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

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

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

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

Abstract

The *restricted likelihood* is introduced as a method to address concerns about the miss-specification of the likelihood in Bayesian models. To apply, the data is summarized through a set of insufficient statistics, targeting inferential quantities of interest, and the prior is updated with only the summary statistics. By a careful choice of the summary, 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. Examples specifically address reducing sensitivity to outliers, where classical robust estimators (e.g., M-estimators) are natural choices for conditioning statistics. With these choices, the method can be thought of as a blend of classical robust estimation and Bayesian methods.

A Markov chain Monte Carlo (MCMC) algorithm, applicable for the linear model and a wide range of choices for summary statistics, is developed to overcome implementation challenges. The method is demonstrated with several examples including simulations and messy insurance agency data with many outliers. Success is manifested in better predictive performance for data points of interest as compared to competing methods.

*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.

KEYWORDS: Approximate Bayesian computation, Markov chain Monte Carlo, M-estimation, Robust regression.

1 Introduction

Bayesian methods are useful to a wide range of scientific problems, with their value having been demonstrated both empirically and theoretically. However, these methods run into difficulties for two major and prevalent classes of problems: handling data sets with outliers and dealing with model misspecification. From Bayesian perspective the model refers to the prior, the likelihood, as well as the loss function when formal inference is required. While optimality of Bayesian methods is unquestioned if one accepts the validity of all three of these elements, healthy skepticism encourages us to question each. Concern about the prior distribution has been addressed through objective Bayesian methods (Berger, 2006) prior elicitation methods (Garthwaite et al., 2005; O'Hagan et al., 2006). Concern about the loss function is reflected in the extensive literature on Bayesian hypothesis tests (Kass and Raftery, 1995). Despite the common use of flat-tailed distributions (Berger, 1994) or nonparametric techniques References?, the sampling density has been given less attention from a specifically Bayesian view. The work on predictive diagnostics (Box, 1980) departs from classical traditions.

The focus of this work is the development of techniques to handle imperfections in the sampling density. These imperfections often show themselves through the presence of outliers–cases not reflecting the phenomenon under study. There are three main solutions to Bayesian outlier-handling. The first is to replace the basic sampling density with a mixture model which includes one component for the "good" data and a second component for the "bad" data. The second approach replaces the basic sampling density with a thick-tailed density in an attempt to discount outliers. The third approach fits a flexible (typically nonparametric) model to the data, producing a Bayesian version of a density estimate for both good and bad data. In recent development, inference is made through the use of robust inference functions (Lee and MacEachern, 2014, in press).

The traditional strategies for handling outliers all have their drawbacks. While we view the sampling density for the good data as stable, the outlier-generating processes may be transitory in nature, constantly shifting as the source of bad data changes. This prevents us from appealing to large-sample arguments to claim that, with enough data, we can nail down a model for both good and bad data combined. Instead of attempting to model both good and bad data, we propose a new strategy. In a nutshell, we begin with a complete model as if all of the data are good. But rather than driving the move from prior distribution to posterior distribution by the entire likelihood, we use only the likelihood driven by a few summary statistics that target inferential quantities of interest. We call this the *restricted likelihood*. The update is a formal update from prior distribution to posterior distribution, based on the sampling density of the summary statistics.

While the focus of this work is handling outliers, model miss-specification is more general. By focusing on robustly estimating quantities of interest and conditioning on these estimates, the restricted likelihood offers promiss in handling more general miss-specification issues in model building. The remainder of the paper.....develops Bayesian restricted likelihood (Section 2), shows how it can be applied to a Bayesian linear model (Section 4), illustrates its use on an insurance agencies data set with a novel twist on model evaluation (Section 5), and wraps up with a discussion (Section 6). A major contribution of this work is the computational strategy whose legitimacy is established in Section 4. The technical proofs are in the appendix.

2 The Restricted Likelihood

This sections describes the restricted likelihood using a pair of simple examples and the relates the method to the current literature. For the examples assume $\boldsymbol{y} = (y_1, \ldots, y_n)$ is a random sample of size n from a continuous distribution indexed by a parameter vector $\boldsymbol{\theta}$, with pdf $f(y|\boldsymbol{\theta})$. The full-data, likelihood is $L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$.

For the first example, consider that a known subset of the data is not informative about $\boldsymbol{\theta}$. This case mimics the setting where outlying observations are identified and discarded

before formal analysis. Labeling 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 $L(\boldsymbol{\theta}|y_1,\ldots,y_{n-k}) = \prod_{i=1}^{n-k} f(y_i|\boldsymbol{\theta})$. Equivalently, the full likelihood is rewritten as

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

The second piece, which is independent of θ, is dropped to provide better inference on $\boldsymbol{\theta}$.

A second example involves deliberate censoring of small and large observations. This is sometimes done as a precursor to the analysis of reaction time experiments (e.g., Ratcliff, 1993) where very short and long reaction times are usually explained as anticipation and inattention, respectively, and do not reflect the process being studied. 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 $y \in (t_1,t_2)$. Adjusting y_i, with $c(y_i)$ where $c(y_i) = t_1$ 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 $L(\boldsymbol{\theta}|c(\boldsymbol{y}))$. With slightly non-standard notation, we let $g(t_1|\boldsymbol{\theta}) = F(t_1|\boldsymbol{\theta})$, $g(t_2|\boldsymbol{\theta}) = 1-F(t_2|\boldsymbol{\theta})$, and $g(y|\boldsymbol{\theta}) = f(y|\boldsymbol{\theta})$ for $y \in (t_1,t_2)$. The full likelihood as the product of two pieces,

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

and only the first for the formal update.

Further example abound with more given in Lewis (2014). To generalize, we write the full likelihood in two pieces

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

where $T(\boldsymbol{y})$ is any statistic. 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 (1), the conditioning statistic

is $T(\boldsymbol{y}) = (y_1, \ldots, y_{n-k})$ and in (2), 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.

$T(\boldsymbol{y})$ is a well-defined random variable with a probability distribution indexed by $\boldsymbol{\theta}$ and the update from prior to posterior can be made using $f(T(\boldsymbol{y})|\boldsymbol{\theta})$ rather than $f(\boldsymbol{y}|\boldsymbol{\theta})$. This leads to the *restricted likelihood posterior*

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

where $\pi(\boldsymbol{\theta})$ is the prior distribution and $m(T(\boldsymbol{y}))$ is the marginal distribution of $T(\boldsymbol{y})$ under prior. The associated predictive distribution for a single observation $\tilde{y}$ is

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

Direct use of restricted likelihood appears in many areas of the literature. The motivation is often similar to ours: concern about outliers or, more generally, model misspecification. 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 robust regression estimators for $T(\boldsymbol{y})$. 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). These asymptotic results asserts that under certain assumptions, posterior distribution resembles the asymptotic sampling distribution of the conditioning statistic.

Restricted likelihoods have 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 used in Approximate Bayesian Computation (ABC), which is related to our method. ABC 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. Consequently, effort

is made to choose an approximately sufficient $T(\boldsymbol{y})$ and update to the ABC posterior by using the likelihood $L(\boldsymbol{\theta}|\mathcal{B}(\boldsymbol{y}))$, where $\mathcal{B}(\boldsymbol{y}) = \{\boldsymbol{y}^*|\rho(T(\boldsymbol{y}), T(\boldsymbol{y}^*)) < \epsilon\}$, ρ is a metric, and ϵ is a tolerance level. This is the likelihood conditioned on the collection of data sets close to $\boldsymbol{y}$ as measured by the summary statistic. By choosing an approximately sufficient $T(\cdot)$ and a small enough ϵ the claim $L(\boldsymbol{\theta}|\mathcal{B}(\boldsymbol{y})) \approx L(\boldsymbol{\theta}|T(\boldsymbol{y})) \approx L(\boldsymbol{\theta}|\boldsymbol{y})$. Consequently, the ABC posterior approximates the full data posterior. 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 conditioning is on something other than the full data. For technical reasons, the calculation proceeding from $\pi(\boldsymbol{\theta})$ to $\pi(\boldsymbol{\theta}|\mathcal{B}(\boldsymbol{y}))$ is not a formal conditional update. Simply put, if an external observer told $\mathcal{B}(\boldsymbol{y})$ can supply $T(\boldsymbol{y})$ and vice-versa, the information contained in $\mathcal{B}(\boldsymbol{y})$ is exactly that contained in $T(\boldsymbol{y})$ and proper conditioning will lead to identical posterior distributions. This is not the case with a typical implementation of ABC.

