Bayesian Restricted Likelihood Methods: Conditioning on Insufficient Statistics in Bayesian Regression

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 a new method as an alternative. When working with the new method, 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 careful choice of conditioning statistics, we retain the main benefits of Bayesian methods while reducing the sensitivity of the

*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.

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. With these choices, the method can be thought of as a blend of classical robust estimation and Bayesian methods.

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 wide range of choices for summary statistics. 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; 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. These imperfections often show themselves through the presence of outliers–cases not reflecting the phenomenon under study. 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 mixture model likelihood to obtain the complete-data posterior distribution and 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).

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. 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 because conditioning is done on a restricted set of data; the set which satisfies the observed summary statistics. This is a formal update of the prior distribution based on the sampling density of the summary statistics. 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. Second, 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.

The remainder of the paper develops....

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 are 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:

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

We wish to keep the first piece 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

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

Only the first part is retained 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:

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

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

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

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

Direct use of restricted likelihood appears in many areas of the literature. The motivation is often similar to ours: concern about outliers or, more generally, 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). Lewis et al. (2012) make use order statistics and robust estimators 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; Marjoram et al., 2003; Fearnhead and Prangle, 2012). 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 $\mathcal{B}(\boldsymbol{y}) = \{\boldsymbol{y}^*|\rho(T(\boldsymbol{y}), T(\boldsymbol{y}^*)) < \epsilon\}$, ρ is a metric, and ϵ is a tolerance level. This is the likelihood "conditioned" on the collection of data sets which result in a $T(\cdot)$ within ϵ of the observed $T(\boldsymbol{y})$. With an approximately sufficient $T(\cdot)$ and a small enough ϵ, 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). ABC is related to our method in that the "conditioning" is on something other than the data $\boldsymbol{y}$. However, we specifically seek to condition on an insufficient statistic to guard against misspecification in parts of the likelihood. Additionally, we develop methods where the conditioning is exact (i.e. $\epsilon = 0$).

This work extends the development of Bayesian restricted likelihood by arguing that deliberate choice of $T(\boldsymbol{y})$ is sound practice and also by expanding the class of conditioning statistics in which exact conditioning can be achieved. Our methods do not rely on asymptotic properties, nor do they rely on approximate conditioning.

3 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 4.

The first example is an analysis of Simon Necomb's 66 measurements of the speed of light; two of which are significant outliers in the lower tail. The full model is a standard location-scale Bayesian model:

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

where y_i denotes the i^{th} measurement of the passage time of light. β is interpreted as the passage time of light with σ^2 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. Associated tuning parameters for the M-estimators are chosen to achieve 95% efficiency under normality Huber (1964) and for comparability, roughly 5% of the residuals are trimmed for LTS. Additionally, two other common approaches to outlier handling are fit: 1) replacing the normal distribution with a t-distribution and, 2) replacing the normal distribution 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 σ^2 is $IG(5, \frac{\nu-2}{\nu}10)$ so the prior on the variance is the same as the other models. The mixture takes the form: $y_i \overset{iid}{\sim} pN(\beta, \sigma^2) + (1-p)N(\beta, 10\sigma^2)$ assuming the prior $p \sim beta(20, 1)$ on the probability of belonging to the 'good' component.

The posteriors of β under each model appear in top of Figure 1. As expected, the posterior under the normal model is pulled downward by the two outliers while the heavy tailed model provides robustness against them. The restricted likelihood methods using the M-estimators and LTS statistics also achieve robustness against the outliers. Conditioning on LMS however, results in a posterior similar to the one under the normal model. The M-estimators provide the most precise posteriors in this case. This is reflected in more precise predictions than the heavy-tailed and mixture model as illustrated by the predictive distributions displayed in the bottom of Figure 1.

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 units on which calls were recorded for part of the dataset. The full model is a standard normal Bayesian linear regression:

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

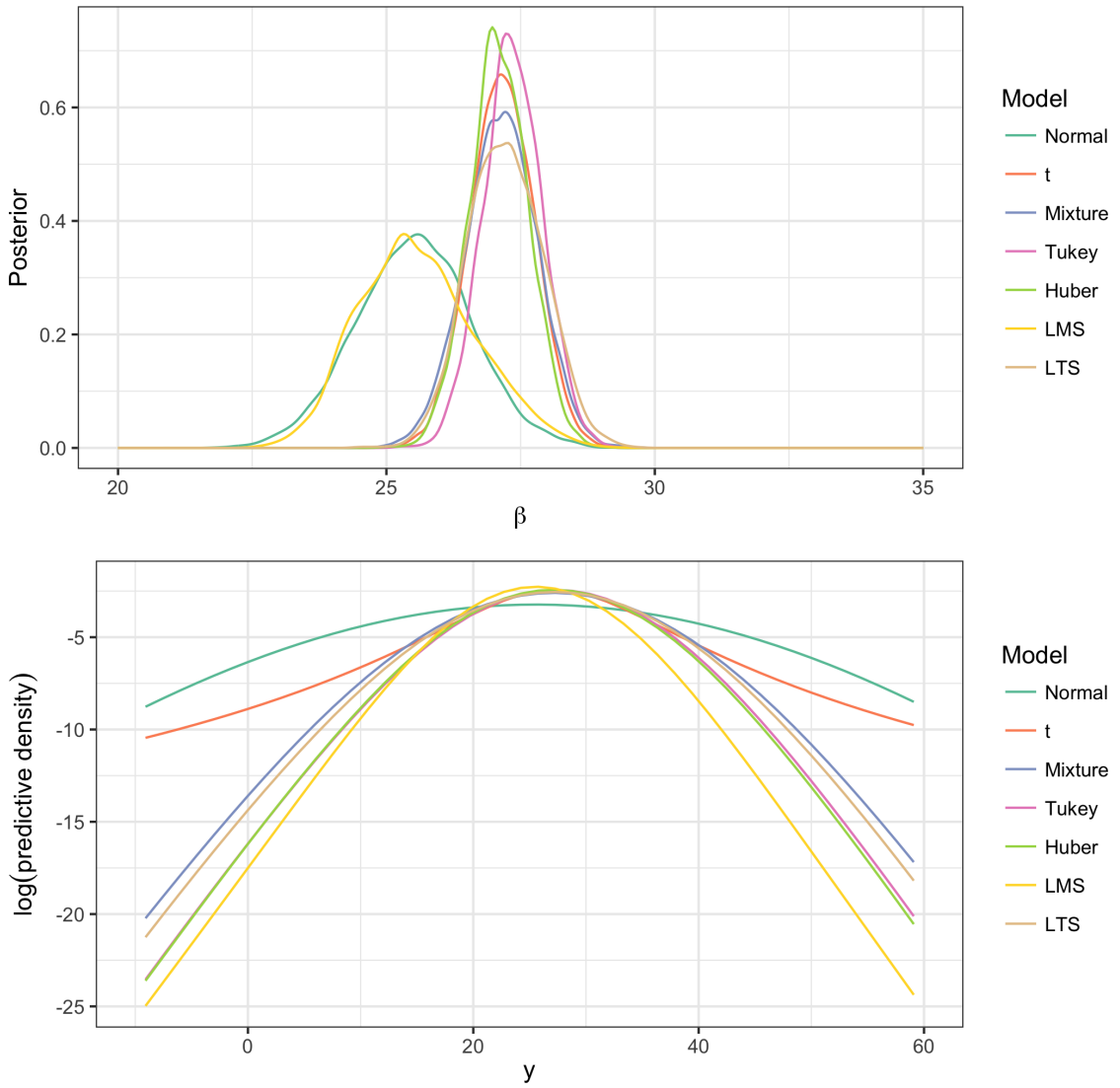


Figure 1: Top: Posteriors of the location parameter β under the normal theory model, the t-model, and four restricted likelihood models for Newcomb's speed of light measurements. Bottom: Posterior predictive distribution of the speed of light under the normal theory model, the t-model, and four restricted likelihood models.

