Bayesian Analysis (2021) **16**, Number 4, pp. 1393–1462

# Bayesian Restricted Likelihood Methods: Conditioning on Insufficient Statistics in Bayesian Regression (with Discussion)*

John R. Lewis†, Steven N. MacEachern‡, and Yoonkyung Lee§

**Abstract.** Bayesian methods have proven themselves to be successful across a wide range of scientific problems and have many well-documented advantages over competing methods. However, these methods run into difficulties for two major and prevalent classes of problems: handling data sets with outliers and dealing with model misspecification. We outline the drawbacks of previous solutions to both of these problems and propose a new method as an alternative. When working with the new method, the data is summarized through a set of insufficient statistics, targeting inferential quantities of interest, and the prior distribution is updated with the summary statistics rather than the complete data. By careful choice of conditioning statistics, we retain the main benefits of Bayesian methods while reducing the sensitivity of the analysis to features of the data not captured by the conditioning statistics. For reducing sensitivity to outliers, classical robust estimators (e.g., M-estimators) are natural choices for conditioning statistics. A major contribution of this work is the development of a data augmented Markov chain Monte Carlo (MCMC) algorithm for the linear model and a large class of summary statistics. We demonstrate the method on simulated and real data sets containing outliers and subject to model misspecification. Success is manifested in better predictive performance for data points of interest as compared to competing methods.

**Keywords:** Markov chain Monte Carlo, M-estimation, robust regression.

## 1 Introduction

Bayesian methods have provided successful solutions to a wide range of scientific problems, with their value having been demonstrated both empirically and theoretically. Bayesian inference relies on a model consisting of three elements: the prior distribution, the loss function, and the likelihood or sampling density. While formal optimality of Bayesian methods is unquestioned if one accepts the validity of all three of these elements, a healthy skepticism encourages us to question each of them. Concern about the

*This research has been supported by Nationwide Insurance Company and by the NSF under grant numbers DMS-10-07682, DMS-12-09194, DMS-15-13566, DMS-16-13110, SBE 19-21523, DMS-20-15490, and DMS-20-15552. The views in this paper are not necessarily those of Nationwide Insurance or the NSF.

†Department of Statistics, The Ohio State University, Columbus, Ohio 43210, lewis.865@buckeyemail.osu.edu

‡Department of Statistics, The Ohio State University, Columbus, Ohio 43210, snm@stat.osu.edu

§Department of Statistics, The Ohio State University, Columbus, Ohio 43210, yklee@stat.osu.edu

 https://doi.org/10.1214/21-BA1257

prior distribution has been addressed through the development of techniques for subjective elicitation (Garthwaite et al., 2005; O'Hagan et al., 2006) and objective Bayesian methods (Berger, 2006). Concern about the loss function is reflected in, for example, the extensive literature on Bayesian hypothesis tests (Kass and Raftery, 1995). The focus of this work is the development of techniques to handle imperfections in the likelihood $f(\boldsymbol{y}|\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\boldsymbol{y})$. Concern for imperfections in the likelihood are reflected in work considering minimally informative likelihoods (Yuan and Clarke, 1999), sensitivities of inferences to perturbations in the model (Zhu et al., 2011), the specification of a class of models and the use of Bayesian model averaging over the class (Clyde and George, 2004), and considerations of such averaging when the specified class may not contain the so-called true data generating model (Bernardo and Smith, 2000; Clyde and Iversen, 2013; Clarke et al., 2013).

Imperfection in the likelihood has also been widely discussed in the classical robustness literature. Hampel (1971), writing on the motivation for studies of robustness, provides a concise description of three mismatches between data and the model that purportedly gives rise to the data: "(i) rounding of the observations; (ii) the occurrence of gross errors; (iii) the model itself may only be an approximation to the underlying chance mechanism". In a Bayesian setting, the first is easily handled with MCMC methods through the introduction of a latent, unrounded variable into the model. We do not consider it here. The second and third are duals. Misspecification of the model (iii) will often make observations appear to be outliers (ii). The literature on robust methods is replete with examples described in terms of "outliers" where the central problem is model misspecification. In the sequel, we follow the tradition of referring to cases that are discordant with the stated model as "outliers", whether this discordance is due to gross error or a consequence of model misspecification.

In practice, the imperfections in a proposed likelihood often show themselves through the presence of outliers – whether due to local misspecification of the model or due to gross error. There are three main solutions to Bayesian outlier-handling. The first is to replace the basic sampling density with a mixture model which includes one component for the "good" data and a second component for the "bad" data. With this approach, the good component of the sampling density is used for prediction of future good data. The second approach replaces the basic sampling density with a thick-tailed density in an attempt to discount outliers, yielding techniques that often provide solid estimates of the center of the distribution but do not easily translate to predictive densities for further good data. The third approach fits a flexible (typically nonparametric) model to the data, producing a Bayesian version of a density estimate for both good and bad data. In recent development, inference is made through the use of robust inference functions (Lee and MacEachern, 2014).

These traditional strategies all have their drawbacks. The outlier-generating processes may be transitory in nature, constantly shifting as the source of bad data changes. This prevents us from appealing to large-sample arguments to claim that, with enough data, we can nail down a model for both good and bad data combined. Instead of attempting to model both good and bad data, we propose a novel strategy for handling outliers. In a nutshell, we begin with a complete model as if all of the data are good.

Rather than driving the move from prior to posterior by the full likelihood, we use only the likelihood driven by a few summary statistics which typically target inferential quantities of interest. We call this likelihood a restricted likelihood because conditioning is done on a restricted set of data; the set which satisfies the observed summary statistics. This restricted likelihood leads to a formal update of the prior distribution based on the sampling density of the summary statistics.

The advantages and disadvantages of the method are detailed throughout the paper using simulated and real data. One conceptual advantage of our method is that inferences and predictions are less sensitive to features of the data not captured by the conditioning statistics than are methods based on the complete likelihood. Choosing statistics targeting the main features of interest allows for inference that focuses on these features. The analysis can help to better understand other features which may not be captured by the conditioning statistics, such as outliers.

The examples in the paper provide a Bayesian analog of classical robust estimators. The main disadvantage of our methods relative to the classical estimators is computational. In Section 3 we detail a data-augmentation MCMC algorithm to fit the models proposed in this paper. The advantages are those of Bayesian methods. As is standard for Bayes-classical comparisons, the Bayesian method requires greater computational effort while providing better inference. As a referee notes, asymptotically, the Bayesian and classical parameter estimates are often very close and have the same limiting posterior variance / sampling variance. In situations where asymptotic approximation suffices, there is no need to use the computational techniques developed in this paper.

The remainder of the paper is as follows: Section 2 introduces the Bayesian restricted likelihood, provides context with previous work, and demonstrates some advantages of the methods on simple examples. Section 3 details an MCMC algorithm to apply the method to Bayesian linear models. This computational strategy is a major contribution to the work, providing an approach to apply the method on realistic examples. Many of the technical proofs are in the Supplementary Material (Lewis et al., 2021) with `R` code available from the authors. Sections 4 and 5 illustrate the method with simulated data and a real insurance industry data set containing many outliers with a novel twist on model evaluation. A discussion (Section 6) provides some final commentary on the new method. An R package `brlm` to implement our methods is available at `github.com/jrlewi/brlm`. Additionally all data and code for the examples in this paper are available at `https://github.com/jrlewi/brlm_paper/`.

# 2 Restricted Likelihood

## 2.1 Examples

To describe the use of the restricted likelihood, we begin with a pair of simple examples for the one-sample problem. For both, the model takes the data $\boldsymbol{y} = (y_1, \ldots, y_n)$ to be a random sample of size $n$ from a continuous distribution indexed by a parameter vector $\boldsymbol{\theta}$, with pdf $f(y|\boldsymbol{\theta})$. The standard, or full, likelihood is $L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$.

The first example considers the case where a known subset of the data is known to be bad in the sense of not informing us about $\boldsymbol{\theta}$. This case mimics the setting where outliers are identified and discarded before doing a formal analysis. Without loss of generality, we label the good cases 1 through $n-k$ and the bad cases $n-k+1$ through $n$. The relevant likelihood to be used to move from prior distribution to posterior distribution is clearly $L(\boldsymbol{\theta}|y_1,\ldots,y_{n-k}) = \prod_{i=1}^{n-k} f(y_i|\boldsymbol{\theta})$. For an equivalent analysis, we rewrite the full likelihood as the product of two pieces:

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \left(\prod_{i=1}^{n-k} f(y_i|\boldsymbol{\theta})\right)\left(\prod_{i=n-k+1}^{n} f(y_i|\boldsymbol{\theta})\right), \tag{1}$$

where the second factor may not actually depend on $\boldsymbol{\theta}$. We wish to keep the first factor and drop the second for better inference on $\boldsymbol{\theta}$.

The second example involves deliberate censoring of small and large observations. This is sometimes done as a precursor to the analysis of reaction time experiments (e.g., Ratcliff, 1993) where very small and large reaction times are physiologically implausible; explained by either anticipation or lack of attention of the subject. With lower and upper censoring times at $t_1$ and $t_2$, the post-censoring sampling distribution is of mixed form, with masses $F(t_1|\boldsymbol{\theta})$ at $t_1$ and $1-F(t_2|\boldsymbol{\theta})$ at $t_2$, and density $f(y|\boldsymbol{\theta})$ for $y \in (t_1,t_2)$. We adjust the original data $y_i$, producing $c(y_i)$ by defining $c(y_i) = t_1$ if $y_i \leq t_1$, $c(y_i) = t_2$ if $y_i \geq t_2$, and $c(y_i) = y_i$ otherwise. The adjusted update is performed with $L(\boldsymbol{\theta}|c(\boldsymbol{y}))$. Letting $g(t_1|\boldsymbol{\theta}) = F(t_1|\boldsymbol{\theta})$, $g(t_2|\boldsymbol{\theta}) = 1-F(t_2|\boldsymbol{\theta})$, and $g(y|\boldsymbol{\theta}) = f(y|\boldsymbol{\theta})$ for $y \in (t_1,t_2)$, we may rewrite the full likelihood as the product of two pieces

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \left(\prod_{i=1}^{n} g(c(y_i)|\boldsymbol{\theta})\right)\left(\prod_{i=1}^{n} f(y_i|\boldsymbol{\theta},c(y_i)).\right), \tag{2}$$

$\prod_{i=1}^{n} f(y_i|\boldsymbol{\theta},c(y_i))$ is the likelihood of the data conditioned on parameters and the summary statistic $c(\cdot)$ and recovers the piece of the full likelihood not in $\prod_{i=1}^{n} g(c(y_i)|\boldsymbol{\theta})$. Only the first part is retained in the analysis. Several more examples are detailed in Lewis (2014).

## 2.2 Generalization

To generalize the approach in (1) and (2), we write the full likelihood in two pieces with a conditioning statistic $T(\boldsymbol{y})$, as indicated below:

$$L(\boldsymbol{\theta}|\boldsymbol{y}) \quad = \quad f(T(\boldsymbol{y})|\boldsymbol{\theta})\ f(\boldsymbol{y}|\boldsymbol{\theta},T(\boldsymbol{y})). \tag{3}$$

Here, $f(T(\boldsymbol{y})|\boldsymbol{\theta})$ is the conditional pdf of $T(\boldsymbol{y})$ given $\boldsymbol{\theta}$ and $f(\boldsymbol{y}|\boldsymbol{\theta},T(\boldsymbol{y}))$ is the conditional pdf of $\boldsymbol{y}$ given $\boldsymbol{\theta}$ and $T(\boldsymbol{y})$. In the dropped case example, the conditioning statistic is $T(\boldsymbol{y}) = (y_1,\ldots,y_{n-k})$. In the censoring example, the conditioning statistic is $T(\boldsymbol{y}) = (c(y_1),\ldots,c(y_n))$. We refer to $f(T(\boldsymbol{y})|\boldsymbol{\theta})$ as the restricted likelihood and $L(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta})$ as the full likelihood.

Bayesian methods can make use of a restricted likelihood since $T(\boldsymbol{y})$ is a well-defined random variable with a probability distribution indexed by $\boldsymbol{\theta}$. This leads to the restricted likelihood posterior

$$\pi(\boldsymbol{\theta}|T(\boldsymbol{y})) \quad = \quad \frac{\pi(\boldsymbol{\theta})f(T(\boldsymbol{y})|\boldsymbol{\theta})}{m(T(\boldsymbol{y}))}, \tag{4}$$

where $m(T(\boldsymbol{y}))$ is the marginal distribution of $T(\boldsymbol{y})$ under the prior distribution. Predictive statements for further (good) data rely on the model. For another observation, say $y_{n+1}$, we would have the predictive density

$$f(y_{n+1}|T(\boldsymbol{y})) = \int f(y_{n+1}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|T(\boldsymbol{y}))\, d\boldsymbol{\theta}. \tag{5}$$

## 2.3 Literature Review

Our motivation for the use of summary statistics in Bayesian inference is concern about outliers or, more generally, model misspecification. Specifically, the likelihood is not specified correctly and concentrating on using well chosen parts of the data can help improve the analysis (e.g., Wong and Clarke, 2004). Direct use of restricted likelihood for this reason appears in many areas of the literature. For example, the use of rank likelihoods is discussed by Savage (1969), Pettitt (1983, 1982), and more recently by Hoff et al. (2013). Lewis et al. (2012) make use of order statistics and robust estimators as choices for $T(\boldsymbol{y})$ in the location-scale setting. Asymptotic properties of restricted posteriors are studied by Doksum and Lo (1990), Clarke and Ghosh (1995), Yuan and Clarke (2004), and Hwang et al. (2005). The tenor of these asymptotic results is that, for a variety of conditioning statistics with non-trivial regularity conditions on prior, model, and likelihood, the posterior distribution resembles the asymptotic sampling distribution of the conditioning statistic.

Restricted likelihoods have also been used as practical approximations to a full likelihood. For example, Pratt (1965) appeals to heuristic arguments regarding approximate sufficiency to justify the use of the restricted likelihood of the sample mean and standard deviation. Approximate sufficiency is also appealed to in the use of Approximate Bayesian Computation (ABC), which is related to our method. ABC is a collection of posterior approximation methods which has recently experienced success in applications to epidemiology, genetics, and quality control (see, for example, Tavaré et al., 1997; Pritchard et al., 1999; Beaumont et al., 2002; Marjoram et al., 2003; Fearnhead and Prangle, 2012; Drovandi et al., 2015). Interest typically lies in the full data posterior and ABC is used for computational convenience as an approximation. Consequently, effort is made to choose an approximately sufficient $T(\boldsymbol{y})$ and update to the ABC posterior by using the likelihood $L(\boldsymbol{\theta}|\mathcal{B}(\boldsymbol{y}))$, where $\mathcal{B}(\boldsymbol{y}) = \{\boldsymbol{y}^*|\rho(T(\boldsymbol{y}), T(\boldsymbol{y}^*)) \leq \epsilon\}$, $\rho$ is a metric, and $\epsilon$ is a tolerance level. This is the likelihood conditioned on the collection of data sets that result in a $T(\cdot)$ within $\epsilon$ of the observed $T(\boldsymbol{y})$. With an approximately sufficient $T(\cdot)$ and a small enough $\epsilon$, heuristically $L(\boldsymbol{\theta}|\mathcal{B}(\boldsymbol{y})) \approx L(\boldsymbol{\theta}|T(\boldsymbol{y})) \approx L(\boldsymbol{\theta}|\boldsymbol{y})$. Consequently, the ABC posterior approximates the full data posterior and efforts have been made to formalize what is meant by approximate sufficiency (e.g., Joyce and Marjoram, 2008).

Our method can be viewed as ABC with $\epsilon = 0$ and it is natural to compare it to ABC. This paper develops sampling methods for fitting Bayesian linear models conditioning exactly on a set of summary statistics ($\epsilon = 0$), even when the statistics follow a continuous distribution. Traditional ABC sampling methods are flexible and will, in general, apply to a broader class of models. The basic sampling method for ABC is the rejection sampling algorithm (Pritchard et al., 1999) which proposes a sample $\boldsymbol{\theta}^*$ from the prior, then new data $\boldsymbol{y}^*$ from the data-model given $\boldsymbol{\theta}^*$. The value $\boldsymbol{\theta}^*$ is accepted as a draw from the ABC posterior if $\rho(T(\boldsymbol{y}), T(\boldsymbol{y}^*)) \leq \epsilon$. Acceptance rates of this algorithm can be intolerably low and several extensions have been proposed to improve efficiency (see, for example, Beaumont et al., 2009; Turner and Van Zandt, 2012). The inefficiency of ABC algorithms is especially problematic in high-dimensional settings since generating high-dimensional statistics that are close to the observed values is difficult. Recently, Turner and Van Zandt (2014) developed the Gibbs ABC method which improves efficiency in the hierarchical setting by making use of conditional independence of the model to make accept/reject decisions at the individual group-level, effectively reducing the dimension of the problem to the number of parameters within each group. We revisit this approach in our comparisons to ABC in Section 5.2, finding that, for a modest increase in computational cost, we obtain an algorithm with better convergence and mixing properties. We also retain the desired posterior distribution – the posterior, having conditioned exactly on the summary statistics.

This work extends the development of Bayesian restricted likelihood by arguing that deliberate choice of an insufficient statistic $T(\boldsymbol{y})$ guided by targeted inference is sound practice. We also expand the class of conditioning statistics for which a formal Bayesian update can be achieved. Our methods do not rely on asymptotic properties, nor do they rely on approximate conditioning.

## 2.4 Illustrative Examples

Before discussing computational details, the method is applied to two simple examples on well known data sets to demonstrate its effectiveness in situations where outliers are a major concern. The full model in each case fits into the Bayesian linear regression framework discussed in Section 3. The first is an example (so far as we know) of gross error; the second is an example of model misspecification for a subset of the observations. The first example is an analysis of Simon Newcomb's 66 measurements of the passage time of light (Stigler, 1977); two of which are significant outliers in the lower tail. The full model is a standard location-scale Bayesian model also used in Lee and MacEachern (2014):

$$\beta \sim N(23.6, 2.04^2),\ \sigma^2 \sim IG(5, 10),\ y_i \overset{iid}{\sim} N(\beta, \sigma^2), i = 1, 2, \ldots, n = 66, \tag{6}$$

where $y_i$ denotes the $i^{th}$ (recorded) measurement of the passage time of light. $\beta$ is interpreted as the passage time of light with the deviations $y_i - \beta$ representing measurement error. Four versions of the restricted likelihood are fit with conditioning statistics: 1) Huber's M-estimator for location with Huber's 'proposal 2' for scale 2) Tukey's M-

estimator for location with Huber's 'proposal 2' for scale 3) LMS (least median squares) for location with associated estimator of scale and 4) LTS (least trimmed squares) for location with associated estimator of scale. Details of these estimators can be found in many places, including (Huber and Ronchetti, 2009). We return to the two M-estimators throughout this paper as we have found them to offer good default choices for practitioners dealing with outliers. A short review of these estimators is provided in the Supplementary Material. The tuning parameters for the M-estimators are chosen to achieve 95% efficiency under normality (Huber and Ronchetti, 2009) and, for comparability, roughly 5% of the residuals are trimmed for LTS. Two additional approaches to outlier handling are considered: 1) the normal distribution is replaced with a t-distribution and, 2) the normal distribution is replaced with a mixture of two normals. The t-model assumes $y_i \overset{iid}{\sim} t_\nu(\beta, \sigma^2)$ with $\nu = 5$. The prior on $\sigma^2$ is $IG(5, \frac{\nu-2}{\nu}10)$ and ensures that the prior on the variance is the same as the other models. The mixture takes the form: $y_i \overset{iid}{\sim} pN(\beta, \sigma^2) + (1-p)N(\beta, 10\sigma^2)$ with the prior $p \sim \text{beta}(20, 1)$ on the probability of belonging to the 'good' component.

The posterior of $\beta$ under each model appears in Figure 1. The posteriors group into two batches. The normal model and restricted likelihood with LMS do not discount the outliers and have posteriors centered at low values of $\beta$. These posteriors are also quite diffuse. In contrast, the t-model, mixture model, and the other restricted likelihood methods discount the outliers and have posteriors centered at higher values. There is modest variation among these centers. Posteriors in this second group have less dispersion than those in the first group. The pattern for predictive distributions differs (see bottom plot in Figure 1). The normal and t-models have widely dispersed predictive distributions. The other predictive distributions show much greater concentration. The restricted likelihood fits based on M-estimators (Tukey's and Huber's) are centered appropriately and are concentrated. The restricted likelihood based on LTS and the mixture model results are also centered appropriately, but comparatively less concentrated. The LMS predictive is concentrated, but it is poorly centered.

As a second example, a data set measuring the number of telephone calls in Belgium from 1950–1973 is analyzed. The outliers in this case are due to a change in measurement units on which calls were recorded for part of the data set. Specifically, for years 1964–1969 and parts of 1963 and 1970, the length of calls in minutes were recorded rather than the number of calls (Rousseeuw and Leroy, 1987). The full model is a standard normal Bayesian linear regression:

$$\boldsymbol{\beta} \sim N_2(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0),\ \sigma^2 \sim IG(a, b),\ \boldsymbol{y} \sim N(X\boldsymbol{\beta}, \sigma^2 I), \tag{7}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$, $\boldsymbol{y}$ is the vector of the logarithm of the number of calls, and $X$ is the $n \times 2$ design matrix with a vector of 1's in the first column and the year covariate in the second. In reality, the model should include a different piece for the part of the data with different units. The outliers are really just a manifestation of model misspecification. Prior parameters are fixed via a maximum likelihood fit to the first 3 data points. In particular, the prior covariance for $\boldsymbol{\beta}$ is set to $\Sigma_0 = g\sigma_0^2(X_p^\top X_p)^{-1}$, with $X_p$ the $3 \times 2$

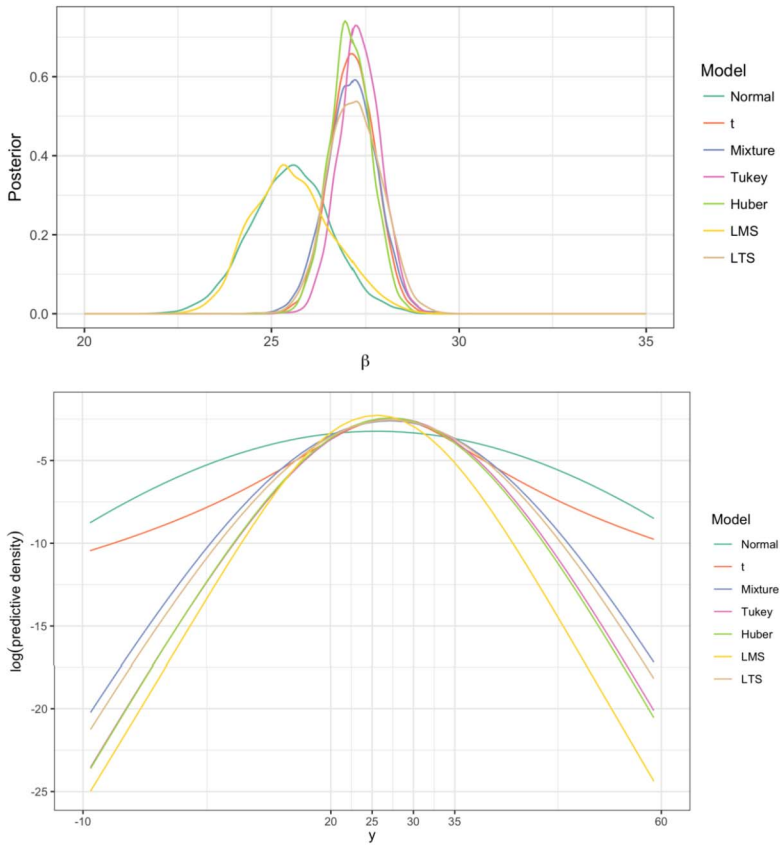


Figure 1: Results from the analysis of the speed of light data. Top: Posterior distributions of $\beta$ under each model. Bottom: Log posterior predictive distributions under each model. The differences in the tails are emphasized in the bottom plot. The horizontal axis is strategically labeled to help compare the centers of the distributions in each of the plots.

design matrix for the first 3 data points, $g = n = 21$, $\sigma_0 = 0.03$ and $\boldsymbol{\mu}_0 = (1.87, 0.03)^\top$. This has the spirit of a unit information prior (Kass and Wasserman, 1995) but uses a design matrix for data not used in the fit. Finally $a = 2$ and $b = 1$.

