Bayesian Analysis (0000) **00**, Number 0, pp. 1

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

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

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

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

*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.

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

 DOI: 0000

1 Introduction

Bayesian methods have provided successful solutions to a wide range of scientific problems, with their value having been demonstrated both empirically and theoretically. Bayesian inference relies on a model consisting of three elements: the prior distribution, the loss function, and the likelihood or sampling density. While formal optimality of Bayesian methods is unquestioned if one accepts the validity of all three of these elements, a healthy skepticism encourages us to question each of them. Concern about the prior distribution has been addressed through the development of techniques for subjective elicitation (Garthwaite et al., 2005; O'Hagan et al., 2006) and objective Bayesian methods (Berger, 2006). Concern about the loss function is reflected in, for example, the extensive literature on Bayesian hypothesis tests (Kass and Raftery, 1995). The focus of this work is the development of techniques to handle imperfections in the likelihood $f(\boldsymbol{y}|\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\boldsymbol{y})$. Concern for imperfections in the likelihood are reflected in work considering minimally informative likelihoods (Yuan and Clarke, 1999), sensitivities of inferences to perturbations in the model (Zhu et al., 2011), the specification of a class of models and the use of Bayesian model averaging over the class (Clyde and George, 2004), and considerations of such averaging when the specified class may not contain the so-called true data generating model (Bernardo and Smith, 2000; Clyde and Iversen, 2013; Clarke et al., 2013). In practice, the imperfections in a proposed likelihood 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 good component of the sampling density is used for prediction of future good data. The second approach replaces the basic sampling density with a thick-tailed density in an attempt to discount outliers, yielding techniques that often provide solid estimates of the center of the distribution but do not easily translate to predictive densities for further good data. The third approach fits a flexible (typically nonparametric) model to the data, producing a Bayesian version of a density estimate for both good and bad data. In recent development, inference is made through the use of robust inference functions (Lee and MacEachern, 2014).

These traditional strategies all have their drawbacks. The outlier-generating processes may be

transitory in nature, constantly shifting as the source of bad data changes. This prevents us from appealing to large-sample arguments to claim that, with enough data, we can nail down a model for both good and bad data combined. Instead of attempting to model both good and bad data, we propose a novel strategy for handling outliers. In a nutshell, we begin with a complete model as if all of the data are good. Rather than driving the move from prior to posterior by the full likelihood, we use only the likelihood driven by a few summary statistics which typically target inferential quantities of interest. We call this likelihood a restricted likelihood because conditioning is done on a restricted set of data; the set which satisfies the observed summary statistics. This restricted likelihood leads to a formal update of the prior distribution based on the sampling density of the summary statistics.

The advantages and disadvantages of the method are detailed throughout the paper using simulated data and real data. One conceptual advantage is that inferences and predictions are less sensitive to features of the data not captured by the conditioning statistics. Choosing statistics targeting main features of interest allows for more targeted inference on these features. The analysis can help to better understand other features, such as outliers, not capture by the conditioning statistics. The examples in the paper suggest advantages in situations where outliers are concern and there is significant prior information for the non-outlying portion of the model that is not outweighed by the data. The main disadvantage over traditional robust estimate techniques are mainly computational. In Section 3 we detail a data-augmented MCMC algorithm to fit the models proposed in this paper. This requires an additional computational step for each iteration of the chain. Details of the additional burden are given and one must weigh the advantages of these methods with this additional burden. Since it is typically that the restricted likelihood posterior converges to the same expected value as the conditioning statistics as the sample size grows, it may not be practical to implement our method in all situations.

The remainder of the paper is as follows: Section 2 introduces the Bayesian restricted likelihood, provides context with previous work, and demonstrates some advantages of the methods on simple examples. Section 3 details an MCMC algorithm to apply the method to Bayesian linear models. This computational strategy is a major contribution to the work, providing an approach to apply

the method on realistic examples. Many of the the technical proofs are in the Appendix 7 with `R` code available from the authors. Sections 4 and 5 illustrate the method with simulated data and a real insurance industry data set containing many outliers with a novel twist on model evaluation. A discussion (Section 6) provides some final commentary on the new method. An R package `brlm` is to implement our methods in available at `github.com/jrlewi/brlm`. Additionally all data and code for the examples in this paper are available at `https://github.com/jrlewi/brlm_paper/revision_2`.

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:

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

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

The second example involves deliberate censoring of small and large observations. This is sometimes done as a precursor to the analysis of reaction time experiments (e.g., Ratcliff, 1993) where very small and large reaction times are physiologically implausible; explained by either anticipation

or lack of attention of the subject. With lower and upper censoring times at t_1 and t_2, the post-censoring sampling distribution is of mixed form, with masses $F(t_1|\boldsymbol{\theta})$ at t_1 and $1-F(t_2|\boldsymbol{\theta})$ at t_2, and density $f(y|\boldsymbol{\theta})$ for $y \in (t_1, t_2)$. We adjust the original data y_i, producing $c(y_i)$ by defining $c(y_i) = t_1$ if $y_i \leq t_1$, $c(y_i) = t_2$ if $y_i \geq t_2$, and $c(y_i) = y_i$ otherwise. The adjusted update is performed with $L(\boldsymbol{\theta}|c(\boldsymbol{y}))$. Letting $g(t_1|\boldsymbol{\theta}) = F(t_1|\boldsymbol{\theta})$, $g(t_2|\boldsymbol{\theta}) = 1 - F(t_2|\boldsymbol{\theta})$, and $g(y|\boldsymbol{\theta}) = f(y|\boldsymbol{\theta})$ for $y \in (t_1, t_2)$, we may rewrite the full likelihood as the product of two pieces

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

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

2.2 Generalization

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

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

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

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

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

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

the predictive density

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

2.3 Literature review

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

Restricted likelihoods have also been used as practical approximations to a full likelihood. For example, Pratt (1965) appeals to heuristic arguments regarding approximate sufficiency to justify the use of the restricted likelihood of the sample mean and standard deviation. Approximate sufficiency is also appealed to in the use of Approximate Bayesian Computation (ABC), which is related to our method. ABC is a collection of posterior approximation methods which has recently experienced success in applications to epidemiology, genetics, and quality control (see, for example, Tavaré et al., 1997; Pritchard et al., 1999; 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 $\mathcal{B}(\boldsymbol{y}) = \{\boldsymbol{y}^*|\rho(T(\boldsymbol{y}), T(\boldsymbol{y}^*)) \leq \epsilon\}$, ρ is a metric, and ϵ is a tolerance level. This is the likelihood conditioned on the collection of data sets that 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 an insufficient statistic $T(\boldsymbol{y})$ guided by targeted inference is sound practice. We also expand the class of conditioning statistics for which a formal Bayesian update can be achieved. Our methods do not rely on asymptotic properties, nor do they rely on approximate conditioning.

2.4 Illustrative Examples

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

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

where y_i denotes the i^{th} (recorded) measurement of the passage time of light. β is interpreted as the passage time of light with the deviations $y_i - \beta$ representing measurement error. Four versions of the restricted likelihood are fit with conditioning statistics: 1) Huber's M-estimator for location with Huber's 'proposal 2' for scale 2) Tukey's M-estimator for location with Huber's 'proposal 2' for scale 3) LMS (least median squares) for location with associated estimator of scale and 4) LTS (least trimmed squares) for location with associated estimator of scale. Details of these estimators can be found in many places, including (Huber and Ronchetti, 2009). We return to the two M-estimators throughout this paper as we have found them to offer good default choices for practitioners

dealing with outliers. A short review of these estimators is provided in the Supplementary Material. The tuning parameters for the M-estimators are chosen to achieve 95% efficiency under normality (Huber and Ronchetti, 2009) and, for comparability, roughly 5% of the residuals are trimmed for LTS. Two additional approaches to outlier handling are considered: 1) the normal distribution is replaced with a t-distribution and, 2) the normal distribution is replaced with a mixture of two normals. The t-model assumes $y_i \overset{iid}{\sim} t_\nu(\beta, \sigma^2)$ with $\nu = 5$. The prior on σ^2 is $IG(5, \frac{\nu-2}{\nu}10)$ and ensures that the prior on the variance is the same as the other models. The mixture takes the form: $y_i \overset{iid}{\sim} pN(\beta, \sigma^2) + (1-p)N(\beta, 10\sigma^2)$ with the prior $p \sim \text{beta}(20, 1)$ on the probability of belonging to the 'good' component.

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

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

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

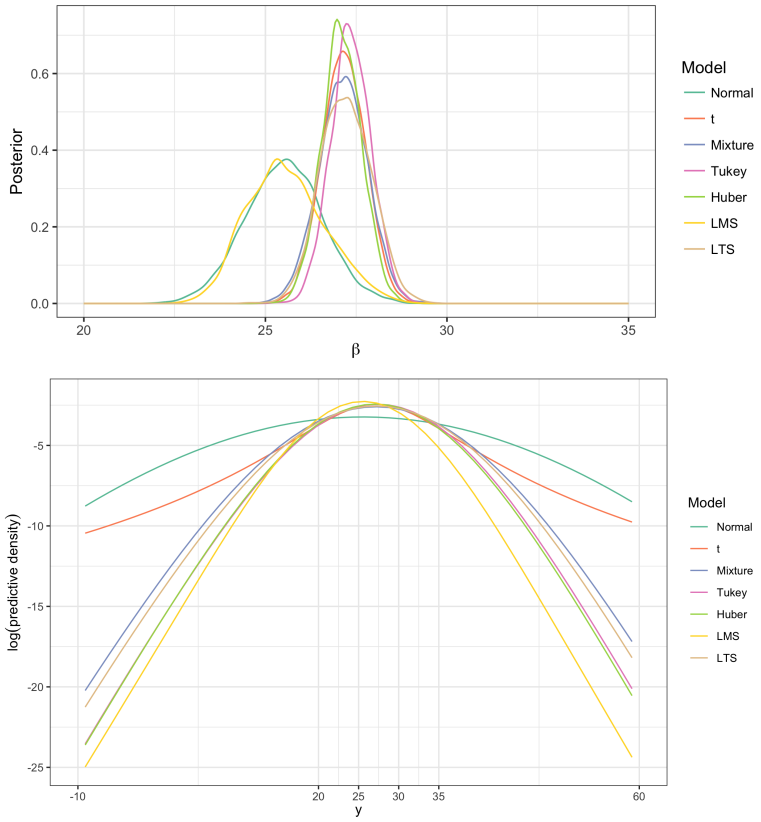

Figure 1: Results from the analysis of the speed of light data. Top: Posterior distributions of β under each model. Bottom: Log posterior predictive distributions under each model. The differences in the tails are emphasized in the bottom plot. The horizontal axis is strategically labeled to help compare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

where $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$, $\boldsymbol{y}$ is the vector of the logarithm of the number of calls, and X is the $n \times 2$ design matrix with a vector of 1's in the first column and the year covariate in the second. In reality, the model should include a different piece for the part of the data with different units. The outliers are really just a manifestation of model misspecification. Prior parameters are fixed via a maximum likelihood fit to the first 3 data points. In particular, the prior covariance for $\boldsymbol{\beta}$ is set to $\Sigma_0 = g\sigma_0^2(X_p^\top X_p)^{-1}$, with X_p the 3×2 design matrix for the first 3 data points, $g = n = 21$, $\sigma_0 = 0.03$ and $\boldsymbol{\mu}_0 = (1.87, 0.03)^\top$. This has the spirit of a unit information prior (Kass and Wasserman, 1995) but uses a design matrix for data not used in the fit. Finally $a = 2$ and $b = 1$.

