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 is ?? this restricted likelihood leads to a formal update of the prior distribution based on the sampling density of the summary statistics.

The remainder of the paper is as follows: Section 2 introduces the Bayesian restricted likelihood and provides context with previous work, Section 3 demonstrates some advantages of the methods on simple examples, and Section 4 details ~~a~~ 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 8 with R code available from the authors. Sections 5 and 6 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 7) provides some final commentary on the new method.

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:

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

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 $y_i \le t_1$, $c(y_i) = t_2$ if $y_i \ge t_2$, and $c(y_i) = y_i$ otherwise. The adjusted update is performed with $L(\boldsymbol{\theta}|c(\boldsymbol{y}))$. Letting $g(t_1|\boldsymbol{\theta}) = F(t_1|\boldsymbol{\theta})$, $g(t_2|\boldsymbol{\theta}) = 1-F(t_2|\boldsymbol{\theta})$, and $g(y|\boldsymbol{\theta}) = f(y|\boldsymbol{\theta})$ for $y \in (t_1,t_2)$, we may rewrite the full likelihood as the product of two pieces

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

Only the first part is retained $_{\wedge}$ the analysis. Several more examples are detailed in Lewis (2014).
in

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:

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

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 insufficient statistic $T(\boldsymbol{y})$ ~~on which to~~ guide d by targeted inference is sound practice. We also expand the class of conditioning statistics in ? which a formal Bayesian update can be achieved. Our methods do not rely on asymptotic properties, nor do they rely on approximate conditioning.

3 Illustrative Examples

Before discussing computational details, the method is applied to two simple examples on well known data sets to demonstrate its effectiveness in situations where outliers are a major concern. The full model in each case fits into the Bayesian linear regression framework discussed in Section 4.

The first example is an analysis of Simon 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} (reco r ded) 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. 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 ~~fit~~ 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.

1987). The full model is a standard normal Bayesian linear regression:

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

where $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$, $\boldsymbol{y}$ is the vector of the logarithm of the number of calls, and X is the $n \times 2$ design matrix with a vector of 1's in the first column and the year covariate in the second. Prior parameters are fixed via a maximum likelihood fit to the first 3 data points. In particular, the prior covariance for β bold 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 μ_0 bold $= (1.87, 0.03)^\top$. 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

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

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

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

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

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

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

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

Properties C5 and C6 of $\boldsymbol{b}$ are called *regression* and *scale equivariance*, respectively. Properties C7 and C8 of s are called *regression invariance* and *scale equivariance*. Many estimators satisfy the above properties, including simultaneous M-estimators (Huber and Ronchetti, 2009; Maronna et al., 2006) for which the R package `brlm` (`github.com/jrlewi/brlm`) is available to implement the MCMC described here. Further software development is required to extend the MCMC implementation beyond these M-estimators. The package also implements the direct computational methods described in Lewis (2014). These methods are effective in lower dimensional problems and were used in both examples in Section 3.

4.2 Computational strategy

The general style of algorithm we present is a data augmented MCMC targeting $f(\boldsymbol{\theta}, \boldsymbol{y}|T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs}))$, the joint distribution of $\boldsymbol{\theta}$ and the full data given the summary statistic $T(\boldsymbol{y}_{obs})$. The Gibbs sampler (Gelfand and Smith, 1990) iteratively samples from the full conditionals 1) $\pi(\boldsymbol{\theta}|\boldsymbol{y}, T(\boldsymbol{y}) = T(\boldsymbol{y}_{obs}))$ 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 ~~this~~ generation.

of θ 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

Do we need these?

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

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 appendix.

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

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

Using the theorem, the general idea is to first start with an initial vector $\boldsymbol{z}^*$ drawn from a known distribution, say $p(\boldsymbol{z}^*)$, and transform via $h(\cdot)$ to $\boldsymbol{y} \in \mathcal{A}$. The proposal density $p(\boldsymbol{y}|\boldsymbol{\theta})$ is then a change-of-variables adjustment on $p(\boldsymbol{z}^*)$ derived from $h(\cdot)$. In general however, the mapping $h(\cdot)$ is many-to-one: for any $\boldsymbol{v} \in \mathbb{R}^n$ and any $c \in \mathbb{R}^+$, $c\boldsymbol{z}^* + X\boldsymbol{v}$ map to the same $\boldsymbol{y}$. This makes the change-of-variables adjustment difficult. We handle this 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 8.1 in the appendix. 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 on $\mathbb{S}$.

2. Set $\boldsymbol{y} = h(\boldsymbol{z}^*)$ and calculate the Jacobian of this transformation in two steps.

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.