Four models are compared: 1) the normal theory base model 2) a two component normal mixture model, 3) a t-model, and 4) a restricted likelihood model conditioning on Tukey's M-estimator for the slope and intercept with Huber's 'proposal 2' for scale. Each model is fit to the remaining 21 data points. The normal theory model is also fit a second time after removing observations 14–21 (years 1963–1970). The omitted cases consist of the obvious large outliers as well as the two smaller outliers at the beginning and end of this sequence of points caused by the change in measurement units. The mixture model allows different mean regression functions and variances for each component. Both components have the same, relatively vague priors. The probability

of belonging to the first component is given a beta$(5, 1)$ prior. The heavy-tailed model fixes the degrees of freedom at 5 and uses the same prior on $\boldsymbol{\beta}$. The prior on $\sigma^2$ is adjusted by a scale factor of 3/5 to provide the same prior on the variance.

The data and 95% credible bands for the posterior predictive distribution under each model are displayed in Figure 2. The normal model fit to all cases results in a very wide posterior predictive distribution due to an inflated estimate of the variance. The t-model provides a similar predictive distribution. The pocket of outliers from 1963 to 1970 overwhelms the natural robustness of the model and leads to wide prediction bands. The outliers, falling toward the end of the time period, lead to a relatively high slope for the regression. In contrast, the normal theory model fit to only the good data results in a smaller slope and narrower prediction bands. The predictive distribution under the restricted likelihood approach is much more precise and is close to that of the normal theory fit to the non-outlying cases. The two component mixture model provides similar results, where the predictive distribution is formulated using only the good component. For these data, the large outliers are easily identified as following a distinct regression, leaving the primary component of the mixture for non-outlying data. In a more complex situation where the outlier generating mechanism is transient (i.e., ever changing and more complex than for these data), modeling the outliers is more difficult. As in classical robust estimation, the restricted likelihood approach avoids explicitly modeling the outliers.

# 3 Restricted Likelihood for the Linear Model

The simple examples in the previous section highlight the beneficial impact of a good choice of $T(\boldsymbol{y})$ with the use of the restricted likelihood. This work focuses on robustness in linear models where natural choices include many used above: M-estimators in the tradition of Huber (1964), least median squares (LMS), and least trimmed squares (LTS). For these choices the restricted likelihood is not available in closed form, making computation of the restricted posterior a challenge. For low-dimensional statistics $T(\boldsymbol{y})$ and parameters $\boldsymbol{\theta}$, the direct computational strategies described in Lewis (2014) can be used to estimate the restricted posterior conditioned on essentially any statistic. These strategies rely on estimation of the density of $f(T(\boldsymbol{y})|\theta)$ using samples of $T(\boldsymbol{y})$ for many values of $\boldsymbol{\theta}$; a strategy which breaks down in higher dimensions. This section outlines a data augmented MCMC algorithm that can be applied to the Bayesian linear model when $T(\boldsymbol{y})$ consists of estimates of the regression coefficients and scale parameter.

## 3.1 The Bayesian Linear Model

We focus on the use of restricted likelihood for the Bayesian linear model with a standard formulation:

$$\begin{aligned} \boldsymbol{\theta} &= (\boldsymbol{\beta}, \sigma^2) \sim \pi(\boldsymbol{\theta}) \\ y_i &= x_i^\top \boldsymbol{\beta} + \epsilon_i, \text{ for } i = 1, \ldots, n \end{aligned} \tag{8}$$

where $x_i$ and $\boldsymbol{\beta} \in \mathbb{R}^p$, $\sigma^2 \in \mathbb{R}^+$, and the $\epsilon_i$ are independent draws from a distribution with center 0 and scale $\sigma$. $X$ denotes the design matrix whose rows are $x_i^\top$.

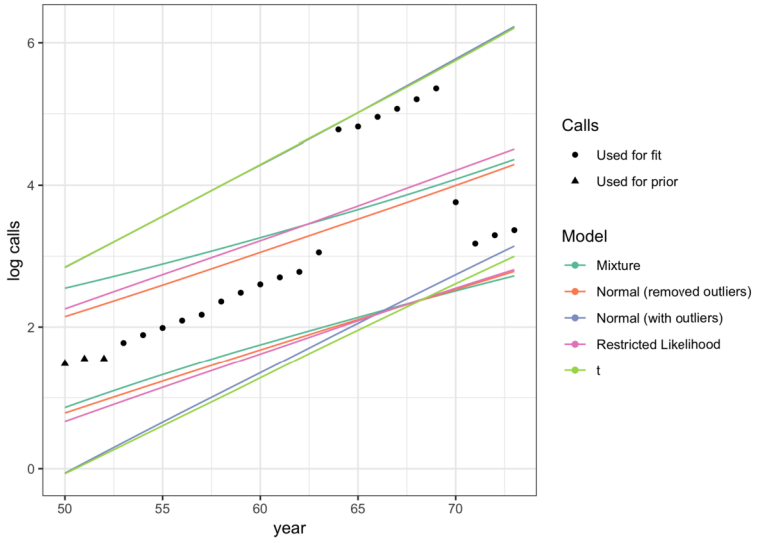


Figure 2: Pointwise posterior predictive intervals of log(calls) under the normal theory model fit to the non-outliers, the restricted likelihood model with Tukey's M-estimator for the slope and intercept with Huber's 'proposal 2' for scale, and a heavy-tailed t-distribution model. The first three data points were used to specify the prior with each model using the remaining 21 for fitting. The normal theory model was also fit after removing observations 14–20 (years 1963–1970).

For the restricted likelihood model, conditioning statistics are assumed to be of the form $T(\boldsymbol{y}) = (\boldsymbol{b}(X, \boldsymbol{y}), s(X, \boldsymbol{y}))$ where $\boldsymbol{b}(X, \boldsymbol{y}) = (b_1(X, \boldsymbol{y}), \ldots, b_p(X, \boldsymbol{y}))^\top \in \mathbb{R}^p$ is an estimator for the regression coefficients and $s(X, \boldsymbol{y}) \in \{0\} \cup \mathbb{R}^+$ is an estimator of the scale. Throughout, observed data and summary statistic is denoted by $\boldsymbol{y}_{obs}$ and $T(\boldsymbol{y}_{obs}) = (\boldsymbol{b}(X, \boldsymbol{y}_{obs}), s(X, \boldsymbol{y}_{obs}))$, respectively. Several conditions are imposed on the model and statistic to ensure validity of the MCMC algorithm:

**C1**. The $n \times p$ design matrix, $X$, whose $i^{th}$ row is $x_i^\top$, is of full column rank.

**C2**. The $\epsilon_i$ are a random sample from some distribution which has a density with respect to Lebesgue measure on the real line and for which the support is the real line.

**C3**. $\boldsymbol{b}(X, \boldsymbol{y})$ is almost surely continuous and differentiable with respect to $\boldsymbol{y}$.

**C4**. $s(X, \boldsymbol{y})$ is almost surely positive, continuous, and differentiable with respect to $\boldsymbol{y}$.

**C5**. $\boldsymbol{b}(X, \boldsymbol{y} + X\boldsymbol{v}) = \boldsymbol{b}(X, \boldsymbol{y}) + \boldsymbol{v}$ for all $\boldsymbol{v} \in \mathbb{R}^p$.

**C6**. $\boldsymbol{b}(X, a\boldsymbol{y}) = a\boldsymbol{b}(X, \boldsymbol{y})$ for all constants $a$.

**C7**. $s(X, \boldsymbol{y} + X\boldsymbol{v}) = s(X, \boldsymbol{y})$ for all $\boldsymbol{v} \in \mathbb{R}^p$.

**C8**. $s(X, a\boldsymbol{y}) = |a|s(X, \boldsymbol{y})$ for all constants $a$.

Properties C5 and C6 of $\boldsymbol{b}$ are called *regression* and *scale equivariance*, respectively. Properties C7 and C8 of $s$ are called *regression invariance* and *scale equivariance*. Many estimators satisfy the above properties, including several traditional simultaneous M-estimators (Huber and Ronchetti, 2009; Maronna et al., 2006) for which the `R` package `brlm` (`github.com/jrlewi/brlm`) is available to implement the MCMC described here. These M-estimators satisfy C3 and C4 since they are optimizers of continuous and differentiable objective functions. Constraints C5–C8 are often satisfied by location and scale estimators but should be checked on a case by case basis. More software development is required to extend the MCMC implementation beyond the M-estimators discussed here. The current version of the `R` package also implements the direct methods described in Lewis (2014). These methods are effective in lower dimensional problems and were used in both examples in Section 2.4.

## 3.2 Computational Strategy

The general style of algorithm we present is a data augmented MCMC targeting $f(\boldsymbol{\theta}, \boldsymbol{y}|T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs}))$, the joint distribution of $\boldsymbol{\theta}$ and the full data given the summary statistic $T(\boldsymbol{y}_{obs})$. The Gibbs sampler (Gelfand and Smith, 1990) iteratively samples from the full conditionals 1) $\pi(\boldsymbol{\theta}|\boldsymbol{y}, T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs}))$ and 2) $f(\boldsymbol{y}|\boldsymbol{\theta}, T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs}))$. When $\boldsymbol{y}$ has the summary statistic $T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})$, the first full conditional is the same as the full data posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. In this case, the condition $T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})$ is redundant. This allows us to make use of conventional MCMC steps for generation of $\boldsymbol{\theta}$ from the first full conditional. For typical regression models, algorithms abound. Details of the recommended algorithms depend on details of the prior distribution and sampling density and we assume this can be done (see e.g., Liu, 1994; Liang et al., 2008).

For a typical model and conditioning statistic, the second full conditional $f(\boldsymbol{y}|\boldsymbol{\theta}, T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs}))$ is not available in closed form. We turn to Metropolis-Hastings (Hastings, 1970), using the strategy of proposing full data $\boldsymbol{y} \in \mathcal{A} := \{\boldsymbol{y} \in \mathbb{R}^n | T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})\}$ from a well defined distribution with support $\mathcal{A}$ and either accepting or rejecting the proposal. Let $\boldsymbol{y}_p, \boldsymbol{y}_c \in \mathcal{A}$ represent the proposed and current full data, respectively. Denote the proposal distribution for $\boldsymbol{y}_p$ by $p(\boldsymbol{y}_p|\boldsymbol{\theta}, T(\boldsymbol{y}_p) = T(\boldsymbol{y}_{obs})) = p(\boldsymbol{y}_p|\boldsymbol{\theta}, \boldsymbol{y}_p \in \mathcal{A}) = p(\boldsymbol{y}_p|\boldsymbol{\theta})$. The last equality follows from the fact that our $p(\cdot|\boldsymbol{\theta})$ assigns probability one to the event $\{\boldsymbol{y}_p \in \mathcal{A}\}$. These equalities still hold if the dummy argument $\boldsymbol{y}_p$ is replaced with $\boldsymbol{y}_c$. The conditional density is

$$f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{y} \in \mathcal{A}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta}) I(\boldsymbol{y} \in \mathcal{A})}{\int_{\mathcal{A}} f(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{y}} = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})}{\int_{\mathcal{A}} f(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{y}}$$

for $\boldsymbol{y} \in \mathcal{A}$ and $I(\cdot)$ the indicator function. This includes both $\boldsymbol{y}_p$ and $\boldsymbol{y}_c$. The Metropolis-Hastings acceptance probability is the minimum of 1 and $R$, where

$$R = \frac{f(\boldsymbol{y}_p|\boldsymbol{\theta}, \boldsymbol{y}_p \in \mathcal{A})}{f(\boldsymbol{y}_c|\boldsymbol{\theta}, \boldsymbol{y}_c \in \mathcal{A})} \frac{p(\boldsymbol{y}_c|\boldsymbol{\theta}, \boldsymbol{y}_c \in \mathcal{A})}{p(\boldsymbol{y}_p|\boldsymbol{\theta}, \boldsymbol{y}_p \in \mathcal{A})} \tag{9}$$

$$= \frac{f(\boldsymbol{y}_p|\boldsymbol{\theta})}{\int_{\mathcal{A}} f(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{y}} \frac{\int_{\mathcal{A}} f(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{y}}{f(\boldsymbol{y}_c|\boldsymbol{\theta})} \frac{p(\boldsymbol{y}_c|\boldsymbol{\theta})}{p(\boldsymbol{y}_p|\boldsymbol{\theta})} \tag{10}$$

$$= \frac{f(\boldsymbol{y}_p|\boldsymbol{\theta})}{f(\boldsymbol{y}_c|\boldsymbol{\theta})} \frac{p(\boldsymbol{y}_c|\boldsymbol{\theta})}{p(\boldsymbol{y}_p|\boldsymbol{\theta})}. \tag{11}$$

For the models we consider, evaluation of $f(\boldsymbol{y}|\boldsymbol{\theta})$ is straightforward. Therefore, the difficulty in implementing this Metropolis-Hastings step manifests itself in the ability to both simulate from and evaluate $p(\boldsymbol{y}_p|\boldsymbol{\theta})$ – the well defined distribution with support $\mathcal{A}$. We now discuss such an implementation method for the linear model in (8).

### Construction of the Proposal

Our computational strategy relies on proposing $\boldsymbol{y}$ such that $T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})$ where $T(\cdot) = (\boldsymbol{b}(X,\cdot), s(X,\cdot))$ satisfies the conditions C3–C8. It is not a simple matter to do this directly, but with the specified conditions, it is possible to scale and shift any $\boldsymbol{z}^* \in \mathbb{R}^n$ which generates a positive scale estimate to such a $\boldsymbol{y}$ via the following theorem, whose proof is in the Supplementary Material.

**Theorem 3.1.** *Assume that conditions C4–C8 hold. Then, any vector $\boldsymbol{z}^* \in \mathbb{R}^n$ with conditioning statistic $T(\boldsymbol{z}^*)$ for which $s(X,\boldsymbol{z}^*) > 0$ can be transformed into $\boldsymbol{y}$ with conditioning statistic $T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})$ through the transformation*

$$\boldsymbol{y} = h(\boldsymbol{z}^*) := \frac{s(X,\boldsymbol{y}_{obs})}{s(X,\boldsymbol{z}^*)}\boldsymbol{z}^* + X\left(\boldsymbol{b}(X,\boldsymbol{y}_{obs}) - \boldsymbol{b}(X, \frac{s(X,\boldsymbol{y}_{obs})}{s(X,\boldsymbol{z}^*)}\boldsymbol{z}^*)\right).$$

Using the theorem, the general idea is to first start with an initial vector $\boldsymbol{z}^*$ drawn from a known distribution, say $p(\boldsymbol{z}^*)$, and transform via $h(\cdot)$ to $\boldsymbol{y} \in \mathcal{A}$. The proposal density $p(\boldsymbol{y}|\boldsymbol{\theta})$ is then a change-of-variables adjustment on $p(\boldsymbol{z}^*)$ derived from $h(\cdot)$. In general however, the mapping $h(\cdot)$ is many-to-one: for any $\boldsymbol{v} \in \mathbb{R}^n$ and any $c \in \mathbb{R}^+$, $c\boldsymbol{z}^* + X\boldsymbol{v}$ map to the same $\boldsymbol{y}$. This makes the change-of-variables adjustment difficult. We handle this by first noticing that the set $\mathcal{A}$ is an $n-p-1$ dimensional space: there are $p$ constraints imposed by the regression coefficients and one further constraint imposed by the scale. Hence, we restrict the initial $\boldsymbol{z}^*$ to an easily understood $n-p-1$ dimensional space. Specifically, this space is the unit sphere in the orthogonal complement of the column space of the design matrix: $\mathbb{S} := \{\boldsymbol{z}^* \in \mathcal{C}^{\perp}(X) \mid ||\boldsymbol{z}^*|| = 1\}$, where $\mathcal{C}(X)$ and $\mathcal{C}^{\perp}(X)$ are the column space of $X$ and its orthogonal complement, respectively. The mapping $h : \mathbb{S} \to \mathcal{A}$ is one-to-one and onto. A proof is provided by Theorem 1.1 of the Supplementary Material. The one-to-one property makes the change of variables more feasible. The onto property is important so that the support of the proposal distribution (i.e. the range of $h(\cdot)$) contains the support of the target $f(\boldsymbol{y}|\theta, y \in \mathcal{A})$, a necessary condition for convergence of the Metropolis-Hastings algorithm (in this case the supports are both $\mathcal{A}$).

Given the one-to-one and onto mapping $h : \mathbb{S} \to \mathcal{A}$, the general proposal strategy is summarized as follows:

1. Sample $\boldsymbol{z}^*$ from a distribution with known density whose support is the entirety of $\mathbb{S}$.
2. Set $\boldsymbol{y} = h(\boldsymbol{z}^*)$ and calculate the Jacobian of this transformation in two steps.
   (a) Scale from $\mathbb{S}$ to the set $\Pi(\mathcal{A}) := \{\boldsymbol{z} \in \mathbb{R}^n | \ \exists\, \boldsymbol{y} \in \mathcal{A}\ s.t.\ \boldsymbol{z} = Q\boldsymbol{y}\}$ with $Q = I - XX^\top$.[1] $\Pi(\mathcal{A})$ is the projection of $\mathcal{A}$ onto $\mathcal{C}^\perp(X)$ and, by condition C7, every element of this set has $s(X, \boldsymbol{z}) = s(X, \boldsymbol{y}_{obs})$. Specifically, set $\boldsymbol{z} = \frac{s(X,\boldsymbol{y}_{obs})}{s(X,\boldsymbol{z}^*)}\boldsymbol{z}^*$. There are two pieces of this Jacobian: one for the scaling and one for the mapping of the sphere onto $\Pi(\mathcal{A})$. The latter piece is given in equation (12).
   (b) Shift from $\Pi(\mathcal{A})$ to $\mathcal{A}$: $\boldsymbol{y} = \boldsymbol{z} + X\left(\boldsymbol{b}(X, \boldsymbol{y}_{obs}) - \boldsymbol{b}(X, \boldsymbol{z})\right)$. This shift is along the column space of $X$ to the unique element in $\mathcal{A}$. The Jacobian of this transformation is given by equation (13).

The final proposal distribution including the complete Jacobian is given in equation (14) with details in the next section. Before giving these details we provide a visualization in Figure 3 of each of the sets described above using a notional example to aid in the understanding of the strategy we take. In the figure, $n = 3$, $p = 1$, and the conditioning statistic is $T(\boldsymbol{y}) = (\min(\boldsymbol{y}), \sum(y_i - \min(\boldsymbol{y}))^2)$. The set $\mathcal{A}$ is depicted for $T(\boldsymbol{y}_{obs}) = (0, 1)$ which we describe as a "warped triangle" in light blue, with each side corresponding to a particular coordinate of $\boldsymbol{y}$ being the minimum value of zero. The other two coordinates are restricted by the scale statistic to lie on the quarter circle of radius one in the positive orthant. In this example, the column vector $X = \mathbf{1}$ (shown as a reference) spans $\mathcal{C}(X)$ and $\mathbb{S}$ is a unit circle on the orthogonal plane (shown in red). $\Pi(\mathcal{A})$ is depicted as the bowed triangle in dark blue. We will come back to this artificial example in the next section in an attempt to visualize the Jacobian calculations.

### Evaluation of the Proposal Density

We now explain each step in computing the Jacobian described above.

**Scale from $\mathbb{S}$ to $\Pi(\mathcal{A})$** The first step is constrained to $\mathcal{C}^\perp(X)$ and scales the initial $\boldsymbol{z}^*$ to $\boldsymbol{z} = \frac{s(X,\boldsymbol{y}_{obs})}{s(X,\boldsymbol{z}^*)}\boldsymbol{z}^*$. For the Jacobian, we consider two substeps: first, the distribution on $\mathbb{S}$ is transformed to that along a sphere of radius $r = \|\boldsymbol{z}\| = s(X, \boldsymbol{y}_{obs})/s(X, \boldsymbol{z}^*)$. By comparison of the volumes of these spheres, this transformation contributes a factor of $r^{-(n-p-1)}$ to the Jacobian. For the second substep, the sphere of radius $r$ is deformed onto $\Pi(\mathcal{A})$. This deformation contributes an attenuation to the Jacobian equal to the ratio of infinitesimal volumes in the tangent spaces of the sphere and $\Pi(\mathcal{A})$ at $\boldsymbol{z}$. Restricting to $\mathcal{C}^\perp(X)$, this ratio is the cosine of the angle between the normal vectors of the two sets at $\boldsymbol{z}$. The normal to the sphere is its radius vector $\boldsymbol{z}$. The normal to $\Pi(\mathcal{A})$ is given in the following lemma with proof provided in the Supplementary Material. Gradients denoted by $\nabla$ are with respect to the data vector.

[1] We have used condition C1 to assume without loss of generality that the columns of $X$ form an orthonormal basis for $\mathcal{C}(X)$ (i.e., $X^\top X = I$).

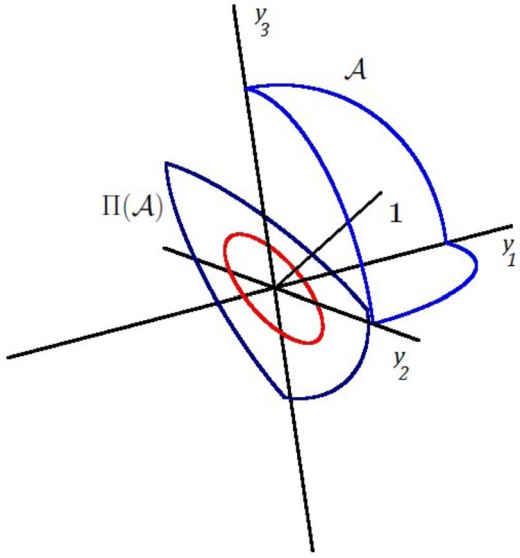


Figure 3: A depiction of $\mathcal{A}$, $\Pi(\mathcal{A})$, and the unit circle for the illustrative example where $b_1(\mathbf{1}, \boldsymbol{y}) = \min(\boldsymbol{y}) = 0$ and $s(\mathbf{1}, \boldsymbol{y}) = \sum(y_i - b_1(\mathbf{1}, \boldsymbol{y}))^2 = 1$. $\mathcal{A}$ is the combination of three quarter circles, one on each plane defined by $y_i = 0$. The projection of this manifold onto the deviation space is depicted by the bowed triangular shape in the plane defined by $\sum y_i = 0$. The circle in this plane represents the sample space for the intermediate sample $\boldsymbol{z}^*$. Also depicted is the vector $\mathbf{1}$, the design matrix for the location and scale setting.