Four models are compared: 1) the normal theory base model 2) a two component normal mixture model, 3) a t-model, and 4) a restricted likelihood model conditioning on Tukey's M-estimator for the slope and intercept with Huber's 'proposal 2' for scale. Each model is fit to the remaining 21 data points. The normal theory model is also fit a second time after removing observations 14-21 (years 1963 - 1970). The omitted cases consist of the obvious large outliers as well as the two smaller outliers at the beginning and end of this sequence of points caused by the change in measurement units. The mixture model allows different mean regression functions and variances for each component. Both components have the same, relatively vague priors. The probability of belonging to the first component is given a beta$(5, 1)$ prior. The heavy-tailed model fixes the degrees of freedom at 5 and uses the same prior on $\boldsymbol{\beta}$. The prior on σ^2 is adjusted by a scale factor of $3/5$ to provide the same prior on the variance.

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

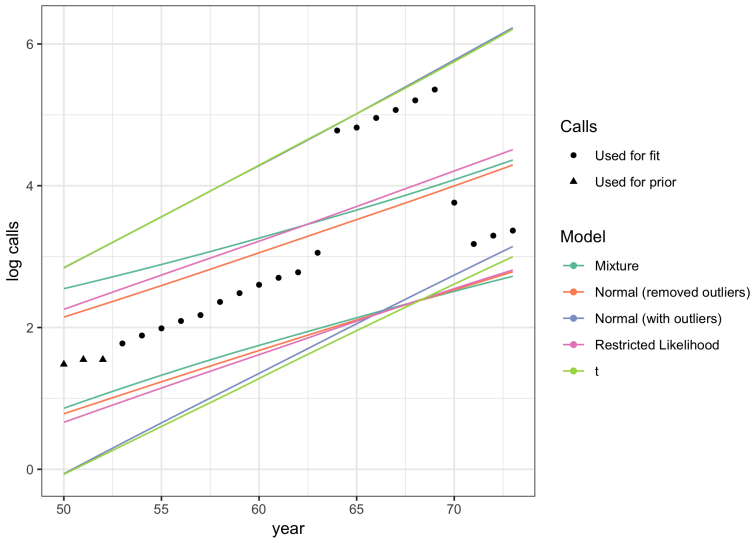

Figure 2: Pointwise posterior predictive intervals of log(calls) under the normal theory model fit to the non-outliers, the restricted likelihood model with Tukey's M-estimator for the slope and intercept with Huber's 'proposal 2' for scale, and a heavy-tailed t-distribution model. The first three data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

3 Restricted Likelihood for the Linear Model

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

3.1 The Bayesian linear model

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

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

where x_i and $\boldsymbol{\beta} \in \mathbb{R}^p$, $\sigma^2 \in \mathbb{R}^+$, and the 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 several traditional simultaneous M-estimators (Huber and Ronchetti,

2009; Maronna et al., 2006) for which the `R` package `brlm` (`github.com/jrlewi/brlm`) is available to implement the MCMC described here. These M-estimators satisfy C3-C4 since they are optimizers of (almost surely) continuous and differentiable objective functions. Constraints C5-C8 are often satisfied by location and scale estimators but should be checked on a case by case basis. More software development is required to extend the MCMC implementation beyond the M-estimators discussed here. The current version of the `R` package also implements the direct computational methods described in Lewis (2014). These methods are effective in lower dimensional problems and were used in both examples in Section 2.4.

3.2 Computational strategy

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

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

$\boldsymbol{y}_c$. The conditional density is

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

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

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

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).

Construction of the proposal

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

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

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

Using the theorem, the general idea is to first start with an initial vector $\boldsymbol{z}^*$ drawn from a known distribution, say $p(\boldsymbol{z}^*)$, and transform via $h(\cdot)$ to $\boldsymbol{y} \in \mathcal{A}$. The proposal density $p(\boldsymbol{y}|\boldsymbol{\theta})$ is then a

change-of-variables adjustment on $p(\boldsymbol{z}^*)$ derived from $h(\cdot)$. In general however, the mapping $h(\cdot)$ is many-to-one: for any $\boldsymbol{v} \in \mathbb{R}^n$ and any $c \in \mathbb{R}^+$, $c\boldsymbol{z}^* + X\boldsymbol{v}$ map to the same $\boldsymbol{y}$. This makes the change-of-variables adjustment difficult. We handle this by first noticing that the set $\mathcal{A}$ is an $n-p-1$ dimensional space: there are p constraints imposed by the regression coefficients and one further constraint imposed by the scale. Hence, we restrict the initial $\boldsymbol{z}^*$ to an easily understood $n-p-1$ dimensional space. Specifically, this space is the unit sphere in the orthogonal complement of the column space of the design matrix: $\mathbb{S} := \{\boldsymbol{z}^* \in \mathcal{C}^\perp(X) \mid ||\boldsymbol{z}^*|| = 1\}$, where $\mathcal{C}(X)$ and $\mathcal{C}^\perp(X)$ are the column space of X and its orthogonal complement, respectively. The mapping $h : \mathbb{S} \to \mathcal{A}$ is one-to-one and onto. A proof is provided by Theorem 7.1 in the Supplementary Material. The one-to-one property makes the change of variables more feasible. The onto property is important so that the support of the proposal distribution (i.e. the range of $h(\cdot)$) contains the support of the target $f(\boldsymbol{y}|\theta, y \in \mathcal{A})$, a necessary condition for convergence of the Metroplis-Hastings algorithm (in this case the supports are both $\mathcal{A}$).

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

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

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

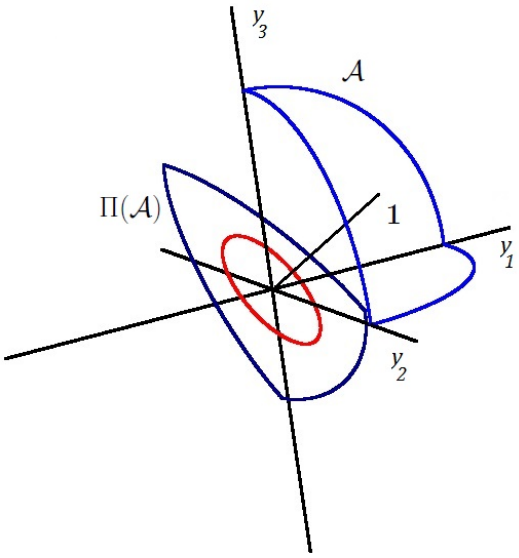

Figure 3: A depiction of $\mathcal{A}$, $\Pi(\mathcal{A})$, and the unit circle for the illustrative example where $b_1(\mathbf{1}, \boldsymbol{y}) = \min(\boldsymbol{y}) = 0$ and $s(\mathbf{1}, \boldsymbol{y}) = \sum(y_i - b_1(\mathbf{1}, \boldsymbol{y}))^2 = 1$. $\mathcal{A}$ is the combination of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

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

Evaluation of the proposal density

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

Scale from $\mathbb{S}$ to $\Pi(\mathcal{A})$

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

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

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

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

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

Shift from $\Pi(\mathcal{A})$ to $\mathcal{A}$

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

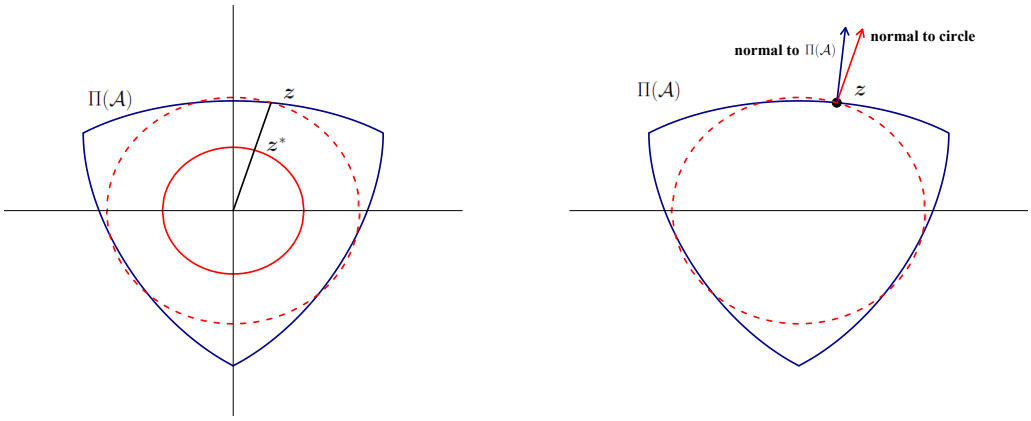

Figure 4: Visualization of the scaling from z^* to z. Left: the first substep scales z^* on the unit circle to the circle of radius $r = ||z||$, resulting in a change-of-variables transformation for the unit circle to a circle of radius r. The contribution to the Jacobian of this transformation is $r^{-(n-p-1)}$.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

is deformed by the shift. The contribution of this deformation to the Jacobian is, again, the ratio of the infinitesimal volumes along $\Pi(\mathcal{A})$ at $\boldsymbol{z}$ to the corresponding volume along $\mathcal{A}$ at $\boldsymbol{y}$. The ratio is calculated by considering the volume of the projection of a unit hypercube in the tangent space of $\mathcal{A}$ at $\boldsymbol{y}$ onto $\mathcal{C}^{\perp}(X)$. Computational details are given in the following lemmas and subsequent theorem. Proofs of the lemmas are given in the appendix and the theorem is a direct result of the lemmas. Throughout, let $\mathcal{T}_y(\mathcal{A})$ and $\mathcal{T}_y^{\perp}(\mathcal{A})$ denote the tangent space to $\mathcal{A}$ at $\boldsymbol{y}$ and its orthogonal complement, respectively.

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

The lemma describes construction of a basis for $\mathcal{T}_y^{\perp}(\mathcal{A})$, leading to a basis for $\mathcal{T}_y(\mathcal{A})$. Both of these bases can be orthonormalized. Let $A = [a_1, \ldots, a_{n-p-1}]$ and $B = [b_1, \ldots, b_{p+1}]$ denote the matrices

whose columns contain the orthonormal bases for $\mathcal{T}_y(\mathcal{A})$ and $\mathcal{T}_y^{\perp}(\mathcal{A})$, respectively. The columns in A define a unit hypercube in $\mathcal{T}_y(\mathcal{A})$ and their projections onto $\mathcal{C}^{\perp}(X)$ define a parallelepiped. We defer construction of A until later.