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. Prior parameters are fixed via a fit to the first 3 data points. In particular, $\Sigma_0 = g\sigma_0^2(X^\top X)^{-1}$, with $\sigma_0 = 0.03$ and $\mu_0 = (1.87, 0.03)^\top$; the MLEs fit to the first three data points. There are $n = 21$ remaining data points and the parameter g is set to 21 reflecting a unit information prior . Finally $a = 2$ and $b = 1$ for the normal theory and restricted likelihood models.

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. The mixture model assumes different mean regression functions and variances for each component, but keeps the same, relatively non-informative 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 to 5 with the same adjustment to the prior on σ^2 as above.

The data and 95% credible bands of the posterior predictive distribution under each model are displayed in Figure 2. The normal theory model is only fit to the obvious non-outlying points. Since the t-model assumes the data are heavy-tailed, the posterior predictive distribution is much wider. On the other hand, the predictive distribution under the restricted likelihood approach is much more precise and is close to that of the normal theory fit that discards the outliers. It is also close to the two component mixture results where the predictive distribution is formulated using only the good component. The mixture model involves explicitly modeling the outlier generating mechanism. In more more complex situations where the outlier generating mechanism is transient (i.e. ever changing and more complex than just a unit error in the recording), modeling the outliers becomes more difficult. Like classical robust estimation, the restricted likelihood approach avoids explicitly modeling the outliers.

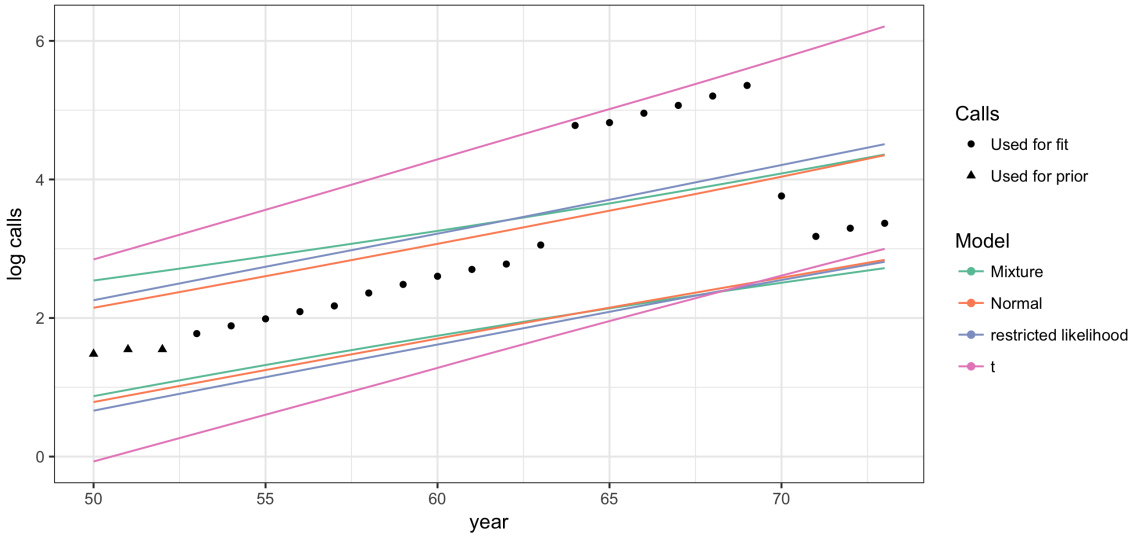


Figure 2: Predictive distribution of log(calls) under the Normal theory model fit to the non-outliers, the restricted likelihood model with Tukey's M-estimator for the slope and intercept with Huber's 'proposal 2' for scale, and a heavy-tailed t-distribution model. The first three data points were used to specify the prior with the remaining points used in the posterior fits. See details in the Appendix.

4 Restricted Likelihood for the Linear Model

The simple examples in the previous section highlight that productive use of the restricted likelihood relies on a good choice of $T(\boldsymbol{y})$. 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 $\boldsymbol{\theta}$, direct computational strategies described in Lewis (2014) can be used to estimate the restricted posterior conditioned on essentially any statistic. These strategies rely on density estimation $f(T(\boldsymbol{y})|\theta)$ using samples of $T(\boldsymbol{y})$ for many values of θ; 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.

4.1 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 ϵ_i are independent draws from a distribution with center 0 and scale σ. X denotes the design matrix whose rows are $x_i^\top$.

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 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.

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 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. Further software development is required to extend the MCMC implementation beyond these M-estimators. The package also implements the direct computational methods described in Lewis (2014). These methods are effective in lower dimensional problems and were used in several of the examples in Section 3.

4.2 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. $\pi(\boldsymbol{\theta}|\boldsymbol{y}, T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs}))$
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 this generation. 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

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

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

$$\begin{aligned} R &= \frac{f(\boldsymbol{y}_p|\boldsymbol{\theta}, \boldsymbol{y}_p \in \mathcal{A})}{f(\boldsymbol{y}_c|\boldsymbol{\theta}, \boldsymbol{y}_c \in \mathcal{A})} \frac{p(\boldsymbol{y}_c|\boldsymbol{\theta}, \boldsymbol{y}_c \in \mathcal{A})}{p(\boldsymbol{y}_p|\boldsymbol{\theta}, \boldsymbol{y}_p \in \mathcal{A})} && (9) \\ &= \frac{f(\boldsymbol{y}_p|\boldsymbol{\theta})}{\int_{\mathcal{A}} f(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{y}} \frac{\int_{\mathcal{A}} f(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{y}}{f(\boldsymbol{y}_c|\boldsymbol{\theta})} \frac{p(\boldsymbol{y}_c|\boldsymbol{\theta})}{p(\boldsymbol{y}_p|\boldsymbol{\theta})} && (10) \\ &= \frac{f(\boldsymbol{y}_p|\boldsymbol{\theta})}{f(\boldsymbol{y}_c|\boldsymbol{\theta})} \frac{p(\boldsymbol{y}_c|\boldsymbol{\theta})}{p(\boldsymbol{y}_p|\boldsymbol{\theta})}. && (11) \end{aligned}$$

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).