**Lemma 3.2.** *Assume that conditions C1–C2, C4, and C7 hold and* $\boldsymbol{y} \in \mathcal{A}$. *Let* $\nabla s(X, \boldsymbol{y})$ *denote the gradient of the scale statistic with respect to the data vector evaluated at* $\boldsymbol{y}$. *Then* $\nabla s(X, \boldsymbol{y}) \in \mathcal{C}^{\perp}(X)$ *and is normal to* $\Pi(\mathcal{A})$ *at* $\boldsymbol{z} = Q\boldsymbol{y}$ *in* $\mathcal{C}^{\perp}(X)$.

As a result of the lemma, the contribution to the Jacobian of this attenuation is

$$\cos(\gamma) = \frac{\nabla s(X, \boldsymbol{y})^{\top} \boldsymbol{z}}{\|\nabla s(X, \boldsymbol{y})\| \|\boldsymbol{z}\|}, \tag{12}$$

where $\gamma$ is the angle between the two normal vectors. This step is visualized in Figure 4 for the notional location-scale example. The figure pictures only $\mathcal{C}^{\perp}(X)$, which in this case is a plane. The unit sphere (here, the solid circle) is stretched to the dashed sphere, contributing $r^{-(n-p-1)}$ to the Jacobian as seen in panel (a). In panel (b), the dashed circle is transformed onto $\Pi(\mathcal{A})$, contributing $\cos(\gamma)$ to the Jacobian. The normal vectors in panel (b) are orthogonal to the tangent vectors of $\Pi(\mathcal{A})$ and the circle.

**Shift from $\Pi(\mathcal{A})$ to $\mathcal{A}$** The final piece of the Jacobian comes from the transformation from $\Pi(\mathcal{A})$ to $\mathcal{A}$. This step involves a shift of $\boldsymbol{z}$ to $\boldsymbol{y}$ along the column space of $X$.

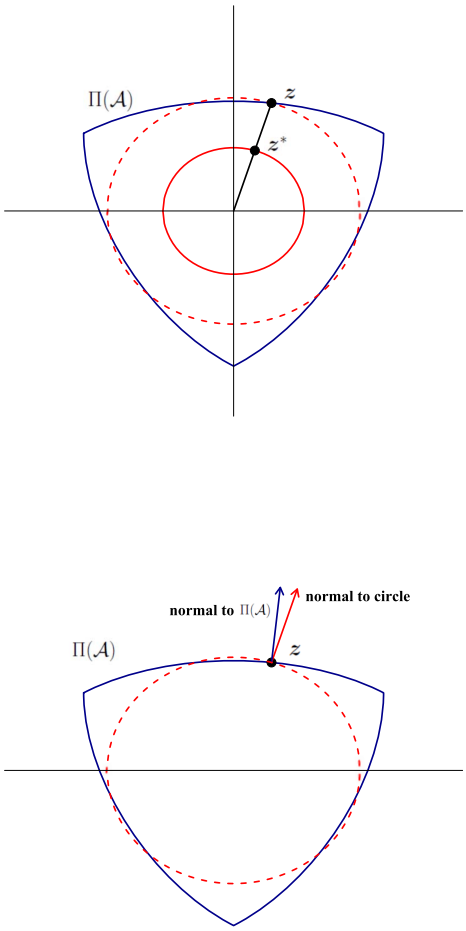


Figure 4: Visualization of the scaling from $z^*$ to $z$. Top: the first substep scales $z^*$ on the unit circle to the circle of radius $r = ||z||$, resulting in a change-of-variables transformation for the unit circle to a circle of radius $r$. The contribution to the Jacobian of this transformation is $r^{-(n-p-1)}$. Bottom: The second substep accounts for the change-of-variables transformation from the circle of radius $r$ to $\Pi(\mathcal{A})$. The normal vectors to these two sets are used to calculate the contribution to the Jacobian of this part of the transformation are shown in the figure.

Since the shift depends on $\boldsymbol{z}$, the density on the set $\Pi(\mathcal{A})$ is deformed by the shift. The contribution of this deformation to the Jacobian is, again, the ratio of the infinitesimal volumes along $\Pi(\mathcal{A})$ at $\boldsymbol{z}$ to the corresponding volume along $\mathcal{A}$ at $\boldsymbol{y}$. The ratio is calculated by considering the volume of the projection of a unit hypercube in the tangent space of $\mathcal{A}$ at $\boldsymbol{y}$ onto $\mathcal{C}^{\perp}(X)$. Computational details are given in the following lemmas and subsequent theorem. Proofs of the lemmas are given in the Supplementary Material.

The theorem is a direct result of the lemmas. Throughout, let $\mathcal{T}_y(\mathcal{A})$ and $\mathcal{T}_y^{\perp}(\mathcal{A})$ denote the tangent space to $\mathcal{A}$ at $\boldsymbol{y}$ and its orthogonal complement, respectively.

**Lemma 3.3.** *Assume that conditions C1–C5 and C7–C8 hold. Then the* $p+1$ *gradient vectors* $\nabla s(X, \boldsymbol{y}), \nabla b_1(X, \boldsymbol{y}), \ldots, \nabla b_p(X, \boldsymbol{y})$ *form a basis for* $\mathcal{T}_y^{\perp}(\mathcal{A})$ *with probability one.*

The lemma describes construction of a basis for $\mathcal{T}_y^{\perp}(\mathcal{A})$, leading to a basis for $\mathcal{T}_y(\mathcal{A})$. Both of these bases can be orthonormalized. Let $A = [a_1, \ldots, a_{n-p-1}]$ and $B = [b_1, \ldots, b_{p+1}]$ denote the matrices whose columns contain the orthonormal bases for $\mathcal{T}_y(\mathcal{A})$ and $\mathcal{T}_y^{\perp}(\mathcal{A})$, respectively. The columns in $A$ define a unit hypercube in $\mathcal{T}_y(\mathcal{A})$ and their projections onto $\mathcal{C}^{\perp}(X)$ define a parallelepiped. We defer construction of $A$ until later.

**Lemma 3.4.** *Assume that conditions C1–C5 and C7–C8 hold. Then the* $n \times (n-p-1)$ *dimensional matrix* $P = QA$ *is of full column rank.*

As a consequence of this lemma, the parallelepiped spanned by the columns of $P$ is not degenerate (it is $n-p-1$ dimensional), and its volume is given by

$$\text{Vol}(P) := \sqrt{\det(P^{\top}P)} = \prod_{i=1}^{r} \sigma_i \tag{13}$$

where $r = \text{rank}(P) = n-p-1$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ are the singular values of $P$ (e.g., Miao and Ben-Israel (1992)). Combining Lemmas 3.3 and 3.4 above leaves us with the following result concerning the calculation of the desired Jacobian.

**Theorem 3.5.** *Assume that conditions C1–C5 and C7–C8 hold. Then the Jacobian of the transformation from the distribution along* $\Pi(\mathcal{A})$ *to that along* $\mathcal{A}$ *is equal to the volume given in* (13).

**The Proposal Density** Putting all the pieces of the Jacobian together we have the following result. Any dependence on other variables, including current states in the Markov chain, is made implicit.

**Theorem 3.6.** *Assume that conditions C1–C8 hold. Let* $\boldsymbol{z}^*$ *be sampled on the unit sphere in* $\mathcal{C}^{\perp}(X)$ *with density* $p(\boldsymbol{z}^*)$. *Using the transformation of* $\boldsymbol{z}^*$ *to* $\boldsymbol{y} \in \mathcal{A}$ *described in Theorem 3.1, the density of* $\boldsymbol{y}$ *is*

$$p(\boldsymbol{y}) = p(\boldsymbol{z}^*) r^{-(n-p-1)} \cos(\gamma)\, Vol(P) \tag{14}$$

*where* $r = s(X, \boldsymbol{y}_{obs})/s(X, \boldsymbol{z}^*)$, *and* $\cos(\gamma)$ *and* $Vol(P)$ *are as in equations* (12) *and* (13), *respectively.*

The proposal is governed by the choice of $p(z^*)$ and a poor choice could lead to an inefficient MCMC algorithm. For all examples in this paper we defined $p(z^*)$ to be the uniform distribution on $\mathbb{S}$. The advantage of this choice is that it requires no further tuning parameters. We have noticed good mixing in terms of the ability of the chain to generate new data $\mathbf{y}$ that is accepted with a reasonable probability. To implement

the method in practice, we generate an n-dimensional independent standard normal $\mathbf{y}^*$ for the proposal and transform this via $h(\cdot)$. Theoretically, the random normal vector would be projected onto $\mathcal{C}^{\perp}(X)$ and scaled to unit norm to generate the uniform on $\mathbb{S}$. Using simple algebra and conditions C5–C8, one can show that $h(\cdot)$ is invariant to this projection and scaling. Another option for the proposal suggested by a reviewer that the authors have yet to study is generating a random walk. As we are proposing values on a complicated manifold, it might be possible to implement this by conducting the random walk on $\mathbf{y}^*$ before transforming via $h(\cdot)$. This could provide advantages in some situations, though we have yet to run into any serious issues with convergence using the independence proposal we utilize here.

Some details for computing the needed quantities are worth further explanation. Computing $\mathrm{Vol}(P)$ involves finding an orthornormal matrix $A$ whose columns span $\mathcal{T}_y(\mathcal{A})$. This matrix can be found by supplementing $B$ with a set of $n$ linearly independent columns on the right, and applying Gram-Schmidt orthonormalization. The computational complexity of this step is $\mathcal{O}(n^3)$. This is infeasibly slow when $n$ is large because it must be repeated at each iterate of the MCMC when a complete data set is drawn. However, using results related to *principal angles* found in Miao and Ben-Israel (1992) the volume (13) can be computed using only $B$. $B$ is constructed by Gram-Schmidt orthogonalization of $\nabla s(X, \boldsymbol{y}), \nabla b_1(X, \boldsymbol{y}), \ldots, \nabla b_p(X, \boldsymbol{y})$, reducing the computational complexity to $\mathcal{O}(np^2)$ – a considerable reduction in computational burden when $n \gg p$. The following corollary formally states how computation of $A$ can be circumvented.

**Corollary 3.7.** *Let $U$ be a matrix whose columns form an orthonormal basis for $\mathcal{C}(X)$ and set $Q = WW^{\top}$ where the columns of $W$ form an orthonormal basis for $\mathcal{C}^{\perp}(X)$. Then the non-unit singular values of $U^{\top}B$ are the same as the non-unit singular values of $W^{\top}A$.*

The lemma implies that $\mathrm{Vol}(P)$ is the product of the singular values of $U^{\top}B$.

Second, the gradients of $\nabla s(X, \boldsymbol{y}), \nabla b_1(X, \boldsymbol{y}), \ldots, \nabla b_p(X, \boldsymbol{y})$ are easily computed in many cases. For example, below we consider M-estimators defined by the estimating equations:

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - x_i^{\top}\boldsymbol{b}(\boldsymbol{y}, X)}{s(\boldsymbol{y}, X)}\right) x_{ij} = 0, \ \sum_{i=1}^{n} \chi\left(\frac{y_i - x_i^{\top}\boldsymbol{b}(\boldsymbol{y}, X)}{s(\boldsymbol{y}, X)}\right) = 0, \tag{15}$$

for $j = 1, 2, \ldots, p$, $x_{ij}$ are the components of $x_j$ and $\psi$ and $\chi$ are almost surely differentiable. The gradients can be found by differentiating this system of equations with respect to each $y_i$. In theory, finite differences could also be used as an approximation if needed.

Finally, it is clear the estimators themselves must be computed for every iteration of the Markov Chain. We have found this burden to be marginal relative to computation of the needed Jacobian. In the simulations and real data analyses presented below, we will see that the additional computational expense needed to fit the Bayesian model is often worthwhile, leading to better performance compared to traditional, non-Bayesian

robust regression estimators. This is most evident when substantive prior information is available and information in the data is limited.

# 4 Simulated Data

We study the performance of restricted likelihood methods in two simulation settings. The first is a hierarchical setting. The second is a variable selection setting where there are several potential covariates but only a few have non-zero effect sizes.

## 4.1 Simulation 1

The first is a hierarchical setting where the data are contaminated with outliers. Specifically, simulated data come from the following model:

$$\begin{aligned} &\theta_i \sim N(\mu, \tau^2),\ i = 1, 2, \ldots, 90 \\ &y_{ij} \sim (1 - p_i)N(\theta_i, \sigma^2) + p_i N(\theta_i, m_i\sigma^2),\ j = 1, 2, \ldots, n_i \end{aligned} \tag{16}$$

with $\mu = 0, \tau^2 = 1, \sigma^2 = 4$. The values of $p_i, m_i$, and $n_i$ depend on the group and are formed using 5 replicates of the full factorial design over factors $p_i, m_i, n_i$ with levels $p_i = .1, .2, .3$, $m_i = 9, 25$, and $n_i = 25, 50, 100$. This results in 90 groups that have varying levels of outlier contamination and sample size. We wish to build models that offer good prediction for the good portion of data within each group. The full model for fitting is a corresponding normal model without contamination:

$$\begin{aligned} &\theta_i \sim N(\mu, \tau^2),\ \sigma_i^2 \sim IG(a_s, b_s),\ i = 1, 2, \ldots, 90, \\ &y_{ij} \sim N(\theta_i, \sigma_i^2),\ j = 1, 2, \ldots, n_i. \end{aligned} \tag{17}$$

For the restricted likelihood versions we condition on robust M-estimators of location and scale in each group: $T_i(y_{i1}, \ldots, y_{in_i}) = (\hat{\theta}_i, \hat{\sigma}_i^2), i = 1, 2, \ldots, 90$. These estimators are solutions to equation (15) (where $x_i \equiv 1$) with user specified $\psi$ and $\chi$ functions designed to discount outliers. The two versions use Huber's and Tukey's $\psi$ function, while both versions use Huber's $\chi$ function. The tuning parameters associated with these functions are chosen so that the estimators are 95% efficient under normally distributed data. These classical M-estimators are commonly used in robust regression settings (Huber and Ronchetti, 2009).

To complete the specification of model (17), the hyperparameters $\mu, \tau^2, a_s$, and $b_s$ must be given priors or fixed. The joint prior density for $\mu$ and $\tau^2$ is improper and proportional to $\tau^{-2}$. The pair $a_s$ and $b_s$ are fixed to a variety of values representing different levels of prior knowledge. For each pair, we set $b_s = 4a_s c$ resulting in a prior mean for each $\sigma_i^2$ of $\frac{4ca_s}{a_s - 1}$, $a_s > 1$. The precision is $\frac{(a_s-1)^2(a_s-2)}{(4ca_s)^2}$, meaning larger $a_s$ and smaller $c$ result in a more informative prior. With $c = 1$ the shrinkage (for large $a_s$) is to the true value of $\sigma^2 = 4$. We consider $a_s = 1.25, 5, 10$ and $c = 0.5, 1, 2$ for a total of nine different priors.

$K = 30$ data sets are generated from (16). For each data set and each pair $(a_s, c)$, the Bayesian models are fit using MCMC. The MCMC for the restricted likelihood

version requires no computational details other than those described for the traditional Bayesian model in Section 3. This is because there are conditioning statistics for each group and the model's conditional independence between the groups allows the data augmentation described earlier to be performed independently within each group. That is, there is a separate Gibbs step for each group to generate the group level data matching the statistics for that group. The acceptance rates for newly generated data across all groups and simulations ranged from 0.57 to 0.68.

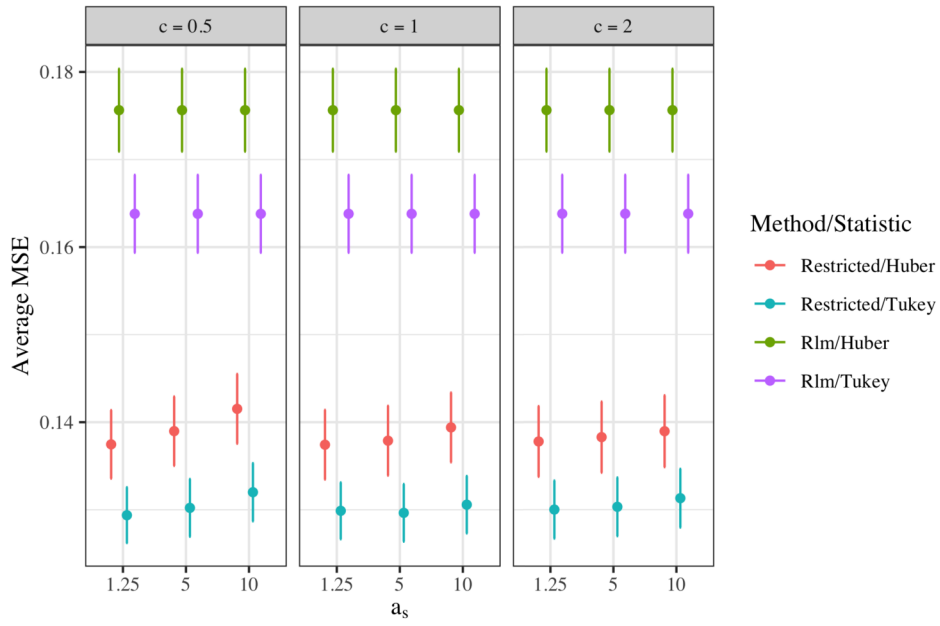


Figure 5: Average MSE plus/minus one standard error for each value of $a_s$ and $c$. Smaller values represent better fits. The panels correspond to $c = 0.5$ (left), $c = 1$ (middle), and $c = 2$ (right), with the values of $a_s$ on the horizontal axis. The average MSE for the normal theory model ranges from 0.24 to 0.25 and is left out of the figure.

The performance of the methods can be evaluated in many ways. For these simulations, we know the true data generating mechanism, and this allows us to make direct comparisons between the fitted model and the truth. The Bayesian methods provide a full predictive distribution for the response, given group, while the classical methods provide only point estimates of parameters. Our comparisons have focused on two main summaries. One summary, not presented here, is the average Kullback-Liebler divergence from the good portion of the true distribution of $Y$ given group to the predictive distribution (Bayes) or to the distribution with point estimates plugged in (classical). For the Bayesian models, the divergence does not have a closed form and must be evaluated numerically. The second summary, preferred by a referee, is the mean squared error (MSE), averaged across groups. Results are presented in Figures 5 and 6. Formally, with the superscript $M$ indicating the method and the additional subscript $k$ indexing

the data set,

$$\text{MSE}_{ik}^{(M)} = (\hat{\theta}_{ik}^{(M)} - \theta_{ik})^2, \tag{18}$$

$$\text{MSE}^{(M)} = \frac{1}{90K} \sum_{k=1}^{K} \sum_{i=1}^{90} \text{MSE}_{ik}^{(M)}. \tag{19}$$

The Bayesian restricted likelihood methods show superior performance under both summaries, but especially for MSE. Figure 5 displays the MSE grouped by pairs $a_s$ and $c$ with error bars plus/minus one standard error within the group. The values of $a_s$ and $c$ do not affect the classical robust linear models. The average MSEs for the normal theory models ranges from 0.24 to 0.25 and are left out of the figure. The results uniformly favor the Bayesian restricted likelihood methods, as seen by substantially lower values of MSE. For both classical and restricted likelihood methods, Tukey's $\psi$ function leads to better performance than does Huber's $\psi$ function.

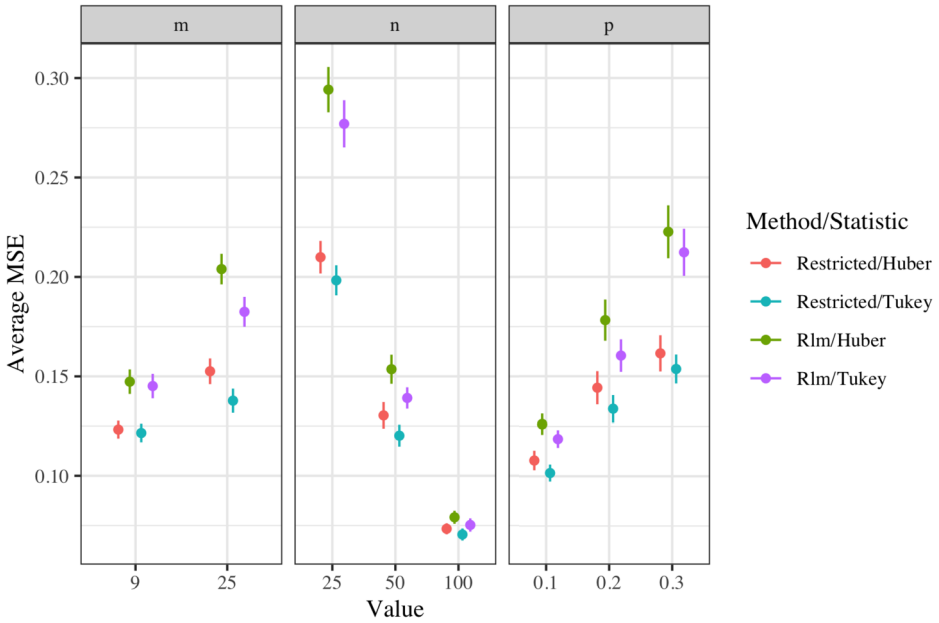


Figure 6: Average MSE plus/minus one standard error grouped by the factors $m$ (left), $n$ (middle), and $p$ (right). These results are for the single prior with $a_s = 5$ and $c = 1$.

It is also interesting to consider the effects of factors $n$, $p$, and $m$. We present the results for a single prior ($a_s = 5$ and $c = 1$). For each simulation $k$, the main effect average MSE is found for each factor $n$, $p$, and $m$. Figure 6 displays the average MSE over the $K = 30$ simulations along with error bars plus/minus one standard error. For each group $n$, $p$, and $m$, the Bayesian restricted likelihood versions have better average loss than do the classical methods. As expected, the average MSE gets larger as the

contamination gets more severe (larger $m$ or larger $p$) and tends to get smaller as the sample size $n$ grows. The advantage of the Bayesian methods is greater for smaller sample sizes and for more severe contamination.

This simulation shows the potential of the restricted likelihood methods to dramatically improve estimation. The simulation also conveys some cautions that are apparent from consideration of KL divergence but not MSE. Specifically, the choice of summary statistic along with the corresponding tuning parameters is important. For the tuning parameters for the $\psi$ functions, we applied the default choice of 95% efficiency at the normal. Under the simulation model here, this choice results in bias in the scale estimation which affects the performance of the method. Tuning parameters must be set when using both classical and Bayesian methods. The Bayesian approach encourages use of a hierarchical model structure and allows one to incorporate prior information in the analysis. These features can improve predictive performance substantially. If poorly handled, they can, of course, harm performance.

