Bayesian Analysis (2021) **TBA**, Number TBA, pp. 1

Statistics, Models, and Likelihoods: Comments on a paper by Lewis, MacEachern, & Lee

Bertrand Clarke

1 Likelihood Selection

Arguably, the chief contribution of this paper is the computational technique given in Subsec. 3.2. This new technique is effective in the context of the factorization (3) given at the beginning of Subsec. 2.2. Secs. 1 and 2 provide the motivation for (3). Both the new computational technique and its motivation merit discussion. Here, we focus on the latter since the examples in the paper show that the computing technique is effective.

The motivation for (3) focuses on a treatment of outliers. Updating a prior using data that has outliers is a challenge to our standard conceptualization of simply choosing a model and prior to form a posterior because model selection is so much harder. There are standard techniques such as using a heavier tailed model that accommodates the outliers. The problem with this is that the model then reflects all the data including the data we don't trust. As a generality this weakens inference. Another standard technique is to isolate the outliers in the 'bad' component of a mixture distribution. The problem with this is that often it is not clear whether the outliers are indeed outlying. They may not fit comfortably with the other data but this cannot in general be distinguished from not fitting the proposed model for the 'good' component because it is mis-specified. A generalization of this technique, not as standard as it perhaps should be, is called cherry-picking introduced in House and Banks (2004) and developed in Banks et al. (2009). The idea is to construct a mixture model by fitting a model to a subset of the data that are in conformity with it, remove the data, and repeat the procedure until all the data is assigned to a model. The resulting mixture of models should be robust. One benefit of this strategy is that the models are used to cluster the data and the result can be investigated with standard model validation methods. The problem with this (in the view of some) is that the models are used as data summarization rather than proposed representations for the data generator (DG).

By contrast, Lewis et al. (2021) proposes to replace model selection treatments of outliers with a statistic selection treatment of outliers. This naturally necessitates a likelihood selection as well. One way to see the proposed procedure is as a generalization of sufficiency. Instead of writing

$$f(y \mid \theta) = g(T(y) \mid \theta)h(y) \tag{1}$$

for a density f, a parameter θ, a random variable Y, a statistic $T(y)$, a function g summarizing the dependence of T on θ, and function of the data $h(\cdot)$, write

$$f(y \mid \theta) = f(T(y) \mid \theta)f(y \mid \theta, T(y)). \tag{2}$$

*

(Unless otherwise specified we use the same notation as in Lewis et al. (2021).) The function $h(y)$ is obviously a special case of $f(y \mid \theta, T(y))$. Otherwise put, when T is sufficient $(Y \mid \theta, T) = (Y \mid T)$ i.e., $(Y \mid \theta, T)$ does not involve θ.

The idea behind using (2) rather than (1) is that T no longer has to be sufficient and therefore can be chosen to reduce the influence of outliers. Indeed, using an insufficient statistic may be better than using a sufficient statistic if the model cannot be assumed accurate to arbitrary precision, a situation that is typical not exceptional. In Lewis et al. (2021), Figs. 1 and 2, the authors give a variety of examples that condition parameters or future outcomes on several non-sufficient statistics and give better inference than using certain 'natural' models that have sufficient or asymptotically sufficient statistics. Since the focus in the paper is on outliers, using statistics that are robust may be more important than using statistics that are sufficient – even if they exist. Indeed, being able to drop θ as in (1) – sufficiency – may only be appropriate in models that are wrong since the true model if it exists need not have a sufficient statistic.

In this sense, the authors' proposal is to choose a conditioning statistic to compensate for inadequate model selection because statistics that are sufficient with respect to it may not encapsulate the inferential information in the data due to model bias. Indeed, the inferential information in the data may be model dependent. That is, some data may be outliers with respect to one model but not another.

2 Likelihoods vs. Models

Taking this one step further, there is no rule that says a likelihood has to come from a model that can be taken as true. A likelihood is simply a function of the parameter holding the data fixed. Techniques such as estimating equations take this line of thinking even further by proposing an optimization problem that may or may not be related to any model that might be taken as true. So, the authors' proposal should properly be termed likelihood selection as opposed to model selection or objective function selection. Otherwise put, the authors are proposing to choose a likelihood for a conditioning statistic (that they have also chosen) in the hope that it will extract the most important information in the data. This seems overall neither more nor less subjective than choosing a model class, prior, loss function, etc.