4.2.1 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 $\boldsymbol{z}^*$ which generates a positive scale estimate to such a $\boldsymbol{y}$ via the following Theorem, whose proof is in the appendix.

Theorem 4.1. *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 point 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. With $\boldsymbol{z}^* \in \mathbb{S}$, $c\boldsymbol{z}^* + X\boldsymbol{v}$ is not (unless $c = 1$ and $\boldsymbol{v} = \boldsymbol{0}$); the scaling by c and/or the affine transformation in the direction of $\mathcal{C}(X)$ takes the point off $\mathbb{S}$. The mapping $h : \mathbb{S} \to \mathcal{A}$ is one-to-one making the change of variables more feasible.

With the domain of $h(\cdot)$ restricted to $\mathbb{S}$, the range is still the entirety of $\mathcal{A}$. This is important so that the support of the proposal distribution (which is 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 Metroplis-Hastings algorithm (in this case the supports are both $\mathcal{A}$). To see that the range of $h(\cdot)$ is $\mathcal{A}$, consider any $\boldsymbol{y} \in \mathcal{A}$ and its projection onto $\mathcal{C}^\perp(X)$: $Q\boldsymbol{y}$ where $Q = I - XX^\top$. [1] It is easy to show that $\boldsymbol{z}^* = Q\boldsymbol{y}/||Q\boldsymbol{y}|| \in \mathbb{S}$ and $h(\boldsymbol{z}^*) = \boldsymbol{y}$.

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 on $\mathbb{S}$.

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

[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., $X^\top X = I$).

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 to aid in the understanding of the strategy we are taking. 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.

4.2.2 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.

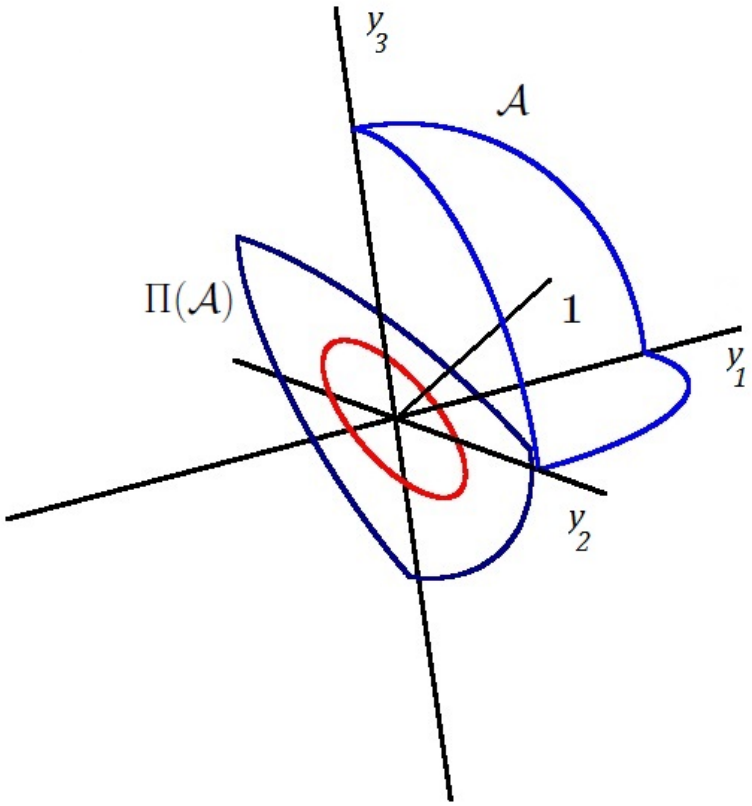


Figure 3: A depiction of $\mathcal{A}$, $\Pi(\mathcal{A})$, and the unit circle for the illustrative example where $b_1(\mathbf{1}, \boldsymbol{y}) = \min(\boldsymbol{y}) = 0$ and $s(\mathbf{1}, \boldsymbol{y}) = \sum(y_i - b_1(\mathbf{1}, \boldsymbol{y}))^2 = 1$. $\mathcal{A}$ is the combination of 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 4.2. *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 γ is the angle between the two normal vectors. This step is visualized in Figure 4 for the artificial location-scale example. The figure pictures only the $\mathcal{C}^{\perp}(X)$, 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. 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. 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. All gradients denote with ∇ are with respect to the data vector.

Lemma 4.3. *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 4.4. *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 4.3 and 4.4 above leaves us with the following result concerning the calculation of the desired Jacobian.

Theorem 4.5. *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 4.6. *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 4.1, 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.*

A few details for computingthe needed quantities are worth further explanation. Computing 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 apply Gram-Schmidt orthonormalization. This is $\mathcal{O}(n^3)$ and 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 $\nabla s(X, \boldsymbol{y}), \nabla b_1(X, \boldsymbol{y}), \ldots, \nabla b_p(X, \boldsymbol{y})$, which is $\mathcal{O}(np^2)$; a considerable reduction in computational burden when $n \gg p$. The following corollary formally states how computation of A can be circumvented.

Corollary 4.7. *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 the 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. For example, below we consider M-estimators defined by the estimating equations:

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

where ψ and χ are almost surely differentiable. Differentiating this system of equa-

tions with respect to each y_i can be used to find the gradients. In theory, finite differences could also be used.

5 Simulated Data

We study the performance of the restricted likelihood in a hierarchical setting contaminated with outliers. Specifically, simulated data come from the following data generating model:

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

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:

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

For the restricted likelihood versions we condition on robust M-estimators of location and scale in each group: $T_i(\boldsymbol{y}_i) = (\hat{\theta}_i, \hat{\sigma}_i^2), i = 1, 2, ..., 90$. The two versions use Huber's and Tukey's ψ function, while both versions use Huber's χ function in (15). These estimators are implemented in the R function `MASS::rlm` and `??` provides further

details on them.

To complete the specification of model (17) we must specify a_s and b_s. For this we use a variety of values representing different levels of prior knowledge. For each we set $b_s = 4a_s c_s$ resulting in a prior mean for each σ_i^2 of $\frac{4c_s a_s}{a_s - 1}$, $a_s > 1$. The precision is $\frac{(a_s-1)^2(a_s-2)}{(4c_s a_s)^2}$; meaning the larger a_s, the more informative the prior. With $c_s = 1$ the shrinkage (for large a_s) is to the true value of $\sigma^2 = 4$ of the variance of the good component of data in the data generating model. We consider $a_s = 1.25, 5, 10$ and $c_s = 1/2, 1, 2$.

For each simulated data set, the Bayesian models are fit using MCMC. The MCMC for the restricted likelihood version requires no further computational details other than those described for the traditional Bayesian model in Section 4. 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 generating group level data matching the statistics for that group.