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

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

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

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

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

The proposal density

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

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

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

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

The proposal is governed by the choice of $p(z^*)$ and a poor choice could cause concern about the efficiency of the convergence of the MCMC algorithm. For all the examples in the paper we defined $p(z^*)$ to simply be the uniform distribution on $\mathbb{S}$. The advantage of this choice is that it requires no further tuning parameters and we have noticed good mixing in terms of the ability of the chain to

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

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

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

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

Second, the gradients of $\nabla s(X, \boldsymbol{y}), \nabla b_1(X, \boldsymbol{y}), \ldots, \nabla b_p(X, \boldsymbol{y})$ are easily computed in many cases.

For example, below we consider M-estimators defined by the estimating equations:

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

where ψ and χ are almost surely differentiable. The gradients can be found by differentiating this system of equations with respect to each y_i. In theory, finite differences could also be used as an approximation if needed.

Finally, it is clear the estimators themselves must be computed for every iteration of the Markov Chain. We have found this burden to be marginal in relation with respect to computing the needed Jacobian. In the simulations and real data analyses presented below, we will see that the additional burden is often of added value over traditional robust regression when substantial prior information is available that is not swamped by current data.

4 Simulated Data

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

4.1 Simulation 1

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

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

with $\mu = 0, \tau^2 = 1, \sigma^2 = 4$. The values of p_i, m_i, and n_i depend on the group and are formed using 5 replicates of the full factorial design over factors p_i, m_i, n_i with levels $p_i = .1, .2, .3$, $m_i = 9, 25$, and $n_i = 25, 50, 100$. This results in 90 groups that have varying levels of outlier contamination and sample size. We wish to build models that offer good prediction for the good portion of data within

each group. The full model for fitting is a corresponding normal model without contamination:

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

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

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

$K = 30$ data sets are generated from (16). For each data set and each pair (a_s, c), the Bayesian models are fit using MCMC. The MCMC for the restricted likelihood version requires no computational details other than those described for the traditional Bayesian model in Section 3. This is because there are conditioning statistics for each group and the model's conditional independence between the groups allows the data augmentation described earlier to be performed independently within each group. That is, there is a separate Gibbs step for each group to generate the group level data matching the statistics for that group. The acceptance rates for newly generated data across all groups and simulations ranged from 0.57 to 0.68

To asses the predictive capability, we seek a metric that takes into account both the estimation of θ_i and σ^2. In this model, estimation of θ_i and prediction are analogous. To this end we consider

the expected value of the log likelihood ratio between the true distribution of good data and a model estimate of this distribution

$$E[\log f(y|\theta_i, \sigma^2) - \log f(y|\hat{\theta}_i, \hat{\sigma^2})] = \log(\hat{\sigma}^2) - \log(\sigma^2) + \frac{1}{2\hat{\sigma}^2}(\sigma^2 + (\theta - \hat{\theta}_i)^2) \quad (18)$$

The expectation is taken with respect to the true data distribution of the good data assuming fixed estimates. These estimates are the posterior means for the Bayesian methods and the standard point estimates for the classical methods.

Figure 5 displays the average of loss 18 grouped by pairs of a_s and c with error bars plus/minus on standard error within the group. The values of a_s and c, do not affect the classical robust linear models. The average loss for the normal theory models ranges from 1.19 to 1.42 and are left out of the figure. For $c = 0.5$ and $c = 1$ and the two larger values of a_s the results favor the restricted likelihood methods.

The choice of $c = 2$ corresponds to a particularly poor prior distribution. The prior has substantial mass above $\sigma^2 = 4$, with prior means for σ^2 from 8.9 to 32 as a_s varies. Additionally, the tuning parameters chosen for the location and scale estimators result in an upward bias in the estimate of σ^2. This bias depends on m and p. For example, for $m = 9$ and $p = .1$, Huber's version converges to roughly 4.8 as n grows. The bias is greater for more severe levels of contamination. The alignment of biases in prior distribution and in likelihood from the summary statistic (when applied to the contaminated data) inflates the estimate of scale. Not surprisingly, a poor prior distribution whose weakness matches the weakness in the likelihood results in poorer inference. In this case, poorer than the classical estimators.

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

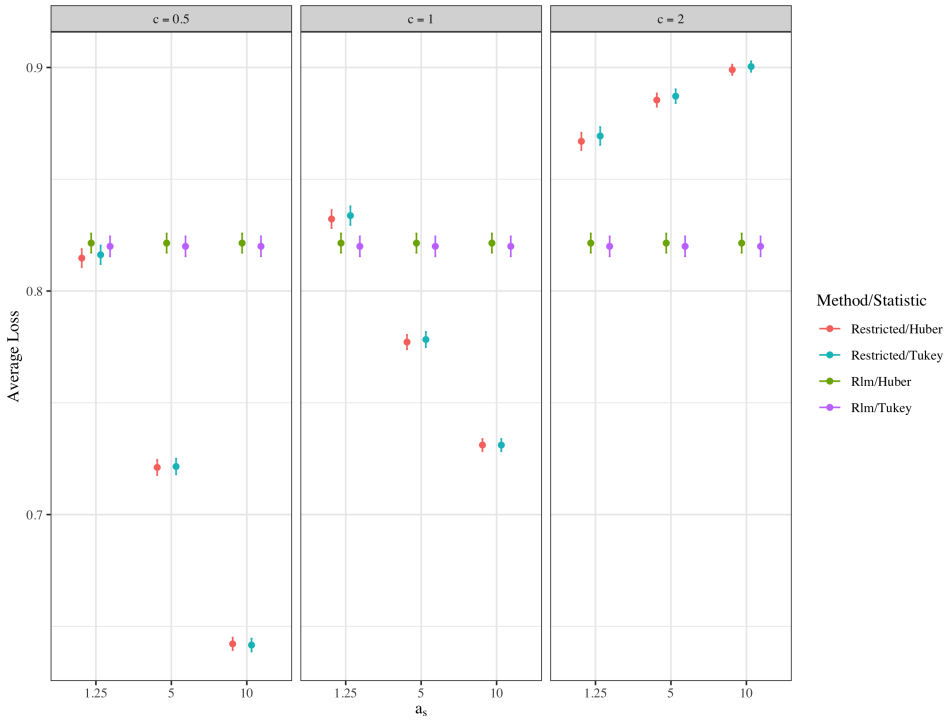

Figure 5: Average Loss plus/minus one standard error for each value of a_s and c. Smaller values represent better fits. The panels correspond to $c = 0.5$ (left), $c = 1$ (middle), and $c = 2$ (right), with the values of a_s on the horizontal axis. The average loss for the normal theory model ranges from 1.19 to 1.42 and is left out of the figure.

method is greater for smaller sample sizes.

This simulation shows the potential of the restricted likelihood and conveys some cautions. Specifically, the choice of summary statistics, along with corresponding tuning parameters is important. For the tuning parameters, we applied the default choice of 95% efficiency at the normal. Under the simulation model here, this choice results in bias in the scale estimation which affects the performance of the method. These choices must be made when using both the classical and Bayesian methods. The Bayesian approach encourages use of a hierarchical model structure and allows one to incorporate prior information in the analysis. These features can improve predictive performance substantially. If poorly handled, they can, of course, harm performance.

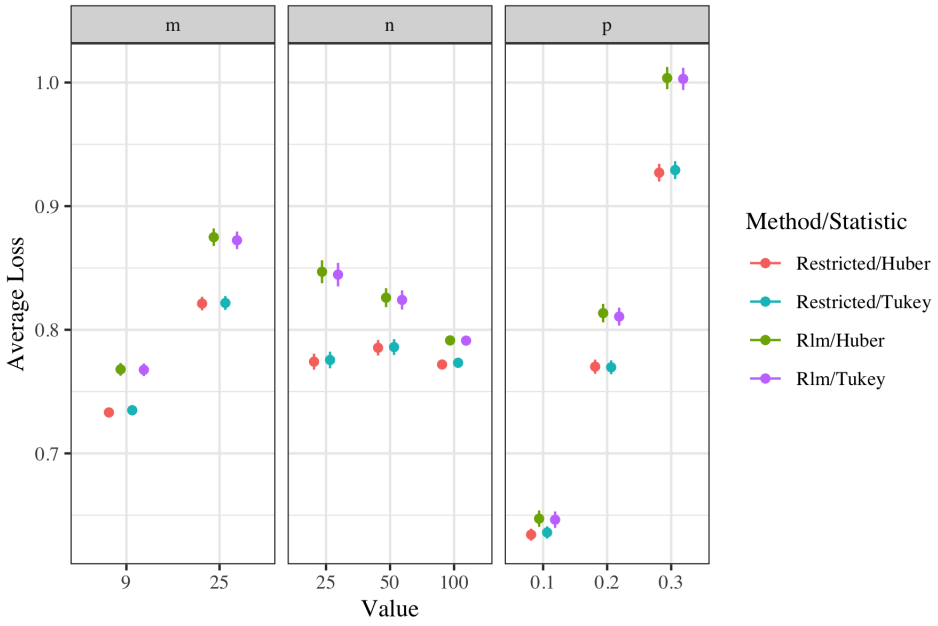

Figure 6: Average loss plus/minus one standard error grouped by the factors m (left), n (middle), and p (right). These results are for the single prior with $a_s = 5$ and $c = 1$.

4.2 Simulation 2

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

only a few govern the data generating mechanism.