This work extends the development of Bayesian restricted likelihood by arguing that deliberate choice of $T(\boldsymbol{y})$ is sound practice and also by expanding the class of conditioning statistics in which exact conditioning can be achieved. Our methods do not rely on asymptotic properties, approximate conditioning, nor do they need to correspond to artificial censoring to create incomplete or missing data (e.g., Albert, 1988; Hoff and Wakefield, 2013). Note - fix references here hoffwakefield for asymptotic; albert for ? .

MOVE THIS PIECE OR REMOVE — The key to productive use of the restricted likelihood is the choice of $T(\boldsymbol{y})$ and the development of computational strategies that allow us to truly condition on the observed $T(\boldsymbol{y})$ and fit the model in formal Bayesian fashion. In this work, we focus on robustness, and natural choices of $T(\boldsymbol{y})$ for the one-sample problem include a set of middling order statistics, a trimmed mean, or a classical robust estimator of location and/or scale. We have previously implemented several of these methods and have found them to perform well (Lewis et al., 2012). Versions most extensible to the linear model include the 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 incomplete posterior conditioned on essentially any statistic. These strategies rely on generation of complete data sets from different values of $\boldsymbol{\theta}$. Each complete data set leads to a statistic $T(\boldsymbol{y})$ under $\boldsymbol{\theta}$, and these generated statistics are used to estimate the density at $T(\boldsymbol{y}_{obs})$, where $\boldsymbol{y}_{obs}$ is the observed complete data. This estimate is fed into Bayes' theorem for the update from prior distribution to posterior distribution.

These direct sampling techniques rely on density estimation and numerical integration. Density estimation becomes difficult for high dimensional $T(\boldsymbol{y})$ and numerical integration breaks down for high dimensional $\boldsymbol{\theta}$. For these settings, an MCMC algorithm is developed in Section 4.2. Software is available from the authors for use with simultaneous M-estimators of location and scale. At the present time, further software development is required to extend beyond simultaneous M-estimators.

3 Examples

Before discussing computational details, the method is applied in a series of simple examples on well known datasets 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 **??**. Specific details are left to the Appendix.

The first example is an analysis of Simon Necomb's 66 measurements of the speed of light; two of which are significant outliers in the lower tail. The full model is a standard location-scale Bayesian model assuming normally distributed errors. 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 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. Additionally, the common approach of assuming a t-distribution for the errors is also fit. The posteriors

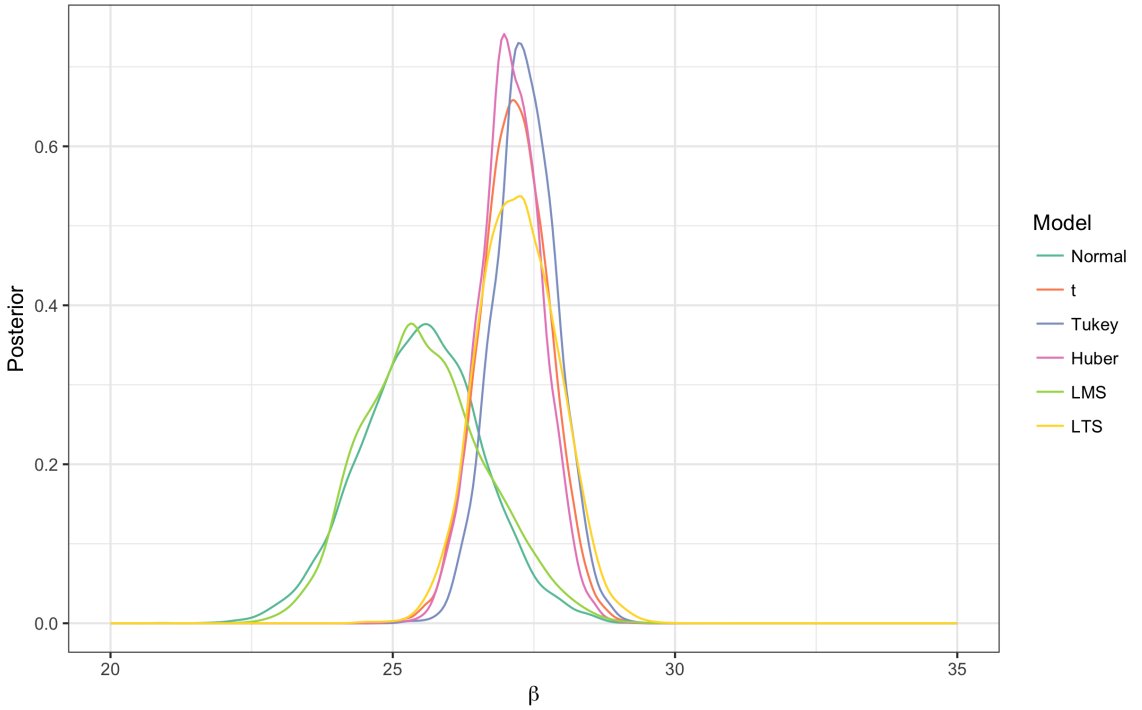

Figure 1: Posteriors un

of β under each model appear in **??**. As expected, the posterior under the normal model is pulled downward by the two outliers while the heavy tailed model provides robustness against them. The restricted likelihood methods using the M-estimators and LTS statistics also achieve robustness against the outliers. Conditioning on LMS however, results in a posterior similar to the one under the normal model. The M-estimators provide the tightest posteriors in this case. They also provide more precise predictions than the heavy-tailed model as illustrated by the predictive distributions displayed in **??** which are also dependent on the posterior for the scale parameter.

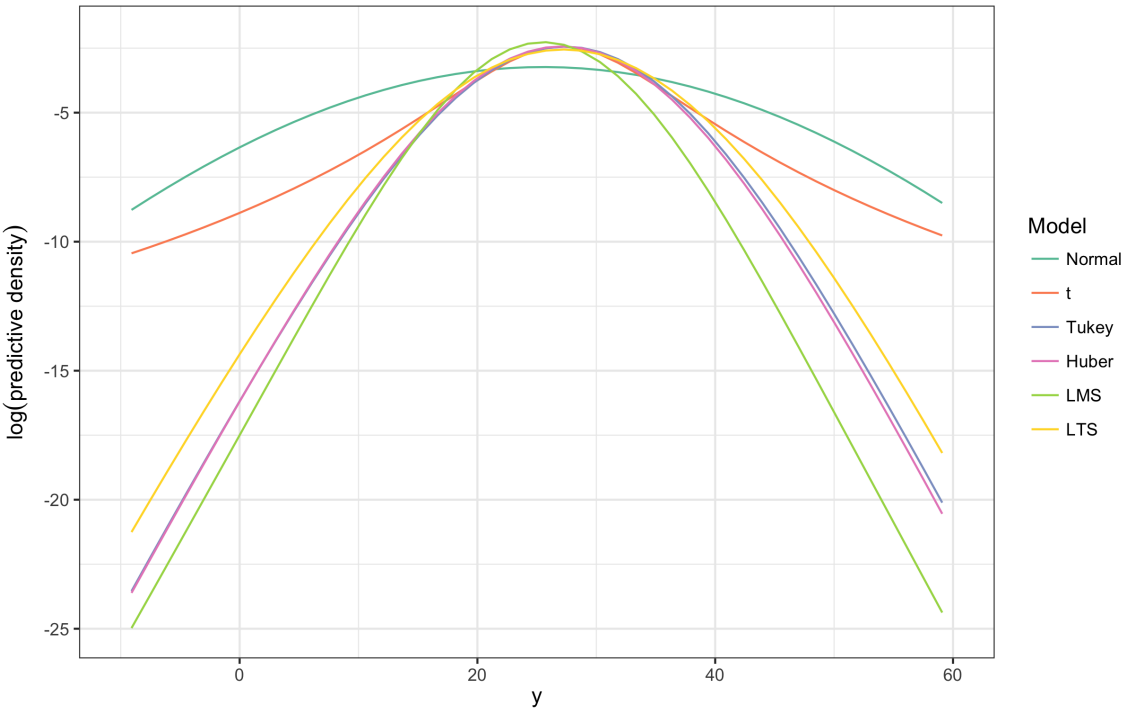

Figure 2: predit

4 Restricted Likelihood for the Linear Model

4.1 The Bayesian linear model

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

$$\begin{aligned} \boldsymbol{\theta} &= (\boldsymbol{\beta}, \sigma^2) \sim \pi(\boldsymbol{\theta}) \\ y_i &= x_i^\top \boldsymbol{\beta} + \epsilon_i, \text{ for } i = 1, \ldots, n \end{aligned} \tag{6}$$