The models are compared using Kullback-Leibler (KL) divergence from the distribution of good data to the posterior predictive distribution in order to assess the predictive capability of the models for the good part of the data. Specifically, for the i^{th} group of the k^{th} simulated data set $\boldsymbol{y}_k$ compute:

$$KL_{ik}^{(M)} = \int \log \frac{f(\tilde{y}|\theta_i, \sigma^2)}{f_i(\tilde{y}|M, \boldsymbol{y}_k)} f(\tilde{y}|\theta_i, \sigma^2)\, dy \tag{18}$$

where M indexes the fitting model, $f(\tilde{y}|\theta_i, \sigma^2)$ is the pdf of a $N(\theta_i, \sigma^2)$ at $\tilde{y}$ with $\sigma^2 = 4$. For the Bayesian models $f_i(\tilde{y}|M, \boldsymbol{y}_k) = \int f(\tilde{y}|\theta_i, \sigma_i^2)\pi(\theta_i, \sigma_i^2|M, \boldsymbol{y}_k)$ where $f(\tilde{y}|\theta_i, \sigma_i^2)$ is $N(\theta_i, \sigma_i^2)$ pdf and $\pi(\theta_i, \sigma_i^2|M, \boldsymbol{y}_k)$ is the posterior for the i^{th} group model parameters under model M for the k^{th} data set. M denotes either the full normal the-

ory model (17) or one of the two restricted likelihood versions discussed above, where conditioning is only on the summary statistics (not the complete $\boldsymbol{y}_k$). For the classical robust fits, we set $f_i(\tilde{y}|M, \boldsymbol{y}_k)$ to a $N(\hat{\theta}_i, \hat{\sigma}_i^2)$ at $\tilde{y}$ as a groupwise plug-in estimator for the predictive distribution. Thus the classical fits are done separately for each group - with no consideration of the hierarchical structure between the groups. The overall mean $\overline{KL}_{\cdot\cdot}^{(M)} = \frac{1}{90K}\sum_{k=1}^{K}\sum_{i=1}^{90} KL_{ik}^{(M)}$ is used to compare the models. Sampling variation is summarized with $SE(\overline{KL}_{\cdot k}^{(M)}) = \sqrt{\frac{1}{K(K-1)}\sum_{k=1}^{K}(\overline{KL}_{\cdot k}^{(M)} - \overline{KL}_{\cdot\cdot}^{(M)})^2}$ where $\overline{KL}_{\cdot k}^{(M)} = \frac{1}{90}\sum_{i=1}^{90} KL_{ik}^{(M)}$.

Figure **??** displays the results for each $a_s = 1.25, 5, 10$ and $c_s = 1/2, 1, 2$ over $K = 30$ data sets. The panels are for $c_s = 1/2, 1, 2$ with $a_s = 1.25, 5, 10$ on the abscissa. For each panel, the KL results for the classical robust linear model fits are the same since the priors have no effect on these fits. The normal theory model results are left out as they perform poorly. Overall, the results are strikingly in favor of the restricted likelihood methods for the range of hyper-parameter values studied. Undoubtably, the most precise and accurate prior studied is $c_s = 1$ and $a_s = 10$ and this results in the lowest (and best) average KL for both the Tukey and Huber restricted likelihood versions. $c_s = 0.5$ performs the worst, but still better than the classical fits. There is evidence that performance starts to degrade as $a_s = 10$ as reflected in a larger average KL for $a_s = 10$. Here, the prior mean and precision are 2.22 and 1.62 and we suspect this is starting to put too much mass on σ_i^2 values much smaller than $\sigma^2 = 4$. For $c_s = 2$, the performance still improves from $a_s = 5$ to $a_s = 10$. Here the mean is 8.89 and precision is only 0.1; not yet large enough to degrade the performance due to an incorrect mean.

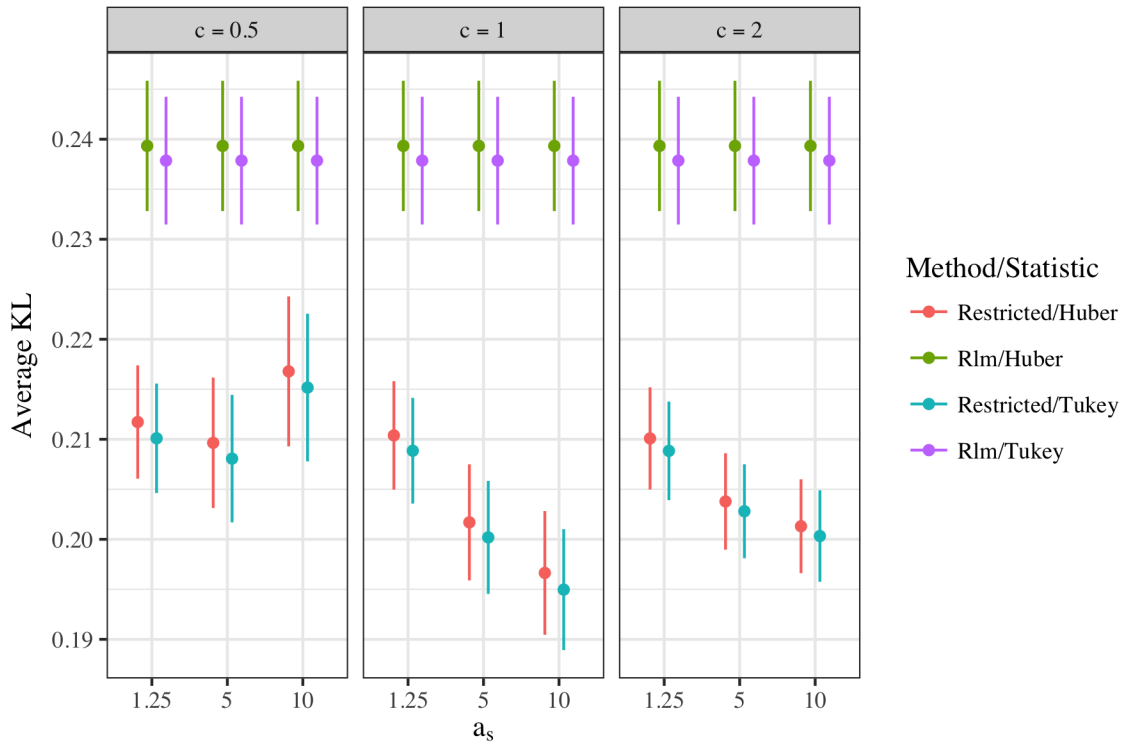


Figure 5: summary of simulation

6 Real Data

We illustrate our methods with a pair of regression models for data from Nationwide Insurance Company, which 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'). We used data from previous years to build a model to forecast future household count. In particular, we use household count as measured during a single month in 2010, to predict household counts in the corresponding month in

2012. The data a grouped by states with a varying number of agencies by state. Identifiers such as agency/agent names are removed. State labels and agency types (identifying the varying contractual agreements) have been made generic to protect the proprietary nature of the data. As an exploratory view, a plot of the square root of household count in 2012, against that in 2010 is shown in Figure 6 for three states. Each state has a varying number of agencies and the different colors represent the varying contractual agreements as they stood in 2010. Among the open agencies, linear correlations exists with strength depending on agency type and state. 'Type 1' agencies are of special interest and one could easily subset the analysis to only these agencies, removing the others. Additionally, a large number of agencies closed sometime before 2012, as represented by the many 0 counts for 2012. The other agency types and the zero counts are 'outliers' and could arguably just be removed prior to analysis. However, we leave them and us 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 restricting likelihood will facilitate prediction for the 'good' part of the data (i.e. open, 'type 1' agencies).