For the simulation, $K = 30$ data sets (including the additional covariates) of size $n = 500$ are generated from the true model with true values $\boldsymbol{\beta} = (1, 1, 1)^\top$ and $\sigma^2 = 2$. We fit the model including all 30 covariates and consider the following methods for the fit 1) classical robust regression with Tukey's estimator of location and Huber's estimator of scale, 2) the corresponding restricted likelihood version 3) A heavy-tailed Bayesian model with a Student-t likelihood with $\nu = 5$ degrees of freedom. For the Bayesian models we take $\boldsymbol{\beta}_{all} \sim N_{20}(\mathbf{0}, \sigma^2_\beta I)$ with $\boldsymbol{\beta}_{all} = (\boldsymbol{\beta}, \boldsymbol{\beta}^*)^\top$ and $\sigma^2 \sim IG(5, 8)$ under the restricted model and $\sigma^2 \sim IG(5, \frac{\nu-2}{\nu}8)$ under the Student-t model. For each data set, we fit the models for $\sigma_\beta = 0.4, 0.6, 0.8, \ldots, 1.4$. The acceptance rates for the restricted likelihood MCMC data-augmentation step range from 0.3 to 0.36 across all the data sets and values of σ_β. To compare performance we first consider the $MSE = (||\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}||^2 + ||\hat{\boldsymbol{\beta}}^*||^2)/30$ for each simulation where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^*$ are point estimates for the fitted model. For the Bayesian models, we use posterior means. The average MSE plus/minus one standard error over the simulations for each σ^2_β are displayed in Figure 7. The classical fit is labeled 'rlm' and is the same for each value of the prior standard deviation σ^2_β. We see for most values of the the prior standard deviation, the Bayesian models ('restricted' and 't') outperform the classical fit. The correlation amongst the covariates causes a certain level of confounding and the prior shrinkage helps to improve estimation. However, too much shrinkage can be detrimental as demonstrated for $\sigma_\beta = 0.4$. While this will help for estimation of $\boldsymbol{\beta}^* = 0$, the estimation of the active parameters $\boldsymbol{\beta}$ can be hindered. The t model seems more sensitive to this effect than the restricted model. The restricted model also has an additional advantage when it comes to prediction of the non-outlying data. To see this, for each simulation we consider the mean negative log-likelihood of the non-outlying data: $MNLL = -\frac{1}{N}\sum \log f(y_i|\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}^*, \hat{\sigma})$ where f is the assumed likelihood and the average is taken over the N non-outlying points y_i. For the classical and restricted fits, f is the normal likelihood and for the 't' it is the heavy-tailed Student-t likelihood. The average MNLL plus/minus one standard error over the simulations for each σ^2_β are displayed in Figure 8. First, the restricted version has a small but consistent improvement over the classical method. Second, it is clear that that the heavy-tailed model suffers when trying to predict the non-outlying data since it assumes the entire data generating mechanism is heavy-tailed.

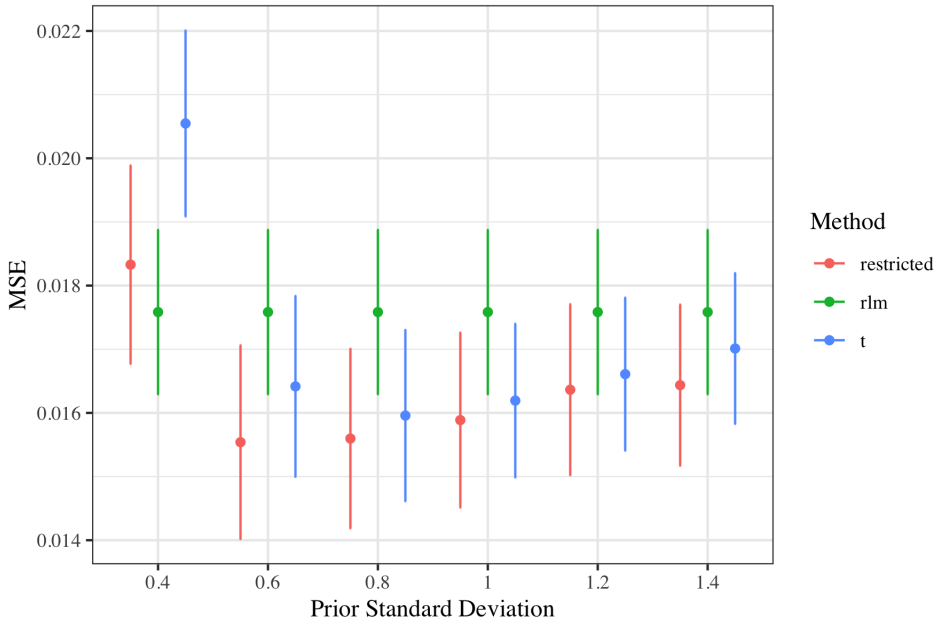

Figure 7: Average MSE plus/minus one standard error over the $K = 30$ simulations for each value of the prior standard deviation (σ_β^2) and each of the fitting methods. 'Restricted' is our method conditioning on Tukey's estimator of location and Huber's estimator of scale. 'rlm' refers to the classical robust linear model fit with the same estimators and 't' is the heavy-tailed Bayesian model with a Student-t likelihood. The 'rlm' results are the same for each σ_β^2.

5 Real Data

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

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

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

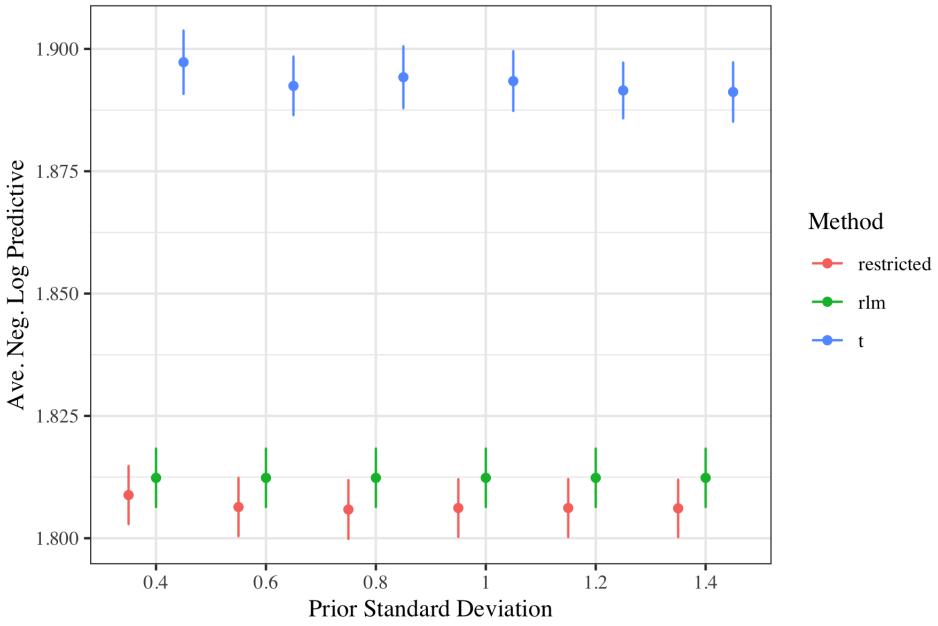

Figure 8: Average MNLL plus/minus one standard error over the $K = 30$ simulations for each value of the prior standard deviation (σ^2_β) and each of the fitting methods. 'Restricted' is our method conditioning on Tukey's estimator of location and Huber's estimator of scale. 'rlm' refers to the classical robust linear model fit with the same estimators and 't' is the heavy-tailed Bayesian model with a Student-t likelihood. The 'rlm' results are the same for each σ^2_β.

Additionally, the counts were scaled to have standard deviation one before analysis.

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

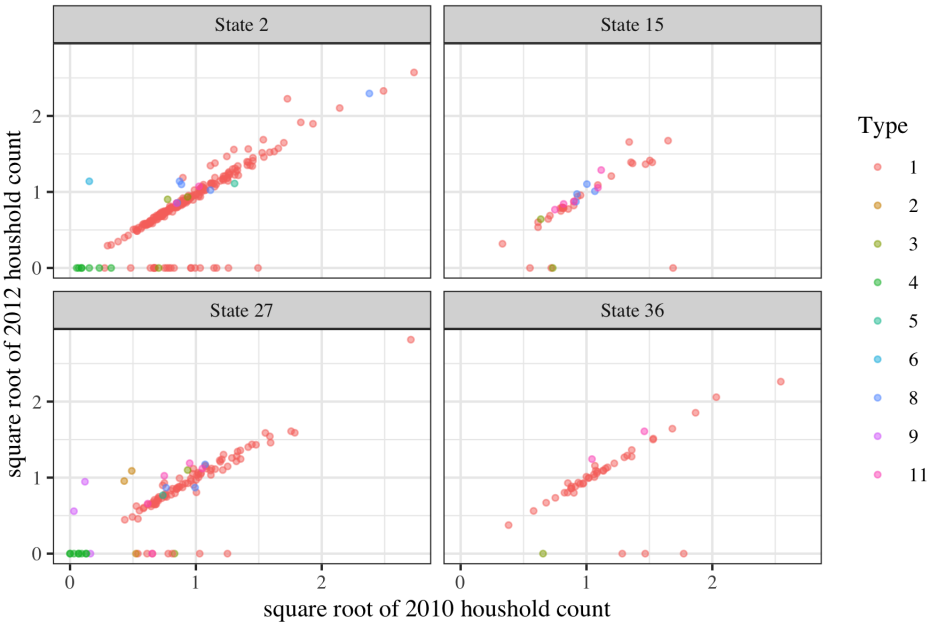

Figure 9: The square root of (scaled) count in 2012 versus that in 2010 for four states. The colors represent the varying contractual agreements as they stood in 2010 ('Type'). Agencies that closed during the 2010-2012 period are represented by the zero counts for 2012.

future performance for agencies that remain open. It is important for planning purposes that the predictions are not overly influenced by a handful of over/underperforming agencies. Our analysis focuses on one aspect of the business problem - the prediction of future performance for agencies, given they remain open.

5.1 State Level Regression model

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

$$\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_0, \Sigma_0); \quad \sigma^2 \sim IG(a_0, b_0); \quad y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \ i = 1, \ldots, n, \tag{19}$$

where $\boldsymbol{\beta}$ is a three dimensional vector ($p = 3$) of regression coefficients for the covariate vector $\boldsymbol{x}_i$ consisting of the square root of household count in 2010, and two different size/experience measures

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

We compare four Bayesian models: the standard Bayesian normal theory model, two restricted likelihood models, both with simultaneous M-estimators, and a heavy-tailed model. For the restricted likelihood methods we use the same simultaneous M-estimators as in the simulation of Section 4 adapted to linear regression. The heavy-tailed model replaces the normal sampling density in (19) with a t-distribution with $\nu = 5$ degrees of freedom. The Bayesian models are all fit using MCMC, with the restricted versions using the algorithm presented in Section 3.2. We also fit the corresponding classical robust regressions and a least squares regression.

Method of model comparison

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

The presence of numerous outliers in the data implies that both training and validation data will contain outliers. For this reason, the evaluation must be robust to a certain fraction of bad data. The two main strategies are to robustify the evaluation function (e.g., Ronchetti et al., 1997) or to retain the desired evaluation function and trim cases (Jung et al., 2014). Here, we pursue the trimming approach with log predictive density for the Bayesian models and log density from plug-in maximum likelihood for the classical fits used as the evaluation function.