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. Two conditions are imposed on the model:

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

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

As summary statistics for the data, we concentrate on robust estimators for the regression coefficients in the linear model and an associated estimator of the scale. In particular, we demonstrate the method using simultaneous M-estimators defined by (**??**) for the one-sample setting and easily extended to the linear model in (6). The estimator of the regression coefficients is denoted by $\boldsymbol{b}(X, \boldsymbol{y}) \in \mathbb{R}^p$ and that of the scale by ~~$s(X, \boldsymbol{y}) \in \mathbb{R}$~~ $s(X, \boldsymbol{y}) \in \{0\} \cup \mathbb{R}^+$. Thus, $T(\boldsymbol{y}) = (\boldsymbol{b}(X, \boldsymbol{y}), s(X, \boldsymbol{y}))$. The observed complete data is denoted by $\boldsymbol{y}_{obs}$ with observed statistic denoted by $T(\boldsymbol{y}_{obs}) = (\boldsymbol{b}(X, \boldsymbol{y}_{obs}), s(X, \boldsymbol{y}_{obs}))$.

4.2 Computational strategy

As mentioned above, direct computational strategies designed to approximate the restricted posterior break down for high-dimensional statistics $T(\boldsymbol{y})$ or high-dimensional parameters $\boldsymbol{\theta}$. However, Markov chain Monte Carlo (MCMC) methods were developed for exactly these situations. We turn to MCMC to fit the model in these circumstances.

The general style of algorithm that we present is a data augmented MCMC algorithm 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 $\pi(\boldsymbol{\theta}|\boldsymbol{y}, T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs}))$ and $f(\boldsymbol{y}|\boldsymbol{\theta}, T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs}))$. Alternative algorithms substitute other generations for the full conditionals, such as the Metropolis-Hastings propose and accept/reject framework. Our algorithms are based on the decomposition of the MCMC algorithm into conventional full data sampling steps where $\boldsymbol{\theta}$ is generated given $\boldsymbol{y}$, and steps that allow us to complete the data set where $\boldsymbol{y}$ is generated given $\boldsymbol{\theta}$ and $T(\boldsymbol{y})$. The latter generation relies on the conditioning set $\mathcal{A} = \{\boldsymbol{y} \in \mathbb{R}^n | T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})\}$. The set $\mathcal{A}$ may contain more than one element (i.e., it is not degenerate) when $T(\boldsymbol{y}) : \boldsymbol{y} \to T(\boldsymbol{y})$ is not a $1-1$ function.

When $\boldsymbol{y}$ has the summary statistic ~~$T(\boldsymbol{y})$ matching $T(\boldsymbol{y}_{obs})$,~~ $T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})$, the first full conditional, $\pi(\boldsymbol{\theta}|\boldsymbol{y}, T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs}))$, is the same as the full data posterior $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. In this case, the condition $T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})$ is redundant. This allows us to make use of conventional MCMC steps for this generation. For ~~In the case of~~ typical regression models, algorithms abound. Details of the recommended algorithms depend on details of the prior distribution and sampling density. As examples, a normal prior distribution and normal likelihood in the regression setting allow one to alternate conditional generations of σ^2 and $\boldsymbol{\beta}$, and blocking the generation of $\boldsymbol{\beta}$ generally leads to quicker convergence and mixing (Liu, 1994); a thick-tailed scale mixture of normal distributions in the style of the hyper-g/n prior (Liang et al., 2008) necessitates an additional stage for the sampler where the scale g/n is generated; a thick-tailed sampling density such as a t distribution can be handled with the addition of a scale parameter for each case and an extra stage where these scale parameters are generated.

For a typical model and conditioning statistic, the second full conditional $f(\boldsymbol{y}|\boldsymbol{\theta}, T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})) = f(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{y} \in \mathcal{A})$ is not available in closed form. As a consequence, we turn to Metropolis-Hastings (Hastings, 1970), using the strategy of proposing full data $\boldsymbol{y}$ ~~in the sample space of the distribution~~ for which $T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})$ and either accepting or rejecting the proposal. Let $\boldsymbol{y}_p$ and $\boldsymbol{y}_c$ represent the proposed and current full data, respectively. In

subsequent algorithms, we propose only from the set $\mathcal{A}$. Denote the proposal distribution for $\boldsymbol{y}_p$ by $p(\boldsymbol{y}_p|\boldsymbol{\theta}, T(\boldsymbol{y}_p) = T(\boldsymbol{y}_{obs}))= p(\boldsymbol{y}_p|\boldsymbol{\theta}, \boldsymbol{y}_p \in \mathcal{A}) = p(\boldsymbol{y}_p|\boldsymbol{\theta})$. The last equality follows from the fact that our $p(\cdot|\boldsymbol{\theta})$ assigns probability one to the event $\{\boldsymbol{y}_p \in \mathcal{A}\}$. These equalities still hold if the dummy argument $\boldsymbol{y}_p$ is replaced with $\boldsymbol{y}_c$. The conditional density is

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

for $\boldsymbol{y} \in \mathcal{A}$. This includes both $\boldsymbol{y}_p$ and $\boldsymbol{y}_c$.

The Metropolis-Hastings acceptance probability is the minimum of 1 and R where,

$$R = \frac{f(\boldsymbol{y}_p|\boldsymbol{\theta}, \boldsymbol{y}_p \in \mathcal{A})}{f(\boldsymbol{y}_c|\boldsymbol{\theta}, \boldsymbol{y}_c \in \mathcal{A})} \frac{p(\boldsymbol{y}_c|\boldsymbol{\theta}, \boldsymbol{y}_c \in \mathcal{A})}{p(\boldsymbol{y}_p|\boldsymbol{\theta}, \boldsymbol{y}_p \in \mathcal{A})} \tag{7}$$

$$= \frac{f(\boldsymbol{y}_p|\boldsymbol{\theta})}{\int_{\mathcal{A}} f(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{y}} \frac{\int_{\mathcal{A}} f(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{y}}{f(\boldsymbol{y}_c|\boldsymbol{\theta})} \frac{p(\boldsymbol{y}_c|\boldsymbol{\theta})}{p(\boldsymbol{y}_p|\boldsymbol{\theta})} \tag{8}$$

$$= \frac{f(\boldsymbol{y}_p|\boldsymbol{\theta})}{f(\boldsymbol{y}_c|\boldsymbol{\theta})} \frac{p(\boldsymbol{y}_c|\boldsymbol{\theta})}{p(\boldsymbol{y}_p|\boldsymbol{\theta})}. \tag{9}$$

Note: Above expression changed. Check argument again.

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 simulate proposals that guarantee $T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})$ and on the evaluation of the proposal density derived from the method used to simulate $\boldsymbol{y}$. We now discuss such an implementation method for the linear model in (6).

