Bayesian Restricted Likelihood Methods

John R. Lewis, Steven N. MacEachern and Yoonkyung Lee

Department of Statistics, The Ohio State University, Columbus, Ohio 43210

lewis.865@osu.edu, snm@stat.osu.edu and yklee@stat.osu.edu *

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 (e.g., use of heavy-tailed likelihoods) and propose the restricted likelihood as an alternative. When working with restricted likelihood, we summarize the data through a set of (insufficient) statistics, targeting inferential quantities of interest, and update the prior distribution with the summary statistics rather than the complete data. By 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 picked up by the conditioning statistics. A major contribution is the development of a data augmented MCMC algorithm for the linear model and a wide range of choices for summary statistics.

*This research has been supported by Nationwide Insurance Company and by the NSF under grant numbers DMS-1007682 and DMS-1209194. The views in this paper are not necessarily those of Nationwide Insurance or the NSF.

We demonstrate the method on an insurance agency data set containing many outliers and subject to model misspecification. Success is manifested in better predictive performance for data points of interest as compared to competing methods.

KEYWORDS: Approximate Bayesian computation, 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. In simple settings, the success of the methods is often attributed to formal optimality properties, sometimes derived through the laws of subjective probability and sometimes through admissibility and the complete class theorems of decision theory. In complex settings, the hierarchical model allows one to create and fit sophisticated models that may, for example, pool information across similar problems.

The development of Bayesian inference relies on a complete Bayesian 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 prior distribution has been addressed through the development of techniques for subjective elicitation (Garthwaite et al., 2005) and the rise of 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 sampling density has been given less attention from a specifically Bayesian view, although the work on predictive diagnostics (Box,

1980) departs from classical traditions.

The focus of this work is the intersection of Bayesian methodology and data analysis. In particular, we develop techniques to handle imperfections in the sampling density. These imperfections often show themselves through the presence of outliers–here taken to be cases not reflecting the phenomenon under study–in the data set. 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 prior distribution is updated with the likelihood from the mixture model to obtain the complete-data posterior distribution. 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, in press).

The traditional strategies for handling outliers all have their drawbacks. While we view the sampling density for the good data as stable, 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: the use of restricted likelihood. In a nutshell, we begin with a complete model as if all of the data are good. But rather than driving the move

from prior distribution to posterior distribution by the entire likelihood, we use only the likelihood of a few summary statistics which typically target inferential quantities of interest. We call this reduced likelihood a restricted likelihood. The update is a formal update from prior distribution to posterior distribution, based on the sampling density of the summary statistics. In our approach, the reader will identify a stream of reasoning which is manifested in classical M-estimation, generalized estimating equations, approximate Bayesian computation, and elsewhere. The novelty of the work is twofold. We make use of classical robust estimators as summary statistics in a formal Bayesian framework, using the sampling density of the estimators as a replacement for the sampling density of the data. We advance the argument that conditioning on an insufficient summary of the data is sound practice, rather than merely being done for computational and modelling convenience.

In addition to outlier-prone data, the sampling density can err due to model misspecification. 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 recommend a common solution for dealing with the problem.

The remainder of the paper develops Bayesian restricted likelihood (Section 2), shows how it can be applied to a Bayesian linear model (Section 3), illustrates its use on an insurance agencies data set with a novel twist on model evaluation (Section 4),

and wraps up with a discussion (Section 5). A major contribution of this work is the computational strategy whose legitimacy is established in Section 3. The technical proofs are in the appendix.

2 Restricted Likelihood

To describe the use of restricted likelihood in a Bayesian framework, we begin with a pair of simple examples for the one-sample problem. In each, 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 $\boldsymbol{\theta}$. The standard, or complete, likelihood would be $L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$.

As a first example, we consider the case where a subset of the data are known to be bad in the sense of not informing us about $\boldsymbol{\theta}$–and where the subset is known. This case mimics the setting where outliers in a data set are identified and discarded before a formal analysis is done. 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 entire 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 term may not actually involve $\boldsymbol{\theta}$. We wish to keep the first piece and drop the second for better inference on $\boldsymbol{\theta}$.

A 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)). 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 x_i by defining $x_i = t_1$ if $y_i \leq t_1$, $x_i = t_2$ if $y_i \geq t_2$, and $x_i = y_i$ otherwise. The adjusted update is performed with $L(\boldsymbol{\theta}|\boldsymbol{x})$. With slightly non-standard notation, we let $g(t_1|\boldsymbol{\theta}) = F(t_1|\boldsymbol{\theta})$, $g(t_2|\boldsymbol{\theta}) = 1 - F(t_2|\boldsymbol{\theta})$, and $g(y|\boldsymbol{\theta}) = f(y|\boldsymbol{\theta})$ for $y \in (t_1, t_2)$. Alternatively, we may rewrite the entire likelihood as the product of two pieces,

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

and retain only the first for the formal update.

Further examples abound. In a completely randomized experimental design, we randomize experimental units to treatment conditions and then ignore the details of the observed randomization (Dean and Voss, 1999); in work with contingency tables, we collapse categories with small counts, coarsening the scale of data (Agresti, 2002); in meta-analysis, we ignore the individual patient-level data and instead work with estimated effects from the studies (O'Rourke, 2007). Further examples are described in Lewis (2014). To describe the approach in (1), (2), and these other settings, we write the complete data likelihood in two pieces, as indicated below:

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

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 $T(\boldsymbol{y}) = (x_1, \ldots, x_n)$. We refer to $f(T(\boldsymbol{y})|\boldsymbol{\theta})$ as the restricted likelihood.

Bayesian methods can make use of restricted likelihood in place of the complete data likelihood since $T(\boldsymbol{y})$ is a well-defined random variable with a probability dis-

tribution indexed by $\boldsymbol{\theta}$. The update from prior distribution to posterior distribution is made on the basis of $f(T(\boldsymbol{y})|\boldsymbol{\theta})$ rather than $f(\boldsymbol{y}|\boldsymbol{\theta})$. This leads to the restricted likelihood posterior

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

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

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

Direct use of restricted likelihood appears in many areas of the literature. The motivation is often similar to ours: concern about model misspecification. For example, the use of rank likelihoods is discussed by Savage (1969), Pettitt (1983, 1982), and more recently by Hoff et al. (2013). 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 the 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 often appealed to when applying approximate Bayesian computation (ABC), a collection of posterior approximation methods which has recently experienced success in applications to epidemiology, ge-

netics, and quality control (see, for example, Tavaré et al., 1997; Pritchard et al., 1999; Marjoram et al., 2003; Fearnhead and Prangle, 2012). ABC is typically used when likelihoods are intractable but simulation of new data from the model is easily done. In the cited references, interest lies in the full data posterior and ABC is used for computational convenience. Consequently, effort is made to choose a $T(\boldsymbol{y})$ such that the likelihood $L(\boldsymbol{\theta}|\boldsymbol{y}) \approx L(\boldsymbol{\theta}|T(\boldsymbol{y}))$. Technical limitations of ABC imply that in realistic (and all non-discrete) settings, the conditioning is not exact, but approximate.

In a related, but distinctly different, approach which also takes advantage of summary statistics, the full data log likelihood is replaced by a loss function (e.g. Bissiri et al., 2013) in an effort to concentrate inference only on parameters of interest. In contrast, the restricted likelihood is formulated using a full probability model, allowing for a formal Bayesian update, while remaining robust to misspecification. This work extends the use of restricted likelihood by arguing that its use is sound practice, and also by expanding the class of conditioning statistics in which exact conditioning can be achieved well beyond ranks. Our methods do not rely on asymptotic properties or approximate conditioning as in previous work (e.g., Albert (1988),Hoff and Wakefield (2013)).