## 4.2 Simulation 2

In this simulation the data are generated from the following mechanism: $y = \boldsymbol{\beta}^\top x + \epsilon$ with $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^\top$ and the error $\epsilon \sim N(0, \sigma^2)$ with probability 0.8 and $\epsilon \sim$ Half-Normal$(0, 25\sigma^2)$ with probability 0.2 (i.e., there is a relatively large amount of one-sided outlier contamination). The components of $x = (x_1, x_2, x_3)$ are correlated with $x_1 \sim N(0, 1)$ and $x_j = x_1 + \eta_j$ with $\eta_j \sim N(0, 4)$ for $j = 2, 3$. This results in a theoretical correlation of $1/\sqrt{5} \approx 0.44$ between $x_1$ and both $x_j, j = 2, 3$. The model used for fitting contains an additional 27 covariates, some of which are also correlated with $x_1$, $x_2$, and $x_3$. Specifically the fitting model is $y = \beta^\top x + \beta^{*\top} x^* + \epsilon$ where $x^*$ and $\beta^*$ are 27 dimensional vectors of extra covariates and slope parameters. Of these 27 covariates, 21 are generated independently from standard normal distributions. Of the remaining 6, two each are generated by adding standard normal noise to $x_1$, $x_2$, and $x_3$. This represents a common situation where several covariates with various levels of correlation amongst them are available for fitting, but only a few govern the data generating mechanism.

For the simulation, $K = 30$ data sets (including the additional covariates) of size $n = 500$ are generated from the true model with true values $\boldsymbol{\beta} = (1, 1, 1)^\top$ and $\sigma^2 = 2$. We fit the model including all 30 covariates and consider the following methods for the fit 1) classical robust regression with Tukey's estimator of location and Huber's estimator of scale, 2) the corresponding restricted likelihood version 3) a heavy-tailed Bayesian model with a Student-t likelihood with $\nu = 5$ degrees of freedom. For the Bayesian models we take $\boldsymbol{\beta}_{all} \sim N_{20}(\mathbf{0}, \sigma^2_\beta I)$ with $\boldsymbol{\beta}_{all} = (\boldsymbol{\beta}, \boldsymbol{\beta}^*)^\top$ and $\sigma^2 \sim IG(5, 8)$ under the restricted model and $\sigma^2 \sim IG(5, \frac{\nu-2}{\nu}8)$ under the Student-t model. For each data set, we fit the models for $\sigma_\beta = 0.4, 0.6, 0.8, \ldots, 1.4$. The acceptance rates for the restricted likelihood MCMC data-augmentation step range from 0.3 to 0.36 across all the data sets and values of $\sigma_\beta$. To compare performance we first consider the $MSE = (||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||^2 + ||\hat{\boldsymbol{\beta}}^* ||^2)/30$ for each simulation where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^*$ are point estimates for the fitted model. For the Bayesian models, we use posterior means. The average MSEs plus/minus one standard

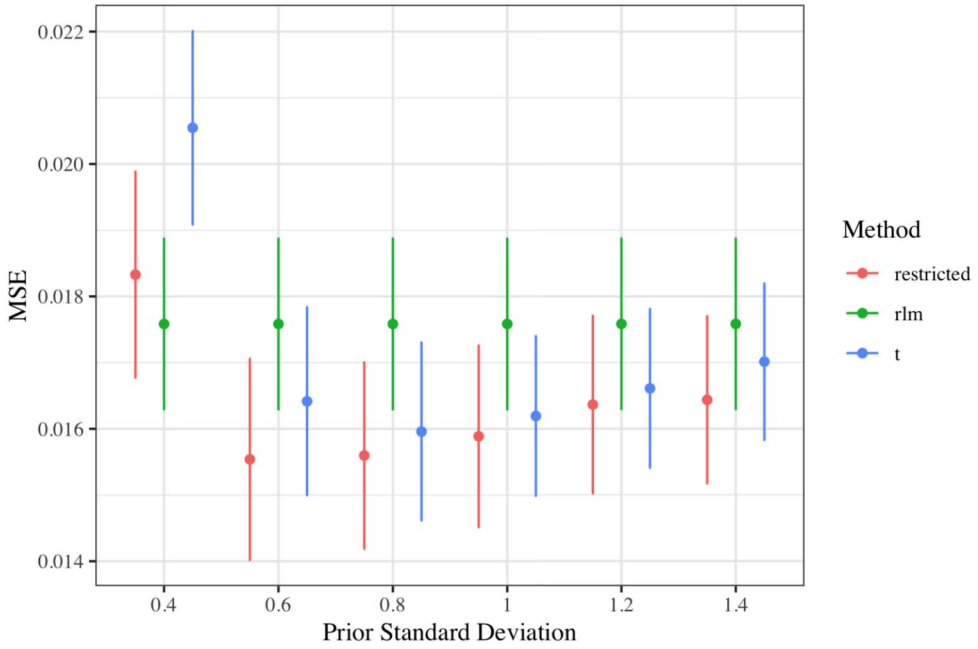


Figure 7: Average MSE plus/minus one standard error over the $K = 30$ simulations for each value of the prior standard deviation ($\sigma_\beta$) and each of the fitting methods. 'Restricted' is our method conditioning on Tukey's estimator of location and Huber's estimator of scale. 'rlm' refers to the classical robust linear model fit with the same estimators and 't' is the heavy-tailed Bayesian model with a Student-t likelihood. The 'rlm' results are the same for each $\sigma_\beta$.

error over the simulations for each $\sigma_\beta^2$ are displayed in Figure 7. The classical fit is labeled 'rlm' and is the same for each value of the prior standard deviation $\sigma_\beta^2$. We see for most values of the prior standard deviation, the Bayesian models ('restricted' and 't') outperform the classical fit. The correlation amongst the covariates causes a certain level of confounding and the prior shrinkage helps to improve estimation. However, too much shrinkage can be detrimental as demonstrated for $\sigma_\beta = 0.4$. While this will help for estimation of $\boldsymbol{\beta}^* = 0$, the estimation of the active parameters $\boldsymbol{\beta}$ can be hindered. The $t$ model seems more sensitive to this effect than the restricted model. The restricted model also has an additional advantage when it comes to prediction of the non-outlying data. To see this, for each simulation we consider the mean negative log-likelihood of the non-outlying data: $MNLL = -\frac{1}{N}\sum \log f(y_i|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}^*, \hat{\sigma})$ where $f$ is the assumed likelihood and the average is taken over the $N$ non-outlying points $y_i$. For the classical and restricted fits, $f$ is the normal likelihood and for the 't' it is the heavy-tailed Student-t likelihood. The average MNLL plus/minus one standard error over the simulations for each $\sigma_\beta^2$ are displayed in Figure 8. First, the restricted version has a small but consistent improvement over the classical method. Second, it is clear

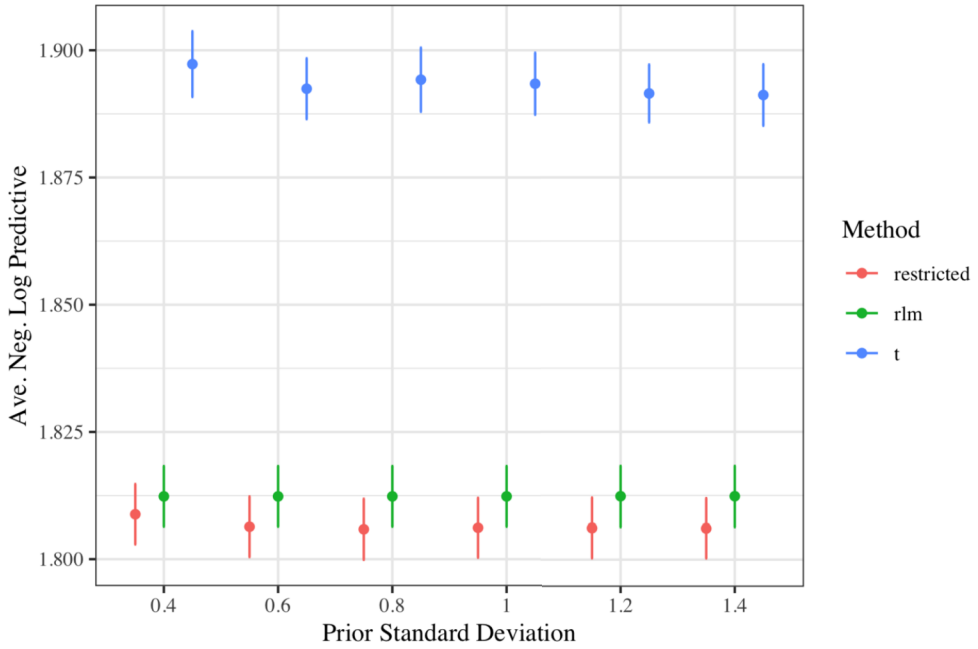


Figure 8: Average MNLL plus/minus one standard error over the $K = 30$ simulations for each value of the prior standard deviation ($\sigma_\beta$) and each of the fitting methods. 'Restricted' is our method conditioning on Tukey's estimator of location and Huber's estimator of scale. 'rlm' refers to the classical robust linear model fit with the same estimators and 't' is the heavy-tailed Bayesian model with a Student-t likelihood. The 'rlm' results are the same for each $\sigma_\beta$.

that the heavy-tailed model suffers when trying to predict the non-outlying data since it assumes the entire data generating mechanism is heavy-tailed.

## 5 Real Data

We illustrate our methods with a pair of regression models for data from Nationwide Insurance Company that concern prediction of the performance of insurance agencies.

Nationwide sells many of its insurance policies through agencies which provide direct service to policy holders. The contractual agreements between Nationwide and these agencies vary. Our interest is the prediction of future performance of agencies where performance is measured by the total number of households an agency services ('household count').

The data are grouped by states with a varying number of agencies by state. Identifiers such as agency/agent names are removed. Likewise, state labels and agency types (identifying the varying contractual agreements) have been made generic to protect the

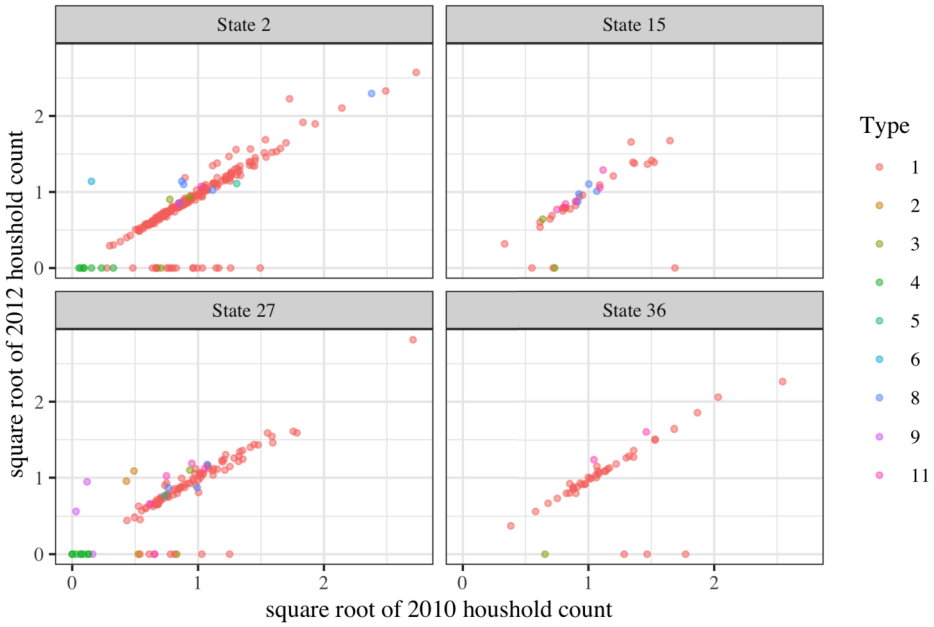


Figure 9: The square root of (scaled) count in 2012 versus that in 2010 for four states. The colors represent the varying contractual agreements as they stood in 2010 ('Type'). Agencies that closed during the 2010–2012 period are represented by the zero counts for 2012.

proprietary nature of the data. Additionally, the counts were scaled to have standard deviation one before analysis.

As an exploratory view, a plot of the square root of (scaled) household count in 2012, against that in 2010 is shown in Figure 9 for four states. The states have varying numbers of agencies and the different colors represent the varying types of contractual agreements as they stood in 2010 ('Type'). A significant number of agencies closed sometime before 2012, as represented by the 0 counts for 2012. Among the open agencies, linear correlations exists with strength depending on agency type and state. 'Type 1' agencies open in 2012 are of special interest. One could easily subset the analysis to only these agencies, removing the others. However, we leave them and use the data as a test bed for our techniques by fitting models that do not account for agency closures or contract type. Our expectation is that the restricted likelihood will facilitate prediction for the 'good' part of the data (i.e., open, 'Type 1' agencies). It is of concern to the company to predict closures and future performance for agencies that remain open. It is important for planning purposes that the predictions are not overly influenced by a handful of over/underperforming agencies. Our analysis focuses on one aspect of the business problem – the prediction of future performance for agencies, given they remain open.

## 5.1 State Level Regression Model

The first analysis is based on individual regressions fit separately within states. The following normal theory regression model is used as the full model for a single state:

$$\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_0, \Sigma_0); \quad \sigma^2 \sim IG(a_0, b_0); \quad y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \ i = 1, \ldots, n, \quad (20)$$

where $\boldsymbol{\beta}$ is a three dimensional vector ($p = 3$) of regression coefficients for the covariate vector $\boldsymbol{x}_i$ consisting of the square root of household count in 2010, and two different size/experience measures related to the number of employees associated with the agency. The response, $y_i$ is the square root of household count in 2012. The hyper-parameters $a_0, b_0, \mu_0$ and $\sigma_0^2$ are all fixed and set from a robust regression fit to the corresponding state's data from the time period two years before. Specifically, let $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ be estimates from the robust linear regression of 2010 counts on 2008 counts. We fix $a_0 = 5$ and set $b_0 = \hat{\sigma}^2(a_0 - 1)$ so the prior mean is $\hat{\sigma}^2$. We set $\boldsymbol{\mu}_0 = \hat{\boldsymbol{\beta}}$ and $\Sigma_0 = n_p \hat{\Sigma}_0$ where $n_p$ is the number of agencies in the prior data set and $\hat{\Sigma}_0$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ derived from the robust regression. This prior is in the spirit of the Zellner's $g$-prior (Zellner, 1986; Liang et al., 2008). In general, scaling the prior variance by a factor $g = n_p$ is analogous to the unit-information prior (Kass and Wasserman, 1995), with the difference that we are using a prior data set, not the current data set, to set the prior. The obvious reason why this model is misspecified is due to omission of the contract type and agency closure information. Closing our eyes to these variables, many of the cases appear as outliers. Additionally, the model assumes equal variance within each state, an assumption whose worth is arguable (see Figure 9).

We compare four Bayesian models: the standard Bayesian normal theory model, two restricted likelihood models, both with simultaneous M-estimators, and a heavy-tailed model. For the restricted likelihood methods we use the same simultaneous M-estimators as in the simulation of Section 4 adapted to linear regression. The heavy-tailed model replaces the normal sampling density in (20) with a $t$-distribution with $\nu = 5$ degrees of freedom. The Bayesian models are all fit using MCMC, with the restricted versions using the algorithm presented in Section 3.2. We also fit the corresponding classical robust regressions and a least squares regression.

### Method of Model Comparison

We wish to examine the performance of the models in a fashion that preserves the essential features of the problem. Since we are concerned with outliers and model misspecification, we understand that our models are imperfect and prefer to use an out-of-sample measure of fit. This leads us to cross-validation. We repeatedly split the data into training and holdout data sets; fitting the model to the training data and assessing performance on the holdout data.

The presence of numerous outliers in the data implies that both training and validation data will contain outliers. For this reason, the evaluation must be robust to a certain fraction of bad data. The two main strategies are to robustify the evaluation function (e.g., Ronchetti et al., 1997) or to retain the desired evaluation function and

trim cases (Jung et al., 2014). Here, we pursue the trimming approach with log predictive density for the Bayesian models and log density from plug-in maximum likelihood for the classical fits used as the evaluation function.

The trimmed evaluation proceeds as follows in our context. The evaluation function for case $i$ in the holdout data is the log predictive density, say $\log(f(y_i))$, with the conditioning on the summary statistic suppressed. The trimming fraction is set at $0 \leq \alpha < 1$. To score a method, we first identify a base method. Denote the predictive density under this method by $f_b(y)$. Under the base method, $\log(f_b(y_i))$ is computed for each case in the holdout sample, say $i = 1, \ldots, M$. Order the holdout sample according to the ordering of $\log(f_b(y_i))$ and denote this ordering by $y^b_{(1)}, y^b_{(2)}, \ldots, y^b_{(M)}$. That is, for $i < j$ $\log(f_b(y^b_{(i)})) < \log(f_b(y^b_{(j)}))$. All of the methods are then scored on the holdout sample with the mean trimmed log marginal pseudo likelihood,

$$TLM_b(A) = (M - [\alpha M])^{-1} \sum_{i=[\alpha M]+1}^{M} \log(f_A(y^b_{(i)})),$$

where $f_A$ corresponds to the predictive distribution under the method "A" being scored. In other words, the $[\alpha M]$ observations with the smallest values of $\log(f_b(y))$ are removed from the validation sample and all of the methods are scored using only the remaining $M - [\alpha M]$ observations. Larger values of $TLM_b(A)$ indicate better predictive performance. This process is advantageous to the base method since the smallest scores from this method are guaranteed to be trimmed. A method that performs poorly when it is the base method is discredited.

### Comparison of Predictive Performance

'Type 1' agencies are of special interest to the company and so the evaluation of the TLM is done on only holdout samples of 'Type 1', whereas the training is done on agencies of all types. This is intended to demonstrate the robustness properties of the various methods. Models are fit to four states labelled State 2, 15, 27, and 36, with $n = 222, 40, 117,$ and 46, representing a range of sample sizes. Fitting is done on $K = 50$ training samples with training sample sizes taken to be $0.25n$ and $0.50n$. Holdout evaluation is done on the remaining ('Type 1') samples. The acceptance rates for the data augmentation step, for all but one training set, range from 0.10 to 0.8 across the states, repetitions, and two versions of the model. The exception was a single training set from State 15 resulting in an usually small acceptance rate under Tukey's version. This case didn't effect the overall results of the simulations but emphasizes the need to check convergence on a case by case basis. The average $TLM_b(A)$ over the $K = 50$ training/holdout samples for the four states and seven methods are shown in Figure 10 where the base model is the Student-t model and $\alpha = 0.3$. Similar results are observed for other base models. The error bars are plus/minus one standard deviation of the average $TLM_b(A)$ over the $K = 50$ training/holdout samples. It is clear that the normal Bayesian model used as the full model (Normal) and the classical ordinary least squares fits (OLS) have poor performance due to the significant amount of outlier contamination in the data. In comparing our restricted methods to their corresponding classical methods, there

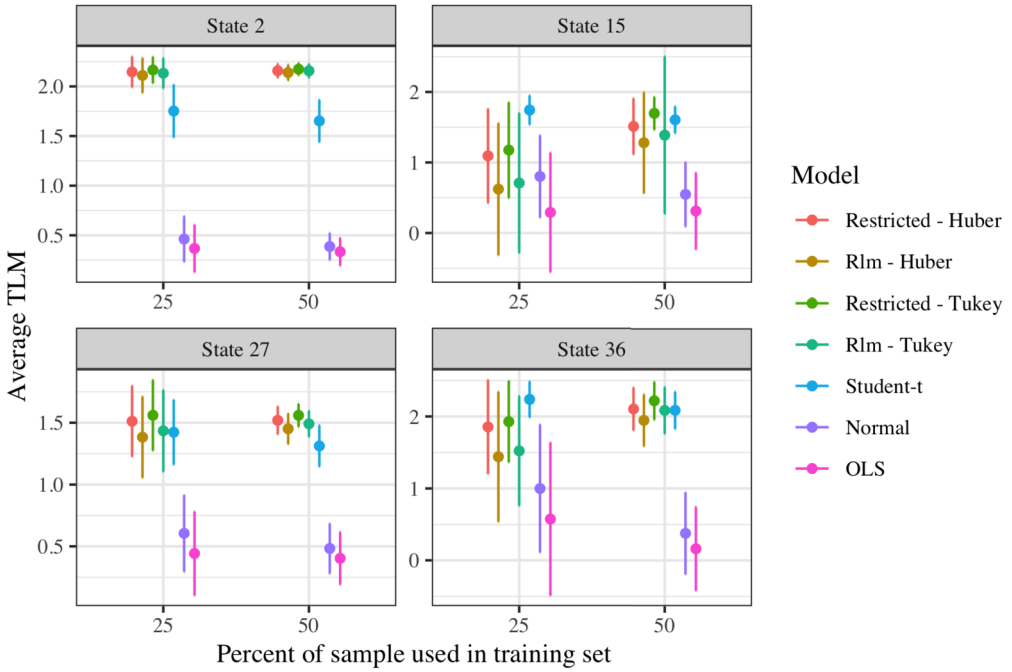


Figure 10: Average TLM plus/minus one standard deviation over $K = 50$ splits into training and holdout samples. The panels are for the different states 2, 15, 27, and 36, with $n = 222, 40, 117$, and 46, respectively. The horizontal axis is the percent of $n$ used in each training set. The color corresponds to the fitting model. Larger values of TLM are better.

is small, but consistent improvement across the states and training sample size. Additionally, variance reduction for the Bayesian versions is evident, especially in State 15, highlighted by the smaller error bars. For state 2, the largest state with $n = 222$, the restricted and classical robust methods have similar performance especially for larger training sample size. This reflects the diminishing effect of the prior as the sample size grows. Notably, the Student-t model performs poorly in comparison for this state. The predictive distribution explicitly accounts for heavy-tailed values, resulting in poorer predictions of the 'good' data (i.e., the 'Type 1' agencies). Likewise, for State 27, another larger state, the Student-t model is outperformed by our restricted methods. For the other states (State 15 and 36), the Student-t performs better than our restricted methods for smaller training sample size (25% of the sample). However, this advantage goes away for the larger training sample size (50% of the sample). Intuitively, as more data is available for fitting, more outliers appear and the heavy-tailed model compensates for them by assuming they come from the tails of the model; an assumption which is detrimental for prediction. Comparisons of the models depend on $\alpha$ as seen in Figure 11 which shows results for different $\alpha$ for training sample size $0.5n$. For smaller $\alpha$ (in this case $\alpha = 0.1$), many outliers are left untrimmed resulting in lower TLM for

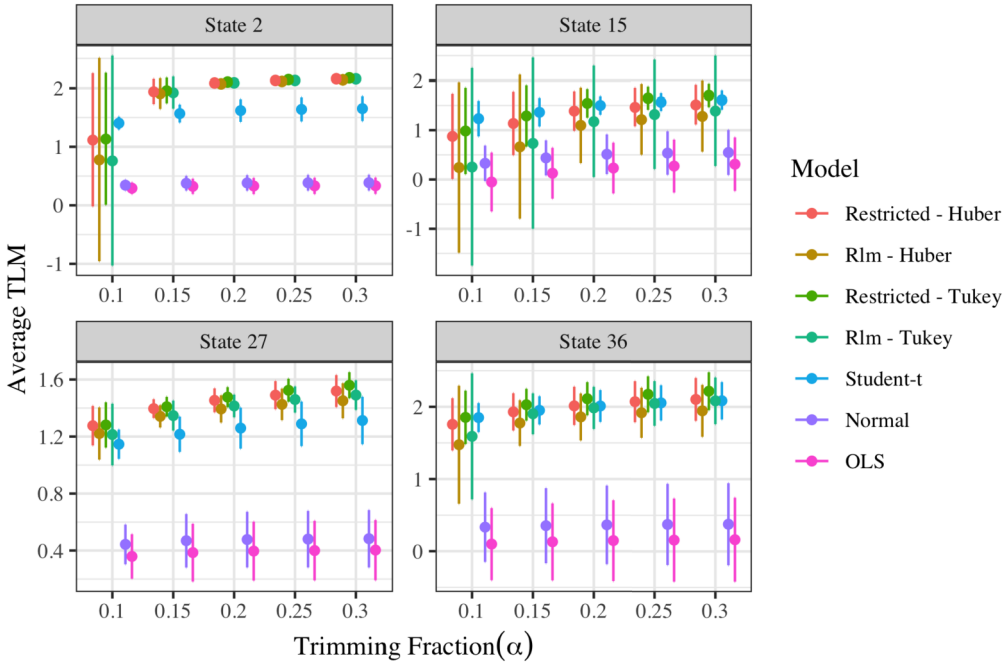


Figure 11: Average TLM plus/minus one standard deviation over $K = 50$ splits into training and holdout samples for several values of the trimming fraction $\alpha$. The training sample size used is $0.5n$. Larger values of TLM are better.

all methods and noticeably larger standard deviation for the classical robust methods and our restricted likelihood. Larger values of $\alpha$ ensure that the predictive performance assessment excludes the majority of outliers. The proportion of 0 counts in the data is roughly 0.14, suggesting that $\alpha$ should be at least this large.