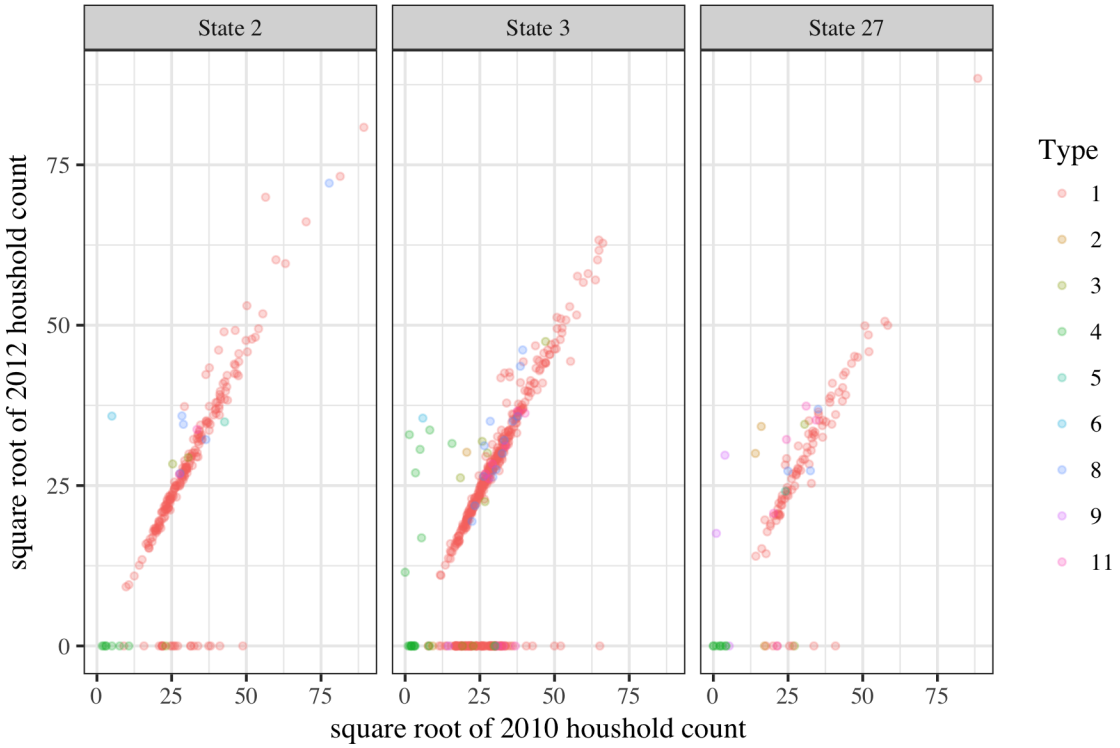


Figure 6: The square root of count in 2012 versus that in 2010 (after centering and scaling). The colors represent the varying contractual agreements as they stood in 2010. Agencies that closed during the 2010-2012 period are represented by the zero counts for 2012. Scalings on the axes are purposely left off for proprietary reasons.

6.1 Regression model

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

$$\beta \sim N(\mu_0, \sigma_0^2); \quad \sigma^2 \sim IG(a_0, b_0); \quad y_i = \beta x_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \ i = 1, \ldots, n, \quad (19)$$

where y_i and x_i are the square rooted household count in 2012 and 2010 for the i^{th} agency, respectively. The hyper-parameters a_0, b_0, μ_0 and σ_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 $\hat{\beta}$ and $\hat{\sigma}^2$ be estimates from the robust linear regression of

2010 counts on 2012 counts. We fix $a_0 = 5$ and set $b_0 = \hat{\sigma}^2(a_0 - 1)$. We set $\mu_0 = \hat{\beta}$ and $\sigma_0^2 = f n_p se(\hat{\beta})^2$ where n_p is the number of agencies in the prior data set and $se(\hat{\beta})$ is the standard error of $\hat{\beta}$ derived from the robust regression. The value of f is varied, controlling the prior certainty on β, with smaller values corresponding to a more certain prior. We take $f = 0.05, 0.1, 0.5, 1$ and note the prior is in the spirit of the Zellner's g-prior (Zellner, 1986; Liang et al., 2008); scaling the prior variance $se(\hat{\beta})^2$ by a factor $g = f n_p$. $f = 1$ is analogous to the unit-information prior (Kass and Wasserman, 1995).

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 Section **??** adapted to linear regression. The heavy-tailed model replaces the normal sampling density in (19) with a t-distribution with $\nu = 3$ degrees of freedom. We also fit the corresponding classical robust regressions and a least squares regression.

6.1.1 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 used as the evaluation function.

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 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 validation sample, say $i = 1, \ldots, M$. Order the validation 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 validation 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. 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.

6.1.2 Comparison of predictive performance

Model performance is assessed using the mean and standard deviation of the TLM across 100 different splits into training and validation samples. 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 was 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 7 for training sample sizes $n = 25, 100,$ and 1000. The t-model is used as the base method to compute TLM. The models pictured are: classical robust regression with Tukey's ψ function (rlm-T), restricted likelihood with Tukey ψ (restr.-T), classical robust regression with Huber's ψ function (rlm-H), restricted likelihood with Huber's ψ (restr.-H), and the thick tailed t-model (t). The normal theory models perform poorly due to the numerous outliers and are left out of the figures. Appearing in the figures are the mean TLM across validations 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 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 likelihood 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 likelihood methods outperform the heavy-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 Bayesian 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 $n = 100$.

6.2 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:

$$\begin{aligned} &\boldsymbol{\beta} \sim N_p(\boldsymbol{\mu}_0, a\Sigma_0); \quad \boldsymbol{\beta}_j \overset{iid}{\sim} N_p(\boldsymbol{\beta}, b\Sigma_0); \quad \sigma_j^2 \sim IG(a_0, b_0); \\ &\boldsymbol{y}_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_j^2), \ i = 1, \ldots, n_j, \ j = 1, \ldots, J \end{aligned} \tag{20}$$

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, $\boldsymbol{\mu}_0$, 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 $\Sigma_\rho = (1-\rho)I + \rho \mathbf{1}\mathbf{1}^\top$. Then we set $\sigma_j^2 = H^{-1}(\Phi(z_j))$ where H is the cdf of an $IG(a_0, b_0)$ and Φ is the cdf of a standard normal. This results in the specified marginal distribution, while introducing correlation via ρ. 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 μ_ρ and ψ_ρ are given beta and gamma distributions, respectively. We fix the parameters of these distributions by again considering fits to individual states from the previous data set. More precise details on setting v_1, v_2 and the the priors on μ_ρ and ψ_ρ are given in the appendix. We note that we tried a range of other fixed hyper-parameters resulting in negligible differences in the results.