The key to productive use of restricted likelihood is the choice of $T(\boldsymbol{y})$ and the development of computational strategies that allow us to truly condition on the observed $T(\boldsymbol{y})$ and fit the model in formal Bayesian fashion. In this work, we focus on robustness, and natural choices of $T(\boldsymbol{y})$ include a set of middling order statistics, a trimmed mean, or a classical robust estimator of location and/or scale. We have previously implemented all of these methods for one-sample problems where we have found them to perform well (e.g., Lewis et al. (2012)). Of these versions, the most extensible to the linear model are the M-estimators, in the tradition of Huber

(1964). The computational strategies we devise in subsequent sections allow us to apply Bayesian restricted likelihood inference well beyond regression models for the mean of a distribution. In particular, quantile regression falls within the framework we develop. The next section develops the necessary computational strategies.

3 Restricted Likelihood for the Linear Model

3.1 The Bayesian linear model

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

$$\begin{aligned} \boldsymbol{\beta} &\sim \pi_1(\boldsymbol{\beta}); \qquad \sigma^2 \sim \pi_2(\sigma^2) \\ y_i &= x_i^\top \boldsymbol{\beta} + \epsilon_i, \text{ for } i = 1, \ldots, n \end{aligned} \tag{5}$$

where x_i and $\boldsymbol{\beta} \in \mathbb{R}^p$, and the ϵ_i are independent draws from a distribution with center 0 and scale σ. Two conditions are imposed on the model:

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

C2. The ϵ_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.

Both conditions can be relaxed, although this would necessitate restating several later results. In the sequel, we specifically consider both normal and t distributions with mean 0 and variance σ^2 for the ϵ_i but note that our methods apply much more widely. The prior distributions π_1 and π_2 can take many forms and may be joined to

form a joint distribution for non-independent $\boldsymbol{\beta}$ and σ^2. The conditionally conjugate normal/inverse gamma pair is a common choice.

The methods we develop apply to the linear model in (5) and to many variations on it. As summary statistics for the data, we consider M-estimators for the coefficients in the linear model and an associated estimator of the scale. These estimates convey information about $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$ while downweighting outliers. The M-estimator of $\boldsymbol{\beta}$ is determined by a ρ function through the minimization

$$\boldsymbol{b}(X, \boldsymbol{y}) \quad = \quad \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} \rho(\frac{y_i - x_i^\top \boldsymbol{\beta}}{\sigma}). \tag{6}$$

The scale estimator $s(X, \boldsymbol{y})$ is determined simultaneously as another M-estimator or is determined through a separate calculation, for example the mean-absolute-deviation from an ℓ_1 regression fit. Continuity of the distribution of the ϵ_i is important as it translates into a continuous distribution for $\boldsymbol{b}$, which is presumed for our computational strategy. The results we derive also apply to estimators that are not M-estimators.

3.2 Computational strategy

For low-dimensional statistics $T(\boldsymbol{y})$ and $\boldsymbol{\theta}$, the direct computational strategies described in Lewis et al. (2012) can be used to evaluate the restricted likelihood posterior. These strategies rely on generation of complete data sets from different values of $\boldsymbol{\theta}$. Each complete data set leads to a statistic $T(\boldsymbol{y}|\boldsymbol{\theta})$, and these generated statistics are used to estimate the density at $T(\boldsymbol{y}_{obs}|\boldsymbol{\theta})$ which is then fed into Bayes theorem for the update from prior distribution to posterior distribution. A variety of variance reduction techniques which exploit properties of the distribution of $\boldsymbol{y}$ improve the performance of these strategies.

For high-dimensional statistics $T(\boldsymbol{y})$ or high-dimensional parameters $\boldsymbol{\theta}$, direct computational strategies break down. When the conditioning statistic is of high-dimension, density estimation becomes difficult and the associated approximate update in (4) is unstable; when $\boldsymbol{\theta}$ is high dimensional, grid-based calculation and other numerical integration strategies fail. However, Markov chain Monte Carlo (MCMC) methods were developed for exactly these situations. We turn to MCMC to fit the model in these circumstances.

The general style of algorithm that we present relies on the decomposition of the sampling density in (3) into one piece involving only $T(\boldsymbol{y})$ and a second piece for the complete data $\boldsymbol{y}$ given $T(\boldsymbol{y})$. Relying on the modularity of MCMC algorithms, we begin with any conventional complete data algorithm. In the case of typical regression models, these algorithms abound. Details of the algorithm depend on details of the prior distribution and sampling density. As examples, a normal prior distribution and normal likelihood in the regression setting allow one to alternate conditional generations of σ^2 and $\boldsymbol{\beta}$, and blocking the generation of $\boldsymbol{\beta}$ generally leads to quicker convergence and mixing (Liu, 1994); a thick-tailed scale mixture of normal distributions in the style of the hyper-g/n prior (Liang et al., 2008) necessitates an additional stage for the sampler where the scale g/n is generated; a thick-tailed sampling density such as a t distribution can be handled with the addition of a scale parameter for each case and an extra stage where these scale parameters are generated. The additional stage needed to implement the restricted likelihood analysis via MCMC is a generation of the complete data given statistic ($T(\boldsymbol{y}_{obs})$) and parameter ($\boldsymbol{\theta}$). Condition C2 facilitates the extension of convergence proofs for the complete data algorithm to those for the incomplete data algorithm. In this subsection, we focus exclusively on the additional stage where we generate $\boldsymbol{y}|\boldsymbol{\theta}, T(\boldsymbol{y})$.

For a typical model and conditioning statistic, the distribution $[\boldsymbol{y}|\boldsymbol{\theta}, T(\boldsymbol{y})]$ is not

available in closed form. As a consequence, we turn to Metropolis-Hastings, using the strategy of proposing the complete data $\boldsymbol{y}$ with summary statistic matching $T(\boldsymbol{y})$ and either accepting or rejecting the proposal. The Metropolis-Hastings acceptance ratio is given by the expression below, truncated above at 1, where $\boldsymbol{y}_p$ represents the proposed complete data and $\boldsymbol{y}_c$ is the current complete data.

$$\begin{aligned} R &= \frac{f(\boldsymbol{y}_p|\boldsymbol{\theta}, T(\boldsymbol{y}_p) = T(\boldsymbol{y}_{obs}))}{f(\boldsymbol{y}_c|\boldsymbol{\theta}, T(\boldsymbol{y}_c) = T(\boldsymbol{y}_{obs}))} \frac{p(\boldsymbol{y}_c|\boldsymbol{\theta}, T(\boldsymbol{y}_c) = T(\boldsymbol{y}_{obs}))}{p(\boldsymbol{y}_p|\boldsymbol{\theta}, T(\boldsymbol{y}_p) = T(\boldsymbol{y}_{obs}))} \quad (7) \\ &= \frac{f(\boldsymbol{y}_p|\boldsymbol{\theta})}{f(\boldsymbol{y}_c|\boldsymbol{\theta})} \frac{p(\boldsymbol{y}_c|\boldsymbol{\theta})}{p(\boldsymbol{y}_p|\boldsymbol{\theta})} \quad (8) \end{aligned}$$

The expression $p(\cdot)$ gives the proposal density. The second line follows from the fact that $T(\boldsymbol{y}_p) = T(\boldsymbol{y}_c) = T(\boldsymbol{y}_{obs})$ for all current and proposed data sets. For the models we consider, evaluation of $f(\boldsymbol{y}|\boldsymbol{\theta})$ is straightforward. We focus on construction of proposals that guarantee $T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})$ for the linear model and on the evaluation of the proposal density.

3.2.1 Construction of proposal

Our computational strategy is easiest to envision in a simple location-scale setting where the design matrix in model (5) consists of a single column of ones. Robust estimation techniques along with model (5) suggest a conditioning statistic $T(\boldsymbol{y})$ which consists of estimates of the scalars β and σ, say $(b(X, \boldsymbol{y}), s(X, \boldsymbol{y}))$. To obtain data $\boldsymbol{y}$ for which $T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})$, we proceed in two steps. First, a vector $\boldsymbol{z}^*$ is generated from a simple manifold with known density, for example, a uniform distribution on the surface of the unit sphere in $\mathbb{R}^{n-1}$ (alternatively from the model with current values of β and σ). This leads to $T(\boldsymbol{z}^*) = (b(X, \boldsymbol{z}^*), s(X, \boldsymbol{z}^*))$. The vector $\boldsymbol{z}^*$ is mapped into a vector $\boldsymbol{y}$ by rescaling and shifting appropriately to match the observed conditioning statistic. For typical estimators, the appropriate scaling is $s(X, \boldsymbol{y}_{obs})/s(X, \boldsymbol{z}^*)$

with the shift trailing along as needed to match $b(X, \boldsymbol{y}_{obs})$. To evaluate the proposal density of $\boldsymbol{y}$, we need to adjust the density of $\boldsymbol{z}^*$ with a Jacobian. In the sequel, we use this artificially low-dimensional example to illustrate the method.