Thus, after choosing a statistic T, the authors choose a likelihood and proceed in the usual way to equip it with a prior, find the posterior given the conditioning statistic, and generate a predictive density. It is then the adequacy of predictions that are the true demonstration of how good a technique is.

One further benefit of this approach is that the main inputs it requires are T and a likelihood. So the authors' method can be seen as a technique for dealing with cases where no model exists. These are termed $\mathcal{M}$-open problems and they are ubiquitous. Recall, $\mathcal{M}$-closed problems are model selection (or predictor selection) problems in which the analyst must choose among finitely many alternatives, implicitly assuming one of them is the DG or objectively 'right' i.e., the selection of the best model/predictor is a source of error far smaller than any other source of errors. $\mathcal{M}$-complete problems are

those in which the analyst must choose among possibly countably many alterantives The assumption is that one of them is right – or at least most right in the sense of introducing negligible errors only – and may be best exhibited as a limit of wrong models (or predictors). In this case, the notion of a true model or best predictor – the two are nearly identical asymptotically, see Theorem 2 in Rissanen (1984)[1] and the discussion following – can be used conceptually but is not available in closed form. M-open problems are those for which there is no true model. This is the typical case because models are rarely (if ever) known to arbitrary precision and there are many problems for which it is implausible to assume a true model. The definition given here are modified from Bernardo and Smith (2000) to be disjoint.

One difference between $\mathcal{M}$-open and $\mathcal{M}$-complete problems is that expectations and convergence are well defined only in $\mathcal{M}$-complete problems. Also, the status of the prior is different in the two classes of problems. In $\mathcal{M}$-open problems we can redefine the prior to be some sort of weighting on 'models' treated as if they were actions giving predictions but expectations and modes of convergence must be replaced, for instance by predictive error. The general prequential approach see Dawid (1984), Dawid and Vovk (1999) and the Shtarkov solution, see Shtarkov (1987), or its Bayes counterpart, Le and Clarke (2016), are other examples of techniques appropriate for $\mathcal{M}$-open settings.

The authors' likelihood selection technique, based on a statistic, may also be useful for a special case at the complex end of $\mathcal{M}$-complete models where there is a true model but we are unble to formulate it in any realistic way, perhaps due to lack of data or other information. An example of this can be seen in one-way ANOVA. Even if the treatments can be regarded as identical, the subjects generally are not. There are subtle differences that may be important and in any realistic problem where we generate subjects we will not be able to identify a 'true model' for each of them, at least not to arbitrary precision. In the classic example of the treatment being a fertilizer and the subjects being plots of land it is easy to imagine small differences in soil composition, moisture, ambient weather, etc. that may be important. The best we can hope to do is to identify a model whose error can be safely assumed smaller than other sources of error. However, this is an assumption we can rarely verify. Taken together this means that although we can imagine a true model for the plots we cannot write it down. Thus, one-way ANOVA is an $\mathcal{M}$-complete problem that we typically approximate by an $\mathcal{M}$-closed problem. So, th authors' approach would apply to these problems as well as $\mathcal{M}$-open problems.

3 Choices, choices...

The most disconcerting aspect of the methodology proposed by Lewis et al. (2021) may be the freedom it seems to give to analysts. After all, it is hard to give general guidance as to how to choose a statistic or a likelihood for it well. On the other hand,

[1] Actually, Rissanen showed that in the ARMA case, the true model is the best predictor in the sense of achieving the minimal variance asymptotically. It not hard to see that this result generalizes readily to other model classes. An exception is that pre-asymptotically a good approximation to a true model may give a predictor that outperforms the predictor from true model because the true model has high variance as a result of its complexity.

adopting a prequential approach removes much of the seeming excess flexibility by imposing a predictive performance criterion. As argued elsewhere, e.g., Sec. 5 in Le and Clarke (2021), a method's predictive success is a measure how much we should trust it. Moreover, there are other efforts to 'square the circle' of merging interpretable modeling with black-box modeling; see Wang and Lin (2021).

With this in mind, suppose we have chosen a statistic that we think extracts the information from the data that we think is most relevant to our inferential goal. The question becomes how to assign a likelihood to it. In their paper Lewis et al. (2021) select a likelihood based on convenience or (coarse) physical modeling. However, it is important to note that the modeling is for the statistic not the data directly. The authors also note that a statistic and its asymptotic distribution could also be used.