## 5.2 Hierarchical Regression Model

The previous analysis treated states independently. A natural extension is to reflect similar business environments between states using a hierarchical regression. The proposed model is:

$$\boldsymbol{\beta} \sim N_p(\boldsymbol{\mu}_0, a\Sigma_0); \quad \boldsymbol{\beta}_j \overset{iid}{\sim} N_p(\boldsymbol{\beta}, b\Sigma_0); \quad \sigma_j^2 \sim IG(a_0, b_0); \tag{21}$$

$$y_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_j^2), \; i = 1, \ldots, n_j, \; j = 1, \ldots, J \tag{22}$$

where $y_{ij}$ is the $i^{th}$ observation of square rooted household count in 2012 in the $j^{th}$ state, $n_j$ is the total number of agencies in state $j$, and $J$ is the number of states. $\boldsymbol{x}_{ij}$ is same three-dimensional covariate vector as before and $\boldsymbol{\beta}_j$ represents the individual regression coefficient vector for state $j$. The parameters $\boldsymbol{\mu}_0$, $\Sigma_0$, $a_0$, and $b_0$ are fixed by fitting the

regression $y_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta} + \epsilon_{ij}$ using Huber's M-estimators to the prior data set from two years before. Using the estimates from this model, we set $\mu_0 = \hat{\beta}$, $\Sigma_0 = n_p \hat{\Sigma}_0$ ($n_p = 2996$ is the number of observations in the prior data set), $a_0 = 5$ and $b_0 = \hat{\sigma}^2(a_0 - 1)$. We constrain $a + b = 1$ in an attempt to partition the total variance between the individual $\boldsymbol{\beta}_j$'s and the overall $\boldsymbol{\beta}$ and take $b \sim \text{beta}(v_1, v_2)$. Using the prior data set, we assess the variation between individual estimates of the $\boldsymbol{\beta}_j$ to set $v_1$ and $v_2$ to allow for a reasonable amount of shrinkage. To allow for dependence across the $\sigma_j^2$ we first take $(z_1, \ldots, z_J) \sim N_J(\mathbf{0}, \Sigma_\rho)$ with $\Sigma_\rho = (1 - \rho)\mathbf{I} + \rho \mathbf{1}\mathbf{1}^\top$. Then we set $\sigma_j^2 = H^{-1}(\Phi(z_j))$ where $H$ is the cdf of an $IG(a_0, b_0)$ and $\Phi$ is the cdf of a standard normal. This results in the specified marginal distribution, while introducing correlation via $\rho$. We assume $\rho \sim \text{beta}(a_\rho, b_\rho)$ with mean $\mu_\rho = a_\rho/(a_\rho + b_\rho)$ and precision $\psi_\rho = a_\rho + b_\rho$. The parameters $\mu_\rho$ and $\psi_\rho$ are given beta and gamma distributions, with fixed hyperparameters. More details on setting prior parameters are given in the Supplementary Material.

Using the same techniques as in the previous section, we fit the normal theory hierarchical model above, a thick-tailed $t$ version with $\nu = 5$ d.f., and two restricted likelihood versions (Huber's and Tukey's) of the model. For the restricted methods, we condition on robust regression estimates fit separately within each state. We also fit classical robust regression counterparts and a least squares regression separately within each state. Additionally, we compare our method to an ABC fit. The ABC version conditions on the Tukey statistics used in our restricted likelihood version. We choose the Tukey version for comparison to ABC since it naturally trims outliers and we expect it to perform the best in this situation. Recall, ABC will approximate the restricted posterior using $\pi(\boldsymbol{\theta} | \rho(T(\boldsymbol{y}_{obs}), T(\boldsymbol{y}^*)) < \epsilon)$. Due to the high-dimension of the parameters and statistics we use the MCMC method called Gibbs ABC developed by Turner and Van Zandt (2014) to obtain samples from the ABC posterior. A brief description of this algorithm is as follows with theoretical details provided by Turner and Van Zandt (2014). Let $\boldsymbol{y}_{j,obs}$ denote the observed data for state $j = 1, \ldots, J$. The shared higher-level parameters are sampled as before since they are, *a posteriori* conditionally independent of the data. The state-level parameters $\boldsymbol{\theta}_j = (\boldsymbol{\beta}_j, \sigma_j), j = 1, 2, \ldots, J$, are sampled using Gibbs ABC. For each iterate of the chain, denote the current state-level parameters and data for state $j$ by $\boldsymbol{\theta}_{j,curr}$ and $\boldsymbol{y}_{j,curr}$. A single update for the state-level parameters loops over $j$ as follows. Propose $\boldsymbol{\theta}_{j,prop}$ from the prior and then propose new data $\boldsymbol{y}_{j,prop}$ from the normal model conditional on $\boldsymbol{\theta}_{j,prop}$. The proposed parameters are accepted with Metropolis-Hastings acceptance probability

$$\alpha_{mh} = \min\{1, \frac{\phi(\rho(T(\boldsymbol{y}_{j,prop}), T(\boldsymbol{y}_{j,obs}))/\delta_{abc})}{\phi(\rho(T(\boldsymbol{y}_{j,curr}), T(\boldsymbol{y}_{j,obs}))/\delta_{abc})}\} \tag{23}$$

with $\phi(\cdot)$ the standard normal pdf and $\delta_{abc}$ a tolerance parameter. Here, we use $\rho(\cdot)$ for the standard Euclidean distance metric to conform to common ABC notation (this is not the $\rho$ of M-estimation). This method makes use of kernel-based ABC (Wilkinson, 2013). Instead of checking whether the distance between the sampled and observed statistics is strictly below some threshold, this method computes the kernel value of the distance which offers a smoother transition between acceptance and rejection. We use a standard Gaussian kernel for this application. The smaller $\delta_{abc}$, the closer the statistics must be for acceptance.

A few notes on our implementation are warranted. First, we are not restricted to sampling the parameters from their prior. A more general proposal distribution can be applied with standard adjustments to $\alpha_{mh}$ to adjust for the proposal (Turner and Van Zandt, 2014). We tried a few different options, including sampling from the full-conditional distributions, but found proposing from the priors was both the easiest to implement and provided the most satisfactory convergence results for this problem. Additionally, the choice of $\delta_{abc}$ is important and can be different for each state. For this we started with a small value for each state ($\delta_{abc} = 0.01$) and iteratively checked the within-state acceptance rate. After each check, if the rate was below 0.1 we increased the $\delta_{abc}$ for that state by a factor of 1.2. These choices were based on some experimentation in order to reach satisfactory convergence in a reasonable number of iterations. To reach satisfactory convergence we had to run the chains for a total of $40,000$ iterations which was 10-fold more than were needed for the restricted likelihood algorithm. In our experimentation, our method takes only about 1.6 times as long as ABC per iteration. We believe that this modest increase in per-iteration computational time is outweighed by apparently better convergence and mixing of the Markov chain. It is quite possible that better choices for the ABC algorithm could help improve its convergence, but we leave this for further research as computational efficiency is not the main focus of this paper.

Hierarchical models naturally require more data and so we include states having at least 25 agencies with sufficient variation within each covariate, resulting in 20 states in total and $n = \sum_j n_j = 3094$ total agencies. For training data we take a stratified (by state) sample of size $3094/2 = 1547$ where the strata sizes are $n_j/2$ (rounded to the nearest integer). The remaining data is used for a holdout evaluation using TLM computed separately within each state: $TLM_b(A)_j = (M_j - [\alpha M_j])^{-1} \sum_{i=[\alpha M_j]+1}^{M_j} \log(f_A(y^b_{(i)j}))$ where $y^b_{(1)j}, y^b_{(2)j}, \ldots, y^b_{(M_j)j}$ is the ordering of the $M_j$ holdout observations within state $j$ according to the log marginals under the base model $b$. For the non-Bayesian models, $f_A(y^b_{(i)j})$ is estimated using plug-in estimators for the parameters for state $j$. $TLM_b(A)_j$ is computed for each state for $K = 50$ splits of training and holdout sets. The Bayesian models are fit using MCMC, with the restricted versions applying the algorithm laid out in Section 3 and adapted to the hierarchical setting as described in Section 4. For the MH-step proposing augmented data, the acceptance rates for the two restricted likelihood models across all states and repetitions ranged from 0.01 to 0.75, with only 7 cases (out of $50 * 20 * 2 = 2000$ chains) with rates below 0.1.

The average over states, $\overline{TLM}_b(A)_\cdot = \frac{1}{22} \sum_{j=1}^{22} TLM_b(A)_j$ for each of the $K$ repetitions is summarized in Figure 12 for several trimming fractions using the Student-t as the base model. The points are the average of the $\overline{TLM}_b(A)_\cdot$ over the $K$ repetitions with error bars plus/minus one standard deviation over $K$ with larger values representing better predictive performance. As the trimming fraction used for the TLM increases, so does TLM since more outliers are being trimmed. Similar patterns were seen in the individual state level regressions in Section 5.1. Despite being used as the base model to compute TLM, the Student-t doesn't perform well in comparison to the robust regressions. We attribute this to the assumption of heavier tails resulting in smaller log marginal values on average; emphasizing again that the t-model will do well to discount

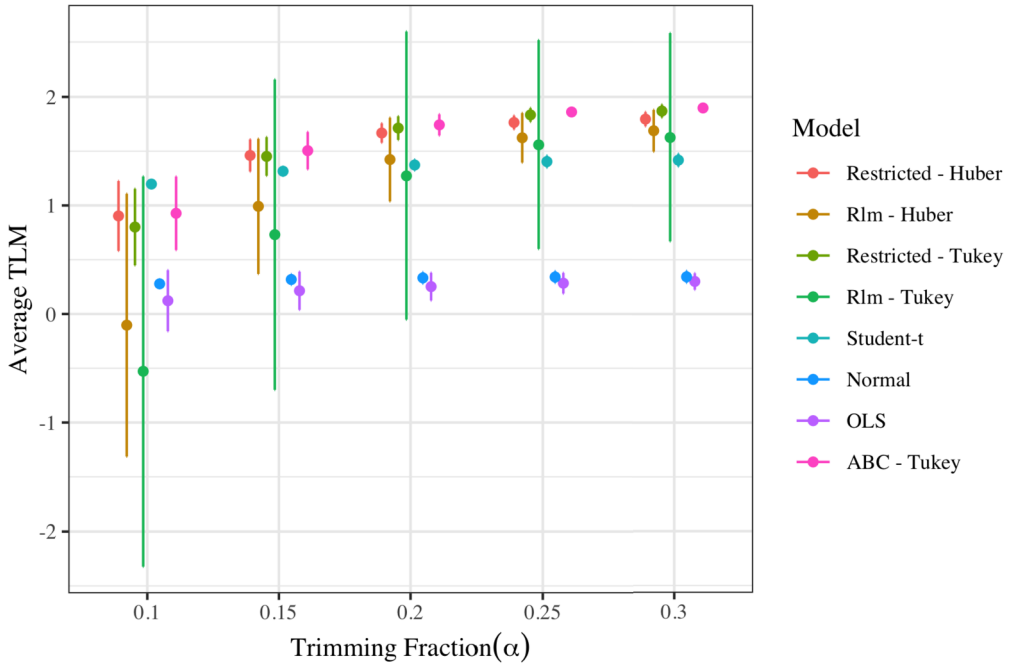


Figure 12: Hierarchical model results: $\overline{TLM}_b(A)$. plus/minus one standard deviation over $K = 50$ splits into training and holdout sets with the Student-t as the base model and several values of the trimming fraction $\alpha$. Larger values of TLM are better.

outlying observations but does not provide a natural mechanism for predicting non-outlying data. For each trimming fraction, our restricted likelihood hierarchical models outperform the classical robust regressions fit separately within each state. The hierarchical model also reduces variance in predictions resulting in smaller error bars. On the surface, ABC performs quite well in comparison to the restricted likelihood. A table of the mean and standard deviation values for trimming fraction 0.3 is provided in Table 1 where we see that the average TLM for ABC is larger than that for restricted likelihood by 0.03. Additionally, the standard deviation of TLM for ABC is half the size of that for restricted likelihood. A closer look at the results shows that the difference in average TLM can be attributed entirely to a single state with a relatively small sample size (see next paragraph).

| Model | Trimming Fraction | mean | std. deviation |
|---|---|---|---|
| Restricted - Tukey | 0.3 | 1.87 | 0.06 |
| ABC - Tukey | 0.3 | 1.90 | 0.03 |

Table 1: Comparison of the average TLM over all states for Restricted and ABC methods with trimming fraction 0.3.

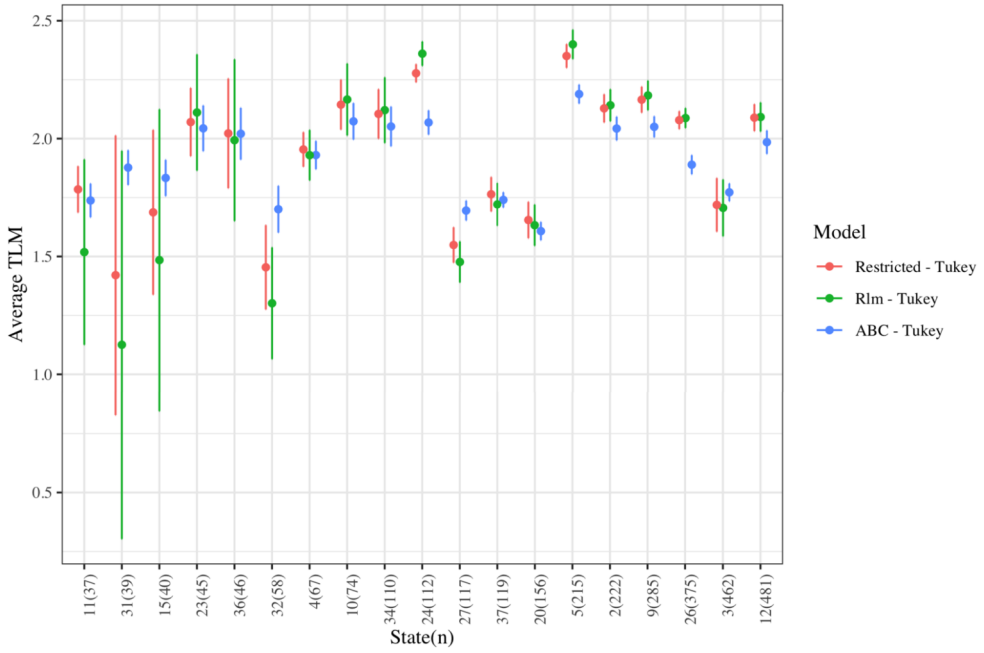


Figure 13: Hierarchical model results: $TLM_b(A)_j$ plus/minus one standard deviation over $K = 50$ repetitions for each state and $\alpha = 0.3$. The states are ordered along the $x$-axis according to number of agencies within the state (shown in parentheses). Results displayed are for the robust models using Tukey's M-estimators. Larger values of TLM are better.

It is also interesting to examine the results within each state. Figure 13 summarizes $TLM_b(A)_j$ with $\alpha = 0.3$ for each state where the points and error bars are the averages plus/minus one standard deviation of $TLM_b(A)_j$ over the $K = 50$ repetitions. The results are only given for the models using Tukey's M-estimators (Huber's version is qualitatively similar). The states are ordered along the $x$-axis according to number of agencies within the state (shown in parentheses). State 28 is removed from the figure as the error bars for the classical robust regression are excessively large and distort the comparison. In several of the smaller states, the restricted hierarchical model performs better than the classical method, with similar performance between the models in most of the larger states, a reflection of the decreased influence of the prior. The hierarchical structure pools information across states, improving performance in the smaller states. The standard deviations are smaller for the hierarchical model in smaller states than they are for the corresponding classical model. In larger states, the standard deviations are virtually identical. Similar benefits are often seen for hierarchical models (e.g., Gelman, 2006). Restricted likelihood performs better when considering the metric at the individual state-level. While there are a few small states that perform much better under ABC (especially state 31), the restricted likelihood average TLM is larger in 14 of the 20 states with a median difference (restricted minus ABC) of 0.04.

# 6 Discussion

This paper develops a Bayesian version of restricted likelihood where posterior inference is conducted by conditioning on a summary statistic rather than the complete data. The framework blends classical estimation with Bayesian methods. Here, we concentrate on outlier-prone settings where natural choices for the conditioning statistic are classical robust estimators targeting the mean of the non-outlying data (e.g., M-estimators). The likelihood conditioned on these estimators is used to move from prior to posterior. The update follows Bayes' Theorem, conditioning on the observed estimators exactly. Computation is driven by MCMC methods, requiring only a supplement to existing algorithms by adding a Gibbs step to sample from the space of data sets satisfying the observed statistic. This step has additional computation costs arising from the need to compute the estimator and an orthonormal basis derived from gradients of the estimator at each iteration. The cost of finding the basis can be reduced by exploiting properties of the geometric space from which the samples are drawn as described in Section 3.2. We have seen good mixing of the MCMC chains across a wide-variety of examples. We have found the benefits of using our Bayesian technique to outweigh the additional computational burden (relative to a classical estimator) in the situation where substantive prior information that will impact the results is available.

The Bayesian restricted likelihood framework can be used to address model misspecification, of which the presence of outliers is but one example. The traditional view is that, if the model is inadequate, one should build a better model. In our empirical work, as data sets have become larger and more complex, we have bumped into settings where we cannot realistically build the perfect model. We ask the question "by attempting to improve our model through elaboration, will the overall performance of the model suffer?" If yes, we avoid the elaboration, retaining a model with some level of misspecification. Acknowledging that the model is misspecified implies acknowledging that the sampling density is incorrect, exactly as we do when outliers are present. In this sense, misspecified models and outliers are reflections of the same phenomenon, and we see restricted likelihood as a method for dealing with this more general problem.

Outside of outlier-prone settings, we might condition on the results of a set of estimating equations designed to enforce a lexical preference for those features of the analysis considered most important, yet still producing inferences for secondary aspects of the problem. This leads to questions regarding the choice of summary statistic to apply. In the literature, great ingenuity has been used to create a wide variety of estimators designed to handle specific manifestations of a misspecified model. The estimators are typically accompanied by asymptotic results on consistency and limiting distribution. These results can be used as a starting point to choose appropriate conditioning statistics in specific settings. For example, a set of regression quantiles may be judged the most important feature of a model. It would then be natural to condition on the estimated regression quantiles and to use a flexible prior distribution to allow for nonlinearities in the quantiles. The computational strategies we have devised allow us to apply our methods in this setting and to make full predictive inference. In general, we recommend a choice of conditioning statistic based on the analyst's understanding of the problem, model, reality, deficiencies in the model, inferences to be made, and the relative importance of various inferences.

The framework we develop here allows us to retain many benefits of Bayesian methods: it requires a complete model for the data; it lets us combine various sources of information both through the use of a prior distribution and through creation of a hierarchical model; it guarantees admissibility of our decision rules among the class based on the summary statistic $T(\boldsymbol{y})$; and it naturally leads us to focus on predictive inference. The work does open a number of questions for further work, including a need to investigate restricted likelihood methods as they relate to model selection, model averaging for predictive performance, and model diagnostics.

## Supplementary Material

Bayesian Restricted Likelihood Methods: Conditioning on Insufficient Statistics in Bayesian Regression – Supplementary Materials (DOI: 10.1214/21-BA1257SUPP; .pdf).

