# Information Conversion, Effective Samples, and Parameter Size

Xiaodong Lin, Jennifer Pittman, and Bertrand Clarke

***Abstract*—Consider the relative entropy between a posterior density for a parameter given a sample and a second posterior density for the same parameter, based on a different model and a different data set. Then the relative entropy can be minimized over the second sample to get a virtual sample that would make the second posterior as close as possible to the first in an informational sense. If the first posterior is based on a dependent dataset and the second posterior uses an independence model, the effective inferential power of the dependent sample is transferred into the independent sample by the optimization. Examples of this optimization are presented for models with nuisance parameters, finite mixture models, and models for correlated data. Our approach is also used to choose the effective parameter size in a Bayesian hierarchical model.**

***Index Terms*—Asymptotic relative efficiency, number of parameters, relative entropy, sample size.**

## I. Introduction

In recent years and in many fields, ever more datasets have been generated for similar purposes. However, because they are from different sources, they are not directly comparable. In an effort to bring the information in different datasets closer to comparability, we want a generic way to express the information content of one dataset relative to another. The information content of interest may be the equivalent independent sample size for a set of dependent data, or the minimum number of parameters required for good data summarization, or may be a representation of the data in term of a different model. In the latter case, we get a virtual sample that is as similar as possible to the original dataset.

To do this, consider the relative entropy between two conditional distributions for the same random variable. Suppose the value of the first conditioning random variable is fixed, but the value of the second remains to be specified. The core idea here is to specify the value of the second conditioning random variable by minimizing the relative entropy. This kind of minimization is embedded in the conventional interpretation of Shannon's Channel Coding Theorem, see Cover and Thomas [12]. We regard the optimal value as (approximately) informationally equivalent to the given value of the first conditioning random variable and call this procedure information conversion. If the conditioning random variables correspond to datasets, then the "optimal value" corresponds to a "virtual" dataset. Interest may focus on the entire virtual dataset or on some feature of it, such as it's (virtual) sample size.

Information conversion is natural in a Bayesian context, which we use for convenience. In this case, we examine the relative entropy between two posterior densities for a parameter $\theta$ formed from two priors, two likelihoods, and two sets of data. If the first dataset is fixed, minimization identifies a new, virtual data set that is (approximately) informationally equivalent to a the first data set. Note that the first dataset is collected under the first likelihood, but the virtual dataset is its representation under the second. Thus, the optimization is over the sample size as well as the actual data values. If desired, the virtual data can be combined with other data from the second likelihood. This form of data pooling is different from merely multiplying the two likelihoods because the pooled data can be regarded as coming from a single conditional distribution.

To crystalize these ideas, suppose the original data $\boldsymbol{x}^n = \{x_1, x_2, \ldots, x_n\}$ is generated from $p(\cdot|\theta)$, where $\theta \in \Omega$, an open set in $\boldsymbol{R}^d$, is the parameter of interest. Consider another dataset $\boldsymbol{y}^m = \{y_1, y_2, \ldots, y_m\}$ generated from $q(\cdot|\theta)$. If $p$ and $q$ are equipped with the same prior $w(\cdot)$ for $\theta \in \Omega$, we want to know what data vector $\boldsymbol{y}^m_{\text{virtual}}$ makes $w_q(\theta|\boldsymbol{y}^m_{\text{virtual}})$ as close as possible to $w_p(\theta|\boldsymbol{x}^n)$. If the $x_i$'s are dependent, and the $y_j$'s are independent, and the Fisher information under $p$ and $q$ are similar, then we expect $m_{\text{virtual}}$, the effective sample size under $q$, to be less than $n$. As the Fisher information of $q$ increases relative to the Fisher information of $p$, then again we expect $m_{\text{virtual}}$ to decrease relative to $n$.

Now our problem is to minimize the relative entropy

$$D(w_p(\theta|\boldsymbol{x}^n)||w_q(\theta|\boldsymbol{y}^m)) = \int w_p(\theta|\boldsymbol{x}^n)\{\log(w_p(\theta|\boldsymbol{x}^n) - \log(w_q(\theta|\boldsymbol{y}^m)))\}d\theta \quad (1)$$

over the vector length $m$ and the entries in the vector $\boldsymbol{y}^m$. For the sake of clarity, we have indicated the arguments of the densities on the left-hand side even though they are integrated out. The main contributions of this paper are to show that solutions to this optimization problem exist, have several intuitively appealing properties, and give interesting results in several important examples.

Manuscript received February 16, 2006; revised June 12, 2007. The material in this paper was presented in part at a Department Seminar, National University of Singapore, June 2005. Part of this work was conducted when X. Lin was a Postdoctoral Research Fellow at the Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC 27709. This work was performed while B. Clarke was on leave at the Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC.

X. Lin is with the Department of Mathematical Sciences, University of Cincinnati, University of Cincinnati, Cincinnati, OH, 45221 USA (e-mail: linxd@math.uc.edu).

J. Pittman is with the Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27705 USA (e-mail: jennifer.pittman@duke.edu).

B. Clarke is with the Department of Statistics, University of British Columbia, Vancouver, BC V6T 1Z2, Canada (e-mail: bertrand@stat.ubc.ca).

Communicated by P. L. Bartlett, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Color version of Figure 1 in this paper is available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2007.909168

Our work contrasts strongly with the classical problem of effective sample size determination which is to choose how big a sample to draw from a population. In this class of problems, one intends to take samples of a random quantity assumed to have a probability distribution within a known family so that the value of the unknown parameter $\theta$ is of interest. Recent contributors include Lee and Zelen [20], Joseph and Belisle [17], Adcock [1], and Lindley [22]. In these papers, a "true" model is assumed and inferences are derived from that one model only. Here, we are transforming actual data to virtual, or effective data, preserving the inferential information content of the data.

One obvious application of a virtual, or effective, sample size $m_{\text{virtual}}$ is to obtain a corrected estimate of precision to account for dependency in the data. For instance, if we have $\boldsymbol{x}^n = \{x_1, x_2, \ldots, x_n\}$ which are known to be correlated, then we can find the effective sample size with respect to the product of marginals and give $s/\sqrt{m_{\text{virtual}}}$ as the corrected estimate of the standard error for the usual sample mean $\bar{\boldsymbol{x}}$, where $s^2$ is the usual estimate of the variance from the data. It is also possible to use the effective sample from our procedure to form $\bar{\boldsymbol{y}}$ and $s(\boldsymbol{y}^m)$. However, unless the minimizing relative entropy is verified to be sufficiently small this may not be accurate.

Our work is also fundamentally different from the problem of statistical calibration. For instance, in the simplest linear-calibration experiment, $x_1, x_2, \ldots, x_n$ are known true values and $y_1, y_2, \ldots, y_n$ are the corresponding readings on a scale. These known pairs of observations $(x_i, y_i)$ are used to calibrate the scale; the calibration problem concerns the estimation of $\hat{x}$, the true but unknown value, by reading $\hat{y}$ on the calibrated scale. At root, this is a kind of inverse regression problem that makes estimates of other values of the independent variable from new observations of the dependent variable, based on previously observed data [19], [28], [24]. This has been extended by Deville and Särndal [13] to a more general calibration setting that permits using different sources with auxiliary information to improve estimates from sample surveys. More recently, Kakade and Foster [18] proposed a natural learning process based on deterministic calibration to approach Nash equilibrium. Gneiting *et al.* [14] applied calibration in a game-theoretic framework for the evaluation of probabilistic forecasts, where the predictive distribution (forecast) is systematically calibrated using incoming observations to maximize the sharpness of the predictive distribution.

These methods, although similar in form to our method, solve different problems. First, they suppose a true model where our method merely converts data from one model to data in terms of another model, neither of which need be true in any objective sense. More important, our goal is not to estimate a true but unknown value on one scale with the help of another. Our goal is to re-express existing data on one scale so it can be used on another scale to estimate the same parameter. These two views are only the same if the parameter can be regarded as a an unknown value from a random variable that will be repeatedly sampled. Regarding the parameter as a random variable is, in essence, the Bayesian approach. However, it is usually assumed that the parameter is constant over the data points, i.e., is not sampled repeatedly, and in practice is rarely measured directly. Thus, outside of a narrow class of slightly unusual calibration problems, our methods answer a different question.

From a Bayesian standpoint, our optimization to find $\boldsymbol{y}^m_{\text{virtual}}$ formalizes informative prior construction. For instance, in the setting of genomic data that we will elaborate in the later section, the posterior may not provide as much precision as we want. So, we may want to pool the data we have with the data from another, earlier, experiment. One could approach the problem with a random effects model or a hierarchical model to provide a combined estimate and standard error. Or, a subjective Bayesian can do this by formulating the old data into the prior. Alternatively, we can use our technique to convert the old data set under an old likelihood into an equivalent data set under the new likelihood. The equivalent data set can be pooled with the new data and used with the new likelihood to update an objective prior. The idea is not to provide more precise estimates, but to pool the data so that more elaborate analyses can be performed on the data. Of course, larger datasets will permit more precision if the inference procedure is unchanged. However, we are more optimistic: A larger dataset can enable more detailed modeling.

Apart from theoretical results to ensure the mathematical feasibility of our optimization, we present results for three examples which illustrate what we think are general principles. First, in a mixture setting, we verify that the ratio of Fisher information identified in our theorem below characterizes how rapidly inferential strength is lost as a consequence of contamination. Second, in a dependence model, we see how correlation reduces inferential strength. This example shows how to find an effective sample size for correct assessment of the precision of estimates. Third, in a nuisance parameter example we see how much information is lost from the existence of a nuisance parameter, even when it is not estimated. Taken together, these results mean that in typical cases, our data is much less informative than we think. Consequently, we suggest that experimental design criteria for sample size selection understates the amount of data needed because they are based on uncontaminated, independent data for a parameter of interest.

In a fourth example, we show how our method extends to determine an effective parameter size in a Bayesian hierarchical model. The central idea here is to redo our optimization on the level of the parameter space rather than on the sample space, in a predictive setting. We find that the sensitivity of inferences to small amounts of correlation in the data appears much higher than might be expected. This dramatizes the importance of keeping the number of parameters to a minimum.

As an application of our considerations here, we use our technique to convert microarray data on a spiked-in gene to virtual new data for the same gene. Improving the baseline calibration by augmenting a small new dataset with converted data from another dataset gives a better basis for evaluating the new data on expression levels for non-spiked-in genes. We use our method on the spiked-in genes because their true activity level is known and their variability level is lower than that of experimental genes. However, our method for converting data applies equally readily to the non-spiked-in genes.

This paper is organized as follows. In Section II, we establish existence and a limited uniqueness for the optimization in (1), and several related quantities. Then we state a theorem showing that the asymptotic relative efficiency characterizes the limiting behavior of the effective sample size with respect to the original

sample. In Section III, we give three examples to illustrate what we think are general properties of the effective sample size. In Section IV, we extend our notion of effective sample size to study effective parameter size. In Section V, we illustrate the validity of our reasoning on data pooling in a microarray setting. The proof of our main theorem is in the Appendix.

## II. Main Properties

We begin by giving sufficient conditions for the conversion of $\boldsymbol{x}^n$ under $p$ into an effective, or virtual, sample $\boldsymbol{y}^m$ under $q$. Then we obtain similar results for the expectation of (1) over $\boldsymbol{y}^m$. This corresponds to seeking an effective sample size without finding an entire effective sample. For completeness, we also examine the effect of taking expectations over $\boldsymbol{x}^n$. The properties in this case are very different because the optimization remains over $\boldsymbol{y}^m$, or $m$.

Once we have established existence, we present some of the asymptotic properties of the solution. Our main theorem requires both $n$ and $m$ to increase. Consequently, the relative entropy between the two posteriors behaves like the relative entropy between their limiting normals. This leads to a Chi-square distribution and the classical relative efficiency.

### A. The Effective Sample for a Given Data Set

Suppose $\boldsymbol{x}^n$ is fixed, and we want to know what data vector $\boldsymbol{y}^m$ would be informationally equivalent to it. The optimization in (1) is to find

$$(\boldsymbol{y}^m)^* = \arg\min_{\boldsymbol{y}^m} D(w_p(\theta|\boldsymbol{x}^n)\|w_q(\theta|\boldsymbol{y}^m)). \tag{2}$$

For a fixed $m$, this distance can be minimized to obtain the optimal sample $\boldsymbol{y}^m$. This is the sample of size $m$ under $q$, which gives inferences for $\theta$ as close as possible to those from using the model $p$ and the sample $\boldsymbol{x}^n$. By minimizing over $m$, we can also obtain the effective sample size $m^*$. Overall, the minimization is over the spaces $\mathcal{R}^m$, where $m$ ranges from 1 to $\infty$. Clearly, a solution $\boldsymbol{y}^m$ to (2) is a function of $\boldsymbol{x}^n$. To see how solutions behave, we assume the regularity conditions used in Clarke [10] and [11], and we follow the convention that random variables are denoted with upper case letters and their outcomes by the corresponding lower case letters. To establish that solutions to the optimization problem exist, we have the following proposition.

*Proposition 1:* The effective sample $\boldsymbol{y}^{m^*}$ exists with probability one, and is unique to the extent that $w_q(\theta|\boldsymbol{y}^{m^*})$ is a one-to-one function of $\boldsymbol{y}^{m^*}$ into the collection of probability densities on the parameter space. Consequently, the effective sample size $m^*$ exists.

*Proof:* The proof of existence has two steps. First, we show that $m^*$ is bounded. Then for any fixed upper bound $M$, we show each $y_i$ is bounded with probability one, $1 \le i \le m^*$.

From Clarke [10, Theorem 4.1], for $\epsilon > 0$ and $K > 0$, there is an $M_\epsilon$ so that when $m > M_\epsilon$, $\{D(w_p(\theta|\boldsymbol{X}^n)\|w_q(\theta|\boldsymbol{Y}^m) > K\}$ has probability at least $1-\epsilon$ in the mixture for $Y^m$.

Observe that $M_\epsilon = M_\epsilon(K)$. We can choose $K$ large and find an appropriate $M_\epsilon$ so that the set $\{D(w_p\|w_q) > K\}$ has probability greater than $1-\epsilon$, and the probability that the minimum over $y^m$ occurs in the set $\{D(w_p\|w_q) < K\}$ is also greater than $1-\epsilon$. Thus, with probability greater than $1-\epsilon$, $m^*$ is less than $M_\epsilon$ so to find the minimum in (2), it is enough to consider $m \in \{1,\dots,M_\epsilon\}$, a compact set.

A similar argument holds for the existence of $y_i$'s, $i = 1,\dots,m^*$. Suppose $y_2,..,y_m$ is fixed and consider varying $y_1$. Since $q$ is regular, for any fixed $\theta$, $q(y_1|\theta) \to 0$ as $y_1 \to \pm\infty$. In this case, the posterior shifts its mass away from $\theta_0$, thereby increasing the relative entropy between $w_p$ and $w_q$. Thus, we can be sure that the minimum in (2) cannot occur for $y_1$ too large or too small, i.e., there is a bound $B_1$ so that the minimum in (2) occurs within $|y_1| < B_1$. The argument for $y_2,\dots,y_m$ is similar to that for $y_1$. Thus, with high probability, the minimum in the right-hand side of (2), which is a continuous function, occurs on $m \le M_\epsilon$ and $|y_i| < B$, for all $i$ and some $B > 0$, a compact set. The minimizing values are the $m^*$ and $y_1,\dots,y_m$ asserted to exist for the right-hand side of (2).

The relative entropy is convex in both its arguments $w_p$ and $w_q$. Thus, the minimum is unique to the extent that $w_q(\cdot|\boldsymbol{y}^{m^*})$ is a one-to-one function of $\boldsymbol{y}^{m^*}$ into the collection of probability densities on the parameter space.

In expression (2), we have treated $\boldsymbol{x}^n$ and $\boldsymbol{y}^m$ as real vectors of length $n$ and $m$ since they are outcomes of random variables. If we take an expectation over $\boldsymbol{Y}^m$ in (2), then we can find an effective sample size without finding an effective sample directly. Now, the problem is to find

$$m_1 = \arg\min_m E_{\boldsymbol{Y}^m} D(w_p(\theta|\boldsymbol{x}^n)\|w_q(\theta|\boldsymbol{y}^m)).$$

The character of the optimization does not otherwise change, as can be seen from the following result.