The strategy described in the previous paragraph extends to full-blown regression models. Robust regression methods lead naturally to a conditioning statistic in the form of a classical M-estimator for $\boldsymbol{\beta}$ and a companion estimator for σ. We denote the resulting estimator which involves the covariates through the design matrix and the response as $T(\boldsymbol{y}) = (\boldsymbol{b}(X, \boldsymbol{y}), s(X, \boldsymbol{y}))$, with $\boldsymbol{b}(X, \boldsymbol{y}) = (b_1(X, \boldsymbol{y}), \cdots, b_p(X, \boldsymbol{y}))^\top$. Simultaneous M-estimators have a number of standard properties C3-C8 which prove useful in the sequel (Huber and Ronchetti, 2009; Maronna et al., 2006).

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 other estimators satisfy these properties, and our subsequent results apply equally well to them. With more cumbersome statements, the upcoming results can be adjusted to handle a relaxation of C4 that $s(X, \boldsymbol{y}_{obs}) > 0$ and $P(s(X, \boldsymbol{y}) > 0) > 0$.

The properties above ensure that any vector $\boldsymbol{y}^* \in \mathbb{R}^n$ can be transformed to another vector, $\boldsymbol{y}$ so that $T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})$. The mechanism by which this happens

is the scaling and shifting presented in the following theorem. The proof of this and other results appear in the appendix.

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

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

The mapping described in Theorem 3.1 is many-to-one in general. The range of the mapping is the sample space restricted to match the observed summary statistic:

$$\mathcal{A} = \{\boldsymbol{y} \in \mathbb{R}^n | T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})\}. \tag{9}$$

$\mathcal{A}$ is an $n-p-1$ dimensional space: there are p constraints imposed by the regression coefficients and one further constraint imposed by the scale. The form of the set is determined by the statistic $T(\boldsymbol{y})$. Figure 1 provides an artificial low-dimensional example of such a set $\mathcal{A}$. In the figure, $n = 3$, and the model is a location-scale model with conditioning statistic $T(\boldsymbol{y}) = (\min(\boldsymbol{y}), \sum(y_i - \min(\boldsymbol{y}))^2)$. The set $\mathcal{A}$ is depicted for $T(\boldsymbol{y}_{obs}) = (0, 1)$ and is a "warped triangle", with each side corresponding to a particular coordinate of $\boldsymbol{y}$ being the minimum. The set may be compact and given by a closed curve, as in the figure, or it may be unbounded.

The set $\mathcal{A}$ typically does not lie in a linear space of dimension $n - p - 1$, and so we must account for both the many-to-one nature of the mapping and a Jacobian when deriving the proposal density. We handle the first point by proposing a vector $\boldsymbol{z}^*$ on an $n - p - 1$ dimensional space which, through a scaling and shifting, maps to a point in $\mathcal{A}$ uniquely. The initial proposal is chosen so that the range of the map is

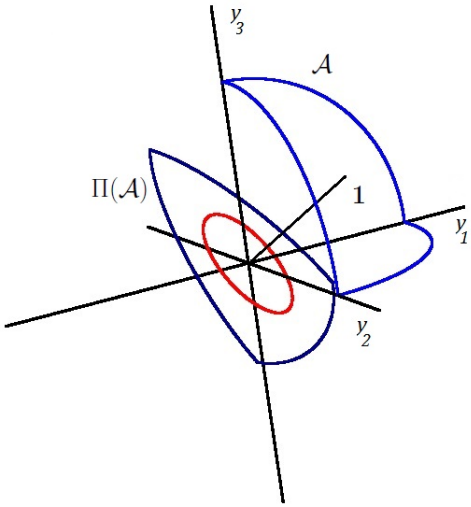

Figure 1: A depiction of $\mathcal{A}$, $\Pi(\mathcal{A})$, and the unit circle for the illustrative example where $b_1(\mathbf{1}, \boldsymbol{y}) = \min(\boldsymbol{y}) = 0$ and $s(\mathbf{1}, \boldsymbol{y}) = \sum(y_i - b_1(\mathbf{1}, \boldsymbol{y}))^2 = 1$. $\mathcal{A}$ is the combination of 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. Also depicted is the vector $\mathbf{1}$, the design matrix for the location and scale setting.

the entirety of $\mathcal{A}$. The Jacobian does not cancel in (7), since the scaling depends on the initial proposal.

Figure 1 shows the type of set from which we draw the initial proposal. Denote the column space of the design matrix X by $\mathcal{C}(X)$ and its orthogonal complement by $\mathcal{C}^{\perp}(X)$. We refer to the latter set as the 'deviation space' as it is the space where the traditional least squares residuals are contained. We define the projection of the set $\mathcal{A}$ onto the deviation space as

$$\Pi(\mathcal{A}) = \{\boldsymbol{z} \in \mathbb{R}^n | \ \exists \ \boldsymbol{y} \in \mathcal{A} \ s.t. \ \boldsymbol{z} = Q\boldsymbol{y}\} \tag{10}$$

where Q is the projection matrix onto $\mathcal{C}^{\perp}(X)$. Explicitly, $Q = I - H$ with $H = XX^{\top}$ where we assume without loss of generality, following condition C1, that the columns of X form an orthonormal basis for $\mathcal{C}(X)$ (i.e., $X^{\top}X = I$). It will also be helpful at

times to write $Q = WW^\top$ where the columns of W form an orthonormal basis for $\mathcal{C}^\perp(X)$.

The initial proposal is drawn from the surface of the unit sphere in the $n - p$ dimensional $\mathcal{C}^\perp(X)$. In the figure, the column vector $\mathbf{1}$ spans $\mathcal{C}(X)$, the triangle with bowed sides is the projection of $\mathcal{A}$ onto $\mathcal{C}^\perp(X)$, and the circle is the set from which the initial proposal is drawn.

For an initial proposal $\boldsymbol{z}^*$ on the surface of the sphere, we move to a point on $\Pi(\mathcal{A})$ through a simple scaling of the point $\boldsymbol{z}^*$. This is followed by undoing the projection with a move from $\Pi(\mathcal{A})$ to its (unique) preimage on $\mathcal{A}$. Together, these two steps correspond to the transformation in Theorem 3.1. The introduction of the initial proposal surface gives us a 1-1 transformation. Properties C5-C8 ensure the mapping described is indeed 1-1. In particular, property C8 ensures the scaling to be unique and C7 implies the scale statistic is unchanged when undoing the projection. Property C5 ensures the uniqueness of undoing the projection.

The general proposal strategy is summarized as follows

1. Sample $\boldsymbol{z}^*$ from a distribution with known density on the unit sphere in $\mathcal{C}^\perp(X)$.

2. Implement the transformation in Theorem 3.1 in two steps

   (a) Scale: $\boldsymbol{z} = \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)} \boldsymbol{z}^*$

   (b) Shift: $\boldsymbol{y} = \boldsymbol{z} + X\left(\boldsymbol{b}(X, \boldsymbol{y}_{obs}) - \boldsymbol{b}(X, \boldsymbol{z})\right)$

3.2.2 Evaluation of the proposal density

Calculation of the appropriate Jacobian of the transformation is absolutely vital and also non-trivial. Writing the transformation from the unit sphere in deviation space to $\mathcal{A}$ in two steps facilitates calculation of the Jacobian in two steps.