5 Simulated Data

We study the performance of restricted likelihood methods in 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:

$$\begin{aligned} &\mu \propto 1,\ \tau^2 \propto \tau^{-2}, \quad \longleftarrow \text{what does "}\propto\text{" sign mean here?} \\ &\theta_i \sim N(\mu, \tau^2),\ \sigma_i^2 \sim IG(a_s, b_s),\ i = 1, 2, \ldots, 90, \\ &y_{ij} \sim N(\theta_i, \sigma_i^2),\ j = 1, 2, \ldots, n_i. \end{aligned} \tag{17}$$

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), a_s and b_s are fixed to a variety of values representing different levels of prior knowledge. For each, we set $b_s = 4a_s c$ resulting in a prior mean for each σ_i^2

group of the k^{th} simulated data set $\boldsymbol{y}_k$ compute:

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

where M indexes the fitting model and $f(\tilde{y}|\theta_i,\sigma^2) = N(\tilde{y}|\theta_i,\sigma^2)$, the normal density function with mean θ_i and variance σ^2, evaluated at $\tilde{y}$. For the Bayesian models $f_i(\tilde{y}|M,\boldsymbol{y}_k) = \int f(\tilde{y}|\theta_i,\sigma_i^2)\pi(\theta_i,\sigma_i^2|M,\boldsymbol{y}_k)d\theta_i d\sigma_i^2$ where $\pi(\theta_i,\sigma_i^2|M,\boldsymbol{y}_k)$ is the posterior for the i^{th} group model parameters under model M for the k^{th} data set. M denotes either the full normal theory model (17) or one of the two restricted likelihood versions, along with specified a_s and c. For the classical robust fits, we set $f_i(\tilde{y}|M,\boldsymbol{y}_k) = N(\tilde{y}|\hat{\theta}_i,\hat{\sigma}_i^2)$ as a groupwise plug-in estimator for the predictive distribution. The classical fits are computed separately for each group with no consideration of the hierarchical structure between the groups. The overall mean $\overline{KL}_{\cdot\cdot}^{(M)} = \frac{1}{90K}\sum_{k=1}^{K}\sum_{i=1}^{90} KL_{ik}^{(M)}$ is used to compare the models, where smaller means correspond to better fits. Sampling variation is summarized with the standard error between the $K = 30$ replicates in the simulation: $SE(\overline{KL}_{\cdot k}^{(M)}) = \sqrt{\frac{1}{K(K-1)}\sum_{k=1}^{K}(\overline{KL}_{\cdot k}^{(M)} - \overline{KL}_{\cdot\cdot}^{(M)})^2}$ where $\overline{KL}_{\cdot k}^{(M)} = \frac{1}{90}\sum_{i=1}^{90} KL_{ik}^{(M)}$.

Figure ~~5~~ 6 displays $\overline{KL}_{\cdot\cdot}^{(M)}$ with error-bars plus/minus one $SE(\overline{KL}_{\cdot k}^{(M)})$ for each $a_s = 1.25, 5, 10$ and $c = 0.5, 1, 2$. The values of a_s and c, do not effect the classical robust linear models. The average KL for the normal theory models ranges from 0.22 to 0.3 which is much worse than the robust methods and hence is left out of the figure. For $c = 0.5$ and $c = 1$, the results favor the restricted likelihood methods with a slight advantage to the use of Tukey's location estimator over Huber's. This is likely due to the fact that Tukey's estimator essentially trims extreme outliers in the estimation procedure while Huber's estimator discounts them (Huber and Ronchetti, 2009).

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

6.1 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:

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

where y_i and x_i are the square rooted household count in 2012 and 2010 for the i^{th} agency, respectively. The hyper-parameters a_0, b_0, μ_0 and σ_0^2 are all fixed and set from a robust regression fit to the corresponding state's data from the time period two years before. Specifically, Let $\hat{\beta}$ and $\hat{\sigma}^2$ be estimates from the robust linear regression of 2010 counts on 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 $\mu_0 = \hat{\beta}$ and $\sigma_0^2 = n_p se(\hat{\beta})^2$ where n_p is the number of agencies in the prior data set and $se(\hat{\beta})$ is the standard error of $\hat{\beta}$ derived from the robust regression. This prior is in the spirit of the Zellner's g-prior (Zellner, 1986; Liang et al., 2008). In general scaling the prior variance $se(\hat{\beta})^2$ by a factor g. $g = n_p$ 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 8).

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 5 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 4.2. We also fit the corresponding classical robust regressions and a least squares regression.

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/holdout samples with training sample sizes taken to be 0 $.25n$ and $0.50n$. Holdout evaluation is done on the remaining (‘Type 1’) samples. For the data augmentation MCMC step under the restricted likelihood models, the acceptance rates range from 0.16 to 0.76 across the states, repetitions, and two versions of the model. The average $TLM_b(A)$ over the $K = 50$ training/holdout samples for the four states and seven methods are shown in Figure 9 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. 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 similarly to our restricted methods for smaller training sample size (25% of the sample). However, the performance is slightly worse 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 10 which shows results for different α for training sample size $0.5n$.