4.2.1 Construction of the proposal

Our computational strategy is easiest to envision in a location-scale setting where the design matrix in model (6) consists of a single column of ones. Robust estimation techniques along with model (6) suggest a conditioning statistic $T(\boldsymbol{y})$ which consists of estimates of the scalars β and σ, say $T(\boldsymbol{y}) = (b(X, \boldsymbol{y}), s(X, \boldsymbol{y}))$. It is not a simple matter to directly sample the required proposal $\boldsymbol{y}$ satisfying $T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})$. However, with equivariance/invariance

properties of $T(\boldsymbol{y})$, we will see in Theorem 4.1 below that it is possible to scale and shift any ~~$\boldsymbol{z}^* \in \mathbb{R}^n$~~ $\boldsymbol{z}^*$ which generates a positive scale estimate to such a $\boldsymbol{y}$. Hence, to obtain $\boldsymbol{y}$ from $\mathcal{A}$, we proceed in two steps. First, a vector $\boldsymbol{z}^*$ is sampled from a known distribution (examples of which are described later). This vector has summary statistic $T(\boldsymbol{z}^*) = (b(X, \boldsymbol{z}^*), s(X, \boldsymbol{z}^*))$ which may not match the observed summary $T(\boldsymbol{y}_{obs})$. Second, $\boldsymbol{z}^*$ is mapped into $\boldsymbol{y} = h(\boldsymbol{z}^*)$ using Theorem 4.1 so that $T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})$. To evaluate the proposal density $p(\boldsymbol{y}|\boldsymbol{\theta})$, we need to adjust the known density of $\boldsymbol{z}^*$, $p(\boldsymbol{z}^*)$, with the Jacobian of the transformation $h(\cdot)$ to $\boldsymbol{y}$. We choose the distribution of the ~~intermediate~~ data vector $\boldsymbol{z}^*$ to have support on a subset of $\mathbb{R}^n$ so that the transformation to $\boldsymbol{y}$ is one-to-one and the Jacobian, $\left|\frac{\partial h^{-1}(\boldsymbol{y})}{\partial \boldsymbol{y}}\right|$, can be computed from first principles. It is the calculation of this Jacobian which is described in the section. We will repeatedly return to an artificial low-dimensional example in the location-scale setting to help visualize this calculation.

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

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

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

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

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

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

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

Properties C5 and C6 of $\boldsymbol{b}$ are called *regression* and *scale equivariance*, respectively. Properties C7 and C8 of s are called *regression invariance* and *scale equivariance*. Many other estimators satisfy these properties, and our subsequent results apply equally well to them. With more cumbersome statements, the upcoming results can be adjusted to handle a relaxation of C4 that $s(X, \boldsymbol{y}_{obs}) > 0$ and $P(s(X, \boldsymbol{y}) > 0) > 0$.

We have stated that any vector $\boldsymbol{z}^* \in \mathbb{R}^n$ can be transformed to a proposal $\boldsymbol{y}$ satisfying $T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})$. The properties above ensure that this can be done through the mechanism presented in the following theorem. The proof of this and other results appear in the appendix.

Theorem 4.1. *Assume that conditions C4-C8 hold. Then, any vector $\boldsymbol{z}^* \in \mathbb{R}^n$ with conditioning statistic $T(\boldsymbol{z}^*)$ for which $s(X, \boldsymbol{z}^*) > 0$ can be transformed into $\boldsymbol{y}$ with conditioning statistic $T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})$ through the transformation*

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

The mapping described in the theorem is many-to-one in general. Its range is the collection of data sets which match the observed summary statistic:

$$\mathcal{A} = \{\boldsymbol{y} \in \mathbb{R}^n | T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs})\}. \tag{10}$$

To see this, note that the theorem shows that any $\boldsymbol{y}$ expressed as in the theorem is an element of $\mathcal{A}$. Further, any $\boldsymbol{y} \in \mathcal{A}$ can be expressed as in the theorem by simply replacing $\boldsymbol{z}^*$ with $\boldsymbol{y}$. That is, any vector already in $\mathcal{A}$ requires no shifting or scaling. 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. The form of the set is determined by the statistic $T(\cdot)$. Figure **??** provides an artificial low-dimensional visualization of such a set $\mathcal{A}$ with a location-scale model. In the figure, $n = 3$, 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 $T(\boldsymbol{y}_{obs}) = (0, 1)$ and is 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 general, this set may be compact and given by a closed curve, as in the figure, or it may be unbounded, depending on the choice of $T(\boldsymbol{y})$. The other sets labeled in the figure will be described shortly.

The set $\mathcal{A}$ typically does not lie in a linear space of dimension $n-p-1$, and so we must account for both the many-to-one nature of the mapping and a Jacobian when deriving the proposal density. We handle the first point by sampling the initial $\boldsymbol{z}^*$ from a reduced $n-p-1$ dimensional space which, through the scaling and shifting of the above theorem, maps uniquely to a point in $\mathcal{A}$. This reduced space is chosen so that the range of the map is the entirety of $\mathcal{A}$. The Jacobian does not cancel in (7), since the scaling depends on the initial proposal.

The reduced space from which the initial sample is taken is simply the unit sphere ~~restricted to~~ in the orthogonal complement of the column space of the design matrix (i.e., the least squares residual space). To help understand the density of the proposal derived from the transformation of this initial sample we introduce the following notation. Denote the column space of the design matrix X by $\mathcal{C}(X)$ and its orthogonal complement by $\mathcal{C}^{\perp}(X)$. ~~This~~ The latter is the least squares residual space which we will often refer to as the 'deviation space'. The projection of the set $\mathcal{A}$ onto the deviation space is

$$\Pi(\mathcal{A}) = \{\boldsymbol{z} \in \mathbb{R}^n | \ \exists \ \boldsymbol{y} \in \mathcal{A} \ s.t. \ \boldsymbol{z} = Q\boldsymbol{y}\} \tag{11}$$

where Q is the projection matrix onto $\mathcal{C}^{\perp}(X)$. Explicitly, $Q = I - H$ with $H = XX^{\top}$ where we assume, without loss of generality following condition C1, that the columns of X form an orthonormal basis for $\mathcal{C}(X)$ (i.e., $X^{\top}X = I$). It will also be helpful at times to write $Q = WW^{\top}$ where the columns of W form an orthonormal basis for $\mathcal{C}^{\perp}(X)$. This set has been introduced because we will first transform the initial distribution on the sphere to the distribution on this set. Then we will transform to the ~~sample space~~ set $\mathcal{A}$.

Returning to the artificial example, Figure **??** depicts $\Pi(\mathcal{A})$ as well as the sphere from

which we obtain the initial sample. The column vector $X = \mathbf{1}$ spans $\mathcal{C}(X)$ and is shown as a reference. The triangle with bowed sides in dark blue is $\Pi(\mathcal{A})$, the projection of $\mathcal{A}$ onto the plane orthogonal to $\mathbf{1}$ (i.e., $\mathcal{C}^{\perp}(X)$). The unit sphere from which the initial sample is taken is depicted as the red circle in this plane. In general, this is the unit sphere in the $n - p$ dimensional space $\mathcal{C}^{\perp}(X)$.

For the transformation of the initial proposal $\boldsymbol{z}^*$ on the surface of the sphere to $\boldsymbol{y} \in \mathcal{A}$, we first move to a point on $\Pi(\mathcal{A})$ through a simple scaling of $\boldsymbol{z}^*$. This is followed by undoing the projection with a move from $\Pi(\mathcal{A})$ to its (unique) preimage on $\mathcal{A}$. Together, these two steps correspond to the transformation in Theorem 4.1. The introduction of the sphere in $\mathcal{C}^{\perp}(X)$ as the initial proposal surface along with properties C5-C8 ensure that the mapping is 1-1. In particular, property C8 ensures the scaling to be unique and C7 implies that the scale statistic is unchanged when undoing the projection. Property C5 ensures the uniqueness of undoing the projection. The general proposal strategy is summarized as follows:

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

2. Calculate the Jacobian of the transformation in Theorem 4.1 in two steps.

   (a) Scale from unit sphere to $\Pi(\mathcal{A})$: $\boldsymbol{z} = \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)} \boldsymbol{z}^*$

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

4.2.2 Evaluation of the proposal density

Calculation of the appropriate Jacobian of the transformation is absolutely vital and also non-trivial. Writing the transformation from the unit sphere in deviation space to $\mathcal{A}$ in two steps facilitates calculation of the Jacobian in two steps as written above. We now explain each step, keeping in mind that the Jacobian of a transformation is simply the ratio of infinitesimal volumes along the tangents of the domain and range of the transformation.