From unit sphere to $\Pi(\mathcal{A})$

The first step is constrained to $\mathcal{C}^{\perp}(X)$ where the unit sphere is transformed to $\Pi(\mathcal{A})$. We further break this piece in two steps: first, the distribution on the unit sphere is transformed to that along a sphere of radius $r = \|\boldsymbol{z}\| = s(X, \boldsymbol{y}_{obs})/s(X, \boldsymbol{z}^*)$. This contributes $r^{-(n-p-1)}$ to the Jacobian. Second, the new sphere is then deformed to $\Pi(\mathcal{A})$. 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 ourselves to the $n-p$ dimensional space $\mathcal{C}^{\perp}(X)$, 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 $\boldsymbol{z}$ and the normal to $\Pi(\mathcal{A})$ is given in the following lemma.

Lemma 3.2. *Assume that conditions C1-C2, C4, and C7 hold. Let $\boldsymbol{y} \in \mathcal{A}$. 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 $\mathcal{C}^{\perp}(X)$.*

The contribution to the Jacobian of this attenuation is

$$\cos(\gamma) = \frac{\nabla s(X, \boldsymbol{y})^{\top} \boldsymbol{z}}{\|\nabla s(X, \boldsymbol{y})\| \|\boldsymbol{z}\|}, \tag{11}$$

where γ is the angle between the two normal vectors. This step is illustrated in Figure 2 for the toy location-scale example.

From $\Pi(\mathcal{A})$ to $\mathcal{A}$

The final piece of the Jacobian comes from the transformation from $\Pi(\mathcal{A})$ to $\mathcal{A}$. For this we return to the full n dimensional space. The second step involves a shift of $\boldsymbol{z}$ to $\boldsymbol{y}$ along the column space of X, but the shift depends on $\boldsymbol{z}$, and so the density on the set $\Pi(\mathcal{A})$ is deformed by the shift. The contribution of this step to the Jacobian is, from first principles, the ratio of the infinitesimal volume along $\Pi(\mathcal{A})$

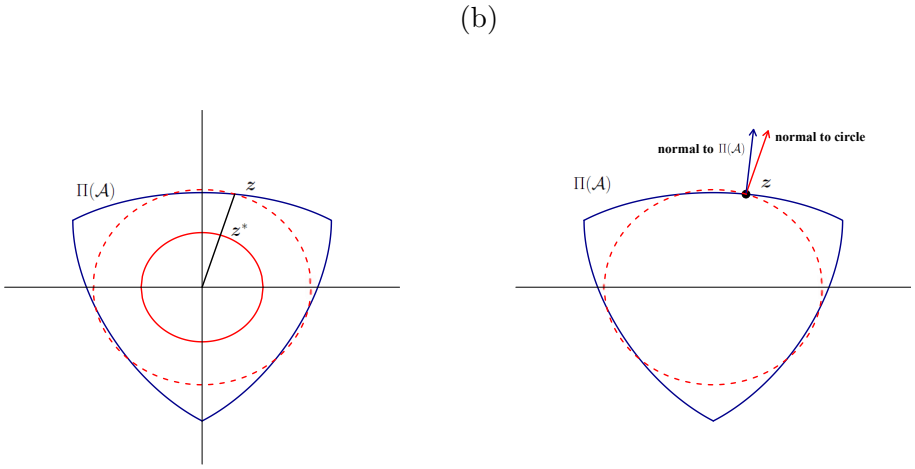

Figure 2: Panel (a) contains a depiction of the stretch from $\boldsymbol{z}^*$ to $\boldsymbol{z}$. The adjustment for the stretch transforms the density along the unit circle to the density along the circle of radius $\|\boldsymbol{z}\|$ (dashed circle). Panel (b) contains a depiction of the deformation from the distribution along the circle to the distribution along $\Pi(\mathcal{A})$. The adjustment can be seen to be the cosine of the angle between the normals to each manifold.

to the corresponding volume along $\mathcal{A}$. 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. 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 $\mathcal{T}_y^{\perp}(\mathcal{A})$. This leads to a basis for $\mathcal{T}_y(\mathcal{A})$. Both of these bases can be orthonormalized. Let $B = [b_1, \ldots, b_{p+1}]$ and $A = [a_1, \ldots, a_{n-p-1}]$ denote the matrices whose columns contain these two orthonormal bases. 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

$$\text{Vol}(P) := \sqrt{\det(P^\top P)} = \prod_{i=1}^{r} \sigma_i \tag{12}$$

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* (12)*.*

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*

$$p(\boldsymbol{y}) = p(\boldsymbol{z}^*) r^{-(n-p-1)} \cos(\gamma)\, Vol(P) \tag{13}$$

where $r = s(X, \boldsymbol{y}_{obs})/s(X, \boldsymbol{z}^*)$*, and* $\cos(\gamma)$ *and* $Vol(P)$ *are as in equations* (11) *and* (12) *respectively.*

In practice, computing A is computationally intensive as it involves orthogonalization of n vectors in n-dimensional space. To find a matrix A, supplement B with a set of n linearly independent columns on the right, and apply Gram-Schmidt orthonormalization to the matrix. This algorithm is slow when n is large, as it is $\mathcal{O}(n^3)$ and A must be found at each iterate of the algorithm when a complete data set is drawn. Fortunately, we can make use of results related to *principal angles* found in Miao and Ben-Israel (1992) to compute the volume in (12) using B and an orthonormal basis for $\mathcal{C}(X)$ (The definition of principal angles can be found in the cited text). Recall, B is constructed by orthogonalization of a basis for $\mathcal{T}_y^{\perp}(\mathcal{A})$. Since this space is of dimension $p+1$, applying Gram-Schmidt to find the orthonormal basis is much faster, the algorithm is $\mathcal{O}(np^2)$, and there is a considerable reduction in computational burden when $n \gg p$. Further, the singular values of P are also the singular values of $W^\top A$, which can be easily obtained through B. 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 $\mathcal{C}(X)$. Then the non-unit singular values of $U^\top B$ are the same as the non-unit singular values of $W^\top A$.*

4 Applications

We illustrate the methods developed in Sections 2 and 3 with a pair of regression models for data from Nationwide Insurance Company, which concern prediction of the performance of insurance agencies. The data contain outliers and are subject to model misspecification. In particular, a group of the data do not follow the same generative process as the data of interest. It would be extremely challenging to model some features of the data. In our analysis, we follow the standard practice

when demonstrating the benefits of robust methods. We work with a naive model for the data which ignores certain features of the problem. We do this both to create a situation where all can agree that the model for the complete data $\boldsymbol{y}$ is imperfect and to preserve the confidentiality of selected aspects of modelling done by Nationwide. We wish to provide inference for the 'good' portion of the data. The two models we fit either treat the analysis as a single regression or as a collection of related regressions. Details of the models, prior distributions, and conditioning statistics are given in the next two subsections.