## References

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M., and Robert, C. P. (2009). "Adaptive Approximate Bayesian Computation." *Biometrika*, 96(4): 983–990. MR2767283. doi: https://doi.org/10.1093/biomet/asp052. 1398

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). "Approximate Bayesian Computation in Population Genetics." *Genetics*, 162: 2025–2035. 1397

Berger, J. (2006). "The Case for Objective Bayesian Analysis." *Bayesian Analysis*, 1: 385–402. MR2221271. doi: https://doi.org/10.1214/06-BA115. 1394

Bernardo, J. M. and Smith, A. (2000). *Bayesian Theory*. John Wiley & Sons Ltd. MR1274699. doi: https://doi.org/10.1002/9780470316870. 1394

Clarke, B. and Ghosh, J. K. (1995). "Posterior Convergence Given the Mean." *The Annals of Statistics*, 23: 2116–2144. MR1389868. doi: https://doi.org/10.1214/aos/1034713650. 1397

Clarke, J. L., Clarke, B., Yu, C.-W., et al. (2013). "Prediction in M-complete Problems with Limited Sample Size." *Bayesian Analysis*, 8(3): 647–690. MR3102229. doi: https://doi.org/10.1214/13-BA826. 1394

Clyde, M. and George, E. I. (2004). "Model Uncertainty." *Statistical Science*, 81–94. MR2082148. doi: https://doi.org/10.1214/088342304000000035. 1394

Clyde, M. A. and Iversen, E. S. (2013). "Bayesian Model Averaging in the M-open Framework." *Bayesian Theory and applications*. MR3221178. doi: https://doi.org/10.1093/acprof:oso/9780199695607.003.0024. 1394

Doksum, K. A. and Lo, A. Y. (1990). "Consistent and Robust Bayes Procedures for Location Based on Partial Information." *The Annals of Statistics*, 18: 443–453. MR1041403. doi: https://doi.org/10.1214/aos/1176347510. 1397

Drovandi, C., Pettitt, A., and Lee, A. (2015). "Bayesian Indirect Inference Using a Para-

metric Auxiliary Model." *Statistical Science*, 30: 72–95. MR3317755. doi: https://doi.org/10.1214/14-STS498. 1397

Fearnhead, P. and Prangle, D. (2012). "Constructing Summary Statistics for Approximate Bayesian Computation: Semi-Automatic Approximate Bayesian Computation." *Journal of the Royal Statistical Society: Series B*, 74: 419–474. MR2925370. doi: https://doi.org/10.1111/j.1467-9868.2011.01010.x. 1397

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). "Statistical Methods for Eliciting Probability Distributions." *Journal of the American Statistical Association*, 100: 680–701. MR2170464. doi: https://doi.org/10.1198/016214505000000105. 1394

Gelfand, A. E. and Smith, A. F. M. (1990). "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association*, 85: 398–409. MR1141740. 1403

Gelman, A. (2006). "Multilevel (Hierarchical) Modeling: What It Can and Cannot Do." *Technometrics*, 48(3): 432–435. MR2252307. doi: https://doi.org/10.1198/004017005000000661. 1424

Hampel, F. R. (1971). "A General Qualitative Definition of Robustness." *The Annals of Mathematical Statistics*, 42: 1887–1896. MR0301858. doi: https://doi.org/10.1214/aoms/1177693054. 1394

Hastings, W. K. (1970). "Monte Carlo Sampling Methods Using Markov Chains and Their Applications." *Biometrika*, 57: 97–109. MR3363437. doi: https://doi.org/10.1093/biomet/57.1.97. 1403

Hoff, P., Fosdick, B., Volfovsky, A., and Stovel, K. (2013). "Likelihoods for Fixed Rank Nomination Networks." *Network Science*, 1: 253–277. 1397

Huber, P. and Ronchetti, E. (2009). *Robust Statistics*. Wiley Series in Probability and Statistics. Hoboken, New Jersey: John Wiley & Sons, Inc, 2nd edition. MR2488795. doi: https://doi.org/10.1002/9780470434697. 1399, 1403, 1410

Huber, P. J. (1964). "Robust Estimation of a Location Parameter." *The Annals of Mathematical Statistics*, 35(1): 73–101. MR0161415. doi: https://doi.org/10.1214/aoms/1177703732. 1401

Hwang, H., So, B., and Kim, Y. (2005). "On Limiting Posterior Distributions." *Test*, 14: 567–580. MR2211395. doi: https://doi.org/10.1007/BF02595418. 1397

Joyce, P. and Marjoram, P. (2008). "Approximately Sufficient Statistics and Bayesian Computation." *Statistical Applications in Genetics and Molecular Biology*, 7(1). MR2438407. doi: https://doi.org/10.2202/1544-6115.1389. 1397

Jung, Y., MacEachern, S., and Lee, Y. (2014). "Cross-validation via Outlier Trimming." In preparation. 1418

Kass, R. E. and Raftery, A. E. (1995). "Bayes Factors." *Journal of the American Statistical Association*, 90: 773–795. MR3363402. doi: https://doi.org/10.1080/01621459.1995.10476572. 1394

Kass, R. E. and Wasserman, L. (1995). "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion." *Journal of the American Statistical Association*, 90(431): 928–934. MR1354008. 1400, 1417

Lee, J. and MacEachern, S. N. (2014). "Inference Functions in High Dimensional Bayesian Inference." *Statistics and Its Interface*, 7(4): 477–486. MR3302376. doi: https://doi.org/10.4310/SII.2014.v7.n4.a5. 1394, 1398

Lewis, J. (2014). "Bayesian Restricted Likelihood Methods." Ph.D. thesis, The Ohio State University. MR3337628. 1396, 1401, 1403

Lewis, J., Lee, Y., and MacEachern, S. (2012). "Robust Inference via the Blended Paradigm." In *JSM Proceedings*, Section on Bayesian Statistical Science, 1773–1786. American Statistical Association. 1397

Lewis, J. R., MacEachern, S. N., and Lee, Y. (2021). "Supplementary Material of "Bayesian Restricted Likelihood Methods: Conditioning on Insufficient Statistics in Bayesian Regression"." *Bayesian Analysis*. doi: https://doi.org/10.1214/21-BA1257SUPP. 1395

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g Priors for Bayesian Variable Selection." *Journal of the American Statistical Association*, 103: 410–423. MR2420243. doi: https://doi.org/10.1198/016214507000001337. 1403, 1417

Liu, J. S. (1994). "The Collapsed Gibbs Sampler in Bayesian Computations with Applications to a Gene Regulation Problem." *Journal of the American Statistical Association*, 89: 958–966. MR1294740. 1403

Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). "Markov Chain Monte Carlo without Likelihoods." *Proceedings of the National Academy of Sciences of the United States of America*, 100: 15324–15328. 1397

Maronna, R., Martin, D., and Yohai, V. (2006). *Robust Statistics: Theory and Methods*. Wiley Series in Probability and Statistics. West Sussex, England: John Wiley & Sons, Ltd. MR2238141. doi: https://doi.org/10.1002/0470010940. 1403

Miao, J. and Ben-Israel, A. (1992). "On Principal Angles Between Subspaces in $\mathbb{R}^n$." *Linear Algebra and its Applications*, 171: 81–98. MR1165446. doi: https://doi.org/10.1016/0024-3795(92)90251-5. 1408, 1409

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons. 1394

Pettitt, A. N. (1982). "Inference for the Linear Model using a Likelihood Based on Ranks." *Journal of the Royal Statistical Society. Series B*, 44: 234–243. MR0676214. 1397

Pettitt, A. N. (1983). "Likelihood Based Inference Using Signed Ranks for Matched Pairs." *Journal of the Royal Statistical Society. Series B*, 45: 287–296. MR0676214. 1397

Pratt, J. W. (1965). "Bayesian Interpretation of Standard Inference Statements." *Journal of the Royal Statistical Society. Series B*, 27: 169–203. MR0196830. 1397

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). "Population Growth of Human Y Chromosomes: A Study of Y Chromosome Microsatellites." *Molecular Biology and Evolution*, 16: 1791–1798. 1397, 1398

Ratcliff, R. (1993). "Methods for Dealing with Reaction Time Outliers." *Psychological Bulletin*, 114: 510. 1396

Ronchetti, E., Field, C., and Blanchard, W. (1997). "Robust Linear Model Selection by Cross-Validation." *Journal of the American Statistical Association*, 92: 1017–1023. MR1482132. doi: https://doi.org/10.2307/2965566. 1417

Rousseeuw, P. J. and Leroy (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons. MR0914792. doi: https://doi.org/10.1002/0471725382. 1399

Savage, I. R. (1969). "Nonparametric Statistics: A Personal Review." *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, 31: 107–144. MR0248950. 1397

Stigler, S. M. (1977). "Do Robust Estimators Work with Real Data?" *The Annals of Statistics*, 5(6): 1055–1098. MR0455205. 1398

Tavaré, S., Balding, D. J., Griffiths, R., and Donnelly, P. (1997). "Inferring Coalescence Times from DNA Sequence Data." *Genetics*, 145: 505–518. 1397

Turner, B. M. and Van Zandt, T. (2012). "A Tutorial on Approximate Bayesian Computation." *Journal of Mathematical Psychology*, 56(2): 69–85. MR2909506. doi: https://doi.org/10.1016/j.jmp.2012.02.005. 1398

Turner, B. M. and Van Zandt, T. (2014). "Hierarchical Approximate Bayesian Computation." *Psychometrika*, 79(2): 185–209. MR3255116. doi: https://doi.org/10.1007/s11336-013-9381-x. 1398, 1421, 1422

Wilkinson, R. D. (2013). "Approximate Bayesian Computation (ABC) Gives Exact Results Under the Assumption of Model Error." *Statistical Applications in Genetics and Molecular Biology*, 12(2): 129–141. MR3071024. doi: https://doi.org/10.1515/sagmb-2013-0010. 1421

Wong, H. and Clarke, B. (2004). "Improvement Over Bayes Prediction in Small Samples in the Presence of Model Uncertainty." *Canadian Journal of Statistics*, 32(3): 269–283. MR2101756. doi: https://doi.org/10.2307/3315929. 1397

Yuan, A. and Clarke, B. (2004). "Asymptotic Normality of the Posterior Given a Statistic." *The Canadian Journal of Statistics*, 32: 119–137. MR2064396. doi: https://doi.org/10.2307/3315937. 1397

Yuan, A. and Clarke, B. S. (1999). "A Minimally Informative Likelihood for Decision Analysis: Illustration and Robustness." *Canadian Journal of Statistics*, 27(3): 649–665. MR1745829. doi: https://doi.org/10.2307/3316119. 1394

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis

with g-prior distributions." In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 233. MR0881437. 1417

Zhu, H., Ibrahim, J. G., and Tang, N. (2011). "Bayesian Influence Analysis: A Geometric Approach." *Biometrika*, 98(2): 307–323. MR2806430. doi: https://doi.org/10.1093/biomet/asr009. 1394

# Invited Discussion*

Christian P. Robert†

*"...the prior distribution, the loss function, and the likelihood or sampling density (...) a healthy skepticism encourages us to question each of them."*

This paper by John Lewis, Steven MacEachern, and Yoonkyung Lee, starts with the great motivation of a misspecified model requiring the use of a (thus necessarily) insufficient statistic and moving to their central concern of simulating the posterior based on that statistic. Indeed, model misspecification sadly remains understudied from a Bayesian perspective and this paper is thus most welcome in addressing the issue. However, when reading through it, one of my reservations is in the authors defining misspecification as equivalent to the presence of outliers in the sample. In my opinion, an outlier model stands as a relatively easy case of misspecification, since the original and hypothetical model remains meaningful for the "good" part of the data. Furthermore, it seems to me that, under this outlying assumption, adding a non-parametric component for the unspecified part of the data would sound like a "more Bayesian" alternative (Robert and Rousseau, 2002). The problem in selecting a statistic $T$ is not really discussed in the paper, while every choice of a statistic $T$ leads to a different answer to what misspecified means and suggests a comparison with Bayesian empirical likelihood (Lazar, 2003; Yang and He, 2012; Mengersen et al., 2013).

I must admit that, when I first reached the Markov chain Monte Carlo (MCMC) component of the paper, I wondered at its relevance for the misspecification issues that sounded central above, before realising this had become the central focus of the paper. I cannot but agree that simulating the observations conditional on a value of the summary statistic $T$ is a true simulation challenge. I remember for instance George Casella mentioning it in association with a Student's $t$ sample in the 1990's. In the same vein, Persi Diaconis has written several papers on the problem Diaconis and Sturmfels (1998) and I am somewhat surprised at the dearth of references on this far from unexplored area, including also the recent papers by Byrne and Girolami (2013); Florens and Simoni (2016); Bornn et al. (2019). In the present case, the linear model assumed as the true model has the rather exceptional feature that it leads to a feasible transform of an unconstrained simulation into a simulation with fixed insufficient statistic $T(y)$, with no ensuing measure theoretic worries if not free from considerable efforts to establish the operation is truly valid. And, while simulating $(\theta, y)$ makes perfect sense in an insufficient setting, the simulation cost is precisely the same as when running a vanilla Approximate Bayesian computation (ABC) (Sisson et al., 2018). This natural comparison with ABC thus begs for the following remark. While taking $\epsilon = 0$ may sound optimal for being "exact", it is not so from an ABC perspective since the convergence rate (in $n$) of the (summary) statistic should be roughly the one of the tolerance (Li and Fearnhead, 2018; Frazier et al., 2018). I also note that, in its practical implementation,

*This work was partly supported by a PaRis AI Research InstitutE (prAIrie) from the Agence Nationale de la Recherche (ANR-19-P3IA-0001).

†CEREMADE, Université Paris Dauphine PSL, University of Warwick, and CREST, xian@ceremade.dauphine.fr

ABC does not suffer from low acceptance rates, since the tolerance is derived (as a quantile) from the simulated distances, i.e., induced by the simulations themselves.

While this may sound irrelevant, let me last mention, if only as a side note for measure-theoretic purists, that the derivation of the conditional distribution of $y$ given $T(y) = T_0$ is usually arbitrary since the conditioning event has probability zero (i.e., the conditioning set is of measure zero). This connects with the Borel-Kolmogorov paradox. The computations in the paper are correct, obviously, but they also rely on one among many choices of a transform.

## References


Bornn, L., Shephard, N., and Solgi, R. (2019). "Moment conditions and Bayesian nonparametrics." *Journal of the Royal Statistical Society, Series B*, 81: 5–43. MR3904778. doi: https://doi.org/10.1111/rssb.12294. 1431

Byrne, S. and Girolami, M. (2013). "Geodesic Monte Carlo on Embedded Manifolds." *Scandinavian Journal of Statistics*, 40(4): 825–845. MR3145120. doi: https://doi.org/10.1111/sjos.12036. 1431

Diaconis, P. and Sturmfels, B. (1998). "Algebraic algorithms for sampling from conditional distributions." *Annals of Statistics*, 26: 363–397. MR1608156. doi: https://doi.org/10.1214/aos/1030563990. 1431

Florens, J.-P. and Simoni, A. (2016). "Regularizing priors for linear inverse problems." *Econometric Theory*, 32(1): 71–121. MR3442503. doi: https://doi.org/10.1017/S0266466614000796. 1431

Frazier, D. T., Martin, G. M., Robert, C. P., and Rousseau, J. (2018). "Asymptotic properties of approximate Bayesian computation." *Biometrika*, 105(3): 593–607. MR3842887. doi: https://doi.org/10.1093/biomet/asy027. 1431

Lazar, N. A. (2003). "Bayesian empirical likelihood." *Biometrika*, 90: 319–326. MR1986649. doi: https://doi.org/10.1093/biomet/90.2.319. 1431

Li, W. and Fearnhead, P. (2018). "On the asymptotic efficiency of approximate Bayesian computation estimators." *Biometrika*, 105(2): 285–299. MR3804403. doi: https://doi.org/10.1093/biomet/asx078. 1431

Mengersen, K., Pudlo, P., and Robert, C. (2013). "Bayesian computation via empirical likelihood." *Proceedings of the National Academy of Sciences*, 110(4): 1321–1326. 1431

Robert, C. and Rousseau, J. (2002). "A Mixture Approach to Bayesian Goodness of Fit." Technical report, Cahiers du CEREMADE, Université Paris Dauphine. 1431

Sisson, S., Fan, Y., and Beaumont, M. (2018). *Handbook of Approximate Bayesian Computation*. New York: Chapman and Hall/CRC. MR3930567. 1431

Yang, Y. and He, X. (2012). "Bayesian empirical likelihood for quantile regression." *Annals of Statistics*, 40: 1102–1131. MR2985945. doi: https://doi.org/10.1214/12-AOS1005. 1431

# Invited Discussion

Bertrand Clarke*

## 1 Likelihood Selection

Arguably, the chief contribution of this paper is the computational technique given in Subsection 3.2. This new technique is effective in the context of the factorization (3) given at the beginning of Subsection 2.2. Sections 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 performs as claimed.

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

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

*bclarke3@unl.edu

(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 $\theta$.

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), Figures 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 $\theta$ 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 alternatives 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. $\mathcal{M}$-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 definitions 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 unable to formulate it in any realistic way, perhaps due to lack of data or other information. An example of this can be seen in standard one-way analysis of variance (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, the 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

[1] Actually, Rissanen showed that in the standard autoregressive moving average case with $p$ autoregressive terms and $q$ model avearge terms (ARMA$(p, q)$), 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 to this result 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.

guidance as to how to choose a statistic or a likelihood for it well. On the other hand, adopting a prequential approach removes much of the seeming excess flexibility by imposing a predictive performance criterion. As argued elsewhere, e.g., Section 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 sequence of *i.i.d.* random variables with common density function $f_\theta(x)$ and distribution function $F_\theta(x)$, $\alpha$ 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 $\Omega$ 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 $\theta_0$, and that $\mu''$ 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 likelihood 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 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 descriptive 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. MR2546391. doi: https://doi.org/10.1098/rsta.2009.0119. 1433

Berk, R. (1970). "Consistency a posteriori." *Annals of Mathematical Statistics*, 41: 894–906. MR0266356. doi: https://doi.org/10.1214/aoms/1177696967. 1436

Bernardo, J. and Smith, A. (2000). *Bayesian Theory*. Chichester: John Wiley & Sons, 2 edition. MR1274699. doi: https://doi.org/10.1002/9780470316870. 1435

Clarke, B., Clarke, J., and Yu, C.-W. (2014). "Statistical Problem Classes and Their Links to Information Theory." *Econometric Reviews*, 33: 337–371. MR3170852. doi: https://doi.org/10.1080/07474938.2013.807190. 1436

Clarke, B. and Ghosh, J. (1995). "Posterior convergence given the mean." *Annals of Statistics*, 23: 2116–2144. MR1389868. doi: https://doi.org/10.1214/aos/1034713650. 1436

Clarke, B. and Le, T. (2021). "Model averaging is asymptotically better than model selection for prediction." Submitted. 1436

Dawid, A. P. (1984). "The prequential approach." *Journal of the Royal Statistical Society*, 147: 287–292. MR0763811. doi: https://doi.org/10.2307/2981683. 1435

Dawid, A. P. and Vovk, V. (1999). "Prequential probability: principles and properties." *Bernoulli*, 5: 125–162. MR1673572. doi: https://doi.org/10.2307/3318616. 1435

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. MR2113610. 1433

Le, T. and Clarke, B. (2016). "Using the Bayesian Shtarkov solution for predictions." *Computational Statistics & Data Analysis*, 104: 183–196. MR3540994. doi: https://doi.org/10.1016/j.csda.2016.06.018. 1435

Le, T. and Clarke, B. (2021). "Interpreting uninterpretable predictors: kernel methods, Shtarkov solutions, and random forests." *To appear: Statistical Theory and Related Fields*. 1436

Lewis, J., MacEachern, S., and Lee, Y. (2021). "Bayesian Restricted Likelihood Methods: Conditioning on Insufficient Statistics in Bayesian Regression." *Bayesian Analysis*, 16: 1–38. MR3337628. 1433, 1434, 1435, 1436, 1437

Rissanen, J. (1984). "Universal Coding, Information, Prediction, and Estimation." *IEEE Transactions on Information Theory*, 30: 629–636. MR0755791. doi: https://doi.org/10.1109/TIT.1984.1056936. 1435

Shtarkov, Y. (1987). "Universal sequential coding of single messages." *Problems in Information Transmission*, 23: 3–17. MR0914346. 1435

Wang, T. and Lin, Q. (2021). "Hybrid Predictive Models: When an Interpretable Model Collaborates with a Black-box Model." *Journal of Machine Learning Research*, 22: 1–38. MR4318493. doi: https://doi.org/10.1080/15502287.2020.1853852. 1436

Yuan, A. and Clarke, B. (1999). "Posterior normality given order statistics." Unpublished manuscript. 1436

Yuan, A. and Clarke, B. (2004). "Asymptotic normality of the posterior given a statistic." *Canadian Journal of Statistics*, 32: 119–137. MR2064396. doi: https://doi.org/10.2307/3315937. 1436

# Invited Discussion

Fabrizio Ruggeri*

## 1 Discussion

I start by congratulating the authors for an original idea and its brilliant implementation and then I move on to a question to which I wish I could have an answer of my own. I am not a deep expert of Bayesian foundations but I grew up thinking that the likelihood function contains all the necessary evidence about the parameters of a given statistical model. The question is: how do you cope with the likelihood principle?

Another curiosity concerns the behaviour of the procedure when there are no outliers: there should be some protection if applied inappropriately. A similar problem arises when there are influential data.

The authors write that the "tuning parameters for the M-estimators are chosen to achieve 95% efficiency under normality". Although used in classical robustness studies, this choice seems very arbitrary to me, especially when data are far from normality. I wonder if there is a way, similar to Akaike Information Criterion (AIC), to choose among different values of those parameters. This could be a relatively simple way to select the parameters, without resorting to more complex approaches such as a prior distribution or a loss function within a decision theoretic framework.

Comparisons have been done with Gaussian and t distributions, or mixtures of the former. I wonder what would have happened if the authors had considered a more robust distribution, such as the family of exponential power-series distributions introduced by Box and Tiao in 1962.

I think the authors should mention another approach meant to deal with outliers: the choice of a (broad) class of statistical models or (neighbourhood of) likelihood functions and the computation of the range (lower and upper bounds) of the Bayesian estimator of the parameter of interest. This approach is (or was?) known as robust Bayesian analysis.

The authors have presented a remarkable approach but it would be more effective if they could provide guidance on which M-estimators are recommended in different situations.

According to the authors, the data augmented Markov chain Monte Carlo algorithm is one of the major contributions of their work. I am impressed by the way they have dealt with it and I expect other discussants, more involved in computational aspects, will comment on it. I have only a concern about the increase in computational complexity when dealing with very high dimensions, both in sample size and parameter space. I would like to know more about it.

*CNR IMATI – Via Alfonso Corti 12 – 20133 Milano – Italy, fabrizio@mi.imati.cnr.it; url: http://www.mi.imati.cnr.it/fabrizio

I have also a suggestion about the comparison of different values of the parameters of the M-estimators. At the beginning of Section 3.2, the authors mention a Gibbs sampler with two full conditionals: the first one is the same as the full data posterior whereas the second one depends on the chosen statistics. I think it could be possible to run many chains in parallel with the first full conditional common to all of them and the second one with different values of the parameters.

It would be interesting to see extensions of the current work in other frameworks. The most obvious one is about generalised linear models, but one could also think of non-homogeneous Poisson processes or other stochastic processes.

I believe the work by Lewis, MacEachern and Lee can stimulate further methodological research and be applied in many practical situations. I commend them one more time for a remarkable paper.

# Contributed Discussion*

Christopher Drovandi†, David J. Nott‡, and David T. Frazier§

## 1 Discussion

We congratulate the authors on a very interesting article. The paper was of interest to us, and perhaps the wider likelihood-free community, for at least two reasons. First, this is another example of how conditioning on summary statistics can be useful outside the intractable likelihood setting. Second, the authors devise a way to condition exactly (for linear regression models) on the observed statistic, avoiding the error from imperfect matching associated with approximate Bayesian computation (ABC).

For inference about unknown parameters $\boldsymbol{\theta}$ based on data $\boldsymbol{y}$, with observed value $\boldsymbol{y}_{\text{obs}}$, the authors make a compelling case for Bayesian inference conditional on a robust summary statistic $T_{\text{obs}} = T(\boldsymbol{y}_{\text{obs}})$ when there is concern about model misspecification. When $T(\boldsymbol{y})$ is an insufficient statistic, the likelihoods $f(\boldsymbol{y}_{\text{obs}}|\boldsymbol{\theta})$ and $f(T_{\text{obs}}|\boldsymbol{\theta})$ differ, and for certain choices of $T(\cdot)$, the "restricted likelihood" can be less sensitive to model misspecification. This leads to Bayesian inference on $\boldsymbol{\theta}$ via the "restricted" posterior, $\pi(\boldsymbol{\theta}|T_{\text{obs}}) \propto f(T_{\text{obs}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. Accessing the restricted likelihood $f(T_{\text{obs}}|\boldsymbol{\theta})$ using conventional methods can be difficult. The authors circumvent this issue by generating samples from, in turn, $\boldsymbol{y} \sim f(\boldsymbol{y}|\boldsymbol{\theta}, T(\boldsymbol{y}) = T_{\text{obs}})$, and $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}|\boldsymbol{y})$. The former sampling step requires that certain Jacobian terms be exactly calculated. This can be achieved in the important case of linear regression models with some specific choices of robust summary statistics.