Indeed, there are many statistics that are robust, asymptotically sufficient, and may provide good inference even if they are not efficient. A natural choice is to use order statistics. If $\dim(\theta) = d$ then one can choose d order statistics, condition on them, and obtain posterior normality. This is possible because any two percentiles are typically asymptotically independent in the $\mathcal{M}$-complete case when the joint distribution of the data is independent. For the special case $d = 1$, we have the following.

Let $X_1, \ldots, X_n, \ldots$ be a seqence of *i.i.d.* random variables with common density function $f_\theta(x)$ and distribution function $F_\theta(x)$, α be a constant, $0 \leq \alpha \leq 1$, and $l = [\alpha n]$, $b_n = l/(n+1), a_n = \sqrt{l(n-l+1)/(n+1)^3}$, and let $\mu(\theta) = F_\theta^{-1}(\alpha)$. Let Ω be a compact set such that $\inf_{\theta\in\Omega} w(\theta) \geq c > 0$, $f_\theta^{(i)}(x)$ be the $i-$th derivative of $f_\theta(x)$ w.r.t. x.

Theorem (Yuan and Clarke, 1999) Assume that $w(\theta)$ is continuous at the true parameter θ_0, and that μ'' exists for $\theta \in \Omega$ and that i) $\inf_{\theta\in\Omega} |\mu'(\theta)| > 0$, ii) $\sup_{\theta\in\Omega} |\mu'(\theta)| < \infty$, and iii) $\exists\ \delta > 0$ so that

$$\sup_{\theta\in\Omega} \sup_{x\in(-\delta,\delta))} |f''_\theta(F_\theta^{-1}(\alpha + x))| < \infty.$$

Then,

$$E_{\theta_0}|w(\theta|X_{l:n}) - N(\theta, \theta_0, \hat{\theta})|d\theta \to 0$$

where $\hat{\theta} = \mu^{-1}(X_{l:n})$, $N(\theta, \theta_0, \hat{\theta})$ is the density of normal distribution with mean $\hat{\theta}$ and variance $\sigma^2(\theta_0)\alpha(1-\alpha)/n(\mu'(\theta_0))^2$ and $\sigma^{-1}(\theta) = f_\theta(F_\theta^{-1}(\alpha))$.

The result and proof are a variation on Clarke and Ghosh (1995) and a special case of Yuan and Clarke (2004). So,if regularity conditions are satisfied and n is large enough, asymptotic normality can be invoked for use in (4) and (5) in Lewis et al. (2021). More generally, if $\dim(\theta) = d$, $w(\theta \mid \ell_1, \ldots, \ell_d) \to M(\theta_T, V)$ (in L^1) where V is a $d \times d$ diagonal matrix that can be given explicitly if desired. This can be extended to some wrong model analyses i.e., certain $\mathcal{M}$-closed or -complete cases because Berk (1970) can be extended as in Clarke and Le (2021) Appendix C.

A separate approach to assigning a likelihood follows from the concept of minimally informative likelihoods (MIL) – a sort of 'dual' concept to reference priors, see Clarke et al. (2014). The idea is, given a statistic, a loss function, and a prior, to choose a

likelihood, or in the parlance of information theory a channel, that provides optimal data compression subject to a distortion constraint i.e., a maximal tolerance on inaccuracy. The MIL achieves the rate distortion function lower bound for a given tolerance. Of course, allowing too large a tolerance means no information is retained and insisting on too small a tolerance means that the data compression will be too little to be helpful. To find the MIL requires the Blahut-Arimoto algorithm but provides a likelihood – a function of the parameter for fixed data – that can be fed into the framework of Lewis et al. (2021). Again, the statistic can be chosen by the analyst – although some statistics are easier to use than others. The MIL in principle loses the least important information in the data or equivalently adds the least information to the data via likelihood selection. The MIL can be generally used although the computing may be unstable in some cases.

Taken together, these two examples illustrate that choosing a statistic may often be enough for inference since the likelhihood can be found automatically, through asymptotics or optimization. Moreover, one can in principle evaluate the robustness of inference to statistic or likelihood selection by comparing asymptotic inference to the MIL and other choices for both the statistic and likelihood. Overall, asserting a model, as opposed to merely identifying a statistic and a likelihood that can be used pragmatically,may make inferences model-driven (and subjective) rather than data driven.