4.1 Nationwide Data

The Nationwide Insurance Company sells many of its insurance policies through agencies which provide direct service to policy holders. The contractual agreements between Nationwide and these agencies vary. Of major interest to Nationwide is the prediction of future performance of agencies where, for our purposes, performance is measured by the total number of households an agency services ('household count'). A serviced household is one in which at least one person living at that residence has at least one policy written through the agency. We used data from previous years to build a model to forecast future household count. In particular, we use agency characteristics, as measured during a single month in 2010, to predict household counts in the corresponding month in 2012. The characteristics used are household count and two measurements of agency size/experience. The two measurements of agency size/experience are, roughly, the number of employed persons at the agency and the length of time the agency has been affiliated with Nationwide. The household counts (response and predictor) have been square rooted to stabilize variance. The data are proprietary, and to mask them all variables have been individually centered and scaled and identifiers (agency/agent names and state labels) have been removed. All

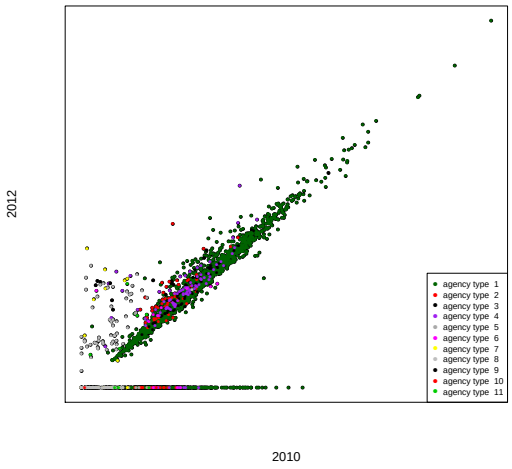

Figure 3: The square root of count in 2012 versus that in 2010 (after centering and scaling). The colors represent the varying contractual agreements as they stood in 2010. Agencies that closed during the 2010-2012 period are represented by the zero counts for 2012.

subsequent analysis is done on this scale. As an exploratory view, a plot of the square root of count in 2012, against that in 2010 is shown in Figure 3. The different colors represent the varying contractual agreements as they stood in 2010. 'Type 1' agencies are of special interest. Among the open agencies, a strong linear correlation exists. The specific linear relationship depends on agency type. The data are characterized by a large number of agencies which were open in 2010 but closed sometime before 2012, as represented by the horizontal band at 0. We use these data as a test bed for our techniques, 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.

4.2 Regression model

The first analysis that we consider is based on a single regression. We use the following standard normal theory regression model

$$\boldsymbol{\beta} \sim N_p(\mu_0, \Sigma_0); \;\; \sigma^2 \sim IG(a_0, b_0); \;\; y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \;\; \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \; i = 1, \ldots, n, \; (14)$$

where $\boldsymbol{\beta}$ is a four dimensional vector ($p = 4$) of regression coefficients for the intercept, square root of count in 2010, and the two size/experience measures, and y_i is the square rooted household count in 2012 for the i^{th} agency with covariate vector $\boldsymbol{x}_i$. Although the mean of covariates and response have been removed, we include the intercept as fitting is done on a holdout set to evaluate predictive performance. The hyper-parameters $a_0, b_0, \boldsymbol{\mu}_0$ and Σ_0 are all fixed and set from a robust regression fit to the data from the time period two years before. μ_0 is set to the estimate of the regression coefficients. Σ_0 is set to $n \cdot \text{var}(\boldsymbol{b})$ where n is taken as the sample size of the prior data set, corresponding to a unit information prior for $\boldsymbol{\beta}$. The hyperparameters for σ^2 are set so that the prior mean is s^2, the estimated variance from the robust regression, and the spread of the prior covers the range of plausible values with high probability. All values are then transformed appropriatly to match the current scale of the data. In the end we take $\mu_0 = (0.18, 0.81, 0.01, -0.02)^\top$ and set the mean of σ^2 to 0.014 and standard deviation to 0.033.

We compare four Bayesian models: the standard Bayesian normal theory model, two restricted likelihood models, both with simultaneous M-estimators as the restriction, and a heavy tailed model. The heavy-tailed model replaces the normal sampling density in (14) with a t-distribution with $\nu = 3$ degrees of freedom. The restricted likelihood methods use standard ρ functions, colloquially known as Huber's ρ and Tukey's ρ. We have used the default tunning parameter settings for the `rlm` function

in the R package `MASS` (Venables and Ripley, 2002). Both use Huber's scale estimator as in the `rlm` implementation. We also fit the corresponding classical robust regressions and a least squares regression.

4.2.1 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 so prefer to use an out-of-sample measure of fit. This leads us to cross-validation. We repeatedly split the data into training and validation sets. We fit the model to the training data and assess its performance on the validation 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 plug-in maximum likelihood for the classical fits.

The trimmed evaluation proceeds as follows in our context. The evaluation function for case i in the hold-out data is the log predictive density, say $\log(f(y_i))$, with the conditioning on the training data suppressed. The trimming fraction is set at $0 \leq \alpha < 1$. To score a method, we first identify a base method. Under the base method, $\log(f(y_i))$ is computed for each case in the validation sample, say $i = 1, \ldots, M$. The $[\alpha M]$ observations with the smallest values of $\log(f(y_i))$ are removed from the validation sample. All of the methods are then scored on the remaining $M - [\alpha M]$ observations in the validation sample with the mean trimmed log

marginal pseudo likelihood, $TLM = (M - [\alpha M])^{-1} \sum \log(f(y_i))$. The sum runs over the remaining observations. This process is advantageous to the base method. A method that performs poorly when it is the base method is discredited. For a complete evaluation, we allow each method to appear as the base method. For brevity, we present only a selection of results in our subsequent analyses.

4.2.2 Comparison of predictive performance

Model performance is assessed using the mean and standard deviation of the TLM across 100 different replicates. First, we include all observations in each validation sample to calculate TLM for each split. We then repeat the evaluation using only certain subsets of the validation sample that are of special interest. Subsets include open agencies, open 'Type 1' agencies, and 'Type 1' agencies. For brevity, we include results for the 'Type 1' agencies only. As noted, assessing model predictions on this set of agencies is of special interest to the company. A range of training sample sizes were used and we include results from $n = 25, 100, 1000$, and 2000 out of a total of 3180 agencies. The trimming fraction, α, ranges from 0 to 0.3. A classical robust regression to the prior data assigns zero weight to around 16% of observations; in essence removing these from the analysis. This informed the range of trimming fractions chosen. In practice, we would set α slightly larger than 0.16.

Model evaluation for 'Type 1' agencies is shown in Figure 4 for training sample sizes $n = 25, 100$, and 1000. The normal theory models perform poorly due to the numerous outliers and are left out. Appearing in the figures are the mean TLM across each validation set for each model and each trimming fraction, α (along the x-axis). The error bars depicted are one standard deviation of the TLM above and below the mean. The models pictured are: classical robust regression with Tukey's ρ (rlm-T), restricted likelihood based on the 'Tukey estimate' (rest.-T), classical robust

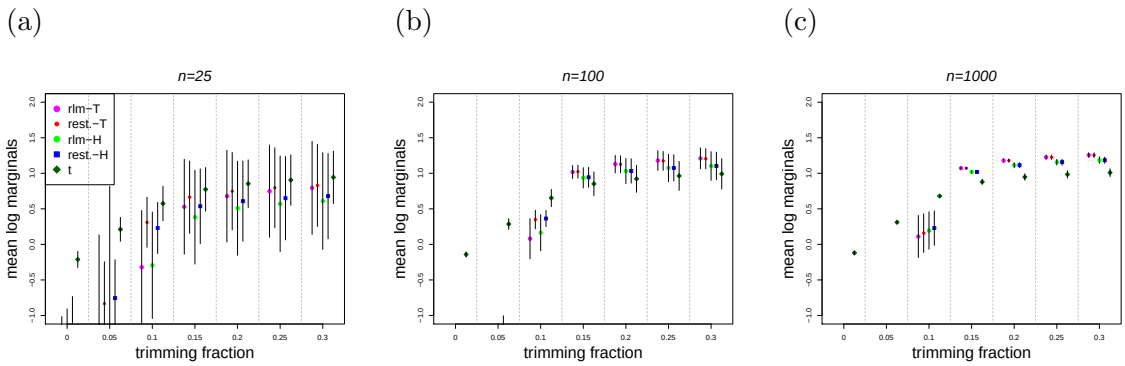

Figure 4: Model evaluation for 'Type 1' agencies for training sample sizes of $n = 25, 100$, and 1000. The t-model is used as the base method to compute TLM. Plotted are the mean TLM for each model against the trimming fraction across the 100 cross-validation samples. Error bars correspond to one standard deviation of TLM above and below the mean. Models are labeled with the following abbreviations: 'rlm' corresponds to a classical robust fit, 'rest.' corresponds to our restricted method, and 't' corresponds to the heavy-tailed t-distribution model. The letters 'T' and 'H' appearing after 'rlm' and 'rest.' correspond to the use of Tukey's and Huber's ρ respectively.

regression with Huber's ρ function (rlm-H), restricted likelihood based on the 'Huber estimate' (rest.-H), and the thick tailed t-model (t). The range of the vertical axis is chosen to enhance important features and as a result, some evaluation measures extend below this range. In particular, the restricted methods perform poorly if no trimming is done; reflecting that these methods are not intended to fit well to outliers. Recall that we expect about 15-16% outliers in the validation sets, thus trimming fractions slightly larger than this amount are needed in order to assess fits to the 'good' data. For $n = 25$, the thick tailed model prevails across trimming fractions, although less so for $\alpha \geq 0.15$. For sample sizes as low as $n = 100$, the restricted methods outperform the thick tailed model with the Tukey version performing the best. The stronger performance of restricted likelihood based on Tukey's method and the t model is to be expected, as many of the residuals are so extreme that trimming is better than winsorizing (as Huber's method effectively does). As expected, with

enough data, the Bayesian methods and their classical counterparts perform similarly, although there is a persistent slight edge in favor of the restricted likelihood methods. We attribute this advantage to the weakly informative prior distribution which pulls the estimates slightly toward better values. The similarity occurs as early as $n = 100$.