*Proposition 2:* Let $\mathcal{N}$ be the density of $\mathcal{N}(\theta^*, (mI(\theta_0))^{-1})$ where $\theta^* = \theta_0 + I(\theta_0)^{-1}l'(\theta_0)$, $I(\theta_0)$ is the usual Fisher information evaluated at $\theta_0$, and $l'(\theta_0)$ is $(1/m)$ times the score function for $q_\theta$. If $\int w_q \log(w_p/\mathcal{N})$ is uniformly integrable, then the effective sample size $m_1$ exists.

*Proof:* It is sufficient to show that for any $\epsilon > 0$ and $K > 0$, there exists an $M_\epsilon$ so that when $m > M_\epsilon$, $E_{\boldsymbol{Y}^m}\{D(w_p(\theta|\boldsymbol{x}^n)\|w_q(\theta|\boldsymbol{y}^m))\} > K$. We have

$$\begin{aligned} &E_{\boldsymbol{Y}^m}\{D(w_p(\theta|\boldsymbol{x}^n)\|w_q(\theta|\boldsymbol{y}^m))\} \\ &\quad = E_{\boldsymbol{Y}^m}\{(\int w_p \log(w_p/\mathcal{N}) - \int w_p \log(w_q/\mathcal{N})\}. \end{aligned} \tag{3}$$

From Clarke [11, Theorem A.1] it is seen that the first term of the right-hand side goes to infinity because as $m$ increases $\mathcal{N}$ concentrates at $\theta_0$. For the second term, the treatment of term $C$ in the Appendix applies to show that the integral goes to zero in probability in the mixture distribution for $\boldsymbol{y}^m$. Now, the uniform integrability implies the expectation also goes to zero, and the overall expression increases to infinity.

As before, the expected relative entropy is convex in its arguments when they are regarded as densities and the only variable we are optimizing over is $m$. Since strictly convex functions have unique minima, moving away from $m_{\text{opt}}$ cannot decrease the relative entropy.

Uniform integrability can be hard to verify. However, the proof only requires the second term on the right to be bounded, so the assumption of uniform integrability is stronger than

needed. We have not given convenient sufficient conditions because solutions to (2) is all we need here.

It is tempting to try to minimize the expectation of (2) over $\boldsymbol{X}^n$ to find an effective sample $\boldsymbol{y}^m$. That is, when $\boldsymbol{x}^n$ is treated as a random sample, the effective sample problem is to find

$$\boldsymbol{y}^{m_2} = \arg\min_{\boldsymbol{y}^m} E_{(\boldsymbol{X}^n)} D(w_p(\theta|\boldsymbol{X}^n)||w_q(\theta|\boldsymbol{y}^m)).$$

This is usually not an interesting optimization, since the meaning of it would be to find a specific data set to match the average behavior of data from an experiment.

Now consider the case when expectations are taken over both $\boldsymbol{X}^n$ and $\boldsymbol{Y}^m$

$$m_{1,2} = \arg\min_{m} E_{(\boldsymbol{Y}^m)} E_{(\boldsymbol{X}^n)} D(w_p(\theta|\boldsymbol{X}^n)||w_q(\theta|\boldsymbol{Y}^m)).$$

Letting $I(\Theta; \boldsymbol{X}^n)$ and $H(\cdot)$ denote the Shannon mutual information between $\Theta$ and $\boldsymbol{X}^n$ and the entropy, respectively, we see from (1), that we have

$$\begin{aligned}
&E_{(\boldsymbol{Y}^m)} E_{(\boldsymbol{X}^n)} D(w_p(\theta|\boldsymbol{X}^n)||w_q(\theta|\boldsymbol{Y}^m)) \\
&\quad = \int m_p(\boldsymbol{x}^n) m_q(\boldsymbol{y}^m) \log \frac{w_p(\theta|\boldsymbol{x}^n)}{w_q(\theta|\boldsymbol{y}^m)} d\theta d\boldsymbol{x}^n d\boldsymbol{y}^m \\
&\quad = \int w(\theta) p(\boldsymbol{x}^n|\theta) \log \frac{w(\theta) p(\boldsymbol{x}^n|\theta)}{m_p(\boldsymbol{x}^n)} d\theta d\boldsymbol{x}^n \\
&\qquad - \int m_q(\boldsymbol{y}^m) w(\theta) p(\boldsymbol{x}^n|\theta) \log w_q(\theta|\boldsymbol{y}^m) d\theta d\boldsymbol{x}^n d\boldsymbol{y}^m \\
&\quad = I(\Theta; \boldsymbol{x}^n) - H(w) - \int w(\theta') q(\boldsymbol{y}^m|\theta') w(\theta) \\
&\qquad \log \left( \frac{w(\theta) q(\boldsymbol{y}^m|\theta)}{m_q(\boldsymbol{y}^m)} \frac{q(\boldsymbol{y}^m|\theta')}{q(\boldsymbol{y}^m|\theta')} \right) d\theta d\theta' d\boldsymbol{y}^m \\
&\quad = I(\Theta; \boldsymbol{X}^n) - I(\Theta', \boldsymbol{Y}^m) \\
&\qquad + m \int w(\theta) w(\theta') D(P(\theta')||P(\theta)) d\theta d\theta'.
\end{aligned}$$

Parallel to the standard results in the reference prior context, this can be approximated by

$$\begin{aligned}
&\left( \frac{d}{2} \log \frac{n}{2\pi e} + \int \frac{1}{2} \log |I_p(\theta)| w(\theta) d\theta + H(w) + o(1) \right) \\
&+ m \int w(\theta) w(\theta') D(P(\theta')||P(\theta)) d\theta d\theta' \\
&- \left( \frac{d}{2} \log \frac{m}{2\pi e} + \int \frac{1}{2} \log |I_q(\theta)| w(\theta) d\theta + H(w) + o(1) \right)
\end{aligned}$$

(see Clarke and Barron [9]), which can be written as

$$mC + \frac{d}{2} \log \frac{n}{m} + \frac{1}{2} \int \log \frac{|I_p(\theta)|}{|I_q(\theta)|} w(\theta) d\theta + o(1)$$

where $C = \int w(\theta) w(\theta') D(P(\theta')||P(\theta)) d\theta d\theta'$ is a finite constant. The dominating term is $mC$, which is linear in $m$ and independent of $n$. Furthermore, by taking the derivative with respect to (w.r.t.) $m$, the minimizing value of $m$ is $d/2C$. When $d = 1$ and $C$ is reasonable, say $C \geq 2$, the optimal $m$ is $m_{\text{opt}} \leq 1/4$, regardless of $n$. Over integer computations, this gives $m_{\text{opt}} = 1$. We have seen from Proposition 2 that minimizing the expectation over $Y$ gives a viable definition. However, when we also take expectations over $\boldsymbol{X}^n$ it is seen that the asymptotic expression increases with $m$. Thus, there is no data from $Y$ that can match $X$, and the more data from $Y$ that is considered the further from the average inferences from $X$ that the approximating posterior would give. This is something of an anomaly, but indicates the delicacy of inferential approximations.

More generally, when $\boldsymbol{x}^n$ is fixed, a plot of $\min_{\boldsymbol{y}^m} D(w_p(\theta|\boldsymbol{x}^n)||w_q(\theta|\boldsymbol{y}^m)$ versus $m$ typically is a convex function with a unique minimum on the interior of its domain. However, taking an expectation over $\boldsymbol{Y}^m$ or $\boldsymbol{X}^n$ often makes the convexity and minimization trivial: The minimum occurs at $m = 1$, and the curve is almost linearly increasing. We attribute this to the fact that the mixture distribution for $\boldsymbol{y}^m$ is stationary in the examples we considered. In fact, the stationary ergodic theorem (see Breiman [6, Sec. 6.7]), shows that a sample mean from a stationary process converges to a nontrivial random variable; here, it would have the distribution of the prior. So, intuitively, the expectation of the relative entropy when conditioning a posterior on a data vector of increasing length $m$ should look like $m$ times the expectation of the limiting random variable. This is consistent with our approximations above.

### *B. The Main Result*

Next, we state the main result which characterizes our intuition about how the distance between two related posteriors behaves asymptotically. It is stated for unidimensional parameters but the proof holds for all finite dimensions.

The use of $P_{\theta_0} \otimes Q_{\theta_0}$-probability, the product of the two probabilities conditioned on $\theta_0$, is necessary in this theorem because we are dealing with the random samples from $P$ and $Q$ simultaneously. However, if one were to solve the optimization problem defined in (2), $\boldsymbol{y}^m$ is regarded as a function of $\boldsymbol{x}^n$. In this case, the cross measure becomes $P$ alone.

*Theorem 3:* Let $\boldsymbol{x}^n = \{x_1, \ldots, x_n\}$ be a random sample from $p(\cdot|\theta)$ and $\boldsymbol{y}^m = \{y_1, \ldots, y_m\}$ be a random sample from $q(\cdot|\theta)$. Assume the following.

I: The maximum-likelihood estimates (MLEs) $\hat{\theta}_1(\boldsymbol{x}^n)$ from $p$ and $\hat{\theta}_2(\boldsymbol{y}^m)$ from $q$ exist and are consistent for the same $\theta_0$ under their respective true distribution.

II: Define

$$H_2(\theta_0, \theta) = \int q(y|\theta_0) \log \frac{1}{q(y|\theta)} dy$$

and

$$\hat{H}_2(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \log \frac{1}{q(y_i|\theta)}$$

and assume $\hat{H}_2(\theta) \to H(\theta_0, \theta)$ uniformly in $\theta$.

III: The prior $w$ has finite second moments, and we have that for any open neighborhood $N$ of $\theta_0$, there is an $r > 0$ and a $\rho > 0$ so that

$$P_{\theta_0} \left( \int_N w(\theta) p(X^n|\theta) < e^{nr} \int_{N^\epsilon} w(\theta) p(X^n|\theta) d\theta \right) = O(e^{-n\rho}).$$

Denote the posterior mean $\tilde{\theta}_1 = E_p(\Theta|\boldsymbol{x}^n)$. When $n, m \to \infty$, where $n/m \to C$ for some $0 < C < \infty$, we have

$$D(w_p(\theta|\boldsymbol{x}^n)||w_q(\theta|\boldsymbol{y}^m)) - \frac{m}{2} I_q(\theta_0)(\tilde{\theta}_1 - \hat{\theta}_2)^2 \to \frac{1}{2}\log\left(C\frac{I_p(\theta_0)}{I_q(\theta_0)}\right) + \frac{1}{2C}\frac{I_q(\theta_0)}{I_p(\theta_0)} - 1/2 \quad (4)$$

in $P_{\theta_0} \otimes Q_{\theta_0}$-probability, the product of the two probabilities, conditioned on $\theta_0$.

The proof, given in the Appendix, rests on techniques derived from standard results in maximum-likelihood (ML) consistency and asymptotic normality of the posterior; see Wald [31], Wolfowitz [33], Walker [32], and Bickel and Yahav [5] for details. Related work includes Chanda [7] and Redner [26]. Nevertheless, the assumptions require explication. First, sufficient conditions for I are well known. Second, when the parameter space is compact, II follows from a uniformization of standard law of large numbers results. For noncompact parameter spaces, assumption II also holds, under mild conditions, for parametric families in exponential form. Third, III holds under relatively mild and verifiable conditions; see Clarke [11]. The appearance of the MLE and the posterior mean is an artifact of our technique of control on certain posterior probabilities in the proof.

*Corollary 4:* Under the assumptions of Theorem 3, choosing $C = I_q(\theta_0)/I_p(\theta_0)$ as the limiting value of $n/m$ gives the smallest constant, zero, in expression (4). In addition, in this case, the minimizing value of $(y)^{m^*}$ in (2) for each $n$ occurs when $\tilde{\theta}_2 = \hat{\theta}_1$.

*Proof:* From Theorem 3, quantity $D(w_p(\theta|\boldsymbol{x}^n)||w_q(\theta|\boldsymbol{y}^m))$ is asymptotically equivalent to

$$\frac{m}{2} I_q(\theta_0)(\tilde{\theta}_1 - \hat{\theta}_2)^2 + \frac{1}{2}\log\left(C\frac{I_p(\theta_0)}{I_q(\theta_0)}\right) + \frac{1}{2C}\frac{I_q(\theta_0)}{I_p(\theta_0)} - 1/2.$$

If we optimize the sum of the last three terms over $C$ in the limit, we find $C = I_q(\theta_0)/I_p(\theta_0)$, and the minimum is 0. When $m$ and $n$ are going to infinity in the ratio of $C$, both $\tilde{\theta}$ and $\hat{\theta}_2$ are going to the true $\theta_0$. The remaining term $\frac{m}{2} I_q(\theta_0)(\tilde{\theta}_1 - \hat{\theta}_2)^2$ goes to zero. Thus, minimizing the relative entropy over $\boldsymbol{y}^m$ for fixed $\boldsymbol{x}^n$ gives zero in the limit of large $n$.

Following this corollary for fixed, but large, $n/m \approx I_q(\theta_0)/I_p(\theta_0)$, the minimum on the right in (2) is close to zero, thereby indicating that the two posteriors are approximately equal so that $(y)^{m^*}$ is mimicking the inferential properties of the given $x^n$. Thus, it is reasonable to regard $\boldsymbol{y}^m$ as the data set that is informationally equivalent to $\boldsymbol{x}^n$ under $p_\theta$, at least in an asymptotic sense.

### C. Informal Chi-Squared Asymptotics

Because the posterior is asymptotically normal, the limiting behavior of $D(w_p(\theta|\boldsymbol{x}^n)||w_q(\theta|\boldsymbol{y}^m))$ should be the same as the limiting behavior of

$$D(\mathcal{N}(\hat{\theta}_1, (nI_p(\theta_0))^{-1})||(\mathcal{N}(\hat{\theta}_2, (m_2 I_q(\theta_0))^{-1}))).$$

Indeed

$$\begin{aligned} &D(\mathcal{N}(\hat{\theta}_1, (nI_p(\theta_0))^{-1})||(\mathcal{N}(\hat{\theta}_2, (mI_q(\theta_0))^{-1}))) \\ &= \frac{1}{2}\log\frac{nI_p(\theta_0)}{mI_q(\theta_0)} \\ &\quad + \frac{mI_q(\theta_0)}{2}\left((\hat{\theta}_1 - \hat{\theta}_2)^2 + \frac{1}{nI_p(\theta_0)} - \frac{1}{mI_q(\theta_0)}\right) \\ &= \frac{1}{2}\log\frac{nk_1}{g(n)} + \frac{I_q(\theta_0)g(n)}{2}(\hat{\theta}_1 - \hat{\theta}_2)^2 + \frac{g(n)}{2nk_1} - \frac{1}{2} \\ &= \frac{1}{2}\log\frac{nk_1}{g(n)} + \frac{g(n)}{2nk_1} - \frac{1}{2} + \frac{I_q(\theta_0)g(n)}{2} \\ &\quad \left((\hat{\theta}_1 - \theta_0)^2 + (\hat{\theta}_2 - \theta_0)^2 - 2(\hat{\theta}_1 - \theta_0)(\hat{\theta}_2 - \theta_0)\right) \end{aligned}$$

in which $k_1 = I_p(\theta_0)/I_q(\theta_0)$ and we have written $m = m_n = g(n)$ for the sample size under $\mathcal{N}(\hat{\theta}_2, (mI_q(\theta_0))^{-1})$.

Now, suppose $g(n) \to \infty$ and that the limit $\lim_{n\to\infty} \frac{g(n)}{n}$ exists in the extended real line. Informally, we may write

$$g(n)(\hat{\theta}_1 - \theta_0)^2 I_q(\theta_0) \longrightarrow_{\mathcal{L}} \chi^2_d \frac{I_q(\theta_0)}{I_p(\theta_0)} \lim_{n\to\infty}\frac{g(n)}{n}.$$

The right-hand side equals

$$\frac{\chi^2_d}{k_1}\lim_{n\to\infty}\frac{g(n)}{n}.$$

Also

$$g(n)(\hat{\theta}_2 - \theta_0)^2 I_q(\theta_0) \longrightarrow_{\mathcal{L}} \chi^2_d.$$

Using these in the relative entropy gives informally

$$\begin{aligned} &D(\mathcal{N}(\hat{\theta}_1, (nI_p(\theta_0))^{-1})||(\mathcal{N}(\hat{\theta}_2, (mI_q(\theta_0))^{-1}))) \\ &\longrightarrow_{\mathcal{L}} \frac{k_1}{2}\lim\log\frac{n}{g(n)} + \frac{\chi^2_d + 1}{2k_1}\lim\frac{g(n)}{n} \\ &\quad - \frac{\mathcal{N}(0,1)\mathcal{N}(0,1)}{\sqrt{k_1}}\lim\frac{g(n)}{n} + \frac{\chi^2_d - 1}{2}. \end{aligned} \quad (5)$$

It is seen that if $g(n)/n \to 0$, the right-hand side of (5) is infinite. When $g(n)/n \to \infty$, the right-hand side of (5) goes to $+\infty$ or $-\infty$, but we rule out the latter because the right-hand side is positive. Thus, the minimum of the left-hand side of (5) will not be achieved at the endpoints of the positive half of the extended real line. This suggests the minimizing value of $D(\mathcal{N}(\hat{\theta}_1, (nI_p(\theta_0))^{-1})||(\mathcal{N}(\hat{\theta}_2, (mI_q)(\theta_0))^{-1})))$ will be on the interior of the positive half line. So, we also expect the minimum of $D(w_p(\theta|\boldsymbol{x}^n)||w_q(\theta|\boldsymbol{y}^m))$ to be on the interior of the positive half-line. Moreover, this suggests a slight weakening of the hypotheses for (4): the existence of the limit of $m/n$ in the extended real line should be enough to ensure minima exist, and parallel to classical theory, this should hold whenever the Fisher information is finite.