4 Two Final Thoughts

A theoretical gap that the authors might want to fill at some point concerns the computing. Specifically, much of the conditioning results in degenerate distributions in the sense that sets such as $\{T(y) = T(y_{obs}\}$ have measure zero in the overall measure space so conditioning on them must be done carefully to ensure the the conditional distributions are compatible from observed value to observed value. Careful conditioning arguments generally come down to the Radon-Nikodym theorem and fortunately are generally common-sense, at least once they are worked out. Can the authors explain their technique in these more formal terms or at least give the intuition to support its theoretical foundation?

A final thought that the authors might want to address is that one of the more valid criticisms of the Bayesian approach as compared to the frequentist approach is that exploratory data analysis (EDA) or initial data analysis (IDA) is much harder – indeed often not feasible – in the Bayesian paradigm. After all, the frequentist doesn't require a likelihood to compute and use meaningful summary statistics. However, the computational methodology in this paper, especially if formalized, amounts to making Bayesian EDA/IDA feasible. One can pick a statistic T (sufficient or not), assign a likelihood through modeling, asymptotics, or MIL's, and then find the posterior or predictive given that statistic. The frequentists can still do EDA/IDA faster (less demanding computationally) but now Bayesian EDA/IDA can be done routinely. So, how can we compare the frequentist EDA/IDA use of summary or desriptive statistics to a Bayesian approach for EDA/IDA based on 'summary' or 'descriptive' posteriors – posteriors based on statistics and likelihoods we can readily choose, at least in principle. Can the authors comment on what sort of results we should expect from a comparison of their Bayesian methodology for EDA/IDA to the established frequentist version?

References

Banks, D., House, L., and Kilhoury, K. (2009). "Cherry-Picking for Complex Data: Robust Structure Discovery." *Philosophical Transactions of the Royal Society, Series A*, 367: 4339–4359. 1

Berk, R. (1970). "Consistency a posteriori." *Ann. Math. Stat.*, 41: 894–906. 4

Bernardo, J. and Smith, A. (2000). *Bayesian Theory*. Chichester: John Wiley & Sons, 2 edition. 3

Clarke, B., Clarke, J., and Yu, C.-W. (2014). "Statistical Problem Classes and Their Links to Information Theory." 33: 337–371. 4

Clarke, B. and Ghosh, J. (1995). "Posterior convergence given the mean." *Ann. Statist.*, 23: 2116–2144. 4

Clarke, B. and Le, T. (2021). "Model averaging is asymptotically better than model selection for prediction." Submitted. 4

Dawid, A. P. (1984). "The prequential approach." *Journal of the Royal Statistical Society*, 147: 287–292. 3

Dawid, A. P. and Vovk, V. (1999). "Prequential probability: principles and properties." *Bernoulli*, 5: 125–162. 3

House, L. and Banks, D. (2004). "Cherry-Picking as a Robustness Tool." In Banks, House, Arabie, McMorris, and Gaul (eds.), *Classification, Cluster Analysis, and Data Mining*, 197–208. Berlin: Springer-Verlag. 1

Le, T. and Clarke, B. (2016). "Using the Bayesian Shtarkov solution for predictions." *Comp. Stat. and Data Analysis*, 104: 183–196. 3

— (2021). "Interpreting uninterpretable predictors: kernel methods, Shtarkov solutions, and random forests." *To appear: Stat. Theory and Related Fields,*. 4

Lewis, J., MacEachern, S., and Lee, Y. (2021). "Bayesian Restricted Likelihood Methods: Conditioning on Insufficient Statistics in Bayesian Regression." *Bayesian Analysis*, 16: 1–38. 1, 2, 3, 4, 5

Rissanen, J. (1984). "Universal Coding, Information, Prediction, and Estimation." *IEEE Trans. Inform. Theory*, 30: 629–636. 3

Shtarkov, Y. (1987). "Universal sequential coding of single messages." *Problems in Information Transmission*, 23: 3–17. 3

Wang, T. and Lin, Q. (2021). "Hybrid Predictive Models: When an Interpretable Model Collaborates with a Black-box Model." *J. Mach. Learning Res.*, 22: 1–38. 4

Yuan, A. and Clarke, B. (1999). "Posterior normality given order statistics." Unpublished manscript. 4

— (2004). "Asymptotic normality of the posterior given a statistic." *Can. J. Statist.*, 32: 119–137. 4