4.3 Hierarchical regression model

Nationwide agencies span many states and insurance regulations and the competitive environment varies between states. A natural extension to the previous analysis is a hierarchical regression model, grouping agencies within each state to reflect similar business environments. Using the same study design with the same training and validation splits, we re-analyze the data using the following hierarchical regression model:

$$\boldsymbol{\beta} \sim N_p(\mu_0, a\Sigma_0); \quad \boldsymbol{\beta}_j \overset{iid}{\sim} N_p(\boldsymbol{\beta}, b\Sigma_0); \quad \sigma_j^2 \sim IG(a_0, b_0);$$
$$\boldsymbol{y}_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_j^2), \ i = 1, \ldots, n_j, \ j = 1, \ldots, J$$

where y_{ij} represents the i^{th} observation in the j^{th} state, n_j is the total number of agencies in each state, and J is the number of states. $\boldsymbol{x}_{ij}$ is a four dimensional vector comprised of the same covariates as above. $\boldsymbol{\beta}_j$ represents the individual regression coefficient vector for state j. We match this model to the non-hierarchical model in several ways. First, μ_0, Σ_0, a_0, and b_0 are fixed as before. We constrain $a + b = 1$ in an attempt to partition the total variance between the individual $\boldsymbol{\beta}_j$'s and the overall $\boldsymbol{\beta}$. We take $b \sim \text{beta}(v_1, v_2)$. Using the previous data set, we assess the variation between individual estimates of the β_j to set v_1 and v_2 to allow for a reasonable amount of shrinkage. To allow for dependence across the σ_j^2 we first take $(z_1, \ldots, z_J) \sim N_J(\mathbf{0}, \Sigma_\rho)$ with $\Sigma_\rho = (1 - \rho)I + \rho J$. Then we set $\sigma_j^2 = H^{-1}(\Phi(z_j))$ where H is the

cdf of an $IG(a_0, b_0)$ resulting in the specified marginal distribution, while introducing correlation via ρ. We assume $\rho \sim \text{beta}(a_\rho, b_\rho)$ with mean μ_ρ and precision $\psi_\rho = a_\rho + b_\rho$. The parameters μ_ρ and ψ_ρ are given beta and gamma distributions respectively, both with fixed hyperparameters. To choose these fixed values we again consider fits to individual states from the previous dataset. Plugging the estimates of z_j into the multivariate normal, the mean of μ_ρ is set to the MLE of ρ and the variance is set to the observed inverse Fisher information matrix, inflated by a factor of 2 to weaken the prior for this parameter. We use the same MLE and inflated information matrix to set the mean for ψ_ρ. Its variance is chosen to cover a range of plausible values. A range of other values for the fixed hyper-parameters was also studied. The differences in results were negligible.

Using the same techniques as in the previous section, we fit the normal theory hierarchical model above, a thick tailed t version with $\nu = 3$ d.f., and two restricted likelihood versions (Huber's and Tukey's) of the model. For the restricted likelihood methods, we condition on robust regression estimates fit separately within each state. Both use Huber's scale estimator. We also fit classical robust regression counterparts and a least squares regression separately within each state.

We digress briefly to note that no additional computational strategies outside of those discussed in Section 3.2 are needed to fit the restricted hierarchical models described here. Since we condition on statistics which are computed within each state, the model's conditional independence between the states allows the data augmentation described earlier to be performed independently within each state. Updates of hyperparameters follow conventional MCMC procedures. We note that different types of statistics could be chosen for each state, if desired, allowing for a large amount of flexibility.

Selected results for the hierarchical fits appear in Figure 5. Hierarchical models

(a) (b)

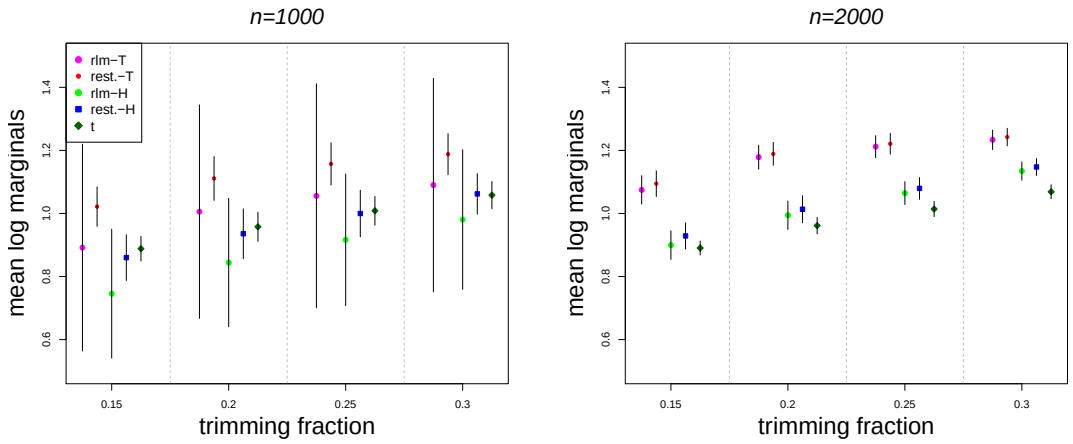

Figure 5: Model evaluation for 'Type 1' agencies under the hierarchical model for $n = 1000$ and 2000. The t-model is used as the base method to compute TLM. Plotted are the mean TLM for each model against the trimming fraction across the 100 cross-validation samples. Error bars correspond to one standard deviation of TLM above and below the mean. Models are labeled using the same notation as the previous figure. Only the relevant trimming fractions ($\alpha \geq .15$) are pictured.

naturally require more data and so we consider only training sizes of $n = 1000$ and 2000. Trimming fractions between 0.15 and 0.3 are displayed as patterns for smaller trimming fractions are similar to those from the non-hierarchical fits. That is, without sufficient trimming, the restricted likelihood fits' evaluation measure is poor. Again, the normal theory fits, both Bayesian and classical, perform poorly and are left out of the figures. We see that Tukey's version of restricted fits performs best in each case (assuming sufficient trimming). Huber's version also tops the thick tailed model for $n = 2000$. The Bayesian restricted fits considerably outperform their respective individual classical robust fits for training size of $n = 1000$. This observation remains, though marginally so, for $n = 2000$. The advantage of the hierarchical models seen here is due to the pooling of information across states, resulting in better predictive performance as compared to both the thick tailed competitor as well the respective classical fits.

4.4 Comparison of hierarchical and non-hierarchical fits

The performance of the methods for the hierarchical and non-hierarchical models can be contrasted through our cross validations studies. We focus on Tukey's and Huber's conditioning statistic and concentrate our evaluation on the 'Type 1' agencies. Table 1 displays the mean TLM for each model and range of trimming fractions. Our summary below focuses exclusively on realistic trimming fractions, $\alpha \geq 0.15$, and Tukey's conditioning statistic.