## III. Examples

In this section, we present three examples to show how to find effective samples and effective sample sizes.

TABLE I
EFFECTIVE SAMPLE SIZE $m$ VERSUS THE ORIGINAL SAMPLE SIZE $n$, WHEN $\theta_0 = 0$. THE ASYMPTOTIC RATIO FROM THE THEOREM IS $0.25$. THESE RESULTS ARE BASED ON TEN RUNS

| $n$ | 50 | 100 | 200 | 500 |
|---|---|---|---|---|
| $\bar{m}$ | 12.0 | 24.4 | 49.0 | 124.0 |
| $\bar{m}/n$ | 0.24 | 0.244 | 0.245 | 0.248 |
| Std. dev. | 0.0331 | 0.0296 | 0.0283 | 0.0212 |

TABLE II
EFFECTIVE SAMPLE SIZE $m$ VERSUS THE ORIGINAL SAMPLE SIZE $n$, WHEN $n = 100$ AND $\theta_0$ CHANGES FROM 1 TO 5

| $\theta_0$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $\bar{m}$ | 27.8 | 29.5 | 35.8 | 44.2 | 48.9 |
| $\bar{m}/n$ | 0.278 | 0.295 | 0.358 | 0.442 | 0.489 |
| Asymptotic ratio | 0.28 | 0.32 | 0.36 | 0.43 | 0.47 |

### A. A Mixture Example

Suppose $\boldsymbol{x}^n = \{x_i, 1 \le i \le n\}$ are independent and identically distributed (i.i.d.) samples drawn from a two component Gaussian mixture distribution $1/2\mathcal{N}(0,1) + 1/2\mathcal{N}(\theta,1)$. Here, $\theta$ is the parameter of interest, and the posterior $w_p(\theta|\boldsymbol{x}^n)$ can be stated in closed form. Because one component in the mixture has no information about the parameter, only some of the data in the sample is useful for inference on $\theta$.

Now consider a different sample $\boldsymbol{y}^m = \{y_i, 1 \le i \le m\}$, drawn independently from $\mathcal{N}(\theta,1)$. Based on $\boldsymbol{y}^m$, a different posterior distribution $w_q(\theta|\boldsymbol{y}^m)$ can be derived. In contrast to the previous model, every data point in $\boldsymbol{y}^m$ has information useful for inferences on $\theta$. Thus, to produce a posterior distribution based on $\boldsymbol{y}^m$ that is close to $w_p(\theta|\boldsymbol{x}^n)$, we expect to need a sample size $m$ smaller than $n$.

Consider a $\mathcal{N}(0,1)$ prior on $\theta$, so that $w_q(\theta|\boldsymbol{y}^m) = \mathcal{N}(m(\bar{y})/(m+1), 1/(m+1))$. Also, we have

$$w_p(\theta|\boldsymbol{x}^n) = \frac{\prod_{i=1}^n\{1+exp\{-((x_i-\theta)^2-x_i^2)/2\}\}exp\{-\theta^2/2\}}{\int\prod_{i=1}^n\{1+exp\{-((x_i-\theta)^2-x_i^2)/2\}\}exp\{-\theta^2/2\}d\theta}.$$

The effective sample and sample size are

$$[(\boldsymbol{y}^m)^*, m^*] = \arg\min_{\boldsymbol{y}^m, m} D(w_p(\theta|\boldsymbol{x}^n)||w_q(\theta|\boldsymbol{y}^m)). \quad (6)$$

Since $\boldsymbol{x}^n$ is fixed, the minimization depends only on $\bar{y}$ and $m$.

To see how the effective sample size in (6) depends on $n$, we have done two studies. In the first study, we have taken the true value $\theta_0$ to be 0 and let the original sample size $n$ range from 50 to 500. From the discussion after the main theorem, the optimal asymptotic value of the ratio $m/n$ is $I_p(\theta_0)/I_q(\theta_0)$, which is $1/4$ for $\theta_0 = 0$. Table I reports the effective sample size we obtain after solving the proposed optimization procedure. From this table, it is clear that as $n$ increases, $m/n$ approaches $1/4$ and its standard deviation around its limiting mean decreases.

Next we study the case when the true $\theta_0$ changes. In Table II, we report the effective sample size and sample ratio we obtained from the optimization procedure for $\theta_0 = 1,2,3,4,5$. The corresponding values of the asymptotic ratio $I_p(\theta_0)/I_q(\theta_0)$ are also given in the table. Again, it is seen that the empirical ratio matches its asymptotic value for all $\theta$'s. As $\theta$ moves away from zero, it takes more and more effective samples up to 50% of the sample size as expected, because the mixing weights on the $\theta$ dependent component is 0.5.

### B. A Dependence Model

Consider the following two models:

$$\text{Model I:} \quad \boldsymbol{x}^n = \{x_1, \ldots, x_n\} \sim \mathcal{N}(\mu\mathbf{1}, \Sigma_1)$$
$$\text{Model II:} \quad \boldsymbol{y}^m = \{y_1, \ldots, y_m\} \sim \mathcal{N}(\mu\mathbf{1}, \Sigma_2)$$

where $\Sigma_1$ is an $n \times n$ covariance matrix and $\Sigma_2$ is an $m \times m$ diagonal covariance matrix $\tau^2\boldsymbol{I}$ and the only parameter of interest is $\mu$. The problem is to find a sample from the second model that achieves the minimum distance between $w_p(\mu|\boldsymbol{x}^n)$ and $w_q(\mu|\boldsymbol{y}^m)$. With the same prior $w(\mu) = \mathcal{N}(\theta,\sigma)$ on both models, the posterior from Model I is $w_p(\mu|\boldsymbol{x}^n) \sim \mathcal{N}(s,t)$, where

$$s = \frac{(\boldsymbol{x}^n)'\Sigma_1^{-1}\mathbf{1} + \frac{\theta}{\sigma^2}}{\mathbf{1}'\Sigma_1^{-1}\mathbf{1} + \frac{1}{\sigma^2}}, \quad \text{and} \quad t = \mathbf{1}'\Sigma_1^{-1}\mathbf{1} + 1/\sigma^2.$$

The posterior from Model II is $w_q(\mu|\boldsymbol{y}^m) \sim \mathcal{N}(\nu,\delta)$ where

$$\nu = \frac{\tau^2/m}{\sigma^2 + \tau^2/m}\theta + \frac{\sigma^2}{\sigma^2 + \tau^2/m}\bar{y} \quad \text{and} \quad \delta = \frac{\sigma^2\tau^2}{m\sigma^2 + \tau^2}.$$

If $\theta = 0$ and $\sigma = \tau = 1$ and we have a sample $\boldsymbol{x}^n$ from Model I, the $s$ and $t$ in the posterior distribution become

$$s = \frac{(\boldsymbol{x}^n)'\Sigma_1^{-1}\mathbf{1}}{\mathbf{1}'\Sigma_1^{-1}\mathbf{1} + 1} \quad \text{and} \quad t = \mathbf{1}'\Sigma_1^{-1}\mathbf{1} + 1.$$

The posterior from Model II becomes $w_q(\mu|\boldsymbol{y}^m) = \mathcal{N}(\nu,\delta)$ with

$$\nu = \frac{m}{m+1}\bar{y} \quad \text{and} \quad \delta = \frac{1}{m+1}.$$

The optimal values of $\boldsymbol{y}^m$ and $m^*$ are

$$[(\boldsymbol{y}^m)^*, m^*] = \arg\min_{\boldsymbol{y}^m, m} D(w_p(\theta|\boldsymbol{x}^n)||w_q(\theta|\boldsymbol{y}^m)). \quad (7)$$

To investigate the effect of correlation in (7), we chose $\Sigma_1$ to be block diagonal with 1's on the main diagonal and a common value of $\rho$ for all off-diagonal elements in each block. The entries outside the blocks are0 . When $\rho$ is near 1, we expect the effective sample size from a data set $\boldsymbol{x}^n$ to be the number of blocks. This is verified by our computations.

Table III shows the results when $n = 100$, $\rho = 0.9$ and $\mu = 1$. We chose block sizes of $10, 5, 2$, and 1 corresponding to 10, 20, 50, and 100 blocks. For one randomly generated value of $\boldsymbol{x}^n$ from $\mathcal{N}(\mathbf{1}, \Sigma_1)$, where $\Sigma_1$ changed from run to run to be consistent with the block structure, the effective sample size and optimal $\bar{y}$'s were found. It is seen that the effective sample size is a random perturbation around the number of blocks. Also, the optimal $\bar{y}$'s are seen to be close to $\mu = 1$. The surfaces shown in Fig. 1 depict the negative relative entropies between the two posteriors as a function of $\bar{y}$ and $m$. The dot indicates the maximizing pair recorded in Table III. Roughly, each block corresponds to a data point.

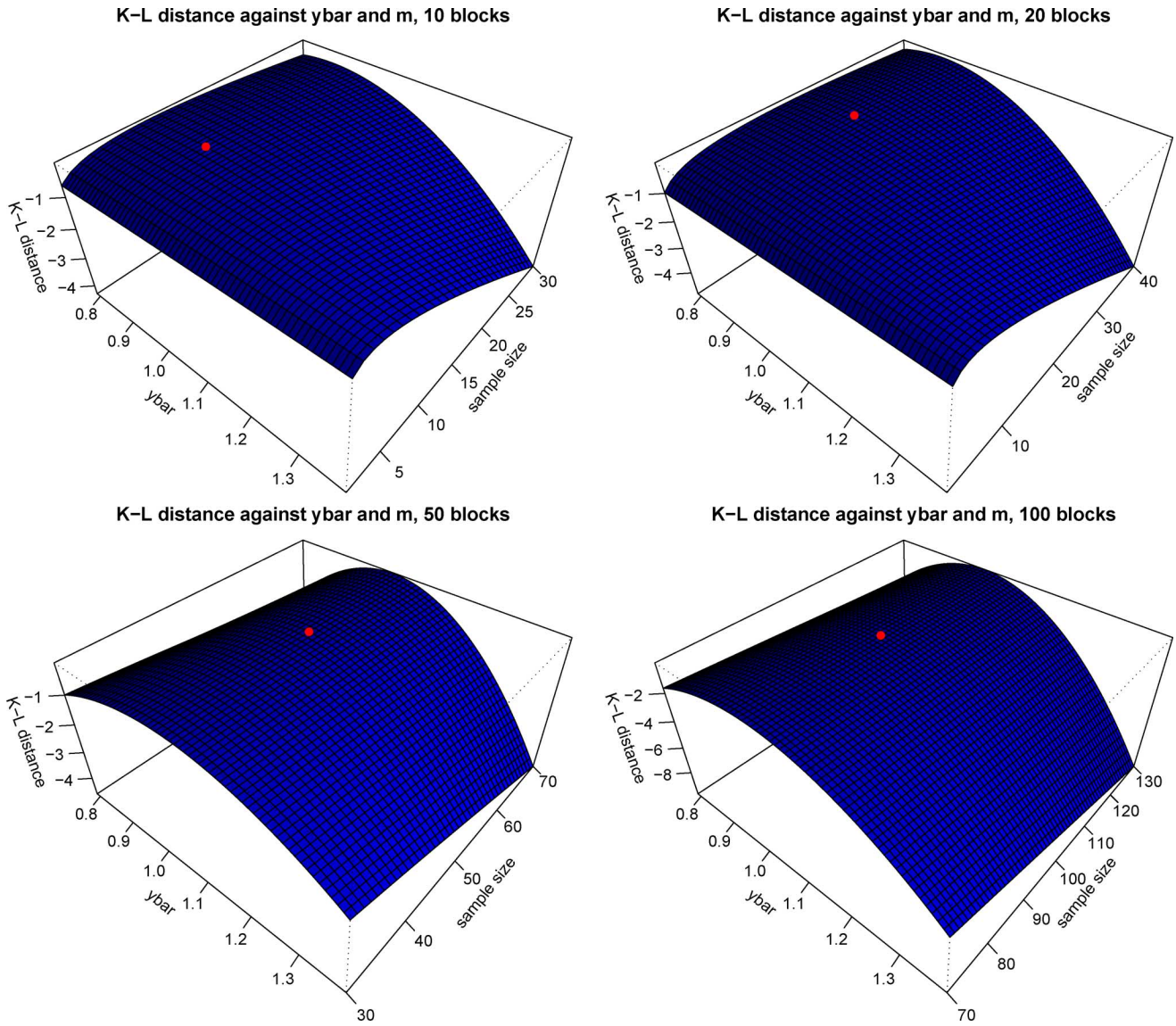


Fig. 1. The plots of negative relative entropy against $\bar{y}$ and $m$ for $\Sigma_1$ with 10, 20, 50, and 100 blocks. The dot indicates the maximum.

TABLE III
EFFECTIVE SAMPLE SIZE AND OPTIMAL $\bar{y}$ VALUES FOR VARIOUS NUMBERS OF BLOCKS

| Number of blocks | 10 | 20 | 50 | 100 |
|---|---|---|---|---|
| Eff. sample size | 10 | 21 | 52 | 98 |
| Optimal $\bar{y}$ | 0.94 | 0.95 | 1.04 | 1.05 |

For completeness, we see how the effective sample size changes as $\rho$ increases from 0.05 to 1. To see this, Fig. 2 shows a plot of the mean and standard deviation of the effective sample size as a function of $\rho$ when there is only one block. Fig. 3 shows the analogous plot when there are ten blocks. It is seen that when $\rho$ is close to 1 the effective sample size is close to the number of blocks. For smaller values of $\rho$, the effective sample size and its standard deviation increase rapidly. This reflects the fact that heightened variability corresponds to less information.