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

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

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

Comparison of predictive performance

'Type 1' agencies are of special interest to the company and so the evaluation of the TLM is done on only holdout samples of 'Type 1', whereas the training is done on agencies of all types. This is intended to demonstrate the robustness properties of the various methods. Models are fit to four states labelled State 2, 15, 27, and 36, with $n = 222, 40, 117$, and 46, representing a range of sample

sizes. Fitting is done on $K = 50$ training samples with training sample sizes taken to be $0.25n$ and $0.50n$. Holdout evaluation is done on the remaining ('Type 1') samples. The acceptance rates for the data augmentation step,for all but one training set, range from 0.10 to 0.8 across the states, repetitions, and two versions of the model. The exception was a single training set from State 15 resulting in an usually small acceptance rate under Tukey's version. This case didn't effect the overall results of the simulations but emphasizes the need to check convergence on a case by case basis. The average $TLM_b(A)$ over the $K = 50$ training/holdout samples for the four states and seven methods are shown in Figure 10 where the base model is the Student-t model and $\alpha = 0.3$. Similar results are observed for other base models. The error bars are plus/minus one standard deviation of the average $TLM_b(A)$ over the $K = 50$ training/holdout samples. It is clear that the normal Bayesian model used as the full model (Normal) and the classical ordinary least squares fits (OLS) have poor performance due to the significant amount of outlier contamination in the data. In comparing our restricted methods to their corresponding classical methods, there is small, but consistent improvement across the states and training sample size. Additionally, variance reduction for the Bayesian versions is evident, especially in State 15, highlighted by the smaller error bars. For state 2, the largest state with $n = 222$, the restricted and classical robust methods have similar performance especially for larger training sample size. This reflects the diminishing effect of the prior as the sample size grows. Notably, the Student-t model performs poorly in comparison for this state. The predictive distribution explicitly accounts for heavy-tailed values, resulting in poorer predictions of the 'good' data (i.e., the Type 1 agencies). Likewise, for State 27, another larger state, the Student-t model is outperformed by our restricted methods. For the other states (State 15 and 36), the Student-t performs better to our restricted methods for smaller training sample size (25% of the sample). However, this advantage goes away for the larger training sample size (50% of the sample). Intuitively, as more data is available for fitting, more outliers appear and the heavy-tailed model compensates for them by assuming they come from the tails of the model; an assumption which is detrimental for prediction. Comparisons of the models depend on α as seen in Figure 11 which shows results for different α for training sample size $0.5n$. For smaller α (in this case $\alpha = 0.1$), many outliers are left untrimmed resulting in lower TLM for all methods and noticeably larger standard deviation for the classical robust methods and our restricted likelihood. Larger values of α

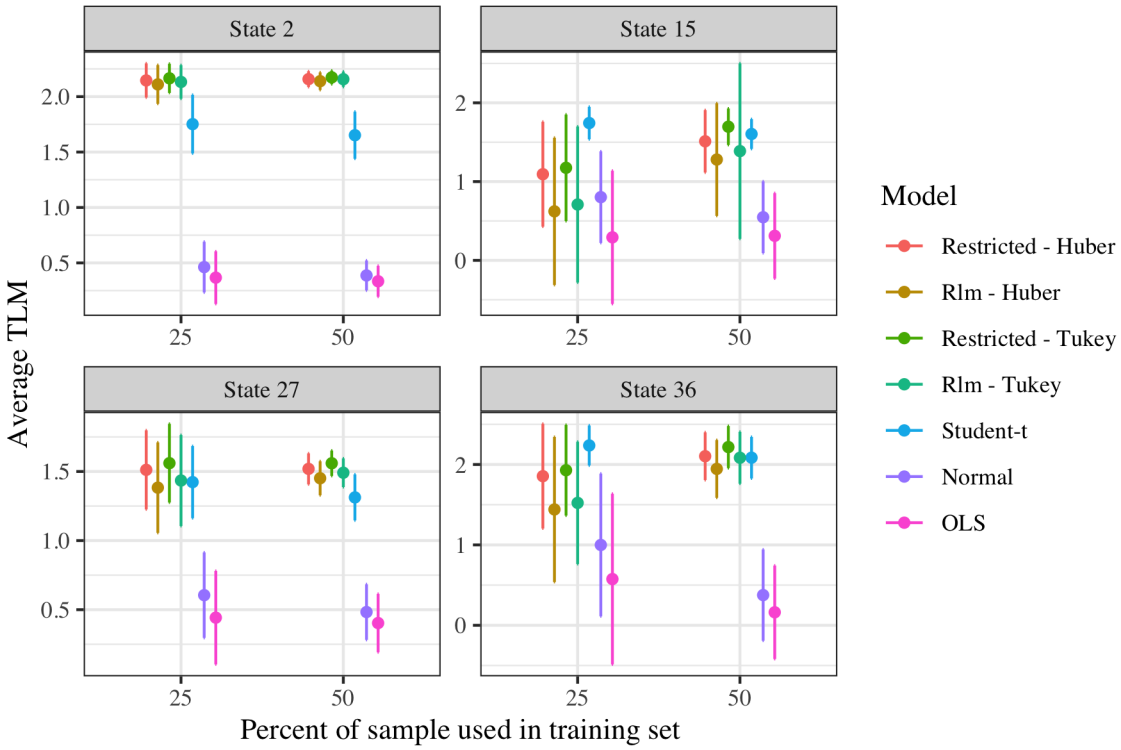

Figure 10: Average TLM plus/minus one standard deviation over $K = 50$ splits into training and holdout samples. The panels are for the different states 2, 15, 27, and 36, with $n = 222, 40, 117$, and 46, respectively. The horizontal axis is the percent of n used in each training set. The color corresponds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

ensure that the predictive performance assessment excludes the majority of outliers. The proportion of 0 counts in the data is roughly 0.14, suggesting that α should be at least this large.

5.2 Hierarchical regression model

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

$$\boldsymbol{\beta} \sim N_p(\boldsymbol{\mu}_0, a\Sigma_0); \quad \boldsymbol{\beta}_j \overset{iid}{\sim} N_p(\boldsymbol{\beta}, b\Sigma_0); \quad \sigma_j^2 \sim IG(a_0, b_0); \tag{20}$$

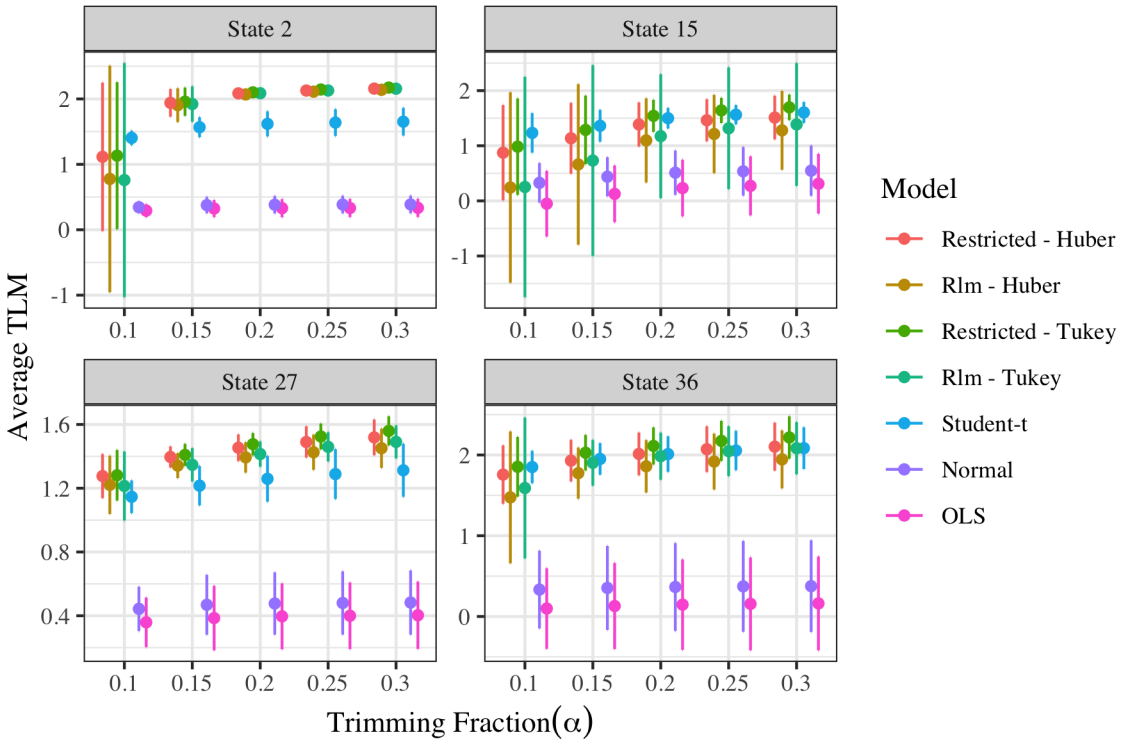

Figure 11: Average TLM plus/minus one standard deviation over $K = 50$ splits into training and holdout samples for several values of the trimming fraction α. The training sample size used is $0.5n$. Larger values of TLM are better.

$$y_{ij} = \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta}_j + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_j^2), \; i = 1, \ldots, n_j, \; j = 1, \ldots, J \tag{21}$$

where y_{ij} is the i^{th} observation of square rooted household count in 2012 in the j^{th} state, n_j is the total number of agencies in state j, and J is the number of states. $\boldsymbol{x}_{ij}$ is same tree-dimensions covariate vector as before and $\boldsymbol{\beta}_j$ represents the individual regression coefficient vector for state j. The parameters $\boldsymbol{\mu}_0$, Σ_0, a_0, and b_0 are fixed by fitting the regression $y_{ij} = \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta} + \epsilon_{ij}$ using Huber's M-estimators to the prior data set from two years before. Using the estimates from this model, we set $\mu_0 = \hat{\beta}$, $\Sigma_0 = n_p\hat{\Sigma_0}$ ($n_p = 2996$ is the number of observations in the prior data set), $a_0 = 5$ and $b_0 = \hat{\sigma}^2(a_0 - 1)$. We constrain $a + b = 1$ in an attempt to partition the total variance between the individual $\boldsymbol{\beta}_j$'s and the overall $\boldsymbol{\beta}$ and take $b \sim \text{beta}(v_1, v_2)$. Using the prior data set, we assess the

variation between individual estimates of the $\boldsymbol{\beta}_j$ to set v_1 and v_2 to allow for a reasonable amount of shrinkage. To allow for dependence across the σ_j^2 we first take $(z_1, \ldots, z_J) \sim N_J(\mathbf{0}, \Sigma_\rho)$ with $\Sigma_\rho = (1-\rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}^\top$. Then we set $\sigma_j^2 = H^{-1}(\Phi(z_j))$ where H is the cdf of an $IG(a_0, b_0)$ and 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, with fixed hyperparameters. More details on setting prior parameters are given in the Supplementary Material.