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

The first step is constrained to $\mathcal{C}^{\perp}(X)$ where the unit sphere is transformed to $\Pi(\mathcal{A})$. We

further break this ~~piece~~ step into ~~in two steps~~ into two substeps: first, the distribution on the unit sphere is transformed to that along a sphere of radius $r = \|\boldsymbol{z}\| = s(X, \boldsymbol{y}_{obs})/s(X, \boldsymbol{z}^*)$. This contributes a factor of $r^{-(n-p-1)}$ to the Jacobian. Second, the new sphere is then deformed to $\Pi(\mathcal{A})$. This deformation contributes an attenuation to the Jacobian equal to the ratio of infinitesimal volumes in the tangent spaces of the sphere and $\Pi(\mathcal{A})$ at $\boldsymbol{z}$. Restricting ourselves to the $n-p$ dimensional space $\mathcal{C}^{\perp}(X)$, this ratio is the cosine of the angle between the normal vectors of the two sets at $\boldsymbol{z}$. The normal to the sphere is its radius vector $\boldsymbol{z}$. The normal to $\Pi(\mathcal{A})$ is given in the following lemma.

Lemma 4.2. *Assume that conditions C1-C2, C4, and C7 hold. Let $\boldsymbol{y} \in \mathcal{A}$. Let $\nabla s(X, \boldsymbol{y})$ denote the gradient of the scale statistic with respect to the data vector evaluated at $\boldsymbol{y}$. Then $\nabla s(X, \boldsymbol{y}) \in \mathcal{C}^{\perp}(X)$ and is normal to $\Pi(\mathcal{A})$ at $\boldsymbol{z} = Q\boldsymbol{y}$ in $\mathcal{C}^{\perp}(X)$.*

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 illustrated in Figure **??** for the toy location-scale example. The figure ~~only~~ pictures only the deviation space which in this case is a plane. The original unit sphere (here, the solid circle) is first 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. As stated above, the Jacobian is the ratio of infinitesimal lengths along these tangent vectors. This is the same as the ratio of the infinitesimal lengths of the normal vectors (i.e., $\cos(\gamma)$). This generalization extends to higher dimensions where the 'lengths' along the tangent vectors become volumes along the tangent spaces.

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

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

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

Lemma 4.3. *Assume that conditions C1-C5 and C7-C8 hold. Then the $p+1$ gradient vectors $\nabla s(X,\boldsymbol{y}), \nabla b_1(X,\boldsymbol{y}), \ldots, \nabla b_p(X,\boldsymbol{y})$ form a basis for $\mathcal{T}_y^{\perp}(\mathcal{A})$ with probability one.*

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

Lemma 4.4. *Assume that conditions C1-C5 and C7-C8 hold. Then the $n \times (n-p-1)$ dimensional matrix $P = QA$ is of full column rank.*

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

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

where $r = \mathrm{rank}(P) = n-p-1$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ are the singular values of P (e.g., Miao and Ben-Israel (1992)). Combining Lemmas 4.3 and 4.4 above leaves us with the following result concerning the calculation of the desired Jacobian.

Theorem 4.5. *Assume that conditions C1-C5 and C7-C8 hold. Then the Jacobian of the transformation from the distribution along $\Pi(\mathcal{A})$ to that along $\mathcal{A}$ is equal to the volume given in* (13).

The proposal density

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

Theorem 4.6. *Assume that conditions C1-C8 hold. Let* $\boldsymbol{z}^*$ *be sampled on the unit sphere in* $\mathcal{C}^{\perp}(X)$ *with density* $p(\boldsymbol{z}^*)$*. Using the transformation of* $\boldsymbol{z}^*$ *to* $\boldsymbol{y} \in \mathcal{A}$ *described in Theorem 4.1, the density of* $\boldsymbol{y}$ *is*

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

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

In practice, computing A directly to find P and $\mathrm{Vol}(P)$ is computationally intensive as it involves orthogonalization of n vectors in n-dimensional space. To find a matrix A, supplement B with a set of n linearly independent columns on the right, and apply Gram-Schmidt orthonormalization to the matrix. This algorithm is infeasibly slow when n is large because it is $\mathcal{O}(n^3)$ and must be repeated at each iterate of the MCMC when a complete data set is drawn. Fortunately, we can make use of results related to *principal angles* found in Miao and Ben-Israel (1992) to compute the volume in (13) using B and an orthonormal basis for $\mathcal{C}(X)$ (The definition of principal angles can be found in the cited text). Recall, B is constructed by orthogonalization of a basis for $\mathcal{T}_y^{\perp}(\mathcal{A})$. Since this space is of dimension $p+1$, applying Gram-Schmidt to find the orthonormal basis is much faster, the algorithm is $\mathcal{O}(np^2)$, and there is a considerable reduction in computational burden when $n \gg p$. Further, the singular values of $P = QA$ are also the singular values of $W^{\top}A$ (recall, $Q = WW^{\top}$), which can be easily obtained through B. The following corollary formally states how computation of A can be circumvented.

Corollary 4.7. *Let* U *be a matrix whose columns form an orthonormal basis for* $\mathcal{C}(X)$*. Then the non-unit singular values of* $U^{\top}B$ *are the same as the non-unit singular values of* $W^{\top}A$*.*

Thus, we can compute $\mathrm{Vol}(P)$ by finding the singular values of $U^{\top}B$, reducing the compu-

tational burden substantially.

5 Applications

We illustrate the methods developed in Sections 2 and 4 with a pair of regression models for data from Nationwide Insurance Company, which concern prediction of the performance of insurance agencies. The data contain outliers and are subject to model misspecification. In particular, a group of the data do not follow the same generative process as the data of interest. It would be extremely challenging to model some features of the data. In our analysis, we follow the standard practice when demonstrating the benefits of robust methods. We work with a naive model for the data which ignores certain features of the problem. We do this both to create a situation where all can agree that the model for the full data $\boldsymbol{y}$ is imperfect and to preserve the confidentiality of selected aspects of modelling done by Nationwide. We wish to provide inference for the 'good' portion of the data. The two models we fit either treat the analysis as a single regression or as a collection of related regressions. Details of the models, prior distributions, and conditioning statistics are given in the next two subsections.

5.1 Nationwide Data

The Nationwide Insurance Company sells many of its insurance policies through agencies which provide direct service to policy holders. The contractual agreements between Nationwide and these agencies vary. Of major interest to Nationwide is the prediction of future performance of agencies where, for our purposes, performance is measured by the total number of households an agency services ('household count'). A serviced household is one in which at least one person living at that residence has at least one policy written through the agency. We used data from previous years to build a model to forecast future household count. In particular, we use agency characteristics, as measured during a single month in 2010, to predict household counts in the corresponding month in 2012. The characteristics used are household count and two measurements of agency size/experience. The

two measurements of agency size/experience are, roughly, the number of employed persons at the agency and the length of time the agency has been affiliated with Nationwide. The household counts (response and predictor) have been square rooted to stabilize variance. The data are proprietary, and to mask them all variables have been individually centered and scaled and identifiers (agency/agent names and state labels) have been removed. All subsequent analysis is done on this scale. As an exploratory view, a plot of the square root of count in 2012, against that in 2010 is shown in Figure **??**. The different colors represent the varying contractual agreements as they stood in 2010. 'Type 1' agencies are of special interest. Among the open agencies, a strong linear correlation exists. The specific linear relationship depends on agency type. The data are characterized by a large number of agencies which were open in 2010 but closed sometime before 2012, as represented by the horizontal band at 0. We use these data as a test bed for our techniques, fitting models that do not account for agency closures or contract type. Our expectation is that ~~the incomplete~~ restricting the likelihood will facilitate prediction for the good part of the data.