We first note that for the non-hierarchical model, there is little difference between mean TLM for $n = 1000$ and $n = 2000$, with the numbers differing only in the third decimal place. This is due to the posterior predictive distributions having stabilized. The mean TLMs for the hierarchical model show a greater change with increases of about 0.05 to 0.07 as the training sample size changes from 1000 to 2000. For calibration, the mean TLM for a normal with mean 0.5 and variance 1 is approximately this size when trimming is done under a standard normal base model. Thus, the increase in mean TLM is substantial. We attribute the change for the hierarchical model to the improvement in fits, particularly for states with fewer agencies.

Direct comparison of the hierarchical and non-hierarchical models shows that, for $n = 1000$, the non-hierarchical model has uniformly (for α of interest) better mean TLM. The differences are substantial, and the summaries primarily reflect greater stability of fits on a state-by-state basis under the non-hierarchical model. To a lesser extent, they reflect variation in the evaluation criterion which stems from modest validation sample size, particularly with larger trimming fractions. The trimmed cases are not proportionally distributed across states. The pattern changes for $n = 2000$, with the hierarchical model showing larger mean TLMs for trimming fractions 0.15 and 0.20. The improvement reflects the ability of the hierarchical model to capture

	Trimming fraction (α)			
	0.15	0.2	0.25	0.3
Tukey ($n = 1000$)				
Non-Hier.	1.072 (0.014)	1.179 (0.022)	1.226 (0.029)	1.255 (0.033)
Hier.	1.021 (0.063)	1.110 (0.070)	1.157 (0.067)	1.187 (0.065)
Tukey ($n = 2000$)				
Non-Hier.	1.068 (0.029)	1.178 (0.007)	1.225 (0.011)	1.254 (0.014)
Hier.	1.094 (0.041)	1.189 (0.036)	1.221 (0.033)	1.242 (0.028)
Huber ($n = 1000$)				
Non-Hier.	1.020 (0.020)	1.114 (0.035)	1.157 (0.041)	1.184 (0.045)
Hier.	0.861 (0.073)	0.937 (0.079)	1.001 (0.074)	1.063 (0.064)
Huber ($n = 2000$)				
Non-Hier.	1.015 (0.021)	1.112 (0.014)	1.154 (0.019)	1.181 (0.023)
Hier.	0.930 (0.041)	1.014 (0.043)	1.080 (0.035)	1.148 (0.027)

Table 1: Mean (standard deviation) of TLM for 'Type 1' agencies for the restricted non-hierarchical and hierarchical models for $n = 1000$ and 2000.

differences in regressions across the states which is realized when the training sample size is large enough. We attribute the better performance of the non-hierarchical model for the largest trimming fractions to variation in the evaluation.

5 Discussion

In this work, we have presented an approach which begins to reconcile Bayesian methods with the practice of data analysis. Many routine choices in an analysis react to the gap between reality and the statistical model, where a bit of set-up work improves inferential performance. Often, these choices can be recast in the framework of restricted likelihood, lending them more formality and facilitating development of theoretical results. But a much greater benefit of our framework is that it leads us to blend classical estimation with Bayesian methods. Here, we use the likelihood from robust regression estimators to move from prior distribution to posterior distribution. Conditioning on the estimator, the update follows Bayes' Theorem exactly. Compu-

tation is driven by MCMC methods, requiring only a modest supplement to existing algorithms. In another context, we might condition on the results of a set of estimating equations, designed to enforce lexical preferences for those features of the analysis considered most important, yet still producing inferences for secondary aspects of the problem. In other settings, we envision conditioning on a mix of estimators and some of the observed data.

The framework we propose allows us to retain many benefits of Bayesian methods: it requires a full and 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.

This same framework retains many of the benefits of classical estimation. Great ingenuity has been used to create a wide variety of estimators in this tradition, many of which are designed to handle specific flaws in the model. The estimators are typically accompanied by asymptotic results on consistency and distribution. Many of these results carry over to our blend of classical and Bayesian methods, although regularity conditions differ. We expect our procedures to have strong large sample performance, especially in settings where pooling of information is of value.

This framework opens a number of questions, including a need to revisit such issues as model selection, model averaging for predictive performance, and the role of diagnostics. Perhaps the biggest question is which summary statistic to choose. For this, we recommend a choice 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.

6 Appendix

Proof of Theorem 3.1.

Proof.

$$
\begin{aligned}
s(X, \boldsymbol{y}) &= \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{y}^*)} s(X, \boldsymbol{y}^*) = s(X, \boldsymbol{y}_{obs}), \qquad \text{and} && (15)\\
\boldsymbol{b}(X, \boldsymbol{y}) &= \boldsymbol{b}\left(X, \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{y}^*)} \boldsymbol{y}^* + X\left(\boldsymbol{b}(X, \boldsymbol{y}_{obs}) - \boldsymbol{b}(X, \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{y}^*)} \boldsymbol{y}^*)\right)\right) && (16)\\
&= \boldsymbol{b}(X, \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{y}^*)} \boldsymbol{y}^*) + \boldsymbol{b}(X, \boldsymbol{y}_{obs}) - \boldsymbol{b}(X, \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{y}^*)} \boldsymbol{y}^*) && (17)\\
&= \boldsymbol{b}(X, \boldsymbol{y}_{obs}) && (18)
\end{aligned}
$$

□

Proof of Lemma 3.2.

Proof. We first show that $\nabla s(X, \boldsymbol{y}) \in \mathcal{C}^{\perp}(X)$. Recall that $H = I - Q$. By the regression invariance property C7 of s, we have

$$
s(X, \boldsymbol{y}) = s(X, Q\boldsymbol{y} + H\boldsymbol{y}) = s(X, Q\boldsymbol{y}). \tag{19}
$$

Thus, by the chain rule $\nabla s(X, \boldsymbol{y}) = Q \nabla s(X, Q\boldsymbol{y}) = Q \nabla s(X, \boldsymbol{z})$. Hence $X^{\top} \nabla s(X, \boldsymbol{y}) = 0$ as desired. From equation (19), all vectors $\boldsymbol{z}' \in \Pi(\mathcal{A})$ satisfy $s(X, \boldsymbol{z}') = s(X, \boldsymbol{y}) = s(X, \boldsymbol{y}_{obs})$, and so all directional derivatives of s along each tangent $\boldsymbol{v}$ to $\Pi(\mathcal{A})$ in $\mathcal{C}^{\perp}(X)$ at $\boldsymbol{z}$ are equal to 0 (i.e., $\nabla s(X, \boldsymbol{z}) \cdot \boldsymbol{v} = 0$). Thus $\nabla s(X, \boldsymbol{z})$ is orthogonal to $\Pi(\mathcal{A})$ at $\boldsymbol{z}$. Since $\Pi(\mathcal{A})$ has dimension $n - p - 1$, $\nabla s(X, \boldsymbol{z})$ gives the unique (up to scaling and reversing direction) normal in the $n - p$ dimensional $\mathcal{C}^{\perp}(X)$. □

Proof of Lemma 3.3

Proof. Without loss of generality, assume the columns of X form an orthonormal basis for $\mathcal{C}(X)$ and likewise the columns of W form and orthonormal basis for $\mathcal{C}^{\perp}(X)$. With earlier notation, $H = XX^{\top}$ and $Q = WW^{\top}$. The set $\mathcal{A}$ is defined by the $p+1$ equations $s(X, \boldsymbol{y}) = s(X, \boldsymbol{y}_{obs})$, $b_1(X, \boldsymbol{y}) = b_1(X, \boldsymbol{y}_{obs}), \ldots, b_p(X, \boldsymbol{y}) = b_p(X, \boldsymbol{y}_{obs})$. Consequently, the gradients are orthogonal to $\mathcal{A}$. Let $\nabla \boldsymbol{b}(X, \boldsymbol{y})$ denote the $n \times p$ matrix with columns $\nabla b_1(X, \boldsymbol{y}), \ldots, \nabla b_p(X, \boldsymbol{y})$. We seek to show the $n \times (p+1)$ matrix $[\nabla \boldsymbol{b}(X, \boldsymbol{y}), \nabla s(X, \boldsymbol{y})]$ has rank $p+1$. Using property C5, we have that

$$\boldsymbol{b}(X, \boldsymbol{y}) = \boldsymbol{b}(X, Q\boldsymbol{y} + H\boldsymbol{y}) = \boldsymbol{b}(X, Q\boldsymbol{y}) + X^{\top}\boldsymbol{y}$$