### C. A Nuisance Parameter Model

When nuisance parameters exist, their existence alone will "use up" data regardless of whether they are estimated. Indeed, consider the posterior for the parameters of interest, after the nuisance parameters have been integrated out. This marginal posterior will typically need more data to achieve the same degree of concentration as a posterior for the parameters of interest formed from a model without nuisance parameters. To see this, consider the following two models

$$\text{Model I:}\quad x_i \sim \mathcal{N}(\mu, a^2), \qquad \mu \sim \mathcal{N}(0,1),\ 1 \le i \le n.$$

$$\text{Model II:}\quad y_j \sim \mathcal{N}(\mu, \lambda),$$
$$(\mu, \lambda) \sim \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\{-\frac{\lambda}{2}(\mu-1)^2\} e^{-\lambda},\ 1 \le j \le m.$$

In Model I, $a$ is a fixed constant and we have a prior only on $\mu$. In Model II, we use a conjugate prior on the pair $(\mu, \lambda)$; the sample is used for inference on both parameters. If $\mu$ is the only

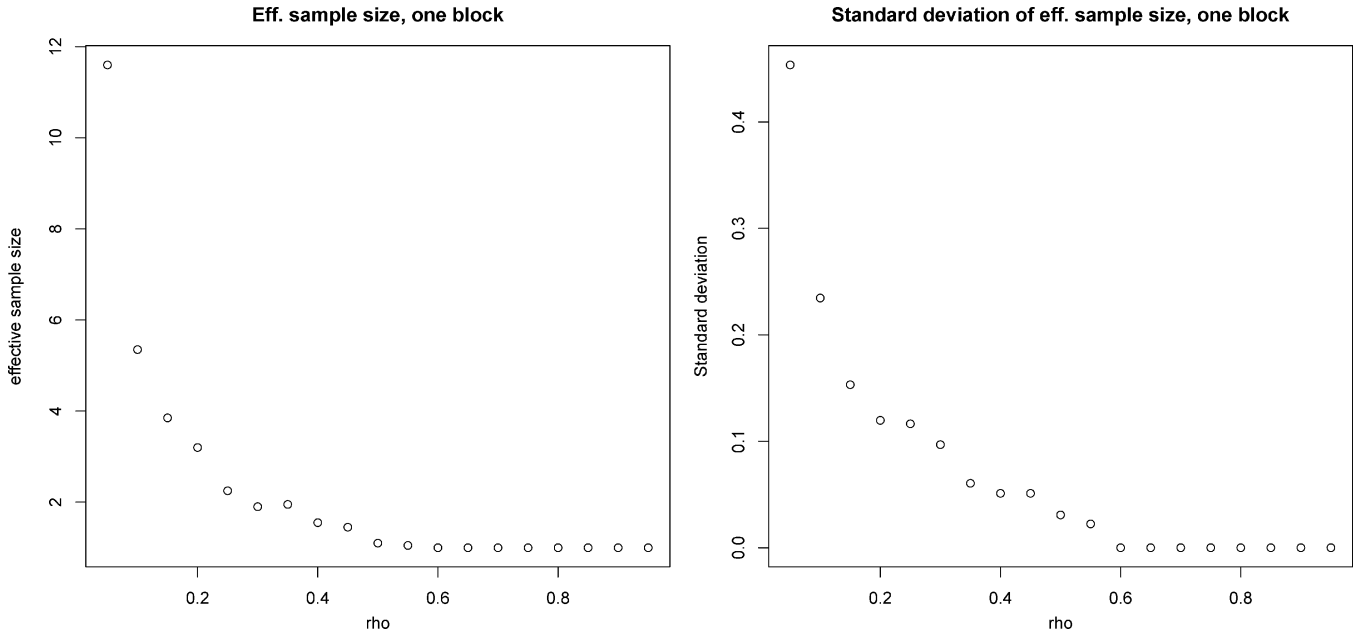


Fig. 2. Change of effective sample size when $\rho$ changes, one block .

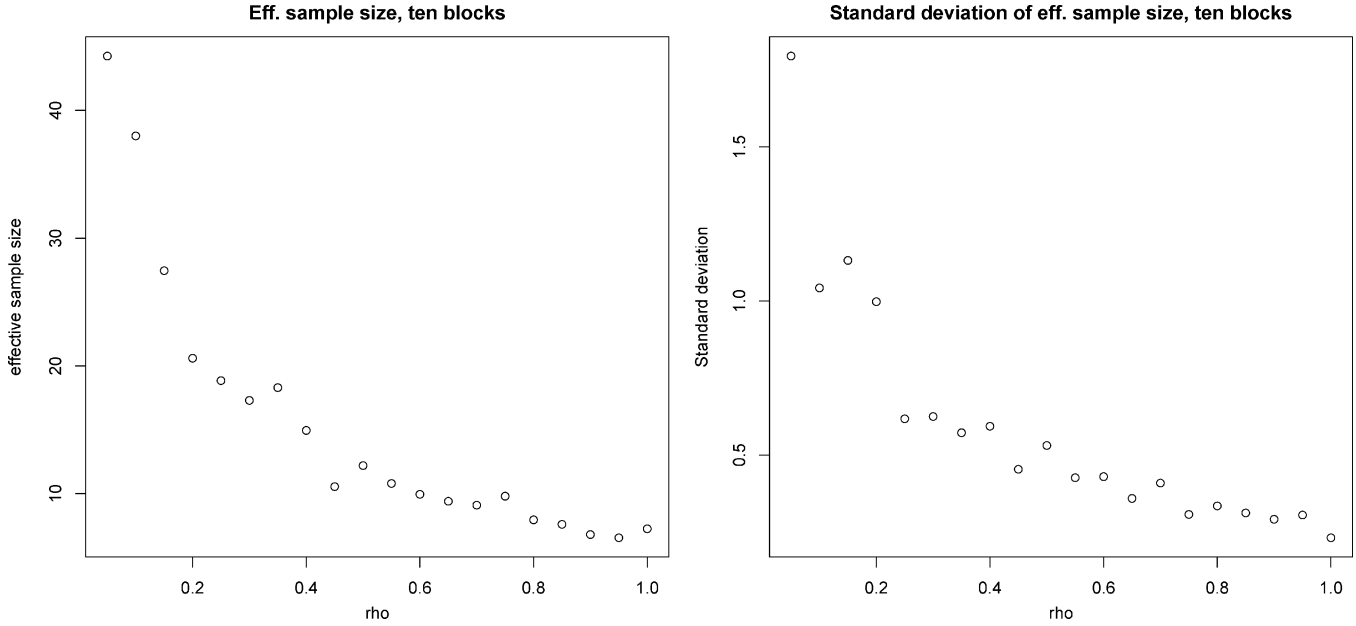


Fig. 3. Change of effective sample size when $\rho$ changes, ten blocks.

parameter of interest, Model II will require a larger sample size than Model I for the same size of credibility intervals.

The posterior distribution of $\mu$ given $\boldsymbol{x}^n$ is

$$w_p(\mu|\boldsymbol{x}^n) = \mathcal{N}\left(\frac{n\bar{x}}{n+a^2}, \frac{a^2}{m+a^2}\right).$$

The posterior distribution of $(\mu, \lambda)$ is

$$w_q(\mu, \lambda|\boldsymbol{y}^m) = B(2\pi)^{-(n+1)/2}\lambda^{(n+1)/2}\exp\{-\lambda C\},$$

where

$$B = \frac{D^{m/2+1}\sqrt{m+1}(2\pi)^{n/2}}{\Gamma(n/2+1)}$$

$$C = n\mu^2/2 + (\mu-1)^2/2 - \mu\sum_{j=1}^{m} y_j + \sum_{j=1}^{m} y_j^2/2 + 1$$

and

$$D = \frac{(\sum_{j=1}^{m} y_j)^2 + (m+1)\sum_{j=1}^{m} y_j^2 + 2\sum_{j=1}^{m} y_j + 3n + 4}{2(n+1)}.$$

To compare the two posteriors, we integrate out the parameter $\lambda$:

$$w_3(\mu|\boldsymbol{y}^m) = \int_0^\infty w_q(\mu, \lambda|\boldsymbol{y}^m)d\lambda.$$

Clearly, $w_3(\mu|\boldsymbol{y}^m)$ depends only on $m$, $\sum_{j=1}^m y_j$, and $\sum_{j=1}^m y_j^2$. Now, the optimization is to find $\boldsymbol{y}^m$ and $m$ such that

$$-\int_{-\infty}^{\infty} w_p(\mu|\boldsymbol{x}^n)\log w_3(\mu|\boldsymbol{y}^m)d\mu$$

is minimized. For $n=200$, $a=1$, and $\bar{x}=0$, the minimum is achieved when $m=230$, $\bar{y}\approx 0$, $\sum_{j=1}^m y_j^2/n\approx 1$. It seems that the mere presence of a nuisance parameter necessitates more data even when the nuisance parameter is not estimated.

Up to this point, we have used the same prior on both likelihoods. However, the mathematics are qualitatively the same when we permit different priors on the two likelihoods. The difference is in the interpretation: Allowing the second model to have a different prior and doing the same optimization asks for the virtual data that would update the second prior by the second likelihood to give inferences as close as possible to the first prior and likelihood. This may be important when the first prior is informative and the second prior is objective.

So, consider a third model where a reference prior under relative entropy is used on the pair of parameters

$$\text{Model III:}\quad z_k\sim\mathcal{N}(\mu,\sigma),\ (\mu,\sigma)\sim\frac{1}{\sigma},\quad 1\le k\le l$$

in which $l$ is the sample size for this third model. With this prior, the posterior is a student's $t$ distribution

$$w_4(\mu|\boldsymbol{z}^l)=(s^2+(\bar{z}-\mu)^2)^{l/2+2}\frac{\Gamma(l/2+2)s^{l-1}}{\Gamma(l/2+3/2)\sqrt{\pi}}.$$

The relative entropy between the posterior from Model III and the posterior $w_p$ from Model I is minimized when $l$ goes to $\infty$. This is so because the reference prior provides so little information on $\mu$ that it takes infinite data points to make the two posteriors match in the tails. In fact, even the tails of the two posteriors differ—one is polynomial, the other exponential. Observe that with the proper choice of $\boldsymbol{z}^l$, the student's $t$ goes to the normal $w_p(\mu|\boldsymbol{x}^n)$, the posterior from the first model, as $l$ goes to infinity.

## IV. Effective Parameter Size

The analysis for effective sample size can be carried over to the context of effective parameter size for Bayesian hierarchical models. We hope that a small sacrifice in prediction accuracy will yield a large reduction in model complexity. In our example below, we start with a model that permits as many parameters as data points. Since the parameters are tied together by a common distribution, we expect stantial reduction in complexity

$$\text{Model I:}\quad x_i\sim\mathcal{N}(\mu_i,1),\ \mu_i\sim\mathcal{N}(\theta,1),\quad \theta\sim\mathcal{N}(0,1),\quad 1\le i\le n.$$

In this model, the $\mu_i$'s are conditionally independent. However, for a sample $\boldsymbol{x}^n=\{x_1,\ldots,x_n\}$, the dependence among the means of the outcomes carries over to the outcomes themselves so that a one-to-one correspondence between the sample and the parameters $\mu_i$ may not be necessary. That is, far fewer parameters than are in the model are actually needed to describe the outcomes.

To see this, consider a hierarchical model with a dependence structure on the $\mu_i$

$$\text{Model II:}\quad x_i\sim\mathcal{N}(\mu_i,1),\ \{\mu_1,\ldots,\mu_n\}\sim\mathcal{N}(\theta\mathbf{1},\Sigma),\quad \theta\sim\mathcal{N}(0,1),\quad 1\le i\le n.$$

For simplicity, assume $\Sigma$ is of the form

$$\Sigma=\begin{pmatrix}1 & & \rho\\ & \cdots & \\ \rho & & 1\end{pmatrix}$$

where the off-diagonal entries have a common value $\rho$ that controls the dependency among the $\mu_i$'s. When $\rho$ goes to 0, Model II reduces to Model I. We limit our attention to these simple covariance matrices even though similar arguments can handle more complex $\Sigma$.

Different from the effective sample derivations in the previous section, here we are interested in how the dependence among the $\mu_i$'s affects the prediction accuracy of $x_{n+1}|\boldsymbol{x}^n$. Let the predictive distribution of Model I given $\boldsymbol{x}^n$ be $m_p(x^{n+1}|\boldsymbol{x}^n)$. In order to define the corresponding $m_q(x^{n+1}|\boldsymbol{x}^n)$ for Model II, we extend the model by assuming stationarity. That is, we assume

$$x_i\sim\mathcal{N}(\mu_i,1),\ \{\mu_1,\ldots,\mu_n,\mu_{n+1}\}\sim\mathcal{N}(\theta\mathbf{1},\Sigma^*),\quad \theta\sim,\mathcal{N}(0,1),\quad 1\le i\le n+1$$

where $\Sigma^*$ extends $\Sigma$ by adding an extra row and column with all entries equal to $\rho$ except for the $((n+1)\times(n+1))$th entry which is 1.

We use a two-step procedure to identify the effective sample size. In the first step, we optimize over $\rho$ to find the maximum level of correlation $\rho^*$ allowed such that the difference between the prediction accuracies of the two models is controlled within a prespecified threshold. Given this $\rho^*$, in step two we perform the optimization procedures for identifying the effective sample size, to obtain the effective parameter size.

More specifically, in step one, we compare the predictive ability of $\boldsymbol{x}^n$ under the two models. If $\rho=0$, the relative entropy between the predictive distributions $m_p(x_{n+1}|\boldsymbol{x}^n)$ and $m_q(x_{n+1}|\boldsymbol{x}^n)$ is zero. For a given tolerance level $\epsilon>0$, we define

$$\rho^*=\arg\min_{\rho}\int m_p(x_{n+1}|\boldsymbol{x}^n)(\log(m_p(x_{n+1}|\boldsymbol{x}^n))-\log(m_q(x_{n+1}|\boldsymbol{x}^n)))dx_{n+1}<\epsilon. \quad (8)$$

Once the $\rho^*$ is obtained, in step two, we can treat the parameters $\boldsymbol{\mu}^n$ as if they were data. That is, they can be effectively replaced by a set of independent parameters $(\boldsymbol{\mu}^m)^*$ using the optimization procedure from the effective sample size problem. Formally, we define the effective parameter size to be

$$m^*=E_{m(\boldsymbol{\mu}^n)}\arg\min_{(\boldsymbol{\mu}^m)^*m}\left[D(w(\theta|\boldsymbol{\mu}^n)\|w(\theta|(\boldsymbol{\mu}^m)^*))\right]. \quad (9)$$

This expression requires explication. First, the inner optimization is the same as what we use for effective sample size, but now it is in terms of parameters. The outer expectation means we take the average of the effective numbers of parameters found as $\mu^n$ varies. This approach does not give a linear function of $n$,

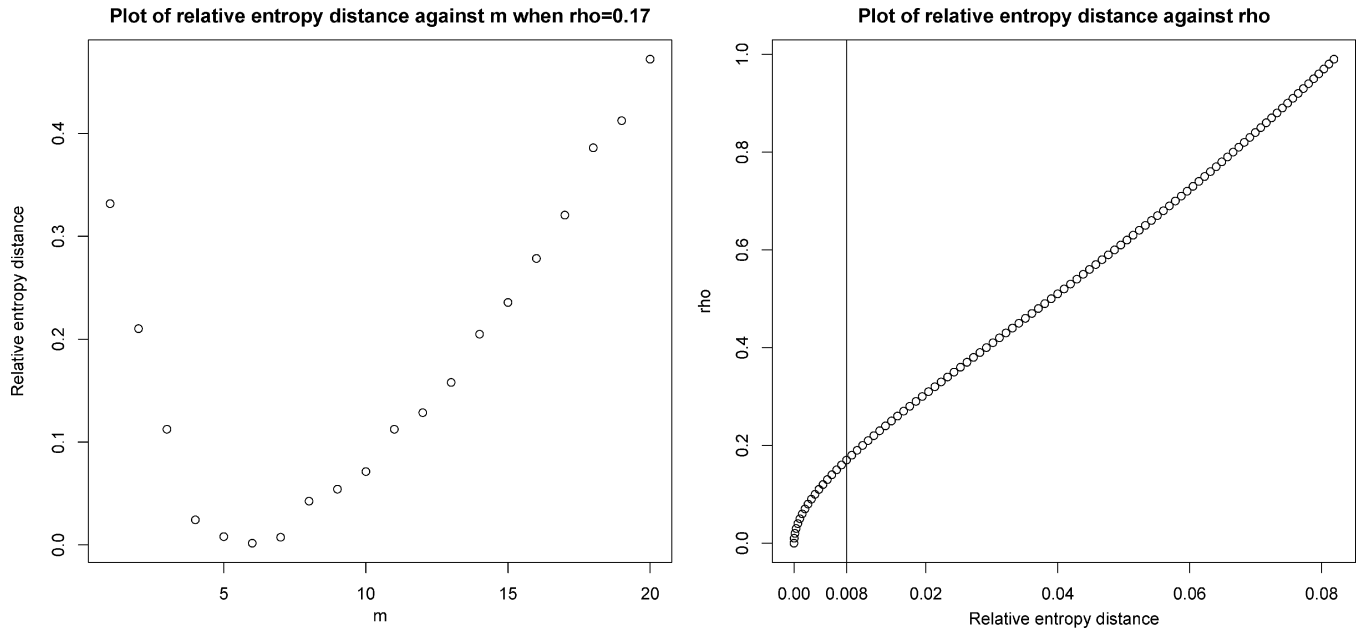


Fig. 4. Determining $\rho^*$ and $m^*$.

unlike the quantity mentioned in Section II, because the inner optimization and the outer expectation are not over the same variable. Expression (9) gives a convex function of $m$ with a unique minimum on the interior of its domain.

In fact, expressions (8) and (9) correspond to minimization of redundancies in source coding. Subject to a distortion bound $\epsilon$, we first find the dependence structure indexed by $\rho$ that achieves the minimum redundancy. Given this optimal $\rho$, we minimize a redundancy in (9) to take advantage of the dependence structure in $\mu^n$ so as to represent the information about $\theta$ in $\mu^n$, the parameters which govern the $x_i$'s, by a smaller set of parameters $\mu^m$. We focus on the information on $\theta$ because it summarizes the location of the $x_i$'s in aggregate and therefore is the information we want to express compactly.