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 incomplete 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.

We digress briefly to note that for the restricted likelihood methods no additional computational strategies outside of those discussed in Section 4.2 are needed to fit

the 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 8. Hierarchical models naturally require more data and so we consider only training sizes of $n = 1000$ and 2000. Again, the t-model is used as the base method for computing TLM. 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 Bayesian 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 the restricted likelihood with Tukey's estimator performs best in each case (assuming sufficient trimming). Huber's version also tops the thick tailed model for $n = 2000$. The Bayesian restricted likelihood 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.

6.3 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 (see rows 1 and 3 of the table). 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 0.08 as the training sample size changes from 1000 to 2000 (see rows 2 and 4 of the table). 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 (rows 1 and 2). 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 $n = 2000$ (rows 3 and 4), 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 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.

	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 Bayesian restricted likelihood non-hierarchical and hierarchical models for $n = 1000$ and 2000.

7 Discussion

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 presented here, 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. Computation 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. For example, the computational strategies we devised here allow us to apply the method to inference

on quantiles of a regression model. 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. In our words, to provide desireable inference, we recommend use of robust and relevant summary statistics in conjunction with Bayesian models.

8 Appendix

8.1 Proofs

Proof of Theorem 4.1.

Proof.

$$
\begin{aligned}
s(X, \boldsymbol{y}) &= s\left(X, \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)}\boldsymbol{z}^* + X\left(\boldsymbol{b}(X, \boldsymbol{y}_{obs}) - \boldsymbol{b}(X, \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)}\boldsymbol{z}^*)\right)\right) && (21)\\
&= \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)} s(X, \boldsymbol{z}^*) = s(X, \boldsymbol{y}_{obs}), \qquad \text{and} && (22)\\
\boldsymbol{b}(X, \boldsymbol{y}) &= \boldsymbol{b}\left(X, \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)}\boldsymbol{z}^* + X\left(\boldsymbol{b}(X, \boldsymbol{y}_{obs}) - \boldsymbol{b}(X, \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)}\boldsymbol{z}^*)\right)\right) && (23)\\
&= \boldsymbol{b}(X, \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)}\boldsymbol{z}^*) + \boldsymbol{b}(X, \boldsymbol{y}_{obs}) - \boldsymbol{b}(X, \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)}\boldsymbol{z}^*) && (24)\\
&= \boldsymbol{b}(X, \boldsymbol{y}_{obs}) && (25)
\end{aligned}
$$

□

Proof of Lemma 4.2.

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

$$
s(X, \boldsymbol{y}) = s(X, Q\boldsymbol{y} + H\boldsymbol{y}) = s(X, Q\boldsymbol{y}). \tag{26}
$$

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 (26), 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 $\mathcal{C}^{\perp}(X)$. $\square$

Proof of Lemma 4.3

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

$$[XX^{\top}, WW^{\top}]^{\top}[\nabla \boldsymbol{b}(X, \boldsymbol{y}), \nabla s(X, \boldsymbol{y})] = \begin{pmatrix} X & \mathbf{0} \\ WW^{\top}\nabla b(X, \boldsymbol{y}) & \nabla s(X, \boldsymbol{y}) \end{pmatrix} \tag{27}$$

The last column comes from Lemma 4.2. 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 (27). This last matrix has rank $p+1$ since $\nabla s(X, \boldsymbol{y}) \neq \mathbf{0}$ by C8, and so does $[\nabla b(X, \boldsymbol{y}), \nabla s(X, \boldsymbol{y})]$. $\square$

Proof of Lemma 4.4

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 4.7

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 8.1. *(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 8.2. *(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 8.1 and Theorem 8.2. 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 8.2, 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 8.1, the non-unit singular values of $W^\top A$ are the same as the non-unit singular values of $U^\top B$. □

8.2 Setting the hierarchical prior values

In setting the priors we use the same previous data set used to set the priors for the non-hierarchical model (Section 6.1) and several heuristic arguments. While the analyses in Section 6.2 set the hyper-parameters using what is described here, the results were not sensitive to these choices. This section describes the heuristics used in setting these prior parameters and is given for completeness. Using the previous data set we fit separate (robust) regressions to each state and a regression to the ~~entire~~ entirety of the data at once. Let the estimates for the fits to each state be $\hat{\beta}_1, \ldots, \hat{\beta}_J, \hat{\sigma}_1, \ldots, \hat{\sigma}_J$ and the estimates from the single regression be $\hat{\beta}$ and $\hat{\sigma}$. These are classical robust estimates using Tukey's regression and Huber's scale. Let n_j denote the number of observations in the j^{th} state and set $n = \sum n_j$.

First, consider v_1 and v_2 in the prior $b \sim \text{beta}(v_1, v_2)$. In the hierarchical model (20), $b = 0$ implies all the $\boldsymbol{\beta}_j's$ are equal (no variation between states) and $b = 1$ implies the $\boldsymbol{\beta}_j's$ vary about μ_0 according to $\Sigma_0 = n \cdot \text{var}(\hat{\beta})$ (see Section 6.1). We seek a prior measure for what we think b should be. In other words, how much prior uncertainty should we allow in $\boldsymbol{\beta}$ as opposed to the uncertainty amongst the $\boldsymbol{\beta}_j's$? Using the prior fit, a measure for uncertainty for $\boldsymbol{\beta}$ is $\Sigma_{\hat{\beta}} = \text{var}(\hat{\beta})$, the estimate of the covariance from the single regression. For the $\boldsymbol{\beta}_j's$, take $\delta_j = \hat{\beta}_j - \hat{\beta}$ and set the prior uncertainty to $\Sigma_\delta = n^{-1} \sum n_j \delta_j \delta_j^\top$. Consider the value $g = \left(|\Sigma_\delta| / |\Sigma_{\hat{\beta}}| \right)^{1/p}$. Heuristically, g is measure of the amount of uncertainty between the $\boldsymbol{\beta}_j's$ to the amount of uncertainty in $\boldsymbol{\beta}$. Now in the prior, we heuristically set the uncertainty in the $\boldsymbol{\beta}_j's$ ($b\Sigma_0$) to be approximately equal to $g \cdot \text{var}(\hat{\beta})$. That is, $b\Sigma_0 \approx g \cdot \text{var}(\hat{\beta}) = \frac{g}{n}\Sigma_0$, suggesting $b \approx \frac{g}{n}$. Hence we set $E[b] = \frac{g}{n}$. The precision, $v_1 + v_2$, is set to be relatively high at 20, completing the specification for the prior on b.