Using the same techniques as in the previous section, we fit the normal theory hierarchical model above, a thick-tailed t version with $\nu = 5$ d.f., and two restricted likelihood versions (Huber's and Tukey's) of the model. For the restricted methods, we condition on robust regression estimates fit separately within each state. We also fit classical robust regression counterparts and a least squares regression separately within each state. Hierarchical models naturally require more data and so we include states having at least 25 agencies with sufficient variation within each covariate, resulting in 20 states in total and $n = \sum_j n_j = 3094$ total agencies. For training data we take a stratified (by state) sample of size $3094/2 = 1547$ where the strata sizes are $n_j/2$ (rounded to the nearest integer). The remaining data is used for a holdout evaluation using TLM computed separately within each state: $TLM_b(A)_j = (M_j - [\alpha M_j])^{-1} \sum_{i=[\alpha M_j]+1}^{M_j} \log(f_A(y^b_{(i)j}))$ where $y^b_{(1)j}, y^b_{(2)j}, ..., y^b_{(M_j)j}$ is the ordering of the M_j holdout observations within state j according to the log marginals under the base model b. For the non-Bayesian models, $f_A(y^b_{(i)j})$ is estimated using plug-in estimators for the parameters for state j. $TLM_b(A)_j$ is computed for each state for $K = 50$ splits of training and holdout sets. The Bayesian models are fit using MCMC, with the restricted versions applying the algorithm laid out in Section 3 and adapted to the hierarchical setting as described in Section 4. For the MH-step proposing augmented data, the acceptance rates for the two restricted likelihood models across all states and repetitions ranged from 0.01 to 0.75, with only 7 cases (out of 50*20*2 = 2000 chains) with rates below 0.1

The average over states, $\overline{TLM}_b(A). = \frac{1}{22}\sum_{j=1}^{22} TLM_b(A)_j$ for each of the K repetitions is summarized in Figure 12 for several trimming fractions using the Student-t as the base model. The points are the average of the $\overline{TLM}_b(A).$ over the K repetitions with error bars plus/minus one standard

deviation over K with larger values representing better predictive performance. As the trimming fraction used for the TLM increases, so does TLM since more outliers are being trimmed. Similar patterns were seen in the individual state level regressions in Section 5.1. Despite being used as the base model to compute TLM, the Student-t doesn't perform well in comparison to the robust regressions. We attribute this to the assumption of heavier tails resulting in smaller log marginal values on average; emphasizing again that the t-model will do well to discount outlying observations but does not provide a natural mechanism for predicting non-outlying data. For each trimming fraction, our restricted likelihood hierarchical models outperform the classical robust regressions fit separately within each state. The hierarchical model also reduces variance in predictions resulting in smaller error bars.

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

6 Discussion

This paper develops a Bayesian version of restricted likelihood where posterior inference is conducted by conditioning on a summary statistic rather than the complete data. The framework blends classi-

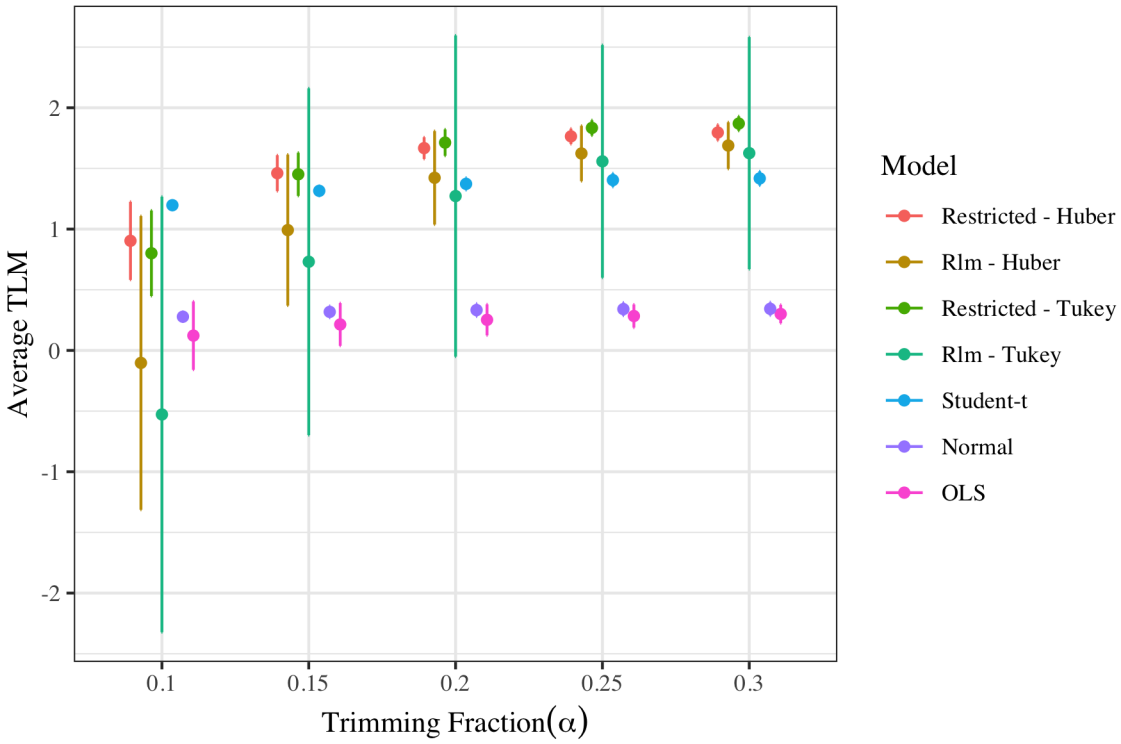

Figure 12: Hierarchical model results: $\overline{TLM}_b(A)$. plus/minus one standard deviation over $K = 50$ splits into training and holdout sets with the Student-t as the base model and several values of the trimming fraction α. Larger values of TLM are better.

cal estimation with Bayesian methods. Here, we concentrate on outlier-prone settings where natural choices for the conditioning statistic are classical robust estimators targeting the mean of the non-outlying data (e.g., M-estimators). The likelihood conditioned on these estimators is used to move from prior to posterior. The update follows Bayes' Theorem, conditioning on the observed estimators exactly. Computation is driven by MCMC methods, requiring only a supplement to existing algorithms by adding a Gibbs step to sample from the space of data sets satisfying the observed statistic. This step has additional computation costs arising from the need to compute the estimator and an orthonormal basis derived from gradients of the estimator at each iteration. The cost of finding the basis can be reduced by exploiting properties of the geometric space from which the samples

are drawn as described in Section 3.2. We have seen good mixing of the MCMC chains across a wide-variety of examples. We have found the additional computational burden to be beneficial in situation where substantial prior information that will impact the results is available.

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

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

The framework we develop here allows us to retain many benefits of Bayesian methods: it requires

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

7 Supplementary Material

7.1 Practical Considerations for Using the Restricted Likelihood

We offer the following guidelines and recommendations for using these methods in practice.

- We have the method useful in outlier prone situations where good prior information can be encoded for the 'good' part of the model. The additional computation burden may be too much when this is not the case. Traditional robust estimation techniques are often quicker and can provide comparable results.

- Currently, we recommend conditioning on classical robust estimators when pursing these methods. Default choices are Huber's and Tukey's estimators as outlined in the paper. We have found default choices (e.g. obtaining 95% efficiency under normally distributed data) for the tuning parameters to work well in many situations. However, depending on the amount of contamination, some adjustment may be necessary.

- The likelihood itself can indeed take many forms and, like the conditioning statistic, can depend on the application. We have developed methods for fitting these models under a large class of conditioning statistics. Currently the default choice in outlier prone situations is to use a normal likelihood model, with the conditioning intended to reduce the sensitivity of the analysis to the outliers.

- Comparisons amongst various choices of conditioning statistics and models can be made via

standard cross-validation techniques like those done in this paper where models are fit to several training and holdout sets. One should take care in the method of evaluation. When the goal is to obtain good estimation/prediction for the non-outlying data, evaluation should take place on non-outlying data or the metric should be adjusted to discount outlying data ins some way (such as the trimmed-log-marginal we introduced TLM).

7.2 Proofs

Proof of Theorem 3.1.

Proof.

$$
\begin{aligned}
s(X,\boldsymbol{y}) &= s\left(X, \frac{s(X,\boldsymbol{y}_{obs})}{s(X,\boldsymbol{z}^*)}\boldsymbol{z}^* + X\left(\boldsymbol{b}(X,\boldsymbol{y}_{obs}) - \boldsymbol{b}(X, \frac{s(X,\boldsymbol{y}_{obs})}{s(X,\boldsymbol{z}^*)}\boldsymbol{z}^*)\right)\right) && (22)\\
&= \frac{s(X,\boldsymbol{y}_{obs})}{s(X,\boldsymbol{z}^*)} s(X,\boldsymbol{z}^*) = s(X,\boldsymbol{y}_{obs}), \qquad \text{and} && (23)\\
\boldsymbol{b}(X,\boldsymbol{y}) &= \boldsymbol{b}\left(X, \frac{s(X,\boldsymbol{y}_{obs})}{s(X,\boldsymbol{z}^*)}\boldsymbol{z}^* + X\left(\boldsymbol{b}(X,\boldsymbol{y}_{obs}) - \boldsymbol{b}(X, \frac{s(X,\boldsymbol{y}_{obs})}{s(X,\boldsymbol{z}^*)}\boldsymbol{z}^*)\right)\right) && (24)\\
&= \boldsymbol{b}(X, \frac{s(X,\boldsymbol{y}_{obs})}{s(X,\boldsymbol{z}^*)}\boldsymbol{z}^*) + \boldsymbol{b}(X,\boldsymbol{y}_{obs}) - \boldsymbol{b}(X, \frac{s(X,\boldsymbol{y}_{obs})}{s(X,\boldsymbol{z}^*)}\boldsymbol{z}^*) && (25)\\
&= \boldsymbol{b}(X,\boldsymbol{y}_{obs}) && (26)
\end{aligned}
$$

□

Theorem 7.1. *The mapping $h : \mathbb{S} \to \mathcal{A}$ with h defined in Theorem 3.1 is one-to-one and onto.*

Proof. One-to-one: Let $z_1, z_2 \in \mathbb{S}$ with $h(z_1) = h(z_2)$. Rearrangement implies $z_1 = cz_2 + Xv$ for known $c \in \mathbb{R}$ and $v \in \mathbb{R}^p$ depending on $\boldsymbol{b}(X,\boldsymbol{y}_{obs})$, $s(X,\boldsymbol{y}_{obs})$, $\boldsymbol{b}(X,z_1)$, $s(X,z_1)$, $\boldsymbol{b}(X,z_2)$, $s(X,z_2)$. Given $z_2 \in \mathbb{S}$, $v \neq 0$ implies $z_1 \notin \mathcal{C}^{\perp}(X)$ and $c \neq 1$ implies $||z_1|| \neq 1$. Thus $z_1 \in \mathbb{S}$ implies $c = 1$ and $v = 0$.