We verify that this procedure gives reasonable results. First we must evaluate (8) to find $\rho^*$ as an increasing function of $\epsilon$. For Model I, we get

$$m_p(x_{n+1}|\boldsymbol{x}^n) = \int p(x_{n+1}|\theta) w_p(\theta|\boldsymbol{x}^n) d\theta.$$

It is easy to obtain $p(x_{n+1}|\theta) = \mathcal{N}(\theta, \sqrt{2})$ so that we find

$$w_p(\theta|\boldsymbol{x}^n) = \mathcal{N}\left(\frac{n(\bar{x})}{n+2}, \sqrt{\frac{2}{n+2}}\right).$$

Thus

$$m_p(x_{n+1}|\boldsymbol{x}^n) = \mathcal{N}\left(\frac{n(\bar{x})}{n+2}, \sqrt{2+\frac{2}{n+2}}\right).$$

For Model II, $w_q(\theta|\boldsymbol{x}^n) = \mathcal{N}(s,t)$, where

$$s = (\boldsymbol{x}^n)' \boldsymbol{I}((\boldsymbol{I}+\Sigma^{-1})^{-1})'\Sigma^{-1}\mathbf{1},$$

and

$$t = \left(\mathbf{1}'\Sigma^{-1}\mathbf{1} - \mathbf{1}(\Sigma^{-1})'((\boldsymbol{I}+\Sigma^{-1})^{-1})'\Sigma^{-1}\mathbf{1} - 1\right)^{-1/2}.$$

Hence, we have

$$m_q(x_{n+1}|\boldsymbol{x}^n) = \mathcal{N}(s, \sqrt{2+t^2}).$$

Using $m_p(x_{n+1}|\boldsymbol{x}^n)$ and $m_q(x_{n+1}|\boldsymbol{x}^n)$, we can compute $\rho^*$ for different values of $\epsilon$. To illustrate, suppose we take $\boldsymbol{x}^n$ to be a sample of size 100 generated from a standard normal distribution. The left panel of Fig. 4 shows the plot of the relative entropy distance between $m_p(x_{n+1}|\boldsymbol{x}^n)$ and $m_q(x_{n+1}|\boldsymbol{x}^n)$ as a function of $\rho$ ranging from $[0,1]$. Since the range of the relative entropy depends on how the data affect the integrand in expression (9), we set the threshold $\epsilon$ as the value below which 10% of the range falls. In this example, it corresponds to a relative entropy distance of $0.008$ as shown in the left panel of Fig. 4. When $\epsilon = 0.008$, we get $\rho^* = 0.17$ which we use in the covariance matrix $\Sigma$ of $(\mu^n)$.

Next, we use the optimization in (9) to obtain the effective parameter size $m^*$. The mixture distribution for $\mu^n$ from Model I is $\mathcal{N}(\mathbf{0}, \Sigma_1)$, where

$$\Sigma_1 = \left(\mathbf{1}'\Sigma^{-1}\mathbf{1}+1\right)\left((\mathbf{1}'\Sigma^{-1}\mathbf{1}+1)\Sigma^{-1} - (\Sigma^{-1})^{-1}\mathbf{1}\mathbf{1}'\Sigma^{-1}\right).$$

We randomly generate 100 $\mu^n$'s from this mixture distribution and use the same procedures as in the dependence model (see Section III-B) to obtain an effective parameter size $m$ for each generated value of $\mu^n$. Averaging the 100 values of $m$ gives $6.23$ as an approximation to the $m^*$ in (9).

In the right panel of Fig. 4, we show a plot of one instance of the relative entropy distance in (9) against the number of parameters $m$. Clearly, the curve is convex with a minimum at $m = 6$. Most plots of the other instances for the 100 $\mu^n$'s resemble this figure.

For completeness, we redid this procedure for a range of values of $\rho$. For each $\rho$ and each of 100 $\mu^n$'s, we got a convex curve as in Fig. 4 which we could minimize. The average of these minima is reported in Table IV. It is seen that even slight

TABLE IV
EFFECTIVE PARAMETER SIZE FOR DIFFERENT VALUES OF $\rho$

| $\rho$ | 0.2 | 0.1 | 0.05 | 0.01 |
|---|---|---|---|---|
| Eff. parameter size | 4.97 | 9.14 | 17.1 | 52.1 |

amounts of correlation among the parameters give a substantial reduction in the number of parameters needed to describe the data with only a small amount of exactitude sacrificed. When the correlation is small but nontrivial, say around 0.2, we readily get good data summarization with 1/20th as many parameters as data points.

## V. APPLICATION TO GENOMIC DATA

The analysis of gene expression data from DNA microarray technology involves the estimation of gene expression levels, an area of active current research in statistics, see Rubinstein *et al.* [27]. Popular estimation approaches for short oligonucleotide array data such as MAS5.0 (see Affymetrix [3]), RMA (see Irizarry *et al.* [16]), GCRMA (see Wu *et al.* [34]), MBEI (see Li and Wong [21]), and PDNN (see Zhang, Miles, and Aldape [35]), where each gene is represented by a set of probe level expression intensities, summarize these intensities to estimate the amount of target transcript present in the sample. These methods use background correction and normalization to reduce systematic variation within and across replicate arrays.

Researchers, however, often obtain the measurements on the same genes at different time points or in different experiments. Typically, these data are very expensive, and some of the experiments cannot be repeated. Thus, it is very useful for practitioners to determine quantitatively whether the data from one batch can be combined with the data from another batch, and if so, how to do so. Unfortunately, the above approaches do not address issues concerning the biases between arrays from various laboratories and/or batches. Methods exist for adjusting the summarized transcript level estimates, see Distance Weighted Discrimination in Benito *et al.* [4] for instance, or for selecting subsets of genes whose summarized expression measures are consistent across batches as in Parmigiani *et al.* [25]. However, with the increasing number of publicly available datasets and the quick maturation of gene expression studies from small-chip-number hit-and-run type projects to those with a more robust study design—The Tumor Analysis Best Practices Working Group [30] is an example—there is a need for careful statistical methods which can adjust for such biases at the probe level for specific genes of interest.

Here we use the technique described and studied in the previous sections to convert one set of data to its equivalent virtual data set in another distribution. This second distribution is chosen to be the distribution of a second data set. Consequently, the converted data and the second data set can be combined. Thus, one analysis can in principle be performed on the unified data set. We suggest that this approach can be applied generally to make simpler analyses feasible.

### A. Data

We have implemented our procedure on data from a series of controlled "spike-in" experiments on Affymetrix gene chips. "Spike-ins" are often used as a baseline comparison for other genes, i.e., researchers examine the spike-in genes to decide whether the whole batch of measurements reflect reality and thus can be used for further analysis. The data we use here come from experiments in which 12 different gene transcript groups, each containing the same set of genes at different concentrations, were spiked into a background consisting of a labeled mixture of mRNA from a human breast tissue source. This mixture was then hybridized onto Affymetrix HU95Av2 GeneChips, see Affymetrix [2]. Twenty distinct probe pairs interrogate each transcript. Finally, each of the measurements for a given gene at a given concentration was replicated on several samples in each of two batches of arrays. One batch of arrays ("Taiwan samples") represent primary breast cancer tumor biopsy samples from the Koo Foundation Sun Yat-Sen Cancer Center in Taipei, while the other batch of arrays ("Duke samples") represent primary breast cancer tumor biopsy samples from Duke University Medical Center in Durham, NC. Samples selected for the analysis detailed in the following were clinically matched on traditional disease risk factors (e.g., axillary lymph node status, estrogen receptor status) and processed at the same laboratory in separate batches and at different time points.

Scans of the chips were performed with an Affymetrix GeneChip scanner and the computed single intensity value for each probe cell was calculated using the Affymetrix Microarray Analysis Suite v5.0, see Affymetrix [2], [3]. This software provides scaling factors which can be used to adjust the data to a common target intensity; the factors for these chips lie within quality control bounds (The Tumor Analysis Best Practices Working Group [30]). The gene expression measures were log base 2 transformed but we performed no normalization of these data before the analysis discussed below. Since the spike-ins were created in the laboratory and capture probe-creation variability as well as hybridization variability, they represent a compromise between experimental laboratory probes (where the amount of transcript is unknown) and factory spike-ins (which capture only hybridization variability).

### B. Analysis

The microarray data are considered to be consistent with having come from a log-normal distribution (Hoyle *et al.* [15]). After log transformation, it is reasonable to assume that the Taiwan data follows a Normal distribution $\mathcal{N}(\mu\mathbf{1}, \Sigma_1)$ and the Duke sample follows $\mathcal{N}(\mu\mathbf{1}, \Sigma_2)$. We also assume that the off-diagonal terms of $\Sigma_1$ and $\Sigma_2$ are the same because the correlation between any two probes on the same chip can be regarded as constant. Thus, the covariance matrix has the following form:

$$\Sigma_i = \sigma_i^2 \begin{pmatrix} 1 & & \rho_i \\ & \cdots & \\ \rho_i & & 1 \end{pmatrix}, \quad \text{for } i = 1, 2.$$

We want to convert one sample of the Taiwan dataset to a "virtual" sample in the Duke dataset. Let $\boldsymbol{x}^n$ denote the Taiwan sample, and $\boldsymbol{y}^m$ denote the corresponding Duke data that we want to optimize.

Since we are converting the Taiwan data to the Duke data, we use the prior knowledge on the mean of the Duke data and set the prior on the Taiwan model as $\pi_p(\mu) \sim \mathcal{N}(\overline{y}_{\text{duke}}, 1)$. Similarly,

we set the prior for the Duke model as $\pi_q(\mu) \sim \mathcal{N}(\bar{x}_{\text{taiwan}}, 1)$. The optimization problem is to find $m$ and $\boldsymbol{y}^m$ to minimize

$$KL(w_p(\mu|\boldsymbol{x}^n)||w_q(\mu|\boldsymbol{y}^m)). \tag{10}$$

Clearly

$$\Sigma_i^{-1} = \sigma_i^{-2}\left(\frac{1}{1-\rho_i}I - \frac{\rho_i \mathbf{1}\mathbf{1}'}{(1+(n-1)\rho_i)(1-\rho_i)}\right)$$

for $i = 1, 2$. Thus, the posterior distribution $w_p(\mu|\boldsymbol{x}^n) = \mathcal{N}(\mu_1, t_1)$, where

$$\mu_1 = \frac{n\bar{x} + \sigma_1^2(1+(n-1)\rho_1)\bar{x}_{\text{taiwan}}}{n + \sigma_1^2(1+(n-1)\rho_1)}$$
$$t_1 = \frac{n}{\sigma_1^2(1+(n-1)\rho_1)} + 1$$

and $w_q(\mu|\boldsymbol{y}^m) = \mathcal{N}(\mu_2, t_2)$, where

$$\mu_2 = \frac{m\bar{y} + \sigma_2^2(1+(m-1)\rho_2)\bar{y}_{\text{duke}}}{m + \sigma_2^2(1+(m-1)\rho_2)}$$
$$t_2 = \frac{m}{\sigma_2^2(1+(m-1)\rho_2)} + 1.$$

Here, $\bar{x}$ is the mean of $\boldsymbol{x}^n$, and $\bar{y}$ is the mean of $\boldsymbol{y}^m$. Clearly, the minimum for the $KL$ distance is achieved when $\mu_1 = \mu_2$, and $t_1 = t_2$. By solving these two equations, we obtain the optima

$$\bar{y}^{\text{optimal}} = \bar{x} + \frac{\sigma_1^2(1+(n-1)\rho_1)(\bar{y}_{\text{duke}} - \bar{x}_{\text{taiwan}})}{n} \tag{11}$$

and

$$m^{\text{optimal}} = \frac{(1-\rho_2)n\sigma_2^2}{\sigma_1^2(1+(n-1)\rho_1) - \rho_2 n\sigma_2^2}.$$

When the value of $\rho_1$ is close to 1, $\bar{y}^{\text{optimal}}$ is approximately $\sigma_1^2(\bar{y}_{\text{duke}} - \bar{x}_{\text{taiwan}})$. When $\rho_1$ is close to 0, $\bar{y}^{\text{optimal}}$ can be approximated by $\sigma_1^2(\bar{y}_{\text{duke}} - \bar{x}_{\text{taiwan}})/n$. The $\sigma_1^2$ and $\rho_1$ jointly determine the corrections we need to make on the Taiwan sample, when converting to the Duke data using its model.

Note that in (10), the parametric families used to form the posteriors are the same, normals with unknown means and variances indexed by the same two parameters, a variance and a correlation. For simplicity of modeling and convenience of forming posteriors, we have equipped each of the parametric families with its conjugate prior, a normal density having known variance, taken to be 1. In fact, given the typical spread of the data, unit variance in the prior is reasonable; the prior can be regarded as objective.

This formulation in (10) is superficially the same as (1) or (2). However, the mathematical form of the posteriors in (10) is the same because they are both based on the normal likelihood. Permitting both posteriors to come from the same likelihood can reduce the problem to triviality by making the two posteriors too similar, possibly the same. For instance

$$\arg\min_{\boldsymbol{y}^m} KL(w_p(\mu|\boldsymbol{x}^n)||w_q(\mu|\boldsymbol{y}^m)) = 0$$

which corresponds to a $\bar{y}^{\text{optimal}} = \bar{x}$. So, straightforward application of the optimization does not achieve the desired data conversion. Consequently, it is seen that the two entries in the relative entropy must be different enough that it is meaningful to compare them.

TABLE V
THE AVERAGE GENE LEVEL EXPRESSION ESTIMATES ACROSS SAMPLES FOR GENE 4, SPIKE-IN 9. THE $\rho$ ARE ESTIMATED USING THE ORIGINAL DATA

| Gene location | 5' | middle | 3' |
|---|---|---|---|
| Taiwan original | 6.607 | 6.381 | 6.723 |
| Taiwan converted | 6.864 | 6.619 | 7.151 |
| Duke original | 6.932 | 6.721 | 7.133 |

One fix for this problem is to treat the two posteriors operationally as representing two inference processes we want to make similar. That is, someone who starts with virtual Duke data and wants to use it with actual Taiwan data should be similar to some one who starts with actual Taiwan data and wants to use it to help inferences based on Duke virtual data. In short, the order of receipt of information should not lead to big differences. More prosaically, updating a prior based on converted data by actual data or updating a prior based on actual data using converted data should give similar inferences.

Mathematically, this leads us to the expression

$$\arg\min_{\boldsymbol{y}^m, m} KL(w_p(\cdot|\boldsymbol{x}^n)||w_q(\cdot|\boldsymbol{y}^m))$$

in which we have used $\pi_p(\mu) \sim \mathcal{N}(\bar{y}_{\text{duke}}, 1)$ in the first entry and $\pi_q(\mu) \sim \mathcal{N}(\bar{x}_{\text{taiwan}}, 1)$ in the second. Note that the data dependence in the prior here does not give incoherence because we are only going to use $\boldsymbol{y}^m$ to make inferences, not the combined data $\boldsymbol{x}^n$ and $\boldsymbol{y}^m$. Relative to the Duke data, the data dependence in the prior here represents a pre-experimental state of information assuming the two samples are unrelated. It is only the optimization that ties $\boldsymbol{x}^n$ and $\boldsymbol{y}^m$ together. Otherwise put, our use of data dependent priors ensures both posteriors are based on the same information and are meaningfully different, so the optimization can be done.

The MLE is used for the estimates of $\sigma_i$ and $\rho_i$, for $i = 1, 2$. Depending on the perspective of the geneticists, the estimates for $\rho$'s can be obtained either by the MLE using the original data sample, or the MLE using the standardized sample (subtracted by the group mean). Estimates obtained using either of these two procedures reflect specific prior beliefs and expectations regarding microarray data that depend on the real applications. If we expect all probe level expression measures of a given gene to be very similar, or if we are supposing that the correlation within a probe set should be larger than the correlation across probe sets, then we should use a large $\rho$. If we believe that each probe within a probe set represents a different sequence of genetic bases and that measured expression varies, often greatly, with different subsequences of the same gene then we should use a small $\rho$. In either case the choice of $\rho$ should be made within the context of the specific application of interest.