In setting the parameters for the beta prior on μ_ρ and gamma prior on ψ_ρ we first take $\hat{z}_j = \Phi^{-1}(H(\hat{\sigma}_j^2))$. As in the prior we assume $(\hat{z}_1, \ldots, \hat{z}_J) \sim N_J(\mathbf{0}, \Sigma_\rho)$ with

$\Sigma_\rho = (1-\rho)I + \rho\mathbf{1}\mathbf{1}^\top$ and find the MLE, $\hat{\rho}_{mle}$, and observed inverse Fisher information, $I^{-1}(\rho_{mle})$. The mean of the beta prior on μ_ρ is set to $\hat{\rho}_{mle}$. Its variance is inflated somewhat and set to $2I^{-1}(\hat{\rho}_{mle})$. Since $\text{var}(\rho|\mu_\rho, \psi_\rho) = \mu_\rho(1-\mu_p)/(\psi_\rho+1)$ we replace μ_ρ with $\hat{\rho}_{mle}$, $\text{var}(\rho|\mu_\rho, \psi_\rho)$ with $2I^{-1}(\hat{\rho}_{mle})$, and set the mean of the gamma prior on ψ_ρ equal to $\hat{\rho}_{mle}(1-\hat{\rho}_{mle})/(2I^{-1}(\hat{\rho}_{mle})) - 1$. Finally, we arbitrarily set the variance of the gamma prior to be approximately the same as the mean.

References

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

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

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.

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.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57:97–109.

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

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.

Joyce, P. and Marjoram, P. (2008). Approximately sufficient statistics and Bayesian computation. Statistical Applications in Genetics and Molecular Biology, 7(1).

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.

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.

Lee, J. and MacEachern, S. N. (2014). Inference functions in high dimensional bayesian inference. Statistics and Its Interface, 7(4):477–486.

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'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.

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.

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

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, page 233.

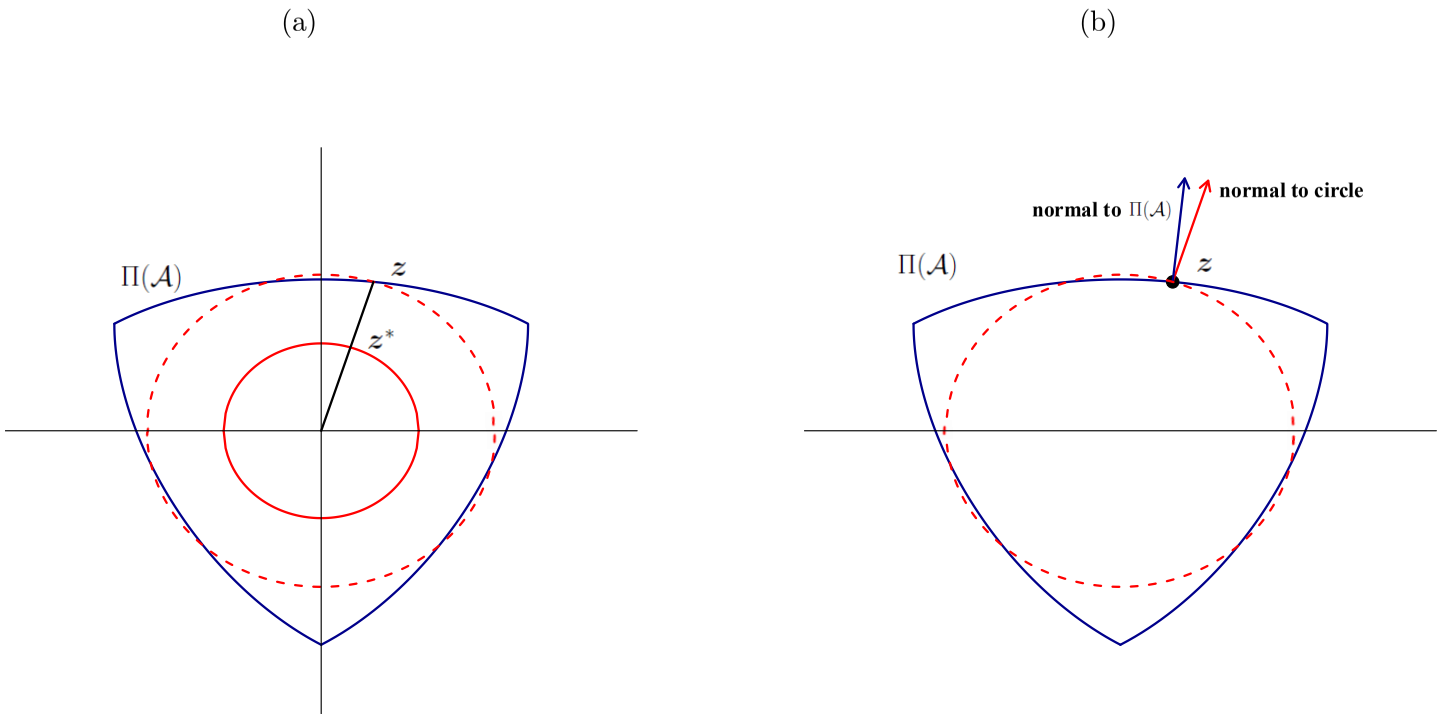
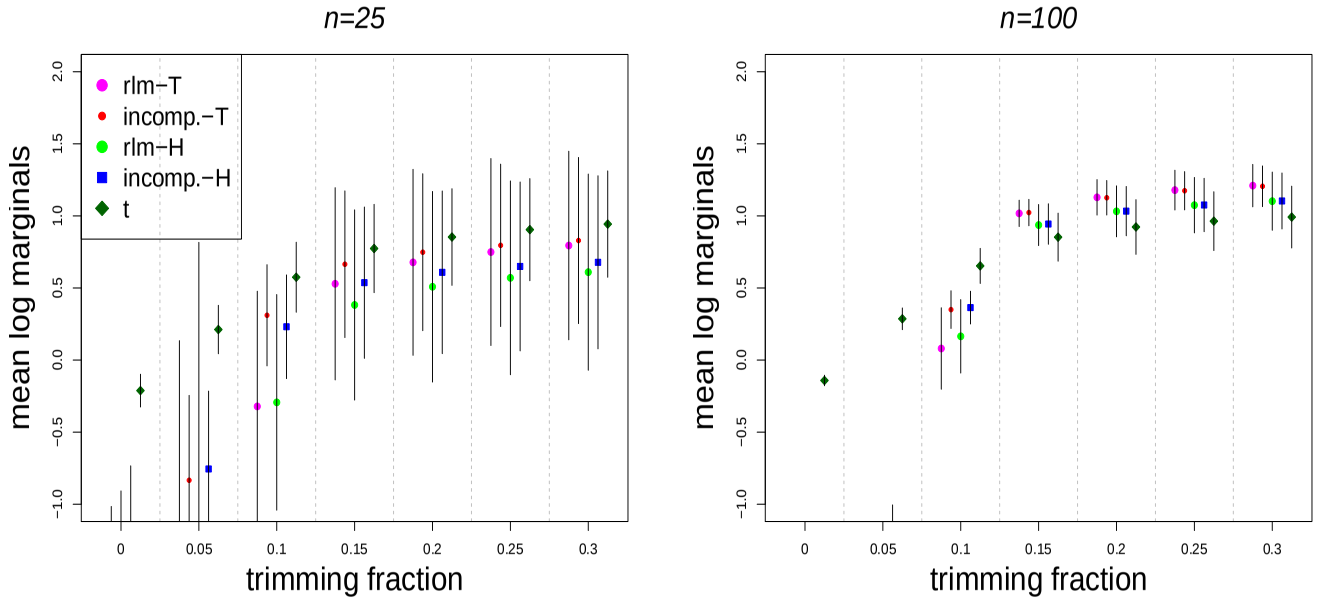