We are interested in the authors opinion about whether this exact conditioning approach can be extended to more complex regression models? In more complex settings, likelihood-free computational methods that the paper avoids may be re-visited.

We would like to bring the authors attention to another useful likelihood-free method called synthetic likelihood (SL) (Wood, 2010; Price et al., 2018), which could be particularly useful in complex regression models where exact conditioning is difficult. SL also targets a posterior based on the restricted likelihood $\pi(\boldsymbol{\theta}|T_{\text{obs}}) \propto f(T_{\text{obs}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. However, SL approximates $f(T(\boldsymbol{y})|\boldsymbol{\theta})$ directly using a multivariate normal density, leading to the "synthetic likelihood" $N\{T_{\text{obs}}; b(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta})\}$, where $b(\boldsymbol{\theta}) = \mathbb{E}[T(\boldsymbol{y})|\boldsymbol{\theta}]$ and $\Sigma(\boldsymbol{\theta}) = \text{Var}\,[T(\boldsymbol{y})|\boldsymbol{\theta}]$ can be estimated from $m$ independent simulated datasets. Combining the estimated SL with a prior yields the Bayesian SL (BSL) posterior $\pi_{\text{BSL}}(\boldsymbol{\theta}|T_{\text{obs}})$. BSL is appealing since it does not require tuning of the tolerance and distance function

*This work was supported by the Australian Research Council and a Singapore Ministry of Education Academic Research Fund Tier 1 grant.

†Centre for Data Science, Queensland University of Technology, Australia, c.drovandi@qut.edu.au

‡Department of Statistics and Applied Probability, National University of Singapore, Singapore, standj@nus.edu.sg

§Department of Econometrics and Business Statistics, Monash University, Australia, david.frazier@monash.edu

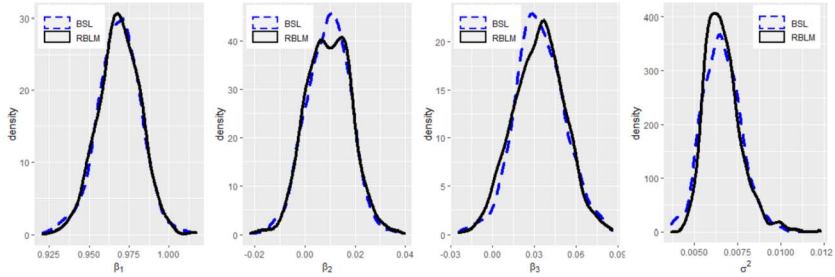


Figure 1: Comparison of BSL and the restricted Bayesian linear model (RBLM).

as in ABC. Further, it has been shown (Price et al., 2018; Frazier et al., 2021) that the BSL posterior depends weakly on $m$, so that it can be chosen to maximise computational efficiency. Third, when the summary statistic is asymptotically Gaussian, Frazier et al. (2021) show that BSL is more computationally efficient than ABC.

The robust summaries considered by the authors are asymptotically Gaussian under mild conditions, and BSL may be of interest for more complex regression models with similar summaries. If $f(T_{\text{obs}}|\boldsymbol{\theta}) \approx N\{T_{\text{obs}}; b(\boldsymbol{\theta}), \Sigma(\boldsymbol{\theta})\}$, then BSL can well approximate $\pi(\boldsymbol{\theta}|T_{\text{obs}})$. Under regularity conditions (Yuan and Clarke, 2004), $\pi(\sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)|T_{\text{obs}})$ is well-approximated by $N\{\sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n); 0, \Sigma(\boldsymbol{\theta}_0)\}$, where $\hat{\boldsymbol{\theta}}_n$ is the restricted maximum likelihood estimate, and converges towards some $\boldsymbol{\theta}_0$. Similarly, Frazier et al. (2021) demonstrate that $\pi_{\text{BSL}}(\sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)|T_{\text{obs}})$ converges to $N\{\sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n); 0, \Sigma(\boldsymbol{\theta}_0)\}$. Hence, in large samples, or when $f(T_{\text{obs}}|\boldsymbol{\theta})$ is approximately Gaussian, $\pi_{\text{BSL}}(\sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)|T_{\text{obs}}) \approx \pi(\sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_n)|T_{\text{obs}})$. Although asymptotic arguments motivate the BSL normal likelihood, its semiparametric extensions often result in better finite sample approximations than a direct use of these asymptotic results.

To illustrate the potential of BSL for robust regression, we run BSL on the insurance agency dataset for state number 27, and use $m = 20$. As shown in Figure 1, the approximate posteriors are almost identical to the approach of the paper.

## References

Frazier, D., Nott, D. J., Drovandi, C., and Kohn, R. (2021). "Bayesian inference using synthetic likelihood: asymptotics and adjustments." *arXiv:1902.04827v4*. 1443

Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. (2018). "Bayesian synthetic likelihood." *Journal of Computational and Graphical Statistics*, 27(1): 1–11. MR3788296. doi: https://doi.org/10.1080/10618600.2017.1302882. 1442, 1443

Wood, S. N. (2010). "Statistical inference for noisy nonlinear ecological dynamic systems." *Nature*, 466(7310): 1102–1104. 1442

Yuan, A. and Clarke, B. (2004). "Asymptotic normality of the posterior given a statistic." *Canadian Journal of Statistics*, 32(2): 119–137. MR2064396. doi: https://doi.org/10.2307/3315937. 1443

# Contributed Discussion

Michael Lavine*

It was a pleasure to read this thoughtful and well-written article by Lewis, MacEachern, and Lee. It seems to us that it makes two main contributions: first, replacing the usual likelihood function with one that is hoped to be more robust to outliers and misspecification and second, providing a Markov Chain Monte Carlo (MCMC) scheme for drawing from the posterior distribution. We focus our remarks on the first contribution.

The authors' purpose in replacing the usual likelihood function is to summarize the data "through a set of insufficient statistics, targeting inferential quantities of interest"[1] out of "[c]oncern for imperfections in the likelihood."[2] To us, it seems the authors' purpose is very similar to that of marginal and conditional likelihoods in Royall (1997). In Royall's words (page 155),

> For example, when $X_1, \ldots, X_n$ are i.i.d. random variables, unless $n$ is small, we need not specify the precise form of the distribution of a single element, $X$, in order to confidently model the marginal distribution of $\bar{X}$ as normal (because of the central limit theorem). When the variance of this marginal distribution is replaced by a consistent estimator, the resulting estimated likelihood function for the mean, $EX$, is valid, in a specific approximate sense, under a wide range of specific parametric models for the distribution of $X$ (Tsou and Royall, 1995).

Indeed, the major contribution of Tsou and Royall (1995) is to

> . . . examine the concept of robustness as it relates to likelihood functions. We note five ways that likelihood functions can be used to represent and interpret statistical data as evidence. These various uses suggest corresponding senses in which one likelihood function can approximate another, and these in turn suggest different senses in which a likelihood function can be 'robust.' We establish some general relationships among these senses of robustness, and examine two general techniques for producing robust likelihoods.[3]

Tsou and Royall (1995) and Royall (1997) write about likelihood functions, not about posterior distributions, so they don't say they are conditioning on insufficient statistics. But that is, in effect, what they are doing, or would be doing if they used their robust likelihood functions to produce posteriors. Further development is in Royall and Tsou (2003) which makes a careful distinction between the object of *inference* and the object of *interest* when the hypothesized model is wrong.

Tsou and Royall (1995), Royall (1997), and Royall and Tsou (2003) consider robust likelihoods for inference about parameters, whereas Lewis, MacEachern, and Lee demonstrate the value of their methods for predictions. The two points of view might be combined by adapting the ideas of the first three papers to the setting of predictive likelihood. The basic idea of predictive likelihood is that predictands can be treated

*Department of Mathematics and Statistics, UMass, Amherst, lavine@math.umass.edu

[1] Quotation from the abstract.

[2] Quotation from page 2.

[3] Quotation from the abstract.

as unknowns similarly to parameters. See Bjørnstad (1990) and Bjørnstad (1996) for further discussion.

Finally, we would like to add to Lewis, MacEachern, and Lee's summary of the literature on concerns for imperfections in the likelihood. An approach they did not mention is that of Lavine (1991b) which "introduces a method for computing ranges of posterior expectations over reasonable classes of sampling distributions that lie 'close to' a given parametric family. By treating the prior as a probability measure on the space of sampling distributions this article also gives a unified treatment to what are usually considered two separate problems—sensitivity to the prior and sensitivity to the sampling model." See Lavine (1991a) for details. In those articles the posterior expectation could also be a predictive expectation.

## References

Bjørnstad, J. F. (1990). "Predictive Likelihood: A Review (with discussion)." *Statistical Science*, 5: 242–265. MR1062578. 1445

Bjørnstad, J. F. (1996). "On the Generalization of the Likelihood Function and the Likelihood Principle." *Journal of the American Statistical Association*, 91(434): 791–806. MR1395746. doi: https://doi.org/10.2307/2291674. 1445

Lavine, M. (1991a). "An Approach to Robust Bayesian Analysis for Multidimensional Parameter Spaces." *Journal of the American Statistical Association*, 86: 400–403. MR1137122. 1445

Lavine, M. (1991b). "Sensitivity in Bayesian Statistics: The Prior and the Likelihood." *Journal of the American Statistical Association*, 86: 396–399. 1445

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*. Monographs on Statistics & Applied Probability. Chapman and Hall/CRC. MR1629481. 1444

Royall, R. and Tsou, T.-S. (2003). "Interpreting Statistical Evidence by Using Imperfect Models: Robust Adjusted Likelihood Functions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65: 391–404. MR1983754. doi: https://doi.org/10.1111/1467-9868.00392. 1444

Tsou, T. and Royall, R. (1995). "Robust Likelihoods." *Journal of the American Statistical Association*, 90: 316–320. MR1325138. 1444

# Contributed Discussion

Jong-Min Kim*

## 1 Summary

It is a great honor to have the chance of congratulating the authors (John R. Lewis, Steven N. MacEachern, and Yoonkyung Lee) on an interesting and valuable paper. This paper develops a Bayesian version of restricted likelihood where posterior inference is conducted by conditioning on a summary statistic rather than the complete data. The authors found the benefits of using our Bayesian technique to outweigh the additional computational burden in the situation where substantive prior information that will impact the results is available. My suggestion to reduce computational burden can be Copula-based models which have received much attention in recent years in various fields because of several attractive properties. First, due to Sklar's theorem (Sklar, 1959), copulas allow us to model the marginal distributions and the joint dependence structure separately (Joe, 1997). Second, they are invariant under increasing and continuous transformations. Third, copulas do not require the normal distribution assumption to find the measure of dependence, unlike Pearson's correlation. Copula models have been widely used to model dependence between macroeconomic and financial time series (Cherubini et al., 2011).

Masarotto and Varin (2012) proposed Gaussian Copula marginal regression (GCMR) which implemented maximum simulated likelihood estimation based on a variant of the GHK algorithm (Geweke, Hajivassiliou and Keane) because Gaussian copula provides a mathematically convenient framework to handle various forms of dependence in regression models arising longitudinal and spatial data.

The extension of the GCMR method is Guolo and Varin (2014) marginal beta regression model exploits the probability integral transformation to relate response $Y_t$ to covariates $x_t$ and to a standard normal error $\epsilon_t$. Kim and Hwang (2017) proposed copula directional dependence by using the Guolo and Varin (2014) marginal extension of the beta regression model for time series analysis and the cumulative distribution function of a normal variable. But the GCMR method has also a computation burden. To reduce the computation cost, Masarotto and Varin (2017) suggested composite likelihoods to reduce the computational effort through convenient likelihood factorizations (Varin et al., 2011) and sparse methods designed to approximate the Gaussian copula correlation matrix with a more manageable block-diagonal matrix. The suggestion by Masarotto and Varin (2017) can be applied to the authors' paper.

Another interesting Copula approach for the authors' proposed method to reduce the computation cost is Wojtyś et al. (2016) sample selection models under the situation in which an outcome of interest is observed for a restricted non-randomly selected sample

*Statistics Discipline, Division of Science and Mathematics, University of Minnesota-Morris, Morris, MN, 56267, USA, jongmink@umn.edu

of the population. The estimation of these models is based on a binary equation, which describes the selection process, and an outcome equation, which is used to examine the substantive question of interest. Once again, I was very impressive to read the authors' paper. I hope my comment to this wonderful research paper will be helpful.

## References

Cherubini, U., Mulinacci, S., Gobbi, F., Romagnoli, S. *Dynamic Copula Methods in Finance*, first ed. Wiley, 2011. 1446

Guolo, A. and Varin, C. Beta regression for time series analysis of bounded data, with application to Canada Google flu trends. *The Annals of Applied Statistics* 2014; **8**(1): 74–88. MR3191983. doi: https://doi.org/10.1214/13-AOAS684. 1446

Joe, H. *Multivariate models and multivariate dependence concepts*. CRC Press, 1997. MR1462613. doi: https://doi.org/10.1201/b13150. 1446

Kim, J.-M., Hwang, S. Directional Dependence via Gaussian Copula Beta Regression Model with Asymmetric GARCH Marginals. *Communications in Statistics: Simulation and Computation* 2017; **46**(10), 7639–7653. MR3764992. doi: https://doi.org/10.1080/03610918.2016.1248572. 1446

Masarotto, G., Varin, C. Gaussian Copula Marginal Regression. *Electronic Journal of Statistics* 2012; **6**, 1517–1549. MR2988457. doi: https://doi.org/10.1214/12-EJS721. 1446

Masarotto, G., and Varin, C. Gaussian Copula Regression in R. *Journal of Statistical Software* 2017; **77**(8), 1–26. 1446

Sklar, A. Fonctions de repartition á n dimensions et leurs marges. *Publ.Inst. Statist. Univ. Paris* 1959; **8**: 229–231. MR0125600. 1446

Varin, C., Reid, N., Firth, D. An Overview of Composite Likelihood Methods. *Statistica Sinica* 2011; **21**, 5–42. MR2796852. 1446

Wojtyś, M., Marra, G., and Radice, R. Copula Regression Spline Sample Selection Models: The R Package SemiParSampleSel. *Journal of Statistical Software* 2016; **71**(6), 1–66. 1446

# Contributed Discussion*

Malay Ghosh† and Debashis Ghosh‡

We congratulate the authors for their novel contribution to the Bayesian literature. As mentioned by the authors, the success of any Bayesian method depends on three components: the likelihood, the prior and the loss. For an applied scientist, specification of the loss is not that critical or is often not explicitly dealt with, as often descriptive measures such as posterior means, medians, variances and quantiles suffice to meet inferential needs.

That leaves one with the likelihood and the prior. The present paper focuses on "imperfect likelihood". They seem also to be linking model misspecification with lack of model robustness. Introduction of M-estimators in this regard is an age-long practice as embraced in this article. The paper does a nice job summarizing the literature dating back to the work of Huber. One obvious extension beyond the current paper is to examine the utility of the restricted likelihood approach in comparison to Bayesian nonparametric methods (Hjort et al., 2010).

Much of the success of the present approach hinges not just on an arbitrary linear model, but specifically on the fixed effects linear regression model where the two basic components for inference are the least squares estimator of the regression parameters as well as the residual error variance. In fact, abusing the terminology, we can even label these estimators as "approximately sufficient" when no distributional assumption is made. Indeed these are minimal sufficient with the added assumption of normality. The very natural question that emerges then is how to extend the present proposal to other linear models, for example, in mixed linear models with unknown regression parameters as well as unknown variance components.

More specifically, the pivotal Theorem 3.1 does not seem to have a natural extension beyond what is given this paper. Particularly, the choice of the $T(\cdot)$ statistic becomes an arduous task as one steps outside the proposed linear model. Even the estimating equations related to M-estimators emanate from the given linear model. Proceeding to generalized linear models as well as nonlinear models, the choice of $T(\cdot)$ becomes formidable, although such a choice may turn out to be somewhat simpler for the former than the latter. A similar comment applies to the development of the proposal density of $y$ as done in Theorem 6, an ingenious derivation in this article. Some insight may be derived by the following heuristic argument. Suppose $\mathbf{z}^*$ is drawn from a distribution $F$ whose mean is $\mu_Z \in R^n$. Then $\mathbf{y} \equiv h(\mathbf{z}^*)$ is approximately equal to

$$\frac{s(X, \mathbf{y}_{obs})}{s(X, \mu_Z)}\mathbf{z}^* + X\left(\mathbf{b}(X, \mathbf{y}_{obs}) - \mathbf{b}\left(X, \frac{s(X, \mathbf{y}_{obs})}{s(X, \mu_Z)}\mu_Z\right)\right).$$

*D. Ghosh would like to acknowledge the support of NSF DMS 1914937.

†Department of Statistics, University of Florida, Gainesville, FL 32611-8545, ghoshm@ufl.edu

‡Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO, 80045, debashis.ghosh@cuanschutz.edu

Thus, we have a linearity condition for $\mathbf{y}|\mathbf{z}^*$ that bears some parallel to the linearity condition given in the sufficient dimension reduction literature (Li and Duan, 1989; Li, 1991; Brillinger, 2012). By drawing on the vast literature for sufficient dimension reduction, summarized recently in Li (2018), this might suggest models to which we can expand the approach to.

There seems to be a bigger issue involved. Many view the likelihood and the prior combined into a single multilevel model. Misspecification can occur in one or in the other, or even in both. Thus model diagnostics have turned out to be a real favorite tool for Bayesians, although there too one lacks a more or less universally accepted procedure.

In summary, we commend the authors for an original and thought-provoking article. But then the question is not unlike that of Alice in Wonderland: "Where do we go from here"?

## References

Brillinger, D. R. (2012). "A generalized linear model with "Gaussian" regressor variables." In *Selected Works of David Brillinger*, 589–606. Springer. MR2906069. doi: https://doi.org/10.1007/978-1-4614-1344-8. 1449

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (2010). *Bayesian nonparametrics*, volume 28. Cambridge University Press. MR2722988. doi: https://doi.org/10.1017/CBO9780511802478.002. 1448

Li, B. (2018). *Sufficient dimension reduction: Methods and applications with R*. CRC Press. MR3838449. doi: https://doi.org/10.1201/9781315119427. 1449

Li, K.-C. (1991). "Sliced inverse regression for dimension reduction." *Journal of the American Statistical Association*, 86(414): 316–327. MR1137117. 1449

Li, K.-C. and Duan, N. (1989). "Regression analysis under link violation." *The Annals of Statistics*, 1009–1052. MR1015136. doi: https://doi.org/10.1214/aos/1176347254. 1449

# Contributed Discussion

Shota Gugushvili* and Carel F. W. Peeters†

Motivated by robustness considerations against model misspecification, in this paper the Authors propose to perform posterior inference by conditioning on (insufficient) summary statistics rather than the full data at hand. Their proposal can be seen as a hybrid Bayesian-frequentist approach, or as a Bayesian take on restricted likelihood estimation. The Authors numerically implement and examine their idea in the linear model setting. We commend the Authors for their highly interesting, well-written contribution, and for expanding the literature at the interface of Bayesianism and frequentism. Below we bring up several points for discussion.

Our first point revolves around the issue of dimension. The Authors (at least partly) justify their approach as feasible in higher dimensions (in terms of the conditioning statistic $T(\mathbf{y}_{obs})$ and parameters $\boldsymbol{\theta}$). The success and efficiency of their Markov chain Monte Carlo (MCMC) sampler depends on the data augmentation step (Section 3.2). The latter happens to be a Metropolis-Hastings step to sample $\mathbf{y}$ conditional on the observed statistic $T(\mathbf{y}_{obs})$ (and the current parameter value $\boldsymbol{\theta}$). However, when the dimension of $\mathbf{y}$, i.e. the sample size $n$, is large, the Metropolis-Hastings steps, and as such the Authors' MCMC sampler too, might run into problems. A mathematical reason for this are the opposing forces of volume and density for probability measures in high-dimensional spaces; see, e.g., Betancourt (2018) and Giraud (2015). In particular, probability measures in high-dimensional spaces tend to concentrate on 'typical sets' that become increasingly singular as the dimension of the space grows. In the present setting, the situation becomes exacerbated with a growing parameter space or feature space dimension $p$ as well: though $\mathbf{y}$ is $n$-dimensional, when conditioned on $T(\mathbf{y}_{obs})$, its density effectively lives on the $(n-p-1)$-dimensional subspace that is much smaller than the original space. Performing naive Metropolis-Hastings steps in such a situation may effectively reduce to looking for a needle in a haystack. We also note that many situations of current interest are characterized by $p$ being (much) larger than $n$. This does not appear to be covered by the Authors' method, that requires $n-p-1>0$. We would appreciate if the Authors could provide some clarification regarding these musings.

Our second point concerns the assessment of the added practical value of the Authors' MCMC sampling method. As one of the referees has pointed out, the MCMC technique developed in the paper is not required when the asymptotic approximation is sufficient. An additional computational burden of the proposed method seems justified when, quoting the Authors, "substantive prior information that will impact the results is available". However, how does one reliably deduce that one indeed has such

*Mathematical & Statistical Methods Group (Biometris), Wageningen University & Research, Wageningen, The Netherlands, shota.gugushvili@wur.nl

†Mathematical & Statistical Methods Group (Biometris), Wageningen University & Research, Wageningen, The Netherlands, carel.peeters@wur.nl

substantive prior information? This question appears particularly difficult to us in the high-dimensional setting (large $p$).

The third point concerns a comparison of the proposed method to approximate Bayesian computation (ABC). The Authors note a similarity of their approach to ABC (Section 2.3), but regrettably do not perform a full comparison, except in the example from Section 5.2. Now both practical experience and some recent theoretical work (see Frazier et al. 2018 and Frazier et al. 2020) tell us that ABC runs into problems in misspecified and high-dimensional model settings. Given a degree of similarity of the Authors' method to ABC, a fuller comparison with the latter would have been welcome.

The fourth point concerns the restricted notion of misspecification employed by the Authors. We realize that covering all the interesting questions is nigh impossible in a single paper. However, model misspecification can hardly be restricted to presence of a certain percentage of outliers in the data. We note that in the context of variational inference, Wang and Blei (2019) examine several relevant examples that go beyond this notion of misspecification. Can the Authors indicate how their methods would fare in the face of other departures from the model assumptions?

The fifth point deals with the question whether conditioning on insufficient statistics leads to adequate uncertainty quantification in inferential conclusions, even if the approximate posterior is roughly centered on the correct parameter value.

To conclude, we enjoyed reading the paper. Once again, we congratulate the Authors on their work, and look forward to their input in the discussion.