These data conversion procedures are applied on the set of three spike-in samples each for the Taiwan and the Duke samples. In Table V, we show the mean of the original Taiwan sample, the converted Taiwan sample, and the original Duke sample, on spike-in # 9 and gene 4. The mean of the converted data for each Taiwan sample, measured at different levels, is adjusted so it is closer to the mean of Duke data. The detailed

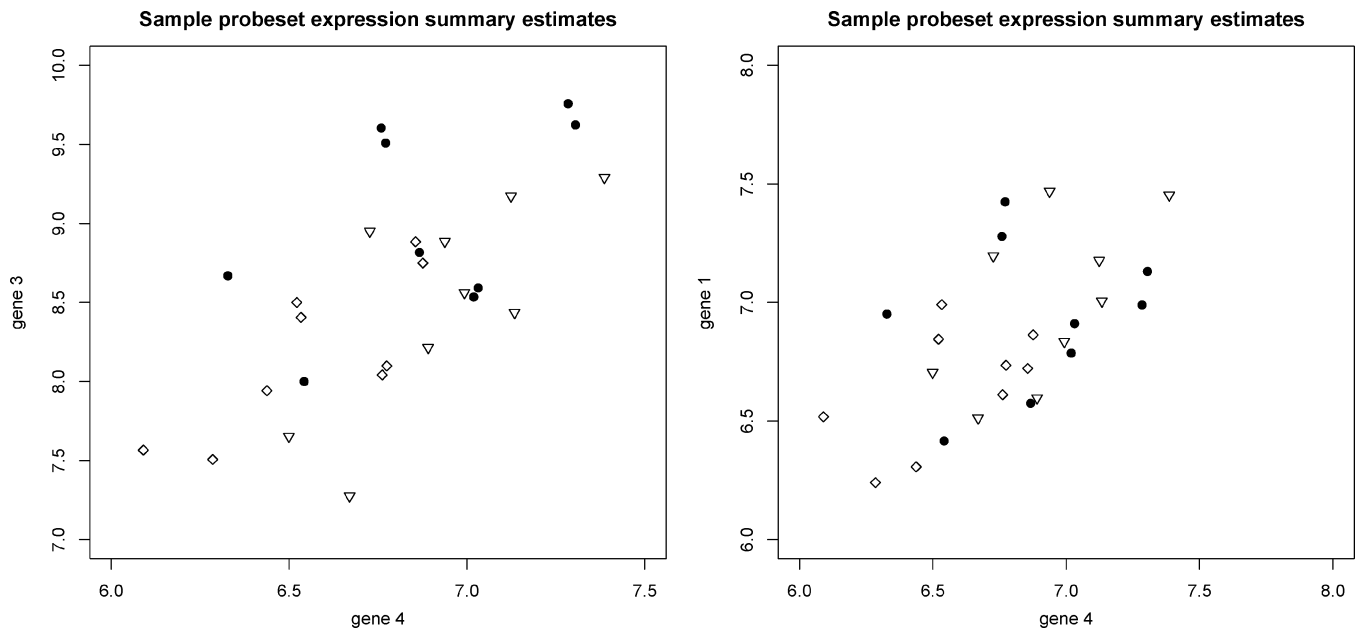


Fig. 5. The plot of probe set level expression for two genes, when $\rho$ estimated from the original data. The left side figure shows the plot for Gene 4 and Gene 3, and the right side figure shows the plot for Gene 4 and Gene 1. $\cdot$ denotes the converted Taiwan data, $\diamond$ denotes the original Taiwan data, and $\triangle$ denotes the original Duke data.

TABLE VI
THE CONVERTED DATA AND THE ORIGINAL TAIWAN SAMPLES FOR GENE 4, SPIKE-IN 9. TAIWAN SAMPLE MEASUREMENTS ARE INSIDE BRACES. THE $\rho$ ARE ESTIMATED USING THE ORIGINAL DATA

| Gene location | 5' | middle | 3' |
|---|---|---|---|
| Taiwansample1 | 6.761(7.018) | 6.758(6.521) | 7.286(6.855) |
| Taiwansample2 | 7.031(6.774) | 6.771(6.533) | 7.304(6.876) |
| Taiwansample3 | 6.542(6.285) | 6.327(6.091) | 6.866(6.438) |

TABLE VII
THE AVERAGE GENE LEVEL EXPRESSION ESTIMATES ACROSS SAMPLE FOR GENE 4, SPIKE-IN 9. THE $\rho$ ARE ESTIMATED USING THE MEAN-ADJUSTED DATA

| Gene location | 5' | middle | 3' |
|---|---|---|---|
| Taiwan original | 6.607 | 6.381 | 6.723 |
| Taiwan converted | 6.624 | 6.396 | 6.746 |
| Duke original | 6.932 | 6.721 | 7.133 |

data conversion between each individual sample is shown in Table VI.

In the last study, the $\rho$'s are estimated from the original data. This results in values close to 1. In practice, sometimes we are really looking within a probe set or at replicate measurements. Under these cases, we use the mean-adjusted data by subtracting the group mean. Estimating $\rho$ from the mean-adjusted data gives us $\rho$ values close to 0. In Table VII, we also show the mean of the original Taiwan sample, the converted Taiwan sample, and the original Duke sample, on spike-in # 9 and gene 4. This time, however, the $\rho$ is estimated from the mean-adjusted data and is close to 0. From this table, we can see that the mean of the converted Taiwan data has been adjusted towards the original mean of the Duke data, but to a smaller degree than those shown in Table V. The detailed data conversion between each individual sample when $\rho$'s are measured using the mean-adjusted data, is shown in Table VIII.

TABLE VIII
THE CONVERTED DATA AND THE ORIGINAL TAIWAN SAMPLES FOR GENE 4, SPIKE-IN 9. TAIWAN SAMPLE MEASUREMENTS ARE INSIDE BRACES. THE $\rho$ ARE ESTIMATED USING THE MEAN-ADJUSTED DATA

| Gene location | 5' | middle | 3' |
|---|---|---|---|
| Taiwansample1 | 6.776(7.018) | 6.536(6.521) | 6.878(6.855) |
| Taiwansample2 | 6.791(6.774) | 6.548(6.533) | 6.899(6.876) |
| Taiwansample3 | 6.302(6.285) | 6.105(6.091) | 6.461(6.438) |

Finally, we show plots of the expression level data for pairs of genes. Fig. 5 demonstrates the shifts from the original Taiwan data to the converted Taiwan data, for two gene pairs: Gene 3 with Gene 4 and Gene 1 with Gene 4. The $\rho$'s are estimated using the original data. We see that the converted data represented by $\cdot$, are shifted closer to the original Duke data, denoted by $\diamond$. Fig. 6 shows the same plot for the data conversion, except that the $\rho$'s are estimated using the mean-adjusted data. Clearly, the conversion from the original Taiwan data to the virtual Taiwan data is fairly small. In fact, these results are quite reasonable. When the estimated $\rho$'s are large, we are more confident that shifting the data from one likelihood to another according to the same pattern is reasonable. However, when the estimated $\rho$'s are small, the data do not suggest that a substantial conversion is appropriate. This results in a minimal adjustment around the original data.

## VI. CONCLUSION

Here we have defined and studied a concept of effective samples, defining an informational sense in which two samples maybe essentially the same. The Bayesian formulation is merely pragmatic: Posteriors are often used for inference, but in fact any conditional distribution is amenable to our technique. Here, we have converted data for the sake of intelligibility or convenience. Formally, our approach is to take two likelihoods indexed by the same parameter. We think of one likelihood as

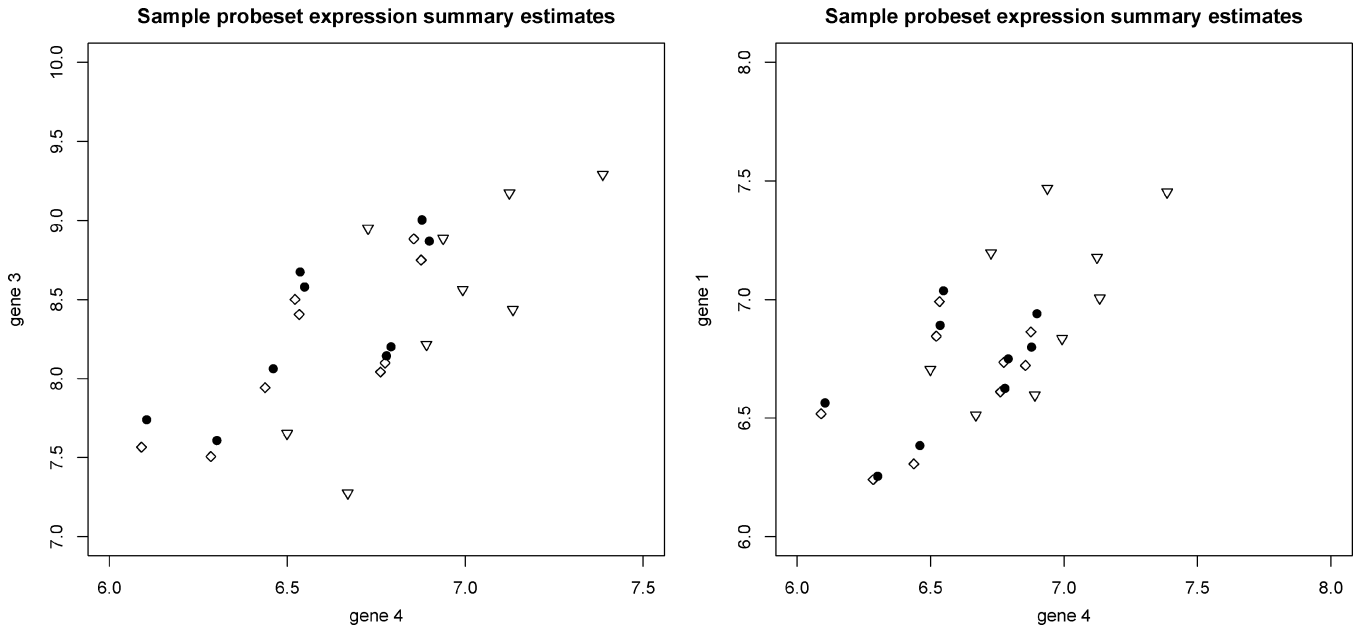


Fig. 6. The plot of probe set level expression for two genes, when $\rho$ estimated from the mean-adjusted data. The left side figure shows the plot for Gene 4 and Gene 3, and the right side figure shows the plot for Gene 4 and Gene 1. $\cdot$ denotes the converted Taiwan data, $\diamond$ denotes the original Taiwan data, and $\triangle$ denotes the original Duke data.

"true" and the other as chosen for convenience. When they are equipped with priors, the two posteriors can be compared. If the data set in the first posterior is fixed and the virtual data set in the second is allowed to vary, the relative entropy between the posteriors is a real-valued function of the vector-valued variable which we have regarded as virtual data. Minimizing this function gives a data set under the second likelihood that gives inferences under the second likelihood that are as close as possible to the first.

As a technique for data compression, in which the second likelihood is chosen to be simpler than the first, we have shown this optimization gives results that are intuitively reasonable and in some cases appealing. Moreover, by treating parameters as if they were data these ideas extend to give a concept of effective parameter size. This led us to the predictive structure we used in searching for an effective parameter size. Our goal there was model reduction: We sought to reduce the number of $\mu_i$'s needed to describe the data set. By automating the selection of $\epsilon$ in a reasonable way, it may be possible to represent our optimization as a function solely of the data. If so, this might be a reasonable choice for the data dependent penalty term in the Deviance Information Criterion (DIC) (Spiegelhalter *et al.* [29]). We have not investigated this in detail here because our focus was on data pooling.

Aside from the direct usefulness of effective samples, sample sizes, and parameter sizes, our approach raises several foundational issues.

First, we have used the relative entropy. If one measures information in a code length sense it is unquestionably the right distance. In fact, the relative entropy is used to define reference priors and we saw in Section III-C that reference priors are noninformative in a strong sense—perhaps so strong that the exquisite sensitivity of the relative entropy to tail behavior militates against its general use. It may be that information should be measured in some cases in a probabilistic sense, thereby leading to an $L1$ distance on posterior densities. As a generality, our intuition would carry over to $L1$, and many other distance measures.

Second, the posterior for the second entry in our definition is relatively unconstrained. Here, we have chosen comparison posteriors for their convenience and meaning. For instance, in the dependence example of Section III-B, the second posterior is formed using the product of marginals from the first likelihood. We justify this on the grounds that the independence model closest to any given model, in relative entropy, is the product of its marginals. In the mixture model of Section III-A we form the second posterior using one of the components of the mixture on the grounds that it would be the most informative. However reasonable these choices are, they are not systematic.

This gap can be filled by choosing a standardized reference likelihood. We leave this as an open problem. It would be intriguing if one could formulate an optimality principle to give likelihoods that could serve as a common yardstick for assessing the amount of information in a data set. One would want such an optimality principle to ensure that the posterior it gave would be close, in relative entropy, to the posterior from the "true" model. This would ensure the informational equivalence of the actual data and the virtual data.

Finally, and more abstractly, our optimization yields a posterior which is intended to be informationally equivalent to a "true" posterior. In effect, the minimization is a nonlinear transformation from one data set to another. It is unclear that there is any important functional relationship between the "true" data and the virtual data. Nevertheless, the two data sets possess the same information. Indeed, our results imply that any data set can be converted to many other informationally equivalent data sets

under different likelihoods. These (data, likelihood) pairs form an equivalence class and knowing the class of the data an experiment generated may be all the experiment can tell you.

## APPENDIX

*Proof of the Theorem:* Let us write

$$\varphi_1(\theta|\boldsymbol{x}^n) = \frac{|nI_p(\theta_0)|^{1/2}}{(2\pi)^{d/2}} \exp\left\{-\frac{n}{2}(\theta-\hat{\theta}_1)' I_p(\theta_0)(\theta-\hat{\theta}_1)\right\}$$

for the asymptotic normal approximation to a posterior. In this expression, $\hat{\theta}_1$ is the MLE from $p$ and $\boldsymbol{x}^n$. We also write

$$\varphi_2(\theta|\boldsymbol{y}^m) = \frac{|mI_q(\theta_0)|^{1/2}}{(2\pi)^{d/2}} \exp\left\{-\frac{m}{2}(\theta-\hat{\theta}_2)' I_q(\theta_0)(\theta-\hat{\theta}_2)\right\}$$

where $\hat{\theta}_2$ is the MLE from $q$ and $\boldsymbol{y}^m$. Consider the decomposition

$$D(w_p(\theta|\boldsymbol{x}^n)||w_q(\theta|\boldsymbol{y}^m)) = A + B + C$$

where

$$A = \int w_p(\theta|\boldsymbol{x}^n) \log \frac{w_p(\theta|\boldsymbol{x}^n)}{\varphi_1(\theta|\boldsymbol{x}^n)} d\theta$$

$$B = \int w_p(\theta|\boldsymbol{x}^n) \log \frac{\varphi_1(\theta|\boldsymbol{x}^n)}{\varphi_2(\theta|\boldsymbol{y}^m)} d\theta$$

$$C = \int w_p(\theta|\boldsymbol{x}^n) \log \frac{\varphi_2(\theta|\boldsymbol{y}^m)}{w_q(\theta|\boldsymbol{y}^m)} d\theta.$$

By Theorem 2.1 in Clarke [11], the first term $A \to 0$ in probability $P(\theta_0)$. Our task is to show that $C$ tends to zero, and then to recognize that $B$ gives the asymptotic form stated in the theorem.

We begin with $C$. Observe that $C$ can be written as $E+F-G$, where

$$E = \int_{\Omega} w_p(\theta|\boldsymbol{x}^n) \log \frac{\varphi_2(\theta|\boldsymbol{y}^m)}{w_q(\theta|\boldsymbol{y}^m)} d\theta$$

$$F = \int_{\Omega^C} w_p(\theta|\boldsymbol{x}^n) \log \varphi_2(\theta|\boldsymbol{y}^m) d\theta$$

$$G = \int_{\Omega^C} w_p(\theta|\boldsymbol{x}^n) \log w_q(\theta|\boldsymbol{y}^m) d\theta.$$

Here, $\Omega$ is a ball $B(\theta_0, r)$ containing $\theta_0$, and $\Omega^C$ is its complement. We show that each of the three terms tends to zero under the stated hypotheses, beginning with $G$ because it represents nonlocal values of $\theta$ and the interaction between the two priors. Note that we have permitted the priors $w_p$ and $w_q$ on $p$ and $q$ to be different, for the sake of greater generality.