Figure 4: Panel (a) contains a depiction of the stretch from $\boldsymbol{z}^*$ to $\boldsymbol{z}$. The adjustment for the stretch transforms the density along the unit circle to the density along the circle of radius $\|\boldsymbol{z}\|$ (dashed circle). Panel (b) contains a depiction of the deformation 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.



(c)

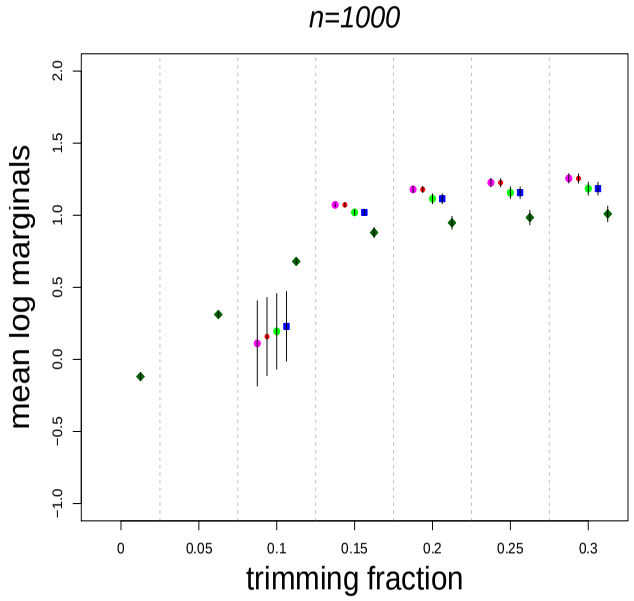


Figure 7: Model evaluation for 'Type 1' agencies for training sample sizes of $n = 25, 100$, and 1000. The t-model is used as the base method to compute TLM. Plotted are the mean TLM for each model against the trimming fraction across the 100 cross-validation samples. Error bars correspond to one standard deviation of TLM above and below the mean. Models are labeled with the following abbreviations: 'rlm' corresponds to a classical robust fit, 'restr.' corresponds to our restricted likelihood method, and 't' corresponds to the heavy-tailed t-distribution model. The letters 'T' and 'H' appearing after 'rlm' and 'restr.' correspond to the use of Tukey's and Huber's ψ respectively.

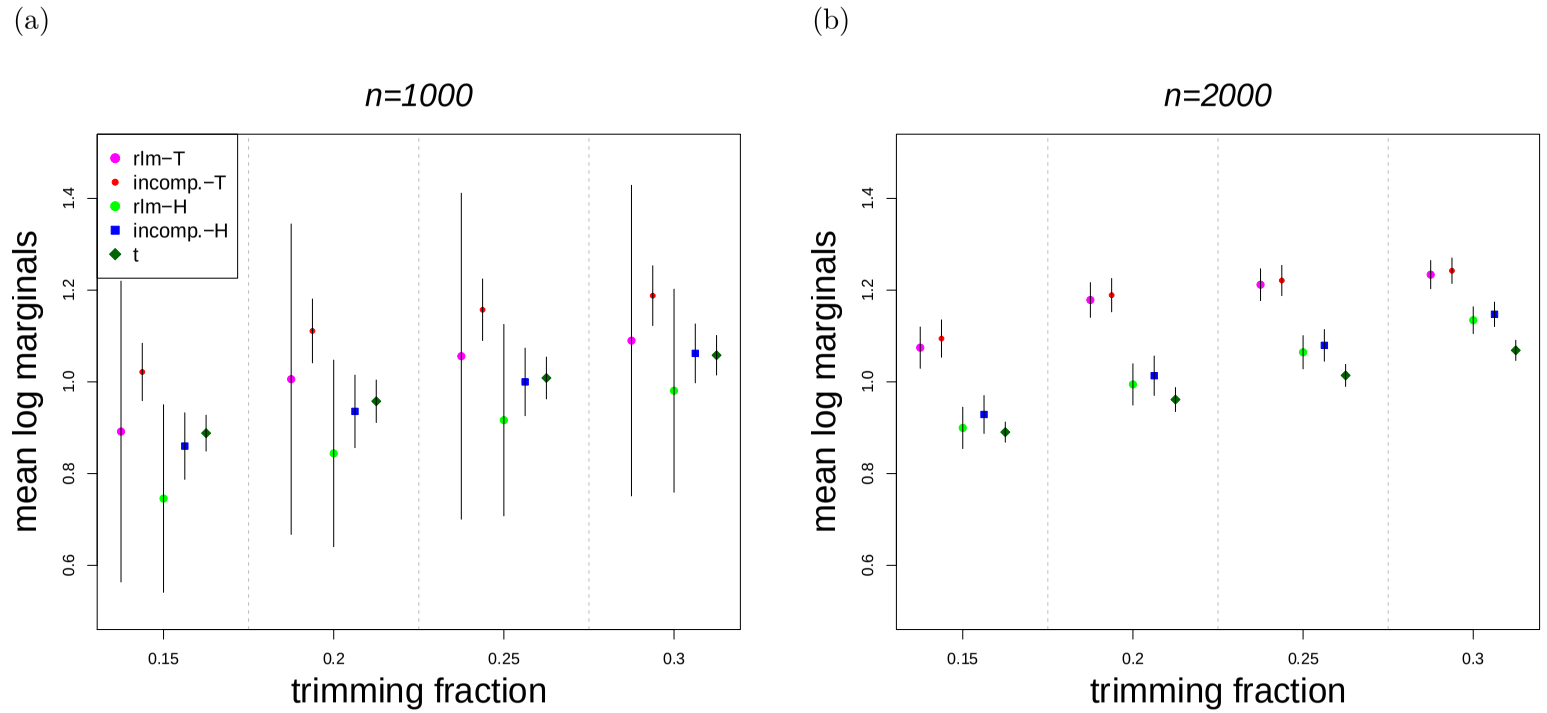


Figure 8: Model evaluation for 'Type 1' agencies under the hierarchical model for $n = 1000$ and 2000. The t-model is used as the base method to compute TLM. Plotted are the mean TLM for each model against the trimming fraction across the 100 cross-validation samples. Error bars correspond to one standard 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.