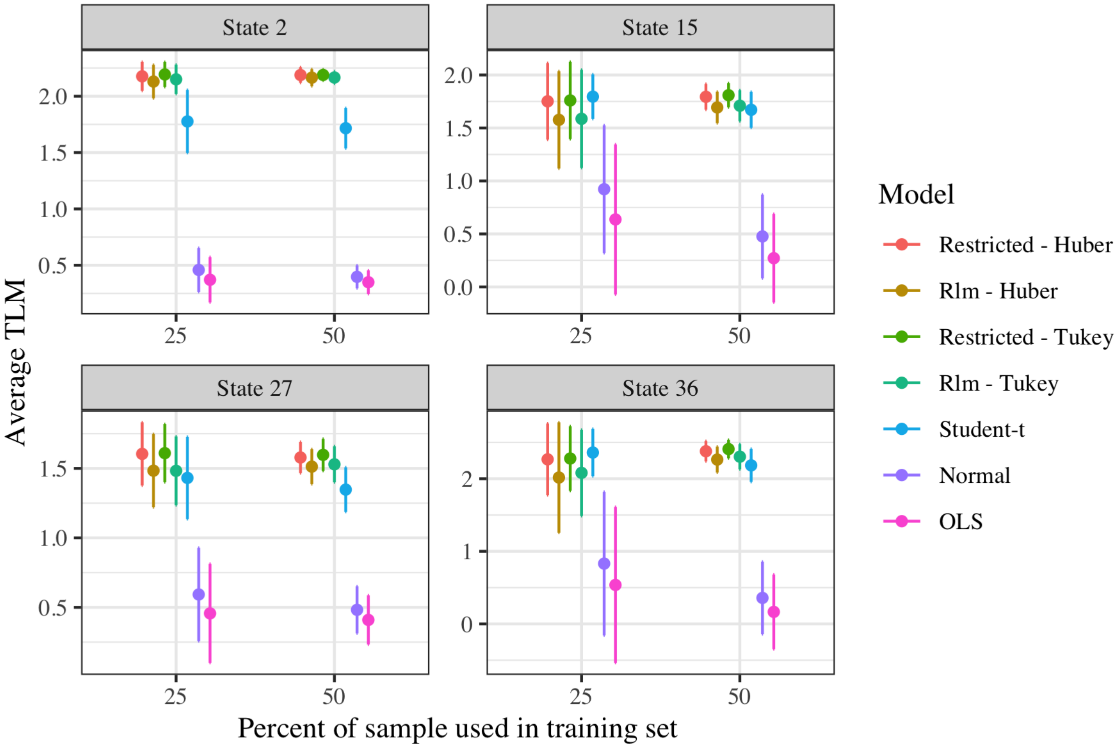

Figure 9: Average TLM plus/minus one standard deviation over $K = 50$ splits into training and holdout samples. The panels are for the different states 2, 15, 27, and 36, with $n = 222, 40, 117$, and 46, respectively. The horizontal axis is the percent of n used in each training set. The color corresponds to the fitting model. Larger values of TLM are better.

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 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.

The parameters μ_0, σ_0^2, a_0, and b_0 are fixed by fitting the regression $y_{ij} = x_{ij}\beta + \epsilon_{ij}$ using Huber's M-estimators to the prior data set from two years before. Using the estimates from this model, we set $\mu_0 = \hat{\beta}$, $\sigma_0^2 = n_p se(\hat{\beta})^2$ ($n_p = 2996$ 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 β_j's and the overall β. We take $b \sim \text{beta}(v_1, v_2)$. Using the prior 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 $\Sigma_\rho = (1 - \rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}^\top$. 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 appendix.

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 resulting in 22 states in total and $n = \sum_j n_j = 3180$ total agencies. For training data we take a stratified (by state) sample of size $3180/2 = 1590$ 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: $TLM_b(A)_j = (M_j - [\alpha M_j])^{-1} \sum_{i=[\alpha M_j]+1}^{M_j} \log(f_A(y_{(i)j}^b))$ where $y_{(1)j}^b, y_{(2)j}^b, ..., y_{(M_j)j}^b$ 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, $f_A(y_{(i)j})^b$ 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 4 and adapted to the hierarchical setting as described in Section 5. For the MH-step proposing augmented data, the acceptance rates for the two restricted likelihood models across all states and repetitions range~~s~~ from 0.24 to 0.74.

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 11 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 errorbars 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 6.1. 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 'good' (i.e., 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 reduce s variance in predictions resulting in smaller error bars. This improvement decreases with α but is still noticeable for $\alpha =$ ~~0.3~~. Both the Tukey and Huber versions perform similarly.

0.2 perhaps?

It is also interesting to examine the results within each state. Figure 12 summarizes $TLM_b(A)_j$ with $\alpha = 0.3$ for each state where the points and errorbars 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 looks similar). The states are ordered along the x-axis according to number of agencies within the state (shown in parentheses). 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).

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. 27

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. 5

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

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. 5

Yuan, A. and Clarke, B. (2004). "Asymptotic Normality of the Posterior Given a Statistic." The Canadian Journal of Statistics, 32: 119–137. 5

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. 26

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