By posterior consistency, we have $\forall \epsilon, r$, $\exists M_\eta$ such that $\forall m > M$

$$\begin{aligned} w_q(\theta|\boldsymbol{y}^m) &= \frac{w_q(\theta) q(\boldsymbol{y}^m|\theta)}{\int w_q(\theta) q(\boldsymbol{y}^m|\theta) d\theta} \\ &= \frac{w_q(\theta) q(\boldsymbol{y}^m|\theta)}{\int_\Omega w_q(\theta) q(\boldsymbol{y}^m|\theta) \left(1 + \frac{\int_{\Omega^C} w_q(\theta) q(\boldsymbol{y}^m|\theta) d\theta}{\int_\Omega w_q(\theta) q(\boldsymbol{y}^m|\theta) d\theta}\right)} \\ &\geq \frac{w_q(\theta) q(\boldsymbol{y}^m|\theta)}{(1+\epsilon)\int_\Omega w_q(\theta) q(\boldsymbol{y}^m|\theta) d\theta} \end{aligned}$$

holds with $P_{\theta_0}$-probability at least $1-\eta$, for any preassigned $\eta > 0$.

Because $\hat{\theta}_2$ is the MLE under $q_{\theta_0}$ we have that

$$\begin{aligned} q(\boldsymbol{y}^m|\theta) \leq q(\boldsymbol{y}^m|\hat{\theta}_2) &= \exp\left\{-\log \frac{1}{q(\boldsymbol{y}^m|\hat{\theta}_2)}\right\} \\ &= \exp\left\{-m\hat{H}(\hat{\theta}_2)\right\}. \end{aligned}$$

Letting $W_2(\Omega) = \int_\Omega w_q(\theta) d\theta$, we have

$$w_q(\theta|\boldsymbol{y}^m) \geq \frac{w_q(\theta) q(\boldsymbol{y}^m|\theta)}{(1+\epsilon)\exp\{-m\hat{H}(\hat{\theta}_2)\} W_2(\Omega)}.$$

Now, we lower-bound $G$ by two terms

$$\begin{aligned} G \geq &\int_{\Omega^C} w_p(\theta|\boldsymbol{x}^n) \log(w_q(\theta) q(\boldsymbol{y}^m|\theta)) d\theta \\ &- \left(\log(1+\epsilon) + \log(w_q(\Omega)) - m\hat{H}(\hat{\theta}_2)\right) w_p(\Omega^C|\boldsymbol{x}^n). \end{aligned}$$

Denote the first term as $G_1$ and the second term as $G_2$. By Assumption III, for any fixed radius $r > 0$ of $\Omega$ there is an $r' > 0$ so that

$$e^{nr'} w_p(\Omega^C|\boldsymbol{x}^n) \to 0 \text{ in probability } P_{\theta_0}.$$

By ML-consistency and Assumption II, we have

$$\hat{H}(\hat{\theta}_2) \to H(\theta_0) \text{ in probability } P_{\theta_0}.$$

Since we assumed that $m/n$ tends to a finite limit, $G_2 \to 0$ in the joint probability $P_{\theta_0} \otimes Q_{\theta_0}$ for $X^n$ and $Y^m$. Note that

$$\log q(\boldsymbol{y}^m|\theta) = m\left(\frac{1}{m}\sum_{i=1}^m \log \frac{1}{q(y_i|\theta)}\right) = -m\hat{H}_2(\theta)$$

so we have

$$G_1 = \int_{\Omega^C} w_p(\theta|\boldsymbol{x}^n) \log w_q(\theta) d\theta - \int_{\Omega^C} w_p(\theta|\boldsymbol{x}^n) m\hat{H}_2(\theta) d\theta.$$

Denote these integrals by $G_{11}$ and $G_{12}$, respectively.

To see $G_{11}$ goes to zero, let

$$G_{13} = \int_{\Omega^C} w_p(\theta|\boldsymbol{x}^n)(1 + |\log w_q(\theta)|) d\theta.$$

Clearly

$$0 \leq |G_{11}| \leq \int_{\Omega^C} w_p(\theta|\boldsymbol{x}^n) |\log w_q(\theta)| d\theta \leq G_{13}.$$

We employ a "change of prior" argument to show that $G_{13} \to 0$ in $P_{\theta_0}$-probability. Write

$$\begin{aligned} G_{13} &= \frac{\int_{\Omega^C} w_p(\theta) p(\boldsymbol{x}^n|\theta)(1+|\log w_q(\theta)|) d\theta}{\int w_p(\theta) p(\boldsymbol{x}^n|\theta)(1+|\log w_q(\theta)|) d\theta} \\ &\quad \times \frac{\int w_p(\theta) p(\boldsymbol{x}^n|\theta)(1+|\log w_q(\theta)|) d\theta}{\int w_p(\theta) p(\boldsymbol{x}^n|\theta) d\theta} \\ &= A(G) \times B(G). \end{aligned}$$

Our "new" prior is

$$w_H(\theta) = \frac{w_p(\theta)(1+|\log w_q(\theta)|)}{\int w_p(\theta)(1+|\log w_q(\theta)|) d\theta}$$

which is well defined because we have assumed second moments are finite. Denote the normalizing constant by

$$C_1 = \int w_p(\theta)(1 + |\log w_q(\theta)|)d\theta.$$

Now we see

$$A(G) = \frac{\int_{\Omega^C} w_H(\theta)p(\boldsymbol{x}^n|\theta)d\theta}{\int w_H(\theta)p(\boldsymbol{x}^n|\theta)d\theta} = w_H(\Omega^C|\boldsymbol{x}^n)$$
$$B(G) = \frac{C_1 \int w_H(\theta)p(\boldsymbol{x}^n|\theta)d\theta}{\int w_p(\theta)p(\boldsymbol{x}^n|\theta)d\theta}.$$

So

$$G_{13} = C_1 w_H(\Omega^C|\boldsymbol{x}^n)\frac{m_H(\boldsymbol{x}^n)}{m(\boldsymbol{x}^n)} = C_1 w_H(\Omega^C|\boldsymbol{x}^n)$$
$$\left(\frac{n^{d/2}|I^*(\hat{\theta})|^{1/2}m_H(\boldsymbol{x}^n)}{p(\boldsymbol{x}^n|\hat{\theta})(2\pi)^{d/2}} \times \frac{p(\boldsymbol{x}^n|\hat{\theta})(2\pi)^{d/2}}{n^{d/2}|I^*(\hat{\theta})|^{1/2}m(\boldsymbol{x}^n)}\right).$$

Proposition 5.1 of Clarke and Barron [8] uses a Laplace integration to show that for any mixture of i.i.d. densities with respect to a smooth prior $w$

$$\frac{p(\boldsymbol{x}^n|\hat{\theta})(2\pi)^{d/2}}{n^{d/2}|I^*(\hat{\theta})|^{1/2}m(\boldsymbol{x}^n)} \to \frac{1}{w(\theta_0)} \text{ in } p_{\theta_0}.$$

(This also follows by examining the proof in Walker [32].) Using this in both factors in parentheses of (12) shows that both are bounded in the limit. Since the posterior factor in (12) goes to 0 (exponentially fast) in $P_{\theta_0}$, it is seen that $|G_{11}|$ goes to 0 in $P_{\theta_0}$ as well.

Now consider the second term

$$G_{12} = -\int_{\Omega^C} w_p(\theta|\boldsymbol{x}^n)m\hat{H}_2(\theta)d\theta.$$

Since $\int w_p(\theta)H_2(\theta,\theta_0)d\theta < \infty$, the "change of prior" argument gives

$$m\left|\int_{\Omega^C} w_p(\theta|\boldsymbol{x}^n)H_2(\theta,\theta_0)d\theta\right| \to 0 \text{ in probability } P_{\theta_0} \otimes Q_{\theta_0}$$

because $m/n$ is controlled. Using Assumption IIin the integral of $G_{12}$ gives that $G_{12} \to 0$ in probability of $P(\theta_0) \otimes Q(\theta_0)$. Thus, $\liminf_{n,m\to\infty} = 0$, in probability of $P_{\theta_0} \otimes Q_{\theta_0}$.

Next, we upper-bound the limit of $G$ by zero as well. By restricting the domain,

$$w_q(\theta|\boldsymbol{y}^m) \le \frac{w_q(\theta)q(\boldsymbol{y}^m|\theta)}{\int_\Omega w_q(\theta)q(\boldsymbol{y}^m|\theta)d\theta}.$$

By Taylor expanding $\log q(\boldsymbol{y}^m|\theta)$ at $\hat{\theta}_2$ on $\Omega$ for $r$ small enough, we have

$$w_q(\theta|\boldsymbol{y}^m) \le \frac{w_q(\theta)q(\boldsymbol{y}^m|\theta)}{(1-o(1))w_q(\theta_0)\int_\Omega \exp\{-\frac{m}{2}(\theta-\hat{\theta}_2)'\hat{I}_2(\theta^{**})(\theta-\hat{\theta}_2)\}d\theta}$$

where $\theta^{**}$ is on the line joining $\hat{\theta}_2$ and $\theta$. So, for any given $\epsilon' > 0$, there is an $M_{\epsilon'}$ so that

$$\int_\Omega \exp\{-\frac{m}{2}(\theta-\hat{\theta}_2)'\hat{I}_2(\theta^{**})(\theta-\hat{\theta}_2)\}d\theta \ge (1-\epsilon')\int_{R^d} \exp\{-\frac{m}{2}(\theta-\hat{\theta}_2)'\hat{I}_2(\theta^{**})(\theta-\hat{\theta}_2)\}d\theta$$

in $Q_{\theta_0}$-probability, when $m > M_{\epsilon'}$. Doing the normal integration gives

$$w_q(\theta|\boldsymbol{y}^m) \le \frac{w_q(\theta)q(\boldsymbol{y}^m|\theta)(2\pi)^{d/2}}{(1-o(1))w_q(\theta_0)|m\hat{I}(\theta^{**})|^{1/2}}.$$

Using the last inequality in the definition of $G$ gives the upper bound

$$G \le \int_{\Omega^C} w_p(\theta|\boldsymbol{x}^n)\log\left(\frac{w_q(\theta)(2\pi)^{d/2}}{(1-o(1))|m\hat{I}(\theta^{**})|^{1/2}}\right)d\theta + \int_{\Omega^C} w_p(\theta|\boldsymbol{x}^n)\log q(\boldsymbol{y}^m|\theta)d\theta$$

for $m$ large enough. Again, the first term goes to 0 in probability by the change of prior argument. (Local supremum assumptions ensure $\hat{I}(\theta^{**}) \to I(\theta_0)$ in probability, see Clarke and Barron [8].) The second term is

$$\int_{\Omega^C} w_p(\theta|\boldsymbol{x}^n)m\hat{H}_2(\theta)d\theta$$

which goes to zero in probability by Assumption II. So, it is seen that, in the limit, $G$ is bounded above by 0 with high probability. Taken together, these results show $G \to 0$ in probability $P_{\theta_0} \bigotimes Q_{\theta_0}$.

Now we show $F$ goes to zero. Recall that

$$F = w_p(\Omega^C|\boldsymbol{x}^n)\log\frac{|mI_q(\theta_0)|^{1/2}}{(2\pi)^{d/2}} + \int_{\Omega^C} w_p(\theta|\boldsymbol{x}^n)(-\frac{m}{2}(\theta-\hat{\theta}_2)'I_p(\theta_0)(\theta-\hat{\theta}_2))d\theta.$$

The first term goes to 0 in probability $P_{\theta_0}$ because $m/n$ converges to a constant and the posterior probability decreases exponentially. It is enough to show the second term goes to zero.

Without loss of generality, consider the univariate case and define $\tilde{\theta_1} = E_p(\Theta|\boldsymbol{x}^n)$. (The multivariate case follows by the same arguments.) The absolute value of the second term is

$$\begin{aligned}|L| &= |\int_{\Omega^C} w_p(\theta|\boldsymbol{x}^n)(-\frac{m}{2}(\theta-\tilde{\theta_1}+\tilde{\theta_1}-\hat{\theta}_2)^2)d\theta| \\ &\le \int_{\Omega^C} w_p(\theta|\boldsymbol{x}^n)(m((\theta-\tilde{\theta_1})^2+(\tilde{\theta_1}-\hat{\theta}_2)^2))d\theta \\ &= mE(I_{\Omega^C}(\theta-\tilde{\theta_1})^2|\boldsymbol{x}^n) + mE(I_{\Omega^C}(\tilde{\theta_1}-\hat{\theta}_2)^2|\boldsymbol{x}^n).\end{aligned}$$

For the first term, we have

$$\begin{aligned}&mE(I_{\Omega^C}(\theta-\tilde{\theta_1})^2|\boldsymbol{x}^n) \\ &\le 2mE(I_{\Omega^C}(\theta^2+\tilde{\theta_1}^2)|\boldsymbol{x}^n)\end{aligned}$$

$$= 2m \int_{\Omega^C} \theta^2 w_p(\theta|\boldsymbol{x}^n) d\theta + 2m\tilde{\theta_1}^2 \int_{\Omega^C} w_p(\theta|\boldsymbol{x}^n) d\theta$$
$$\leq 2m \int_{\Omega^C} (1+\theta^2) w(\theta|\boldsymbol{x}^n) d\theta + 2m\tilde{\theta_1}^2 W_1(\Omega^C|\boldsymbol{x}^n).$$

By the change of prior argument, the first term in the last expression goes to 0 in probability $P_{\theta_0}$. Since $\tilde{\theta_1} \to \theta_0$, the second term in the last expression also goes to 0 in $P_{\theta_0}$-probability because the posterior probability is exponentially small. The second term in $|L|$ is bounded by

$$mE(I_{\Omega^C}(\tilde{\theta_1} - \hat{\theta}_2)^2 \boldsymbol{x}^n) \leq 2m(\tilde{\theta_1}^2 + \hat{\theta}_2^2) E(w_p(\Omega^C|\boldsymbol{x}^n).$$

Since $\hat{\theta}_2 \to \theta_0$ in $Q_{\theta_0}$-probability, the second term goes to 0 in probability of $P_{\theta_0} \bigotimes Q_{\theta_0}$. Combining these, $F \to 0$ in the joint probability of $P\theta_0 \otimes Q\theta_0$.

To finish showing $C$ goes to zero, we consider the term $E$. We begin with a lower bound. By Clarke [11, eq. 4.18] we have

$$\frac{w_q(\theta) q(\boldsymbol{y}^m)}{m_2(\boldsymbol{y}^m)} = w_q(\theta|\boldsymbol{y}^m)$$
$$\leq \frac{w_q(\theta)}{w_q(\hat{\theta}_2)} \exp\{-\frac{m}{2}(\theta - \hat{\theta}_2)' I_q^*(\hat{\theta})(\theta - \hat{\theta}_2)\}$$
$$\times \left(\frac{m(1+\tau)}{2\pi}\right)^{d/2} \frac{|I_q^*(\hat{\theta})|^{1/2}}{1 - e^{-m\epsilon''}}$$

with $Q_{\theta_0}$-probability at least $1 - O(1/m)$ for any positive constant $\epsilon''$ when $m > M_{\epsilon''}$. Thus

$$E \geq \int_\Omega w_p(\theta|\boldsymbol{x}^n) \{\log \frac{w_q(\hat{\theta}_2)}{w_q(\theta)} \frac{|I_q(\theta_0)|^{1/2}}{|I_q^*(\hat{\theta})|^{1/2}} (1 - e^{-m\epsilon''})$$
$$(1+\tau)^{-d/2} - \frac{m}{2}(\hat{\theta}_2 - \theta)'[I_q(\theta_0) - I_q^*(\hat{\theta})](\hat{\theta}_2 - \theta)\} d\theta.$$

First, we remove the dependence on $w_q$. By the continuity of $w_q$ on $\Omega$ we know that $\forall \epsilon$, $\exists r$ such that (s.t.) when $|\theta - \theta_0| < r$, $|w_q(\theta) - w_q(\theta_0)| < \epsilon$. So

$$w_q(\theta_0) - \epsilon \leq \sup_{|\theta - \theta_0| < r} w_q(\theta) \leq w_q(\theta_0) + \epsilon.$$

Since there is an $r$ for any preassigned $\epsilon$, and $\hat{\theta}_2 \to \theta_0$ in $P_{\theta_0}$

$$\int_\Omega w_p(\theta|\boldsymbol{x}^n) \log \frac{w_q(\hat{\theta}_2)}{w_q(\theta)} d\theta \to 0$$

in $P_{\theta_0} \otimes Q_{\theta_0}$.