## References

Betancourt, M. (2018). "A Conceptual Introduction to Hamiltonian Monte Carlo." *arXiv:1701.02434*. MR1699395. doi: https://doi.org/10.1017/CBO9780511470813.003. 1450

Frazier, D. T., Martin, G. M., Robert, C. P., and Rousseau, J. (2018). "Asymptotic properties of approximate Bayesian computation." *Biometrika*, 105(3): 593–607. MR3842887. doi: https://doi.org/10.1093/biomet/asy027. 1451

Frazier, D. T., Robert, C. P., and Rousseau, J. (2020). "Model misspecification in approximate Bayesian computation: consequences and diagnostics." *Journal of the Royal Statistical Society – Series B*, 82(2): 421–444. MR4084170. doi: https://doi.org/10.1111/rssb.12356. 1451

Giraud, C. (2015). *Introduction to High-Dimensional Statistics*. Boca Raton, FL: CRC Press. MR3307991. 1450

Wang, Y. and Blei, D. (2019). "Variational Bayes under Model Misspecification." In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. 1451

# Contributed Discussion

Jack Jewson* and David Rossell†

We congratulate the authors for a thought-provoking paper. The idea is compelling: by conditioning on estimators that are robust to misspecification (e.g. outliers) one robustifies posterior inference. Our main comment is that robust estimators often require hyper-parameters that can critically affect the quality of inference, as pointed out by the authors: "the choice of summary statistic along with the corresponding tuning parameters is important". For illustration, inference using Tukey's loss can be sensitive to the "cut-off' parameter $k$ (Figure 1, left). The authors adopt a default $k = 4.685$ that ensures 95% efficiency if the data were Gaussian. While reasonable, such defaults are sub-optimal under stronger-than-expected contamination (Figure 1, right).

Ideally one would like to learn which hyper-parameter values are most appropriate for the data at hand. For example, one may learn the degrees of freedom for a Student's-$t$ model (which can be seen as a hyper-parameter) via standard inference, but more generally the task can be challenging. Losses such as Tukey's do not define a proper probability model on the data, so likelihood-based methods do not apply. It is nevertheless possible to learn hyper-parameters in such situations using a recent strategy in Jewson and Rossell (2021). The idea is to view Tukey's loss as defining an improper model indexed by $(\beta, \sigma, k)$ that can be embedded into a generalized Bayes framework, and to then find $\hat{k}$ such that the improper model best approximates the data-generating mechanism (in Fisher's divergence). The strategy uses the so-called Hyvärinen score (Hyvärinen, 2005), which can accommodate infinite normalization constants (improper models). Once $\hat{k}$ is obtained, one can apply the author's Bayesian restricted likelihood methods (BRLM). That is, one first learns how robust the summary statistics should be from the observed data, and then applies BRLM. If the contamination is little then one hopes to set large $\hat{k}$ ($k = \infty$ recovers the Gaussian model), whereas under strong contamination one hopes for small $\hat{k}$. Below we extend one of the authors' examples to illustrate that learning $\hat{k}$ can have non-negligible effects on inference. Doing so, robustifies inference to poor hyper-parameter choices, which we believe aligns with the motivation for BRLM.

We reproduce the authors' Simulation 2 where the parameters $\beta$ of a Bayesian linear regression are estimated under a one-sided outlier contamination. Figure 1 (left) illustrates the sensitivity of BRLM to the choice of $k$, for several prior variances $\sigma_\beta$. Learning $k$ from data improves the mean squared error (MSE) for all $\sigma_\beta$, whereas setting different defaults ($k = 2.5$ and $k = 6$) significantly increases MSE. Figure 1 (right) displays the relative MSE vs. the oracle least-squares using only uncontaminated data, showing that the default $k = 4.685$ becomes less efficient as the proportion of outliers increases, relative to learning $k$ via the $\mathcal{H}$-score.

*Department of Business and Economics, Universitat Pompeu Fabra, Barcelona, Spain, jack.jewson@upf.edu

†Department of Business and Economics, Universitat Pompeu Fabra, Barcelona, Spain, david.rossell@upf.edu

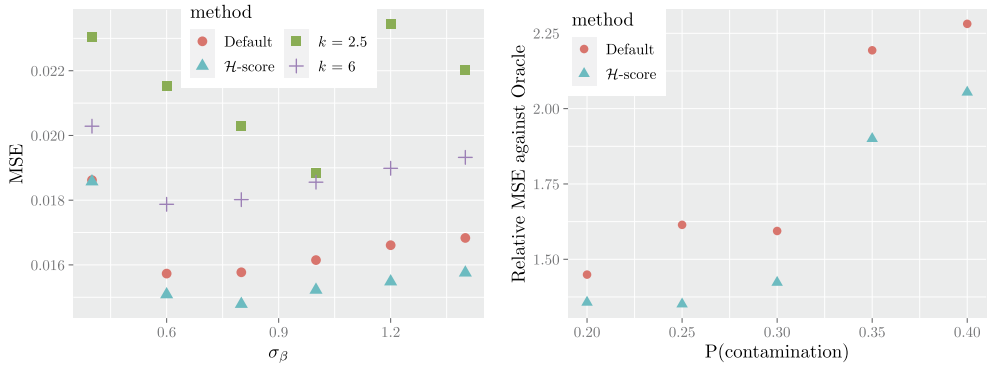


Figure 1: Left: MSE over $K = 30$ simulations under several prior specifications $\sigma_\beta$ and Tukey's hyper-parameter $k$. Right: relative MSE vs. oracle uncontaminated least-squares of default and $\mathcal{H}$-score estimated values of $k$ under increasing outlier contamination.

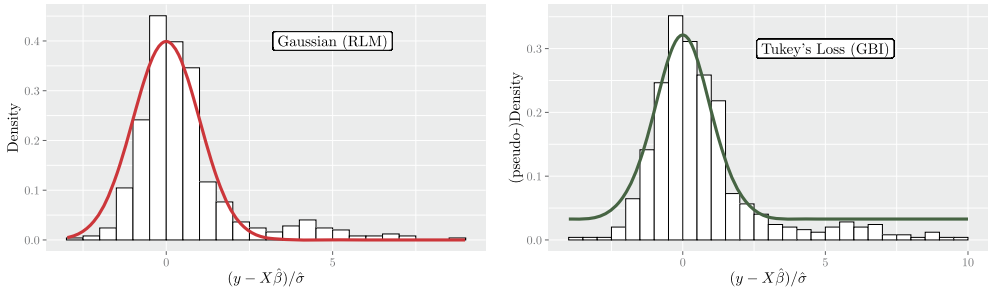


Figure 2: Restricted likelihood Gaussian density (left) and Tukey-based improper density (right) approximation to the fitted residuals produced by those methods.

Our second main comment is on predictive inference. As argued eloquently by the authors, in some settings one wishes to obtain a predictive distribution that represents future non-contaminated data, e.g. the BRLM-estimated Gaussian in Figure 2 (left). In other settings outliers are not a contamination but part of the inherent process, e.g. extreme weather or finance events. Then, the posterior predictive should acknowledge that future extreme events are possible. Such predictive distributions are non-standard when the loss does not define a proper model, e.g. for Tukey's loss (Figure 2 right). However, in a generalized Bayes framework they can still be interpreted as being informative about relative (rather than absolute) probabilities. Figure 2 right has a normal-like central component and flat tails, expressing ignorance on the magnitude of possible outliers. Both views (uncontaminated vs. all data) can be valuable, depending on the application.

## References

Hyvärinen, A. (2005). "Estimation of non-normalized statistical models by score matching." *Journal of Machine Learning Research*, 6(Apr): 695–709. MR2249836. 1452

Jewson, J. and Rossell, D. (2021). "General Bayesian Loss Function Selection and the use of Improper Models." *arXiv preprint arXiv:2106.01214*. 1452

# Contributed Discussion

Arnab Hazra*

I would first congratulate the authors for this thought-provoking paper in the area of Bayesian robust regression. For a random sample $\boldsymbol{y}$ from a continuous distribution indexed by a parameter vector $\boldsymbol{\theta}$, usual Bayesian techniques draw posterior inferences through $\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{\theta})$, where $\pi(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\theta}$ and $\pi(\boldsymbol{y}|\boldsymbol{\theta})$ is the data likelihood. Alternatively, the authors propose to draw inferences about $\boldsymbol{\theta}$ using $\pi(\boldsymbol{\theta}|T(\boldsymbol{y})) = \int \pi(\boldsymbol{\theta}, \boldsymbol{y}|T(\boldsymbol{y}))d\boldsymbol{y}$, where $T(\boldsymbol{y})$ is a robust estimator of $\boldsymbol{\theta}$. Further, this high dimensional integral is approximated using a Gibbs sampler with two full conditionals $\pi(\boldsymbol{\theta}|\boldsymbol{y}, T(\boldsymbol{y}))$ and $\pi(\boldsymbol{y}|\boldsymbol{\theta}, T(\boldsymbol{y}))$. In general, sampling from these full conditionals is non-trivial and the authors develop a computationally intensive but rigorous strategy.

There is a gigantic literature on the choices of $T(\cdot)$. The paper focuses on some more traditional choices like Huber's and Tukey's M-estimators, least median squares, and least trimmed squares. The class of M-estimators is also large; for example, a popular minimum density power divergence estimation (MDPDE) method was proposed by Basu et al. (1998) which has certain advantages over other M-estimators. To implement Bayesian MDPDE or some other classes, allowing user-defined estimating equations in the R package `brlm` would be beneficial. Ghosh and Basu (2013) proposed an MDPDE approach for linear regression in a frequentist setting and their estimating equations satisfy all the conditions C1 through C8 described in the paper. Thus, a Bayesian implementation using the technique developed in this paper is direct.

The method developed in Section 3 does not have any distributional assumption; the condition C2 only assumes the existence of a density with respect to Lebesgue measure on the real line. However, all the examples discussed in the paper assume normality. It would be helpful if some examples for other distributions are discussed.

The limiting posterior/sampling variances in Bayesian/frequentist settings are the same. Thus, for a large sample size, the performance of the proposed method for noninformative/weakly-informative priors and a traditional classical robust estimator would be similar. The sample sizes used in the simulation settings are generally not small. Some examples with smaller sample sizes would be helpful for clarification.

## References

Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. (1998). "Robust and efficient estimation by minimising a density power divergence." *Biometrika*, 85: 549–559. MR1665873. doi: https://doi.org/10.1093/biomet/85.3.549. 1455

Ghosh, A. and Basu, A. (2013). "Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression." *Electronic Journal of Statistics*, 7: 2420–2456. MR3117102. doi: https://doi.org/10.1214/13-EJS847. 1455

*Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, arnab.hazra@kaust.edu.sa

# Contributed Discussion

Kali Chowdhury*

This contribution respectfully wishes to discuss some thoughts and open questions on this work, based on existing literature. The authors outline a method that summarizes "the data through a set of insufficient statistics", where "the prior distribution is updated with the summary statistics rather than the complete data." The major contribution of this construction seems to be "the development of a data augmented" Markov Chain Monte Carlo (MCMC) algorithm for linear models and certain class of summary statistics.

Thus, it is discussed that Bayesian inference considers "the prior distribution, the loss function and the likelihood or sampling density." Further that imperfection in the likelihood could be due to "rounding of the observations", "the occurance of outliers", or the model being misspecified. The first of which may be overcome by MCMC and the "second and third are duals". In regards to the duals, I first kindly draw attention of the reader to Chowdhury (2017) where the author specifically discusses an outlier to be a model specific phenomenon, and thus both gross errors and model mispecifications may give rise to observed "outliers." In regards to the MCMC method, which generally require say ergodicity and/or aperiodicity for convergence, it may be good to better understand under what circumstances the proposed method would attain the true distribution, subject to the tuning parameters constraints.

Accordingly, please note that Chowdhury (2021a) discusses in detail how such a Markov chain may be induced through the conditional distributions of hierarchical parameters and why the distributional convergence results are robust. Chowdhury (2021b) and Chowdhury (2021c) greatly extended this approach to various even more general settings under minimal assumptions, which ensures almost sure convergence of the parameter estimates without the need for a tuning parameter, or insufficient statistics specifically. The algorithm termed Latent Adaptive Hierarchical Expectation-Maximization Like (LAHEML) algorithm, ensures that through a hierarchical model we may ensure almost sure convergence of the parameter estimates, as the link condition holds for all observations. As such, the authors correctly compare their method to other restricted likelihood methods and Approximate Bayesian Computation (ABC) already existing.[1]

This is apropos, since cross-validation under the Bayesian MCMC paradigm may add burdensome computational requirements and may not give unique convergence results. For example, while various MCMC runs may yield better results in particular situations, the problem of unique model specification may be elusive despite the various strong assumptions made (Assumptions C5–C8 need to be checked "on a case by case basis"). As such the LAHEML framework given in Chowdhury (2021a), Chowdhury (2021b), and Chowdhury (2021c) may be an alternative for robust strongly convergent estimators.

*Johns Hopkins University, kchowdh1@jh.edu

[1]Please also see Ren and Gu (1997) and Ray and Chatterjee (2020).

## References

Chowdhury, K. (2017). "Supervised Machine Learning and Heuristic Algorithms for Outlier Detection in Irregular Spatiotemporal Datasets." *Journal of Environmental Informatics*, 33(1). 1456

Chowdhury, K. (2021a). "Functional analysis of generalized linear models under non-linear constraints with applications to identifying highly-cited papers." *Journal of Informetrics*, 15(1): 101–112. 1456

Chowdhury, K. (2021b). "Functional analysis of generalized linear models under non-linear constraints with Artificial Intelligence and Machine Learning Applications to the Sciences." Ph.D. thesis, University of California, Irvine. 1456

Chowdhury, K. e. a. (2021c). "Nonparametric Application of Functional Analysis of Generalized Linear Models Under Nonlinear Constraints." In *Symposium on Data Science and Statistics*. American Statistical Association. 1456

Ray, D. and Chatterjee, N. (2020). "Effect of non-normality and low count variants on cross-phenotype association tests in GWAS." *European Journal of Human Genetics*, 28(3): 300–312. 1456

Ren, J.-J. and Gu, M. (1997). "Regression M-estimators with doubly censored data." *The Annals of Statistics*, 25(6): 2638–2664. MR1604432. doi: https://doi.org/10.1214/aos/1030741089. 1456

# Rejoinder

John R. Lewis*, Steven N. MacEachern†, and Yoonkyung Lee‡

It is a rare opportunity to receive commentary on one's work from thought leaders in the field of Bayesian statistics. We are grateful that the editorial staff selected our paper for discussion and have enjoyed reading the discussions. Many of the comments provide perspective of the kind best suited for a conversation. We look forward to a time when we can have those conversations with the discussants (and others!) in person.

The collection of discussions touches on a wide variety of issues, with most discussants making several points. There is considerable overlap across discussants, and so we have organized this brief rejoinder to non-exhaustively cover the issues that have been raised.

**Model misspecification** Several discussants and referees have viewed our work as perhaps suggesting that the presence of outliers is equivalent to model misspecification or dealing only with the case of outliers. We do not view outliers and model misspecification as equivalent and would describe our work as directly addressing the central question of model misspecification.

We focus our initial presentation (around equation (1) in the paper) on the thought-experiment where a known subset of cases is known to not follow the model under consideration as a device to convey the thinking behind our methods. We believe that there is universal agreement that these cases should be discarded for the analysis. We then move on to more realistic cases. The Belgian call data in Section 2.4 and the Nationwide Insurance data in Section 5 are examples of a more general form of model misspecification – namely where an important covariate is missing. Incorporating the covariate would adjust the model and moderate the misspecification. As is typical in examples used to motivate robust regression, we understand quite a lot about the nature of the misspecification and we could build a better model by using this additional knowledge. The examples are used to motivate techniques for the situation where we do not have this additional knowledge.

The style of analysis we suggest can be used quite broadly. The analyst has great flexibility in the model that is written and in the choice of conditioning statistic (in practice, this may well feel like "conditioning statistics" rather than a single statistic). Many choices lead to straightforward computational implementations. This is particularly true when conditioning on a set of order statistics that lead easily to generation of complete data sets.

*Department of Statistics, The Ohio State University, Columbus, OH, 43210, lewis.865@buckeyemail.osu.edu
†Department of Statistics, The Ohio State University, Columbus, OH, 43210, snm@stat.osu.edu
‡Department of Statistics, The Ohio State University, Columbus, OH, 43210, yklee@stat.osu.edu

**The Borel paradox** Clarke and Robert quite rightly point out that we do not justify our computational strategy at a measure-theoretic level. The paradox is very relevant. The standard approach of considering the limit of a sequence of partitions with non-null probabilities resolves the paradox. The partitions are effectively implicit in the conditions in Section 3 of the paper.

Tied to the notion of conditioning is a caution. In some settings, a seemingly innocuous choice of conditioning statistic may be equivalent to conditioning on the entire data set. Examples include certain discrete problems where there is a single configuration of the data that leads to the observed conditioning statistic; others are more subtle. Darnieder (2011) encounters this phenomenon when using a portion of the likelihood to specify the prior distribution and the remainder to move from prior distribution to posterior distribution.

**Asymptotics** The majority of discussants note that asymptotic arguments suggest that, for large samples, one could substitute a normal distribution for the likelihood to obtain essentially the same results with much quicker computation. The substitution may require an adjustment to the likelihood, as described in the work of Royall and Tsou, and has become commonplace in generalized Bayesian inference. We are in full agreement that, when appropriate, asymptotic approximation of the likelihood coupled with the prior distribution provides a quick and effective means of fitting a Bayesian model. We also believe that there are many situations where sample sizes are too small for the asymptotics to have kicked in. A typical example is a hierarchical model where there is a shortage of data for some portions of the model.

Bayesian statistics (and as a consequence all of statistics) was changed by Markov chain Monte Carlo (MCMC). One key reason for the success of MCMC was that it allowed Bayesian inference in settings where asymptotic approximation failed. The past 30 years show the variety and importance of situations where there is a need to turn to a finite sample fit rather than asymptotic approximation. The dividing line between adequacy and inadequacy of analytic approximation is blurry. Drovandi, Nott and Frazier's example makes this point clear and suggests the possibility of splitting a model into portions with large sample (or other) approximation for some portions and finite-sample evaluation for other portions. The structure of the hierarchical model may provide guidance on where to split.

**Choice of conditioning statistic** Ruggieri raises the issue of whether a particular conditioning statistic $T(\cdot)$ can be selected from a candidate set via data-based choice of a tuning parameter. Jewson and Rossell raise the same issue and implement a data-based tune, showing that inference can be sharpened in this fashion. Hazra suggests an alternative conditioning statistic.

More generally, several discussants raise the issue of choice of the conditioning summary. This is open territory, and it is our belief that the choice should depend upon the goal of the analysis as well as an understanding of potential shortcomings of the working model. While not prescriptive, this is in keeping with the practice of data analysis. The

classical literature on generalized estimating equations that Clarke touches on contains a wealth of information about choice of statistic. We view this literature as a modern classical take on the method of moments where the inferential targets determine the estimating equations. For Bayesians, the literature on approximate Bayesian computation (ABC) methods has primarily focused on ABC as a technique for situations where the likelihood is difficult to evaluate. Nevertheless, the chosen summaries are often closely tied to the goals of the analysis. From our vantage point, the deliberate use of reduced conditioning via data summarization to improve inference for misspecified models in conjunction with ABC should be explored. The recent growth of loss-based replacement of the log-likelihood in Bayesian models implicitly ties inference to the goal of the analysis, although typically by sacrificing Bayes' Theorem.

**Complex problems** Bayesian models are used throughout the academic and corporate worlds and increasingly by governmental and non-profit groups – in short everywhere that data is collected and decisions are to be made. Many of these settings are characterized by the use of complex models that are universally agreed to be very approximate and which are informed by data of questionable quality. The problem may require the use of information from different data sets. In addition, much is often known about certain aspects of the problem. These are precisely the settings where we see the greatest need for Bayesian restricted likelihood methods. The variety of such problems is immense, and we see this as fertile ground for further development of the techniques we advocate.

Ghosh and Ghosh place a spotlight on the mixed model – the hierarchical model for the Nationwide Insurance data is one example of this type of model, but a different mix of information on the individual (state) and the collection of individuals (states) would necessitate different conditioning and perhaps different computation. Drovandi, Nott and Frazier raise the question of more complex regression models. Some will submit to the same strategy that we have used, though straightforward use of our techniques will break down for models with enough complexity. Gugushvili and Peters call attention to one such situation, the challenging $p > n$ problem and to high-dimensional problems more generally. In the absence of strong prior information, we are uncertain how to use our techniques for such problems. Kim describes the use of copula models that would require a different conditioning statistic to match the models and analysis.

**Robust Bayes** Lavine and Ruggieri both make the point that robust Bayesian methods provide a well-developed approach to handling model misspecification. These methods yield a range of posterior summaries, say the smallest set within which the posterior mean is known to lie as the likelihood is varied over a class. Lavine's advances substantially expanded the scope of robust Bayesian analysis. It would be interesting to examine the interplay of restricted likelihood and robust Bayesian methods. A natural approach is to consider the restricted likelihood posterior based on the chosen conditioning statistic as the likelihood ranges over a class. This may well lead to a narrower interval for the posterior mean than does a traditional robust Bayesian analysis.

The question of how restricted likelihood relates to the likelihood principle is unsettling. After summarization through the conditioning statistic, inference is fully likelihood based—it is exactly Bayesian. When viewed before summarization, it appears to violate the likelihood principle in some cases. It can certainly clash with the usual match between sequential updating and updating in one shot. The reason for this is that joint likelihood arising from separate summarization of two sets of data may differ from the likelihood that arises from a single summary of the combined data. With different likelihoods, the two restricted posteriors would differ (Lewis, 2014).

**Prediction** As part of his rich discussion, Clarke connects our work to the philosophy and work on M-open and M-complete inference. The prequential approaches that we associate with Dawid and with Clarke's own move to Bayesianize them are strongly sequential as is natural for many prediction problems. As noted above, as currently formulated, our techniques are better suited to "one-shot" analyses than to sequential analyses.

**Bayesian EDA** Clarke suggests use of restricted likelihood techniques for a Bayesian version of EDA. We had not thought of this possibility, but would be interested to hear more. EDA is one area where Bayesians lag.

**References** The discussants provide numerous references to which many more could be added. In particular, we thank Robert for the references on computation. Though MacEachern had heard of the early Diaconis and Sturmfels paper, he did not make the connection to this setting. We were unaware of the more recent references, had not recently searched for work in the area, and certainly did not intentionally turn a blind eye to existing work. The computational methods we employ have been stable since 2013, widely disseminated in talks, and available on the web.

We thank the discussants for the time they have spent on our work and their sharply written perspectives. While these discussions contain diverse views, we suspect that the full range of views within the Bayesian community on how to handle model misspecification is much greater. As Bayesian applied statistics has matured following the development of MCMC methods, many have asked what the next big challenge is. We would identify model misspecification as one of the two or three biggest challenges currently faced by the Bayesian community.

Our focus has been on methods that are formally Bayesian after summarization of the data through an insufficient statistic. There are, of course, many other approaches that are under development. A few that we see as most closely related to this work are ABC, loss-based generalized Bayesian inference, and the use of fractional likelihoods for Bayesian robustness. The approach of fitting very flexible models and pushing the "model" from specification into inference is also promising. We believe that all of these directions are worth exploration and suspect that we will all be working with a blend of these ideas and others in ten years' time.

## References

Darnieder, W. F. (2011). "Bayesian Methods for Data Dependent Priors." Ph.D. thesis, The Ohio State University. MR2942250. 1459

Lewis, J. (2014). "Bayesian Restricted Likelihood Methods." Ph.D. thesis, The Ohio State University. MR3337628. 1461