Onto: Let $\boldsymbol{y} \in \mathcal{A}$ and consider its projection onto $\mathcal{C}^{\perp}(X)$: $Q\boldsymbol{y}$ where $Q = I - XX^{\top}$. It is easy to show that $\boldsymbol{z}^* = Q\boldsymbol{y}/||Q\boldsymbol{y}|| \in \mathbb{S}$ and $h(\boldsymbol{z}^*) = \boldsymbol{y}$. □

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, we have

$$s(X, \boldsymbol{y}) = s(X, Q\boldsymbol{y} + H\boldsymbol{y}) = s(X, Q\boldsymbol{y}). \tag{27}$$

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 (27), 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

$$[XX^{\top}, WW^{\top}]^{\top}[\nabla \boldsymbol{b}(X, \boldsymbol{y}), \nabla s(X, \boldsymbol{y})] = \begin{pmatrix} X & \mathbf{0} \\ WW^{\top}\nabla b(X, \boldsymbol{y}) & \nabla s(X, \boldsymbol{y}) \end{pmatrix} \tag{28}$$

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 (28). 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 7.2. *(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 7.3. *(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 7.2 and Theorem 7.3. 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 7.3, 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 7.2, the non-unit singular values of $W^{\top}A$ are the same as the non-unit singular values of $U^{\top}B$. □

7.3 Setting the hierarchical prior values

This section describes the how the prior parameters are set in Section 5.2. Using the previous data set from two years prior, we fit separate (robust) regressions to each state and a regression to the entirety of the data at once. Let the estimates for the fits to each state be $\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_J, \hat{\sigma}_1, \ldots, \hat{\sigma}_J$ and the estimates from the single regression be $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$. These are classical robust estimates using Tukey's regression and Huber's scale. For this section, let n_j denote the number of observations in the j^{th} state (of the previous data set) and set $n_p = \sum n_j$.

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 $\boldsymbol{\beta}_j$ are equal (no variation between states) and $b = 1$ implies the $\boldsymbol{\beta}_j$ vary about μ_0 according to $\Sigma_0 = n_p\hat{\Sigma_0}$ (see Section 5.1). We seek a prior measure for what we think b should be. Using the prior fit, a measure for uncertainty for $\boldsymbol{\beta}$ is $\Sigma_{\hat{\beta}} = cov(\hat{\boldsymbol{\beta}})$, the estimate of the covariance from the single regression. For each j, take $\delta_j = \hat{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}$ and set the prior uncertainty to $\Sigma_\delta = n_p^{-1}\sum_j n_j\delta_j\delta_j^\top$. Consider $g = (|\Sigma_\delta|/|\Sigma_{\hat{\beta}}|)^{(1/p)}$ as a measure of the amount of uncertainty between the β_j relative to that of $\boldsymbol{\beta}$. Now in the prior, we heuristically set the uncertainty in the $\beta_j's$ ($b\Sigma_0$) to be approximately equal to $g \cdot \Sigma_{\hat{\beta}}$. That is, $b\Sigma_0 \approx g \cdot \Sigma_{\hat{\beta}} = \frac{g}{n}\Sigma_0$, suggesting $b \approx \frac{g}{n}$. Thus, we set $E[b] = \frac{g}{n}$. The precision, $v_1 + v_2$, is set to 10, completing the specification for the prior on b.

Finally, recall $\rho \sim \text{beta}(a_\rho, b_\rho)$ with mean $\mu_\rho = a_\rho/(a_\rho + b_\rho)$ given a beta prior and precision $\psi_\rho = a_\rho + b_\rho$ given a gamma prior. There is little evidence of any strong correlation amongst estimates of σ_j^2 in the prior data set and we set the prior mean of μ_ρ equal to 0.2 and prior variance to .01. Noting $\text{var}(\rho|\mu_\rho, \psi_\rho) = \mu_\rho(1-\mu_p)/(\psi_\rho + 1)$ we plug in $\mu_\rho = 0.2$ and $\text{var}(\rho|\mu_\rho, \psi_\rho) = 0.01$. Solving for ψ_ρ results in a value of 15. This is taken to be the mean of the gamma prior on ψ_ρ. Finally, we set the rate parameter for to 1 implying the variance of the gamma prior is equal to its the mean. With this specification, the prior on ρ has 80% of the central mass between roughly 0.03 and 0.4 and reflects our prior belief that there is likely only weak positive correlation amongst the σ_j^2's.

7.4 M-estimators

M-estimators offer a natural choice for conditioning in restricted likelihood settings and they can be readily applied when conditioning on these estimators using the method described in this paper. This section gives a brief review

7.5 M-estimators of Location

To begin the review it is easiest to start with the location model

$$y_i = \mu + \epsilon_i, \quad i = 1, \ldots, n, \tag{29}$$

where we assume that ϵ_i are independent identically distributed random variables from the distribution F_0 having mean zero, constant (known) variance, and density f_0 with respect to Lebesgue measure. It follows that $y_i \overset{iid}{\sim} F(y) = F_0(y - \mu)$. The M-estimate for μ is defined by the criterion

$$\hat{\mu} = \arg\min_{\mu} \sum_{i=1}^{n} \rho(y_i - \mu), \tag{30}$$

where ρ is a loss function. It is assumed that $\rho(x)$ is nondecreasing in $|x|$ and $\rho(0) = 0$. If F_0 is standard normal, taking ρ to be squared error loss ($\rho(x) = x^2$) results in the maximum likelihood estimator (MLE). In general, $\rho(x) = -\log f_0(x)$ corresponds to the MLE. In this sense, M-estimators are a generalization of MLEs.

Assuming it exists, the derivative of ρ is denoted by ψ and $\hat{\mu}$ in (30) satisfies the estimating equation

$$\sum_{i=1}^{n} \psi(y_i - \hat{\mu}) = 0. \tag{31}$$

Putting $w_i = \frac{\psi(y_i - \hat{\mu})}{y_i - \hat{\mu}}$, equation (31) can be rewritten as

$$\sum_{i=1}^{n} w_i (y_i - \hat{\mu}) = 0 \tag{32}$$

and an analytical expression for $\hat{\mu}$ is then the weighted average

$$\hat{\mu} = \frac{\sum_i w_i y_i}{\sum_i w_i}. \tag{33}$$

This weighted average cannot be used directly since the w_i depend on $\hat{\mu}$, but it provides a useful insight in how to choose a ψ that downweights outlying observations. Two examples of commonly used ρ functions whose derivatives ψ achieve downweighting are

$$\rho(x) = \begin{cases} x^2, & |x| \leq k \\ 2k|x| - k^2, & |x| \geq k \end{cases} \tag{34}$$

and

$$\rho(x) = \begin{cases} 1 - [1 - (x/k)^2]^3, & |x| \leq k \\ 1, & |x| \geq k. \end{cases} \tag{35}$$

The first of these is known as Huber's loss. It offers a compromise to squared error loss, with the loss becoming linear for absolute residuals larger than k. Huber and Ronchetti (2009) provide a theoretical justification of this choice based the *least informative distribution* in a particular class of contaminated normal distributions. The ρ in (35) is known as Tukey's bisquare.

A notable property of both of the ψ functions here is that they are bounded. This implies a bounded *influence function* (IF). The influence function is a measure of how much an estimator is affected by a small fraction of identical outliers in the sample. Boundedness is desirable because it implies that the effect of a small contamination on a statistical procedure cannot become arbitrarily large. More details on influence functions can be found in Huber and Ronchetti (2009) and Maronna et al. (2006). Tukey's ψ is *re-descending* as it returns to zero outside of $(-k, k)$. As a result, Tukey's estimator downweights large residuals more than does Huber's estimator. Huber's estimator is akin to Winsorising whereas Tukey's is akin to complete trimming.

Note that k is a tuning constant which must be chosen. Under usually satisfied conditions, M-estimators are asymptotically normal and this property can be employed to choose k. The results for the location setting, which extend up to linear regression, are presented here. Again, more detailed results can be found in both Huber and Ronchetti (2009) and Maronna et al. (2006).

For a given distribution F, let μ_0 be the value such that

$$E_F \psi(x - \mu_0) = \int \psi(x - \mu_0)\, dF = 0. \tag{36}$$

Taking $\hat{\mu}$ to be the solution to (31) we have $\hat{\mu} \xrightarrow{p} \mu_0$ and

$$\sqrt{n}(\hat{\mu} - \mu_0) \xrightarrow{\mathcal{D}} N(0, v) \text{ with } v = \frac{E_F(\psi(y-\mu_0)^2)}{(E_F\psi'(y-\mu_0))^2} \tag{37}$$

as $n \to \infty$. The conditions for the above convergence properties to hold depend on ψ and F. In particular, the numerator of v must be finite and the denominator must exist and be nonzero. Redescending ψ functions, such as Tukey's, require slightly stronger conditions than those for monotone ψ.

Under the assumed model, the MLE is optimal in the sense of achieving the minimum asymptotic variance of $v_0 = I(\mu)^{-1}$ where $I(\mu)$ is the Fisher information for μ. Hence, $v \geq v_0$ and k is typically chosen by considering the asymptotic relative efficiency of $\hat{\mu}$

$$\text{Eff}(\hat{\mu}) = \frac{v_0}{v}. \tag{38}$$

The idea is for robust estimators to not perform badly compared to the optimal estimator, under the assumed model. To this end, k is set so that $\text{Eff}(\hat{\mu})$ is reasonably close to one. The standard default is to assume a normal model and set $\text{Eff}(\hat{\mu}) = 0.95$. This leads to $k = 1.345$ for Huber's ρ and $k = 4.685$ for Tukey's. Unless otherwise stated, these values are used for the applications in this dissertation.

7.6 M-estimators of Scale

M-estimation naturally extends the scale model

$$y_i = \sigma\epsilon_i, \quad i = 1, \ldots, n, \tag{39}$$

where ϵ_i are independent identically distributed with density f_0 and $\sigma > 0$ is the unknown parameter. Then y_i come from a scale family with density $f_\sigma(y) = \frac{1}{\sigma} f_0(\frac{y}{\sigma})$. The M-estimator for σ is defined through the estimating equation

$$\sum_{i=1}^{n} \chi\left(\frac{y_i}{\sigma}\right) = 0, \tag{40}$$

for some function χ. Taking $\chi(x) = -x\frac{f_0'(x)}{f_0(x)} - \delta$ for $\delta = 1$ corresponds to the MLE. This reduces to $\chi(x) = x^2 - 1$ for the standard normal yielding the estimate $\hat{\sigma} = \sqrt{\frac{1}{n}\sum y_i^2}$. Using the same idea as

in the location case, a different χ can be chosen in an effort to achieve a robust estimate of σ. Huber (1964) proposed the choice

$$\chi(x) = \begin{cases} x^2 - \delta, & |x| \leq k \\ k^2 - \delta, & |x| > k. \end{cases} \tag{41}$$

For consistency when F is $N(0, \sigma^2)$ we need $E_\Phi[\chi(x)] = \int \chi(x)\ d\Phi = 0$ where Φ is the standard normal distribution. This yields $\delta \approx 0.71016$ for the χ in (41). For convenience, k is often set to 1.345, the value giving 95% efficiency at the normal for Huber's location estimator. This is often referred to as Huber's 'Proposal 2'.

Another popular choice is the step function

$$\chi(x) = I(|x| > c) - \delta. \tag{42}$$

Fixing $\delta = 0.5$ yields the aforementioned *rescaled median absolute deviation*, written here as $\hat{\sigma} = K(\text{Med}(|x|))$ with $K = 1/c$.