5.2 Regression model

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

$$\boldsymbol{\beta} \sim N_p(\boldsymbol{\mu}_0, \Sigma_0); \quad \sigma^2 \sim IG(a_0, b_0); \quad y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \ i = 1, \ldots, n, \tag{15}$$

where $\boldsymbol{\beta}$ is a four dimensional vector ($p = 4$) of regression coefficients for the intercept, square root of count in 2010, and the two size/experience measures, and y_i is the square rooted household count in 2012 for the i^{th} agency with covariate vector $\boldsymbol{x}_i$. Although the mean of covariates and response have been removed, we include the intercept as fitting is done on a holdout set to evaluate predictive performance. The hyper-parameters $a_0, b_0, \boldsymbol{\mu}_0$ and Σ_0 are all fixed and set from a robust regression fit to the data from the time period two years before. Letting $\hat{\boldsymbol{\beta}}$ represent the estimate of the coefficients from this fit we set $\boldsymbol{\mu}_0 = \hat{\boldsymbol{\beta}}$ and $\Sigma_0 = n \cdot \text{var}(\hat{\boldsymbol{\beta}})$ where here n is the sample size of the prior data set. The

hyperparameters for σ^2 are set so that the prior mean is s^2, the estimated variance from the robust regression, and the spread of the prior covers the range of plausible values with high probability. All values are then transformed appropriately to match the current scale of the data. In the end we take $\boldsymbol{\mu}_0 = (0.18, 0.81, 0.01, -0.02)^\top$ and set the mean of σ^2 to 0.014 and standard deviation to 0.033.

We compare four Bayesian models: the standard Bayesian normal theory model, two restricted likelihood models, both with simultaneous M-estimators, and a heavy-tailed model. For the restricted likelihood methods we use the same simultaneous M-estimators as in Section **??** adapted to linear regression. The heavy-tailed model replaces the normal sampling density in (15) with a t-distribution with $\nu = 3$ degrees of freedom. We also fit the corresponding classical robust regressions and a least squares regression.

5.2.1 Method of model comparison

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

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

The trimmed evaluation proceeds as follows in our context. The evaluation function for case i in the hold-out data is the log predictive density, say $\log(f(y_i))$, with the conditioning

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

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

where f_A corresponds to the predictive distribution under the method "A" being scored. In other words, the $[\alpha M]$ observations with the smallest values of $\log(f_b(y))$ are removed from the validation sample and all of the methods are scored using only the remaining $M - [\alpha M]$ observations. This process is advantageous to the base method. A method that performs poorly when it is the base method is discredited. For a complete evaluation, we allow each method to appear as the base method. For brevity, we present only a selection of results in our subsequent analyses.

5.2.2 Comparison of predictive performance

Model performance is assessed using the mean and standard deviation of the TLM across 100 different splits into training and validation samples. First, we include all observations in each validation sample to calculate TLM for each split. We then repeat the evaluation using only certain subsets of the validation sample that are of special interest. Subsets include open agencies, open 'Type 1' agencies, and 'Type 1' agencies. For brevity, we include results for the 'Type 1' agencies only. As noted, assessing model predictions on this set of agencies is of special interest to the company. A range of training sample sizes was used and we include results from $n = 25, 100, 1000$, and 2000 out of a total of 3180 agencies. The trimming fraction, α, ranges from 0 to 0.3. A classical robust regression to the prior

data assigns zero weight to around 16% of observations; in essence removing these from the analysis. This informed the range of trimming fractions chosen. In practice, we would set α slightly larger than 0.16.

Model evaluation for 'Type 1' agencies is shown in Figure **??** for training sample sizes $n = 25, 100$, and 1000. The t-model is used as the base method to compute TLM. The models pictured are: classical robust regression with Tukey's ψ function (rlm-T), restricted likelihood with Tukey ψ (restr.-T), classical robust regression with Huber's ψ function (rlm-H), restricted likelihood with Huber's ψ (restr.-H), and the thick tailed t-model (t). The normal theory models perform poorly due to the numerous outliers and are left out of the figures. Appearing in the figures are the mean TLM across validations set for each model and each trimming fraction, α (along the x-axis). The error bars depicted are one standard deviation of the TLM above and below the mean. The range of the vertical axis is chosen to enhance important features and as a result, some evaluation measures extend below this range. In particular, the restricted likelihood methods perform poorly if no trimming is done; reflecting that these methods are not intended to fit well to outliers. Recall that we expect about 15-16% outliers in the validation sets, thus trimming fractions slightly larger than this amount are needed in order to assess fits to the 'good' data. For $n = 25$, the thick tailed model prevails across trimming fractions, although less so for $\alpha \geq 0.15$. For sample sizes as low as $n = 100$, the restricted likelihood methods outperform the heavy-tailed model with the Tukey version performing the best. The stronger performance of restricted likelihood based on Tukey's method and the t model is to be expected, as many of the residuals are so extreme that trimming is better than winsorizing (as Huber's method effectively does). As expected, with enough data, the Bayesian methods and their classical counterparts perform similarly, although there is a persistent slight edge in favor of the Bayesian restricted likelihood methods. We attribute this advantage to the weakly informative prior distribution which pulls the estimates slightly toward better values. The similarity occurs as early as $n = 100$.

5.3 Hierarchical regression model

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

$$\begin{aligned} &\boldsymbol{\beta} \sim N_p(\boldsymbol{\mu}_0, a\Sigma_0); \quad \boldsymbol{\beta}_j \overset{iid}{\sim} N_p(\boldsymbol{\beta}, b\Sigma_0); \quad \sigma_j^2 \sim IG(a_0, b_0); \\ &\boldsymbol{y}_{ij} = \boldsymbol{x}_{ij}^\top \boldsymbol{\beta}_j + \epsilon_{ij}, \quad \epsilon_{ij} \overset{iid}{\sim} N(0, \sigma_j^2), \ i = 1, \ldots, n_j, \ j = 1, \ldots, J \end{aligned} \tag{16}$$

where y_{ij} represents the i^{th} observation in the j^{th} state, n_j is the total number of agencies in each state, and J is the number of states. $\boldsymbol{x}_{ij}$ is a four dimensional vector comprised of the same covariates as above. $\boldsymbol{\beta}_j$ represents the individual regression coefficient vector for state j. We match this model to the non-hierarchical model in several ways. First, $\boldsymbol{\mu}_0$, Σ_0, a_0, and b_0 are fixed as before. We constrain $a + b = 1$ in an attempt to partition the total variance between the individual $\boldsymbol{\beta}_j$'s and the overall $\boldsymbol{\beta}$. We take $b \sim \text{beta}(v_1, v_2)$. Using the previous data set, we assess the variation between individual estimates of the β_j to set v_1 and v_2 to allow for a reasonable amount of shrinkage. To allow for dependence across the σ_j^2 we first take $(z_1, \ldots, z_J) \sim N_J(\mathbf{0}, \Sigma_\rho)$ with $\Sigma_\rho = (1-\rho)I + \rho\mathbf{1}\mathbf{1}^\top$. Then we set $\sigma_j^2 = H^{-1}(\Phi(z_j))$ where H is the cdf of an $IG(a_0, b_0)$ and Φ is the cdf of a standard normal. This results in the specified marginal distribution, while introducing correlation via ρ. We assume $\rho \sim \text{beta}(a_\rho, b_\rho)$ with mean $\mu_\rho = a_\rho/(a_\rho + b_\rho)$ and precision $\psi_\rho = a_\rho + b_\rho$. The parameters μ_ρ and ψ_ρ are given beta and gamma distributions, respectively. We fix the parameters of these distributions by again considering fits to individual states from the previous data set. More precise details on setting v_1, v_2 and the the priors on μ_ρ and ψ_ρ are given in the appendix. We note that we tried a range of other fixed hyper-parameters resulting in negligible differences in the results.