Then $\nabla \boldsymbol{b}(X, \boldsymbol{y}) = Q\nabla \boldsymbol{b}(X, Q\boldsymbol{y}) + X$ and

$$[XX^{\top}, WW^{\top}]^{\top}[\nabla \boldsymbol{b}(X, \boldsymbol{y}), \nabla s(X, \boldsymbol{y})] = \begin{pmatrix} X & \mathbf{0} \\ WW^{\top}\nabla b(X, \boldsymbol{y}) & \nabla s(X, \boldsymbol{y}) \end{pmatrix} \tag{20}$$

The last column comes from Lemma 3.2. The matrix $[XX^{\top}, WW^{\top}]^{\top}$ is of full column rank (rank n), and so the rank of $[\nabla \boldsymbol{b}(X, \boldsymbol{y}), \nabla s(X, \boldsymbol{y})]$ is the same as the rank of the matrix on the right hand side of (20). This last matrix has rank $p+1$ since $\nabla s(X, \boldsymbol{y}) \neq \mathbf{0}$ by C8, and so does $[\nabla b(X, \boldsymbol{y}), \nabla s(X, \boldsymbol{y})]$. □

Proof of Lemma 3.4

Proof. P is the projection of the columns of A onto $\mathcal{C}^{\perp}(X)$. For this to result in a loss of rank, a subspace of $\mathcal{T}_y(\mathcal{A})$ must belong to $\mathcal{C}(X)$. Following property C5, for an arbitrary vector $X\boldsymbol{v} \in \mathcal{C}(X)$, $\boldsymbol{b}(X, \boldsymbol{y} + X\boldsymbol{v}) = \boldsymbol{b}(X, \boldsymbol{y}) + \boldsymbol{v}$. From the property, we can show that the directional derivative of $\boldsymbol{b}$ along $X\boldsymbol{v}$ with $\boldsymbol{v} \neq \mathbf{0}$ is $\boldsymbol{v}$, which is a nonzero vector. Hence $X\boldsymbol{v} \notin \mathcal{T}_y(\mathcal{A})$. □

Proof of Corollary 3.7

Proof. The corollary relies on a lemma and theorem from Miao and Ben-Israel (1992) which we restate slightly for brevity of presentation. The principal angles between subspaces pluck off a set of angles between subspaces, from smallest to largest. The number of such angles is the minimum of the dimensions of the two subspaces. Miao and Ben-Israel's first result (their Lemma 1) connects these principal angles to a set of singular values, and hence to volumes.

Lemma 6.1. *(Miao, Ben-Israel) Let the columns of $Q_L \in \mathbb{R}^{n\times l}$ and $Q_M \in \mathbb{R}^{n\times m}$ form orthonormal bases for linear subspaces L and M respectively, with $l \leq m$. Let $\sigma_1 \geq \cdots \geq \sigma_l \geq 0$ be the singular values of $Q_M^\top Q_L$. Then $\cos\theta_i = \sigma_i, i = 1, \ldots, l$ where $0 \leq \theta_1 \leq \theta_2 \leq \cdots \leq \theta_l \leq \frac{\pi}{2}$ are the principal angles between L and M.*

Miao and Ben-Israel's second result (their Theorem 3) makes a match between the principal angles between a pair of subspaces and the principal angles between their orthogonal complements.

Theorem 6.2. *(Miao, Ben-Israel) The nonzero principal angles between subspace L and M are equal to the nonzero principal angles between $L^\perp$ and $M^\perp$.*

To establish the corollary, we appeal to Lemma 6.1 and Theorem 6.2. Translating Miao and Ben Israel's notation, we have $M = \mathcal{C}^\perp(X)$, $Q_M = W$, $L = \mathcal{T}_{\boldsymbol{y}}(\mathcal{A})$, and $Q_L = A$. By Theorem 6.2, the nonzero principal angles between $\mathcal{T}_{\boldsymbol{y}}(\mathcal{A})$ and $\mathcal{C}^\perp(X)$ are the same as the nonzero principal angles between $\mathcal{T}_{\boldsymbol{y}}^\perp(\mathcal{A})$ and $\mathcal{C}(X)$. By 6.1, the non-unit singular values of $W^\top A$ are the same as the non-unit singular values of $U^\top B$. □

References

Agresti, A. (2002). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, New Jersey, 2nd edition.

Albert, J. H. (1988). Computational methods using a Bayesian hierarchical generalized linear model.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, 1:385–402.

Bissiri, P., Holmes, C., and Walker, S. (2013). A general framework for updating belief distributions. *arXiv preprint arXiv:1306.6430*.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society. Series A*, 143:383–430.

Clarke, B. and Ghosh, J. K. (1995). Posterior convergence given the mean. *The Annals of Statistics*, 23:2116–2144.

Dean, A. and Voss, D. (1999). *Design and Analysis of Experiments*. Springer Texts in Statistics. Springer Science + Business Media, Inc., New York, New York.

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.

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.

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.

Hoff, P., Fosdick, B., Volfovsky, A., and Stovel, K. (2013). Likelihoods for fixed rank nomination networks. *Network Science*, 1:253–277.

Hoff, P. and Wakefield, J. (2013). Bayesian sandwich posteriors for pseudo-true parameters. *Journal of Statistical Planning and Inference*, 143: 16381642.

Huber, P. and Ronchetti, E. (2009). *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc, Hoboken, New Jersey, 2nd edition.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1): 73–101.

Hwang, H., So, B., and Kim, Y. (2005). On limiting posterior distributions. *Test*, 14:567–580.

Jung, Y., MacEachern, S., and Lee, Y. (2014). Cross-validation via outlier trimming. In preparation.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Lee, J. and MacEachern, S. (2014). Inference functions in high dimensional Bayesian inference. *Statistics and Its Interface*. In press.

Lewis, J. (2014). *Bayesian Restricted Likelihood Methods*. PhD thesis, The Ohio State University.

Lewis, J., Lee, Y., and MacEachern, S. (2012). Robust inference via the blended paradigm. In *JSM Proceedings*, Section on Bayesian Statistical Science, pages 1773–1786. American Statistical Association.

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.

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.

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.

Maronna, R., Martin, D., and Yohai, V. (2006). *Robust Statistics: Theory and Methods*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd, West Sussex, England.

Miao, J. and Ben-Israel, A. (1992). On principal angles between subspaces in $\mathbb{R}^n$. *Linear Algebra and its Applications*, 171:81–98.

O'Rourke, K. (2007). *The Combining of Information: Investigating and Synthesizing What is Possibly Common in Clinical Observations or Studies via Likelihood*. PhD thesis, University of Oxford.

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.

Pettitt, A. N. (1983). Likelihood based inference using signed ranks for matched pairs. *Journal of the Royal Statistical Society. Series B*, 45:287–296.

Pratt, J. W. (1965). Bayesian interpretation of standard inference statements. *Journal of the Royal Statistical Society. Series B*, 27:169–203.

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.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, 114:510.

Ronchetti, E., Field, C., and Blanchard, W. (1997). Robust linear model selection by cross-validation. *Journal of the American Statistical Association*, 92:1017–1023.

Savage, I. R. (1969). Nonparametric statistics: A personal review. *Sankhya: The Indian Journal of Statistics, Series A (1961-2002)*, 31:107–144.

Tavaré, S., Balding, D. J., Griffiths, R., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145:505–518.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition.

Yuan, A. and Clarke, B. (2004). Asymptotic normality of the posterior given a statistic. *The Canadian Journal of Statistics*, 32:119–137.