7.7 Simultaneous M-estimators of Location and Scale

There are many ways to define M-estimates for multiple parameters but for convenience we concentrate on *simultaneous estimators*. We first consider the location and scale setting

$$y_i = \mu + \sigma\epsilon_i, \quad i = 1, \ldots, n, \tag{43}$$

where ϵ_i are independent identically distributed with density f_0, μ and $\sigma > 0$ unknown and y_i has density $f(y) = \frac{1}{\sigma} f_0(\frac{y-\mu}{\sigma})$. Combining the equations in (31) and (40), the simultaneous M-estimators of location and scale are the solutions to

$$\begin{aligned} \sum \psi\left(\frac{y_i - \hat{\mu}}{\hat{\sigma}}\right) &= 0 \\ \sum \chi\left(\frac{y_i - \hat{\mu}}{\hat{\sigma}}\right) &= 0. \end{aligned} \tag{44}$$

The extension of M-estimators to linear regression is straightforward. Consider the standard regression model

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i, \tag{45}$$

where ϵ_i are independent random variables with mean zero and variance σ^2, $\boldsymbol{x}_i \in \mathbb{R}^p$ is a vector of predictors, and $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of regression parameters. The simultaneous M-estimators of $\boldsymbol{\beta}$ and σ are defined by

$$\begin{aligned} \sum \psi\left(\frac{y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)\boldsymbol{x}_i &= \mathbf{0} \\ \sum \chi\left(\frac{y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}}{\hat{\sigma}}\right) &= 0. \end{aligned} \tag{46}$$

Common ψ functions include Huber's and Tukey's mentioned earlier and common χ functions appear in (41) and (42). Assuming the distribution of the errors is F, under standard conditions for ψ, χ, the general result is again $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ and $\hat{\sigma} \xrightarrow{p} \sigma_0$ where $\boldsymbol{\beta}_0$ and σ_0 satisfy

$$\begin{aligned} E_F\psi\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_0}{\sigma_0}\right)\boldsymbol{x}_i &= \mathbf{0} \\ E_F\chi\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_0}{\sigma_0}\right) &= 0. \end{aligned} \tag{47}$$

Further, the estimates are jointly asymptotically normal with a covariance matrix depending on the ψ and χ functions.

References

Berger, J. (2006). "The Case for Objective Bayesian Analysis." Bayesian Analysis, 1: 385–402. 2

Bernardo, J. M. and Smith, A. (2000). Bayesian Theory. John Wiley & Sons Ltd. 2

Clarke, B. and Ghosh, J. K. (1995). "Posterior Convergence Given the Mean." The Annals of Statistics, 23: 2116–2144. 6

Clarke, J. L., Clarke, B., Yu, C.-W., et al. (2013). "Prediction in M-complete Problems with Limited Sample Size." Bayesian Analysis, 8(3): 647–690. 2

Clyde, M. and George, E. I. (2004). "Model uncertainty." Statistical science, 81–94. 2

Clyde, M. A. and Iversen, E. S. (2013). "Bayesian model averaging in the M-open framework." Bayesian theory and applications. 2

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. 6

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. 6

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. 2

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. 13

Gelman, A. (2006). "Multilevel (hierarchical) modeling: what it can and cannot do." Technometrics, 48(3): 432–435. 36

Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications." Biometrika, 57: 97–109. 13

Hoff, P., Fosdick, B., Volfovsky, A., and Stovel, K. (2013). "Likelihoods for Fixed Rank Nomination Networks." Network Science, 1: 253–277. 6

Huber, P. and Ronchetti, E. (2009). Robust Statistics. Wiley Series in Probability and Statistics. Hoboken, New Jersey: John Wiley & Sons, Inc, 2nd edition. 7, 8, 12, 22, 45

Huber, P. J. (1964). "Robust Estimation of a Location Parameter." The Annals of Mathematical Statistics, 35(1): 73–101. 11, 47

Hwang, H., So, B., and Kim, Y. (2005). "On Limiting Posterior Distributions." Test, 14: 567–580. 6

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

Jung, Y., MacEachern, S., and Lee, Y. (2014). "Cross-validation via Outlier Trimming." In preparation. 31

Kass, R. E. and Raftery, A. E. (1995). "Bayes Factors." Journal of the American Statistical Association, 90: 773–795. 2

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. 9, 30

Lee, J. and MacEachern, S. N. (2014). "Inference functions in high dimensional Bayesian inference." Statistics and Its Interface, 7(4): 477–486. 2, 7

Lewis, J. (2014). "Bayesian Restricted Likelihood Methods." Ph.D. thesis, The Ohio State University. 5, 11, 13

Lewis, J., Lee, Y., and MacEachern, S. (2012). "Robust Inference via the Blended Paradigm." In JSM Proceedings, Section on Bayesian Statistical Science, 1773–1786. American Statistical Association. 6

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. 13, 30

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. 13

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. 6

Maronna, R., Martin, D., and Yohai, V. (2006). Robust Statistics: Theory and Methods. Wiley Series in Probability and Statistics. West Sussex, England: John Wiley & Sons, Ltd. 13, 45

Miao, J. and Ben-Israel, A. (1992). "On Principal Angles Between Subspaces in $\mathbb{R}^n$." Linear Algebra and its Applications, 171: 81–98. 19, 20, 42

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. 2

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. 6

— (1983). "Likelihood Based Inference Using Signed Ranks for Matched Pairs." Journal of the Royal Statistical Society. Series B, 45: 287–296. 6

Pratt, J. W. (1965). "Bayesian Interpretation of Standard Inference Statements." Journal of the Royal Statistical Society. Series B, 27: 169–203. 6

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. 6

Ratcliff, R. (1993). "Methods for Dealing with Reaction Time Outliers." Psychological Bulletin, 114: 510. 4

Ronchetti, E., Field, C., and Blanchard, W. (1997). "Robust Linear Model Selection by Cross-Validation." Journal of the American Statistical Association, 92: 1017–1023. 31

Rousseeuw, P. J. and Leroy (1987). Robust regression and outlier detection. John Wiley & Sons. 8

Savage, I. R. (1969). "Nonparametric Statistics: A Personal Review." Sankhya: The Indian Journal of Statistics, Series A (1961-2002), 31: 107–144. 6

Stigler, S. M. (1977). "Do Robust Estimators Work with Real Data?" The Annals of Statistics, 5(6): 1055–1098. 7

Tavaré, S., Balding, D. J., Griffiths, R., and Donnelly, P. (1997). "Inferring Coalescence Times from DNA Sequence Data." Genetics, 145: 505–518. 6

Wong, H. and Clarke, B. (2004). "Improvement over Bayes prediction in small samples in the presence of model uncertainty." Canadian Journal of Statistics, 32(3): 269–283. 6

Yuan, A. and Clarke, B. (2004). "Asymptotic Normality of the Posterior Given a Statistic." The

Canadian Journal of Statistics, 32: 119–137. 6

Yuan, A. and Clarke, B. S. (1999). "A minimally informative likelihood for decision analysis: illustration and robustness." Canadian Journal of Statistics, 27(3): 649–665. 2

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." In Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti, 233. 30

Zhu, H., Ibrahim, J. G., and Tang, N. (2011). "Bayesian influence analysis: a geometric approach." Biometrika, 98(2): 307–323. 2

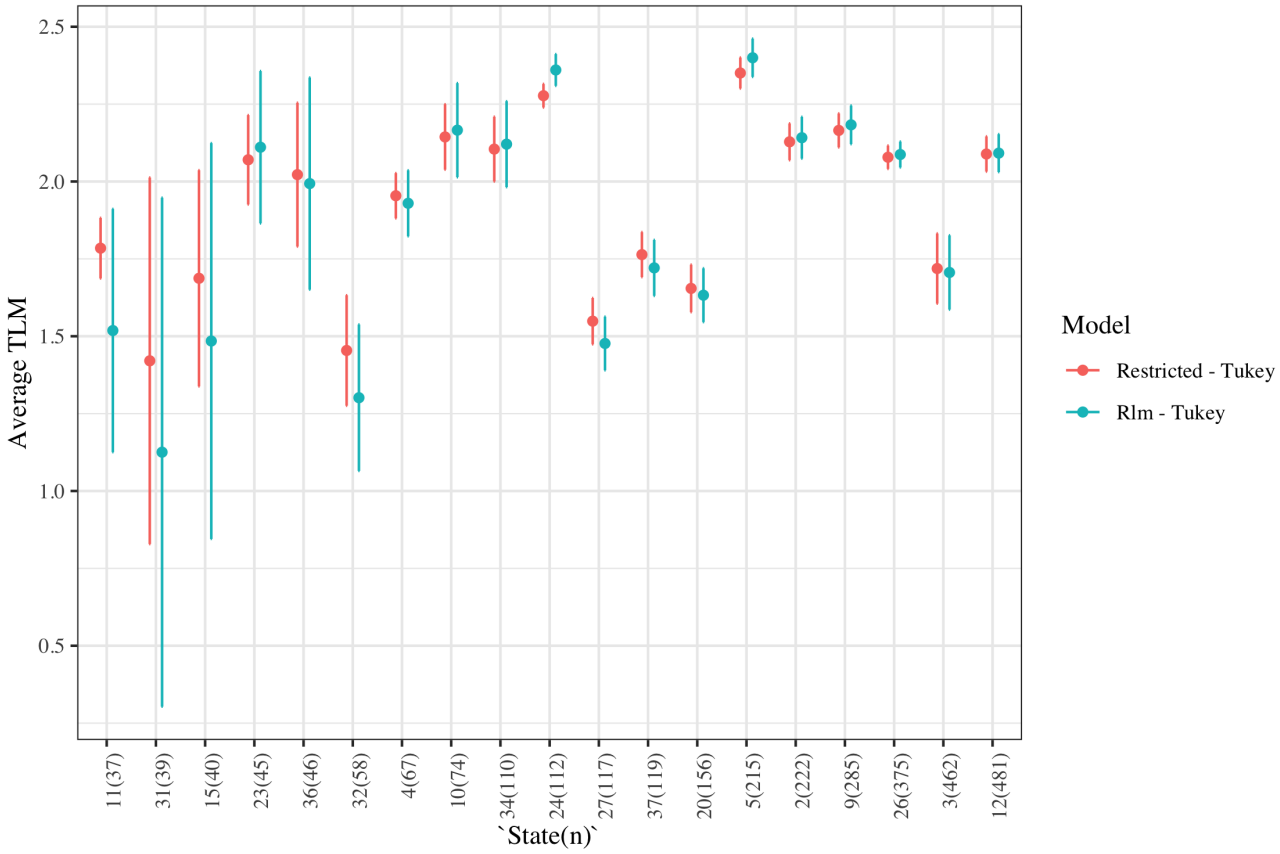

Figure 13: Hierarchical model results: $TLM_b(A)_j$ plus/minus one standard deviation over $K = 50$ repetitions for each state and $\alpha = 0.3$. The states are ordered along the x-axis according to number of agencies within the state (shown in parentheses). Results displayed are for the robust models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