Using the same techniques as in the previous section, we fit the normal theory hierar-

chical model above, a thick tailed t version with $\nu = 3$ d.f., and two restricted likelihood versions (Huber's and Tukey's) of the model. For the incomplete restricted methods, we condition on robust regression estimates fit separately within each state. We also fit classical robust regression counterparts and a least squares regression separately within each state.

We digress briefly to note that for the restricted likelihood methods no additional computational strategies outside of those discussed in Section 4.2 are needed to fit the hierarchical models described here. Since we condition on statistics which are computed within each state, the model's conditional independence between the states allows the data augmentation described earlier to be performed independently within each state. Updates of hyperparameters follow conventional MCMC procedures. We note that different types of statistics could be chosen for each state, if desired, allowing for a large amount of flexibility.

Selected results for the hierarchical fits appear in Figure **??**. Hierarchical models naturally require more data and so we consider only training sizes of $n = 1000$ and 2000. Again, the t-model is used as the base method for computing TLM. Trimming fractions between 0.15 and 0.3 are displayed, as patterns for smaller trimming fractions are similar to those from the non-hierarchical fits. That is, without sufficient trimming, the Bayesian restricted likelihood fits' evaluation measure is poor. Again, the normal theory fits, both Bayesian and classical, perform poorly and are left out of the figures. We see that the restricted likelihood with Tukey's estimator performs best in each case (assuming sufficient trimming). Huber's version also tops the thick tailed model for $n = 2000$. The Bayesian restricted likelihood fits considerably outperform their respective individual classical robust fits for training size of $n = 1000$. This observation remains, though marginally so, for $n = 2000$. The advantage of the hierarchical models seen here is due to the pooling of information across states, resulting in better predictive performance as compared to both the thick tailed competitor as well the respective classical fits.

5.4 Comparison of hierarchical and non-hierarchical fits

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

We first note that for the non-hierarchical model, there is little difference between mean TLM for $n = 1000$ and $n = 2000$, with the numbers differing only in the third decimal place (see rows 1 and 3 of the table). This is due to the posterior predictive distributions having stabilized. The mean TLMs for the hierarchical model show a greater change with increases of about 0.05 to 0.08 as the training sample size changes from 1000 to 2000 (see rows 2 and 4 of the table). For calibration, the mean TLM for a normal with mean 0.5 and variance 1 is approximately this size when trimming is done under a standard normal base model. Thus, the increase in mean TLM is substantial. We attribute the change for the hierarchical model to the improvement in fits, particularly for states with fewer agencies.

Direct comparison of the hierarchical and non-hierarchical models shows that, for $n = 1000$, the non-hierarchical model has uniformly (for α of interest) better mean TLM (rows 1 and 2). The differences are substantial, and the summaries primarily reflect greater stability of fits on a state-by-state basis under the non-hierarchical model. To a lesser extent, they reflect variation in the evaluation criterion which stems from modest validation sample size, particularly with larger trimming fractions. The trimmed cases are not proportionally distributed across states. The pattern changes for $n = 2000$ (rows 3 and 4), with the hierarchical model showing larger mean TLMs for trimming fractions 0.15 and 0.20. The improvement reflects the ability of the hierarchical model to capture differences in regressions across the states which is realized when the training sample size is large enough. We attribute the better performance of the non-hierarchical model for the largest trimming fractions to variation in the evaluation.

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

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

6 Discussion

Many routine choices in an analysis react to the gap between reality and the statistical model, where a bit of set-up work improves inferential performance. Often, these choices can be recast in the framework of restricted likelihood presented here, lending them more formality and facilitating development of theoretical results. But a much greater benefit of our framework is that it leads us to blend classical estimation with Bayesian methods. Here, we use the likelihood from robust regression estimators to move from prior distribution to posterior distribution. Conditioning on the estimator, the update follows Bayes' Theorem exactly. Computation is driven by MCMC methods, requiring only a modest supplement to existing algorithms. In another context, we might condition on the results of a set of estimating equations, designed to enforce lexical preferences for those features of the analysis considered most important, yet still producing inferences for secondary aspects of the problem. For example, the computational strategies we devised here allow us to apply the method to inference on quantiles of a regression model. In other settings, we envision conditioning on a mix of estimators and some of the observed data.

The framework we propose allows us to retain many benefits of Bayesian methods: it requires a full and complete model for the data; it lets us combine various sources of

information both through the use of a prior distribution and through creation of a hierarchical model; it guarantees admissibility of our decision rules among the class based on the summary statistic $T(\boldsymbol{y})$; and it naturally leads us to focus on predictive inference.

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

This framework opens a number of questions, including a need to revisit such issues as model selection, model averaging for predictive performance, and the role of diagnostics. Perhaps the biggest question is which summary statistic to choose. For this, we recommend a choice based on the analyst's understanding of the problem, model, reality, deficiencies in the model, inferences to be made, and the relative importance of various inferences. In our words, to provide desireable inference, we recommend use of robust and relevant summary statistics in conjunction with Bayesian models.

7 Appendix

7.1 Details of Examples in Section 3

7.1.1 Speed of Light

The method is applied to estimate the speed of light using the classical data set of measurements taken by Simon Newcomb available in the `R` package `MASS`. Previous analyses appear in Stigler (1977), Chan and Rhodin (1980), and Gelman et al. (2004). With parameters $\boldsymbol{\theta} = (\beta, \sigma)$, the full model is $\beta \sim N(23.6, 2.04^2)$, $\sigma^2 \sim IG(5, 10)$, $y_i \overset{iid}{\sim} N(\beta, \sigma^2)$ for $i = 1, 2, \ldots, n = 66$ where y_i denotes the i^{th} measurement of the passage time of light. β is interpreted as the passage time of light with σ^2 representing measurement error. Two

outliers on the low end are a clear indication this model is misspecified.

The summary statistics considered for the restricted likelihood each provide an estimate of location and scale: two versions of simultaneous M-estimators (Huber's and Tukey's for location coupled with Huber's 'proposal 2' for scale (**?**)), LMS (least median squares), and LTS (least trimmed squares) estimators. Tunning parameters for the M-estimators are chosen to achieve 95% efficiency under normality and for comparability, roughly 5% of the residuals are trimmed for LTS. For comparison to a traditional Bayesian outlier approach a heavy-tailed model using Student's t likelihood with $\nu = 5$ degrees of freedom is also fit.

7.2 Proofs

Proof of Theorem 4.1.

Proof.

$$
\begin{aligned}
s(X, \boldsymbol{y}) &= s\left(X, \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)}\boldsymbol{z}^* + X\left(\boldsymbol{b}(X, \boldsymbol{y}_{obs}) - \boldsymbol{b}(X, \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)}\boldsymbol{z}^*)\right)\right) && (17)\\
&= \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)} s(X, \boldsymbol{z}^*) = s(X, \boldsymbol{y}_{obs}), \qquad \text{and} && (18)\\
\boldsymbol{b}(X, \boldsymbol{y}) &= \boldsymbol{b}\left(X, \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)}\boldsymbol{z}^* + X\left(\boldsymbol{b}(X, \boldsymbol{y}_{obs}) - \boldsymbol{b}(X, \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)}\boldsymbol{z}^*)\right)\right) && (19)\\
&= \boldsymbol{b}(X, \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)}\boldsymbol{z}^*) + \boldsymbol{b}(X, \boldsymbol{y}_{obs}) - \boldsymbol{b}(X, \frac{s(X, \boldsymbol{y}_{obs})}{s(X, \boldsymbol{z}^*)}\boldsymbol{z}^*) && (20)\\
&= \boldsymbol{b}(X, \boldsymbol{y}_{obs}) && (21)
\end{aligned}
$$

□

Proof of Lemma 4.2.