Now we control the term with the difference in Fisher informations. Consider the integral

$$E_1 = \int_\Omega w_p(\theta|\boldsymbol{x}^n) \frac{m}{2}(\hat{\theta}_2 - \theta)'[I_q(\theta_0) - I_q^*(\hat{\theta})](\hat{\theta}_2 - \theta) d\theta.$$

It is

$$E_1 = [I_q(\theta_0) - I_q^*(\hat{\theta})] \frac{m}{2} \int_\Omega w_p(\theta|\boldsymbol{x}^n)(\hat{\theta}_2 - \theta)^2 d\theta$$
$$\leq [I_q(\theta_0) - I_q^*(\hat{\theta})] \frac{m}{2} \int_\Omega w_p(\theta|\boldsymbol{x}^n)[2(\hat{\theta}_2 - \hat{\theta}_1)^2$$
$$+ 2(\hat{\theta}_1 - \tilde{\theta}_1)^2 + 2(\tilde{\theta}_1 - \theta)^2] d\theta$$
$$\leq [I_q(\theta_0) - I_q^*(\hat{\theta})] \{m(\hat{\theta}_2 - \hat{\theta}_1)^2$$
$$+ m(\hat{\theta}_1 - \tilde{\theta}_1)^2 + m \int_\Omega w_p(\theta|\boldsymbol{x}^n)(\tilde{\theta}_1 - \theta)^2] d\theta\}.$$

(We used the fact that $\int_\Omega w(\theta|\boldsymbol{x}^n) d\theta \leq 1$.) By adding and subtracting $\theta_0$ in $(\hat{\theta}_2 - \hat{\theta}_1)^2$ and using the triangle inequality it is seen that the first term is bounded above by a sum of two $\chi^2$ random variables. When multiplied by $[I_q(\theta_0) - I_q^*(\hat{\theta})]$, a factor going to zero in probability, the corresponding term goes to zero in probability. Observing that $m(\hat{\theta}_1 - \tilde{\theta}_1)^2$ goes to zero in probability as well as $[I_q(\theta_0) - I_q^*(\hat{\theta})]$, again the corresponding term goes to zero. The third term is similar: By standard posterior normality results, $nVar(\Theta|\boldsymbol{x}^n) \to I_p^{-1}(\theta_0)$ in $P_{\theta_0}$-probability. Since $m/n \to C$ and the restriction to $\Omega$ only decreases the integral, again the convergence to zero of $[I_q(\theta_0) - I_q^*(\hat{\theta})]$ ensures the corresponding term goes to zero in probability. Thus, $E$ is lower-bounded, asymptotically, by zero.

We also show $E$ is upper-bounded, asymptotically, by zero. By Assumption III

$$m_2(\boldsymbol{y}^m) \leq (1+\epsilon) \int_\Omega w_q(\theta) q(\boldsymbol{y}^m|\theta) d\theta \text{ in } Q_{\theta_0} \qquad (12)$$

in which we may use any small positive number $\epsilon$. This choice is independent of previous $\epsilon$'s even though we have not used a different notation for it for convenience. Now, we get the expressions at the top of the following page .The form of the inequality above follows from (12) and Taylor expanding $\log q(\boldsymbol{y}^m|\theta)$ at $\hat{\theta}_2$.

To see that the upper bound goes to zero, first note the prior contribution is asymptotically negligible. As before, $w(\theta) \approx w(\theta_0)$ on $\Omega$. That is, if the radius $r$ of $\Omega$ is permitted to shrink slowly, the continuity of $w_q$ ensures that there exists $\delta > 0$ depending on $r$, such that $(1-\delta) w_q(\theta_0) \leq w_q(\theta_0) \leq (1+\delta) w_q(\theta_0)$ on $\Omega$. So the $w_q(\theta)$ in $H_3$ and $H_4$ cancel, leaving $\log \frac{1+\delta}{1-\delta} \to 0$ as $r$ goes to zero.

Next, the integral of $m(\theta - \hat{\theta}_2)'(I(\theta_0) - I^*(\theta^{**}))(\theta - \hat{\theta}_2)$ with respect to the posterior goes to zero because $I(\theta_0) - I^*(\theta^{**}) \to 0$, and the rest of the term is bounded by the posterior variance which converges to the inverse Fisher information.

The remaining two terms are seen to go to zero because the integral is

$$\int_\Omega \exp\left\{-\frac{m}{2}(\theta - \hat{\theta}_2)' I^*(\theta^{**})(\theta - \hat{\theta}_2)\right\} d\theta \approx \frac{|mI^*(\theta^{**})|^{1/2}}{(2\pi)^{d/2}}.$$

Because the normal concentrates on $\Omega$ and $I^*(\theta^{**})$ converge to $I(\theta_0)$, the approximation is exact in the limit. Now, $E$ is upper-bounded by zero, asymptotically.

The remaining task is to identify the limiting behavior of $B$. Using the form of $\phi_1$ and $\phi_2$, we see that $B$ is

$$\int w_p(\theta|\boldsymbol{x}^n) \{\frac{1}{2} \log \frac{nI_p(\theta_0)}{mI_q(\theta_0)} - \frac{n}{2}(\theta - \hat{\theta}_1)' I_p(\theta_0)(\theta - \hat{\theta}_1)$$
$$+ \frac{m}{2}(\theta - \hat{\theta}_2)' I_q(\theta_0)(\theta - \hat{\theta}_2)\} d\theta.$$

Let $\tilde{\theta}_1 = E(\Theta|\boldsymbol{x}^n)$; we have

$$(\theta - \hat{\theta}_1)' I_p(\theta_0)(\theta - \hat{\theta}_1)$$
$$= (\theta - \tilde{\theta}_1 + \tilde{\theta}_1 - \hat{\theta}_1)' I_p(\theta_0)(\theta - \tilde{\theta}_1 + \tilde{\theta}_1 - \hat{\theta}_1).$$

Again, without loss of generality, we consider only the one dimensional case. So, we have

$$2B = \log \frac{nI_p(\theta_0)}{mI_q(\theta_0)} - nI_p(\theta_0) \int w_p(\theta|\boldsymbol{x}^n)(\theta - \tilde{\theta}_1 + \tilde{\theta}_1 - \hat{\theta}_1)^2 d\theta$$
$$+ mI_q(\theta_0) \int w_p(\theta|\boldsymbol{x}^n)(\theta - \tilde{\theta}_1 + \tilde{\theta}_1 - \hat{\theta}_2)^2 d\theta.$$

$$
\begin{aligned}
E &= \int_\Omega w_p(\theta|\boldsymbol{x}^n)\left[\log\frac{\varphi_2(\theta|\boldsymbol{y}^m)m_2(\boldsymbol{y}^m)}{w_q(\theta)q(\boldsymbol{y}^m|\theta)}\right]d\theta \\
&\le \int_\Omega w_p(\theta|\boldsymbol{x}^n)\log H_1 d\theta + \int_\Omega w_p(\theta|\boldsymbol{x}^n)\log H_2 d\theta + \log(1+\epsilon) \\
&= \int_\Omega w_p(\theta|\boldsymbol{x}^n)H_3 d\theta + \int_\Omega w_p(\theta|\boldsymbol{x}^n)\log H_4 d\theta + \log(1+\epsilon)
\end{aligned}
$$

where

$$
\begin{aligned}
H_1 &= \frac{|mI_q(\theta_0)|^{1/2}\exp\{-\frac{m}{2}(\theta-\hat\theta_2)'I(\theta_0)(\theta-\hat\theta_2)\}}{(2\pi)^{d/2}w_q(\theta)q(\boldsymbol{y}^m|\hat\theta_2)\exp\{-\frac{m}{2}(\theta-\hat\theta_2)'I^*(\theta^{**})(\theta-\hat\theta_2)\}} \\
H_2 &= \int_\Omega w_q(\theta)q(\boldsymbol{y}^m|\hat\theta_2)\exp\{-\frac{m}{2}(\theta-\hat\theta_2)'I^*(\theta^{**})(\theta-\hat\theta_2)\} \\
H_3 &= \log\frac{|mI_q(\theta_0)|^{1/2}}{(2\pi)^{d/2}w_q(\theta)} - \frac{m}{2}(\theta-\hat\theta_2)'(I(\theta_0)-I^*(\theta^{**}))(\theta-\hat\theta_2)
\end{aligned}
$$

and

$$
H_4 = \int_\Omega w_q(\theta)\exp\{-\frac{m}{2}(\theta-\hat\theta_2)'I^*(\theta^{**})(\theta-\hat\theta_2)\}.
$$

---

Rearranging and noting cross terms are zero gives

$$
\begin{aligned}
2B = \log\frac{nI_p(\theta_0)}{mI_q(\theta_0)} &- nI_p(\theta_0)(Var(\theta|\boldsymbol{x}^n)+(\tilde\theta_1-\hat\theta_1)^2) \\
&+I_q(\theta_0)\frac{m}{n}n(Var(\theta|\boldsymbol{x}^n)+(\tilde\theta_1-\hat\theta_2)^2).
\end{aligned}
$$

Since $nVar(\theta|\boldsymbol{x}^n)\to I_p^{-1}(\theta_0)$ and $\sqrt{n}(\tilde\theta_1-\hat\theta_1)\to 0$ in $P_{\theta_0)}$

$$
nI_p(\theta_0)(Var(\theta|\boldsymbol{x}^n)+(\tilde\theta_1-\hat\theta_1)^2)\to 1 \text{ in } P_{\theta_0}.
$$

Similarly

$$
I_q(\theta_0)nVar(\theta|\boldsymbol{x}^n)\to\frac{I_q(\theta_0)}{I_p(\theta_0)} \text{ in } P_{\theta_0}.
$$

Thus

$$
B-\frac{m}{2}I_q(\theta_0)(\tilde\theta_1-\hat\theta_2)^2\to\frac{1}{2}\log\left(\frac{CI_p(\theta_0)}{I_q(\theta_0)}\right)+\frac{I_q(\theta_0)}{2CI_p(\theta_0)}-1/2
$$

in $P_{\theta_0}\otimes Q_{\theta_0}$-probability.

Combining the forgoing with the results that $A$ and $C$ go to 0, gives that

$$
\begin{aligned}
D(w_p(\theta|\boldsymbol{x}^n)||w_q(\theta|\boldsymbol{y}^m)) &- \frac{m}{2}I_q(\theta_0)(\tilde\theta_1-\hat\theta_2)^2 \\
&\to\frac{1}{2}\log\left(\frac{CI_p(\theta_0)}{I_q(\theta_0)}\right)+\frac{I_q(\theta_0)}{2CI_p(\theta_0)}-1/2
\end{aligned}
$$

in $P_{\theta_0}\otimes Q_{\theta_0}$.

## ACKNOWLEDGMENT

The authors gratefully acknowledge the insights and suggestions of the members of the Data Mining Program at SAMSI.

## REFERENCES


[1] C. J. Adcock, "Sample size determination: A review," *Statistician*, vol. 46, no. 2, pp. 261–283, 1997.

[2] Affymetrix , Affymetrix Microarry Suite Guide, 5th ed. Santa Clara, CA, Affymetrix, 2001.

[3] Affymetrix , Affymetrix Statistical Algorithms Description Document 5th ed. Santa Clara, CA, Affymetrix, 2002.

[4] M. Benito, J. Parker, Q. Du, J. Wu, D. Xiang, C. Perou, and J. Marron, "Adjustment of systematic microarray data biases," *Bioinformatics*, vol. 20, no. 1, pp. 105–114, 2004.

[5] P. J. Bickel and J. A. Yahav, "Some contributions to the asymptotic theory of Bayes solutions," *Z. Wahrsch. Verw. Geb.*, vol. 11, pp. 257–276, 1969.

[6] L. Breimen, *Probability*. Reading, MA: Addison-Wesley, 1968.

[7] K. C. Chanda, "A note on the consistency and maxima of the roots of the likelihood equations," *Biometrika*, vol. 41, pp. 56–61, 1954.

[8] B. Clarke and A. R. Barron, Information Theoretic Asymptotics of Bayes Methods, Univ. Illinois at Urbana-Champaign, Dept. Statistics, 1989, Tech. Rep.26.

[9] B. Clarke and A. R. Barrron, "Jeffrey's prior is asymptotically least favorable under entropic risk," *J. Statist. Planning and Inference*, vol. 41, pp. 37–60, 1994.

[10] B. Clarke, "Implications of reference priors for prior information and for sample size," *J. Amer. Statist. Assoc.*, vol. 91, no. 433, pp. 173–184, 1996.

[11] B. Clarke, "Asymptotic normality of the posterior in relative entropy," *IEEE Trans. Inf. Theory*, vol. 45, no. 1, pp. 165–176, Jan. 1999.

[12] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[13] J. C. Deville and C. E. Sarndal, "Calibration estimators in survey sampling," *J. Amer. Statist. Assoc.*, vol. 87, pp. 376–382, 1992.

[14] T. Gneiting, F. Balabdaoui, and A. E. Raftery, Probabilistic Forecasts, Calibration and Sharpness. Dept. Statistics, Univ. Washington, Seattle, Tech. Rep., 2005.

[15] D. Hoyle, M. Rattray, R. Jupp, and A. Brass, "Making sense of microarray data distributions," *Bioinformatics*, vol. 18, no. 4, pp. 576–584, 2002.

[16] R. Irizarry, B. Hobbs, F. Collin, Y. Beazer-Barclay, K. Antonellis, U. Scherf, and T. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, pp. 249–264, 2003.

[17] L. Joseph and P. Belisle, "Bayesian sample size determination for normal means and differences between normal means," *Statistician*, vol. 44, pp. 209–226, 1997.

[18] S. M. Kakade and D. P. Foster, "Deterministic calibration and Nash equilibrium," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 2004, vol. 3120, pp. 33–48.

[19] R. G. Krutchkoff, "Classical and inverse methods of calibration," *Technometrics*, vol. 9, pp. 525–539, 1967.

[20] S. Lee and M. Zelen, "Clinical trials and sample size considerations: Another perspective," *Statist. Sci.*, vol. 15, no. 2, pp. 95–110, 2000.

[21] C. Li and W. Wong, "Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection," *PNAS*, vol. 98, no. 1, pp. 31–36, 2001.

[22] D. V. Lindley, "The choice of sample size," *Statistician*, vol. 46, pp. 129–138, 1997.
[23] D. V. Lindley and A. F. M. Smith, "Bayes estimates for the linear model," *J. Roy. Statist. Soc., Ser. B*, vol. 34, pp. 1–41, 1972.
[24] C. Osborne, "Statistical calibration: A review," *Int. Statist.Rev.*, vol. 59, pp. 309–336, 1991.
[25] G. Parmigiani, E. Garrett, R. Anbazhagan, and E. Gabrielson, "A cross-study comparison of gene expression data sets for the molecular classification of lung cancer," *Clinical Cancer Res.*, vol. 10, pp. 2922–2927, 2004.
[26] R. Redner, "Note on the consistency of the maximum likelihood estimate for non-identifiable distributions," *Ann. Statist.*, vol. 9, pp. 225–288, 1981.
[27] B. Rubinstein, J. McAuliffe, S. Cawley, M. Palaniswami, K. Ramamaohanarao, and T. Speed, "Machine learning in low-level microarray analysis," *SIGKDD Explorations*, vol. 5, no. 2, pp. 130–139, 2003.
[28] V. K. Srivastava and N. Singh, "Small-disturbance asymptotic theory for linear-calibration estimators," *Technometrics*, vol. 31, pp. 373–378, 1989.
[29] D. J. Spiegelhalter, N. G. Best, B. P. Clarion, and A. van der Linde, "Bayesian measures of model complexity and fit," *J. Roy. Statist. Soc., Ser. B*, vol. 64, pp. 583–640, 2002.
[30] The Tumor Analysis Best Practices Working Group, "Expression profiling – Best practices for data generation and interpretation in clinical trials," *Nature Reviews Genetics*, vol. 5, pp. 229–237, 2004.
[31] A. Wald, "Note on the consistency of the maximum likelihood estimate," *Ann. Math. Statist.*, vol. 20, pp. 595–601, 1949.
[32] A. M. Walker, "On the asymptotic behaviour of posterior distributions," *J. Roy. Statist. Soc., Ser. B*, vol. 31, pp. 80–88, 1969.
[33] J. Wolfowitz, "Asymptotic efficiency of the maximum likelihood estimator," *Theory of Probab. and Applic.*, vol. 10, pp. 247–260, 1965.
[34] Z. Wu, R. Irizarry, R. Gentleman, R. Murillo, and F. Spencer, Adjustment for Oligonucleotide Expression Arrays. Dept. Biostat., Johns Hopkins Univ., Baltimore, MD, 2003, Dept of Biostat Working Papers.
[35] L. Zhang, M. Miles, and K. Aldape, "A model of molecular interactions on short oligonucleotide microarrays," *Nature Biotechnol.*, vol. 21, no. 7, pp. 818–821, 2003.