w/(|x|\tau_d) \leq \varepsilon_d \to 0$, we can further derive

$$\begin{aligned} \sum_{j=3}^{\infty} b_j w^j &\leq 2(w/x)^2 (cw/(|x|\tau_d))^{j-2} = 2(w/|x|)^2 cw/(|x|\tau_d)/\big[1 - cw/(|x|\tau_d)\big] \\ &= o(|x|^{-2}) w^2. \end{aligned}$$

Combining (14), (15), (16), and (18) yields

$$R_d(x) = \frac{1}{|x|} \int_0^{|x|\Delta_{n,d}} \exp\big[-(-1 + \frac{\tau_d}{|x|})(1+o(1))w - (\frac{w^2}{2|x|^2})(1+o(1))\big] \mathrm{d}w \tag{18}$$

The $o(1)$'s appearing in the last expression are uniform as w varies over the integral interval. Now, the bounded convergence theorem ensures $R_d(x) = (1/\tau_d)(1+o(1))$, which establishes (13). Taking into account the remainder term, we conclude that

$$h_{n,d}(x) = \frac{1}{\tau_d} \phi(x) \exp\big(P_d(x)\big)(1+o(1)) + o(\frac{\sqrt{2\log n}}{n\tau_d}). \tag{19}$$

Our target $(\tau_d/\sqrt{2\log n})E(T_{n,d})$ can now be written as

$$\frac{\tau_d}{\sqrt{2\log n}} E(T_{n,d}) = \frac{\tau_d(n-1)}{\sqrt{2\log n}} E\big[\frac{h_{n,d}(S_{(1)})}{\bar{G}_d(S_{(1)})}\big] = \frac{\tau_d n}{\sqrt{2\log n}} \int_{-\infty}^{\infty} h_{n,d}(x) \bar{G}_d^{n-2}(x) dG_d(x). \tag{20}$$

We decompose the preceding integral into three parts

$$\frac{\tau_d}{\sqrt{2\log n}} E(T_{n,d}) = I_{n,d} + II_{n,d} + III_{n,d} \tag{21}$$

where $I_{n,d}, II_{n,d}$, and $III_{n,d}$ represent the integral of (11) over the intervals $[-\infty, -\varepsilon_d\tau_d]$, $(-\varepsilon_d\tau_d, -(\log n)^{1/4})$, and $[-(\log n)^{1/4}, \infty)$, respectively. The preceding discussion, combined with the approximation $g_d(x) = xG_d(x)(1 + o(1))$ as $x \to -\infty$ and $|x| = o(\tau_d)$, implies that the dominating part is the quantity represented by $II_{n,d}$. We have,

$$\begin{aligned} II_{n,d} &= \frac{n(n-1)}{\sqrt{2\log n}} \int_{-\varepsilon_d\tau_d}^{-(\log n)^{1/4}} xG_d(x)\bar{G}_d^{n-2}(x)dG_d(x)(1+o(1)) \\ &= \frac{1}{\sqrt{2\log n}} \int_{nG_d(-\varepsilon_d\tau_d)}^{nG_d(-(\log n)^{1/4})} G_d^{-1}(w/n)w(1-w/n)^n dw(1+o(1)) \\ &= \frac{1}{\sqrt{2\log n}} \int_{nG_d(-\varepsilon_d\tau_d)}^{nG_d(-(\log n)^{1/4})} \sqrt{-2\log(w/n)}we^{-w}dw(1+o(1)) \\ &= \int_0^\infty we^{-w}dw(1+o(1)) - \frac{1}{\sqrt{2\log n}} \int_0^\infty w\log we^{-w}dw(1+o(1)) \\ &= 1 + o(1). \end{aligned} \tag{22}$$

To arrive at (22) we have used the approximation $G_d^{-1}(z) = \sqrt{-2\log z}(1+o(1))$ for $z \to 0$ in light of Lemma 3.4 and Mill's ratio.

For the remaining two parts, we use Mill's ratio and obtain

$$\begin{aligned} I_{n,d} + III_{n,d} &\leq \frac{\tau_d}{\sqrt{2\log n}}(n-1)\big[P(S_{(1)} \leq -\varepsilon_d\tau_d) + P(S_{(1)} \geq -(\log n)^{1/4})\big] \\ &= \frac{\tau_d}{\sqrt{2\log n}}(n-1)\big[1 - \bar{G}_d^n(-\varepsilon_d\tau_d) + \bar{G}_d^n\big(-(\log n)^{1/4}\big)\big] \\ &\leq \frac{\tau_d}{\sqrt{2\log n}}(n-1)\big[nG_d(-\varepsilon_d\tau_d) + \bar{G}_d^n\big(-(\log n)^{1/4}\big)\big] \\ &\to 0. \end{aligned} \tag{23}$$

Finally, combining (21), (22), and (23), yields the desired result. □

References

[1] Anderson, J. and Anderson, S. (1999). A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review 127*, 2741–2758.

[2] Arulampalam, M., Maskell, S., Gordon, N., and Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions of Signal Processing* **50**, 2, 174–188.

[3] Bengtsson, T., Bickel, P., and Li, B. (2007). *Probability and Statistics: Essays in Honor of David A. Freedman.* IMS Monograph Series. 337–356.

[4] Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society, series B* **68**, 3, 411–436.

[5] Deltuviene, D. and Saulis, L. (2003). Asymptotic expansion of the distribution density function for the sum of random varaibles in the series scheme in large deviations zones. *Acta Applicanda Mathimaticae 78*, 87–97.

[6] Doucet, A., N. F. and Gordon, N., Eds. (2001). *Sequential Monte Carlo Methods in Practice.* Springer-Verlag.

[7] Furrer, R. and Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis* **98**, 2.

[8] Gordon, N., Salmon, D., and Smith, A. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings-F* **140**, 2, 107–113.

[9] Liu, J. (2001). *Monte Carlo strategies in scientific computing.* Springer Series in Statistics. Springer-Verlag, New York. MR1842342 (2002i:65006)

[10] Liu, J. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *J. Amer. Statist. Assoc.* **93**, 443, 1032–1044.

[11] Pitt, M. and Shepard, N. (1999). Filtering via simulation: Auxilliary particle filters. *Journal of American Statistical Association* **94**, 446, 590–599.

[12] Saulis, L. and Statulevicius, V. (2000). *Limit theorems of probability theory.* Springer.

[13] van Leeuwen, P. (2003). A variance minimizing filter for large-scale applications. *Monthly Weather Review 131*, 2071–2084.