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

$$
s(X, \boldsymbol{y}) = s(X, Q\boldsymbol{y} + H\boldsymbol{y}) = s(X, Q\boldsymbol{y}). \tag{22}
$$

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 (22), 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 4.3

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

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

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

$$[XX^\top, WW^\top]^\top [\nabla \boldsymbol{b}(X, \boldsymbol{y}), \nabla s(X, \boldsymbol{y})] = \begin{pmatrix} X & \mathbf{0} \\ WW^\top \nabla b(X, \boldsymbol{y}) & \nabla s(X, \boldsymbol{y}) \end{pmatrix} \tag{23}$$

The last column comes from Lemma 4.2. The matrix $[XX^\top, WW^\top]^\top$ is of full column rank (rank n), and so the rank of $[\nabla \boldsymbol{b}(X, \boldsymbol{y}), \nabla s(X, \boldsymbol{y})]$ is the same as the rank of the matrix on the right hand side of (23). 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 4.4

Proof. P is the projection of the columns of A onto $\mathcal{C}^\perp(X)$. For this to result in a loss of

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

Proof of Corollary 4.7

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

Lemma 7.1. *(Miao, Ben-Israel) Let the columns of $Q_L \in \mathbb{R}^{n\times l}$ and $Q_M \in \mathbb{R}^{n\times m}$ form orthonormal bases for linear subspaces L and M respectively, with $l \leq m$. Let $\sigma_1 \geq \cdots \geq \sigma_l \geq 0$ be the singular values of $Q_M^\top Q_L$. Then $\cos\theta_i = \sigma_i, i = 1, \ldots, l$ where $0 \leq \theta_1 \leq \theta_2 \leq \cdots \leq \theta_l \leq \frac{\pi}{2}$ are the principal angles between L and M.*

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

Theorem 7.2. *(Miao, Ben-Israel) The nonzero principal angles between subspace L and M are equal to the nonzero principal angles between $L^\perp$ and $M^\perp$.*

To establish the corollary, we appeal to Lemma 7.1 and Theorem 7.2. Translating Miao and Ben Israel's notation, we have $M = \mathcal{C}^\perp(X)$, $Q_M = W$, $L = \mathcal{T}_{\boldsymbol{y}}(\mathcal{A})$, and $Q_L = A$. By Theorem 7.2, the nonzero principal angles between $\mathcal{T}_{\boldsymbol{y}}(\mathcal{A})$ and $\mathcal{C}^\perp(X)$ are the same as the nonzero principal angles between $\mathcal{T}_{\boldsymbol{y}}^\perp(\mathcal{A})$ and $\mathcal{C}(X)$. By 7.1, 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

In setting the priors we use the same previous data set used to set the priors for the non-hierarchical model (Section 5.2) and several heuristic arguments. While the analyses in Section 5.3 set the hyper-parameters using what is described here, the results were not sensitive to these choices. This section describes the heuristics used in setting these prior parameters and is given for completeness. Using the previous data set we fit separate (robust) regressions to each state and a regression to the ~~entire~~ entirety of the data at once. Let the estimates for the fits to each state be $\hat{\beta}_1, \ldots, \hat{\beta}_J, \hat{\sigma}_1, \ldots, \hat{\sigma}_J$ and the estimates from the single regression be $\hat{\beta}$ and $\hat{\sigma}$. These are classical robust estimates using Tukey's regression and Huber's scale. Let n_j denote the number of observations in the j^{th} state and set $n = \sum n_j$.

First, consider v_1 and v_2 in the prior $b \sim \text{beta}(v_1, v_2)$. In the hierarchical model (16), $b = 0$ implies all the $\boldsymbol{\beta}_j's$ are equal (no variation between states) and $b = 1$ implies the $\boldsymbol{\beta}_j's$ vary about μ_0 according to $\Sigma_0 = n \cdot \text{var}(\hat{\beta})$ (see Section 5.2). We seek a prior measure for what we think b should be. In other words, how much prior uncertainty should we allow in $\boldsymbol{\beta}$ as opposed to the uncertainty amongst the $\boldsymbol{\beta}_j's$? Using the prior fit, a measure for uncertainty for $\boldsymbol{\beta}$ is $\Sigma_{\hat{\beta}} = \text{var}(\hat{\beta})$, the estimate of the covariance from the single regression. For the $\boldsymbol{\beta}_j's$, take $\delta_j = \hat{\beta}_j - \hat{\beta}$ and set the prior uncertainty to $\Sigma_\delta = n^{-1} \sum n_j \delta_j \delta_j^\top$. Consider the value $g = \left(|\Sigma_\delta| / |\Sigma_{\hat{\beta}}| \right)^{1/p}$. Heuristically, g is measure of the amount of uncertainty between the $\boldsymbol{\beta}_j's$ to the amount of uncertainty in $\boldsymbol{\beta}$. Now in the prior, we heuristically set the uncertainty in the $\boldsymbol{\beta}_j's$ ($b\Sigma_0$) to be approximately equal to $g \cdot \text{var}(\hat{\beta})$. That is, $b\Sigma_0 \approx g \cdot \text{var}(\hat{\beta}) = \frac{g}{n}\Sigma_0$, suggesting $b \approx \frac{g}{n}$. Hence we set $E[b] = \frac{g}{n}$. The precision, $v_1 + v_2$, is set to be relatively high at 20, completing the specification for the prior on b.

In setting the parameters for the beta prior on μ_ρ and gamma prior on ψ_ρ we first take $\hat{z}_j = \Phi^{-1}(H(\hat{\sigma}_j^2))$. As in the prior we assume $(\hat{z}_1, \ldots, \hat{z}_J) \sim N_J(\mathbf{0}, \Sigma_\rho)$ with $\Sigma_\rho = (1-\rho)I + \rho \mathbf{1}\mathbf{1}^\top$ and find the MLE, $\hat{\rho}_{mle}$, and observed inverse Fisher information, $I^{-1}(\rho_{mle})$. The mean of the beta prior on μ_ρ is set to $\hat{\rho}_{mle}$. Its variance is inflated somewhat and set to $2I^{-1}(\hat{\rho}_{mle})$. Since $\text{var}(\rho|\mu_\rho, \psi_\rho) = \mu_\rho(1-\mu_p)/(\psi_\rho + 1)$ we replace μ_ρ with $\hat{\rho}_{mle}$,

$\text{var}(\rho|\mu_\rho, \psi_\rho)$ with $2I^{-1}(\hat{\rho}_{mle})$, and set the mean of the gamma prior on ψ_ρ equal to $\hat{\rho}_{mle}(1-\hat{\rho}_{mle})/(2I^{-1}(\hat{\rho}_{mle})) - 1$. Finally, we arbitrarily set the variance of the gamma prior to be approximately the same as the mean.

References

Albert, J. H. (1988). Computational methods using a Bayesian hierarchical generalized linear model. Journal of the American Statistical Association, 83: 1037–1045.

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

Berger, J. O. (1994). An overview of robust Bayesian analysis. Test, 3:5–124.

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

Chan, L. K. and Rhodin, L. S. (1980). Robust estimation of location using optimally chosen sample quantiles. Technometrics, 22(2):225–237.

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

Doksum, K. A. and Lo, A. Y. (1990). Consistent and robust Bayes procedures for location based on partial information. The Annals of Statistics, 18:443–453.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. Journal of the Royal Statistical Society: Series B, 74:419–474.

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. Journal of the American Statistical Association, 100:680–701.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association, 85:398–409.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). Bayesian Data Analysis. CRC press, 2nd edition.

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

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

Hoff, P. and Wakefield, J. (2013). Bayesian sandwich posteriors for pseudo-true parameters. Journal of Statistical Planning and Inference, 143: 1638–1642.

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

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

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

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

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

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

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

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

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

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. Journal of the American Statistical Association, 103:410–423.

Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. Journal of the American Statistical Association, 89:958–966.

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

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

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

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). Uncertain judgements: eliciting experts' probabilities. John Wiley & Sons.

Pettitt, A. N. (1982). Inference for the linear model using a likelihood based on ranks. Journal of the Royal Statistical Society. Series B, 44:234–243.

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

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

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human Y chromosomes: A study of y chromosome microsatellites. Molecular Biology and Evolution, 16:1791–1798.

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

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

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

Stigler, S. M. (1977). Do robust estimators work with real data? The Annals of Statistics, 5(6):1055–1098.

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

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