Robust Inference via the Blended Paradigm [*]

John Lewis, Yoonkyung Lee, Steven MacEachern

Department of Statistics, The Ohio State University

Abstract

The *blended paradigm* provides a means of achieving robust inference within the Bayesian framework. We seek to improve inference when the sampling distribution posited for the data does not fully capture the data generating process. This paper focuses on the use of robust estimators of location and scale (e.g., Huber estimators) in a Bayesian context. An underlying model induces a distribution for the estimators. This induced distribution yields a likelihood which is used to update from the prior distribution to the posterior distribution. We empirically show that good choices of robust estimators can produce posteriors that are more concentrated around the target value than those based on the full data, reducing both the bias and variance of the posterior mean. The success of the method is illustrated in simulations and in an application for estimation of the speed of light.

1 Introduction

In the traditional Bayesian framework we posit a sampling distribution, $f(\mathbf{X}|\theta)$, for the data vector, $\mathbf{X} = (X_1, \cdots, X_n)$, given the parameter vector $\boldsymbol{\theta}$. The model is completed with a prior distribution, $\pi(\boldsymbol{\theta})$, for the parameter. Given data $\mathbf{X} = \mathbf{x}$, the posterior distribution

$$\pi(\boldsymbol{\theta}|\mathbf{x}) \propto f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{1}$$

is used to conduct inference.

The Bayesian framework offers a philosophically sound way to combine prior knowledge, incorporated through $\pi(\boldsymbol{\theta})$, with the data. However, deficiencies may arise if we do not fully believe that the sampling distribution, $f(\mathbf{X}|\boldsymbol{\theta})$, accurately describes the data generating process for some value of $\boldsymbol{\theta}$. When faced with such a situation, the typical reaction is to try to come up with a better model. For example, in analyses involving outliers, we could model the process generating the outliers using mixture distributions and subsequently try to identify the outliers (e.g., Justel and Peña (1996)). This requires an explicit model for the outliers, but creating such a model can be difficult. In particular, outliers may not occur in any fixed, systematic fashion, an implicit assumption in a typical model for outliers. We can easily imagine an ever changing, transient system where neither the probability of an erratic measurement nor the distribution of erratic measurements is fixed over time.

In the *blended paradigm* we acknowledge the potential inadequacy of our model and seek to drive the Bayesian update with the portion of the data that is directly related to the

[*] This work has been supported by the Nationwide Center for Advanced Customer Insights (NCACI), by the NSA (H98230-10-1-0202), and the NSF (DMS-1007682 and DMS-1209194).

inferences of interest. Specifically, we focus on a statistic, $T(\mathbf{X})$, that provides a 'robust' summary of the data. By 'robust' we mean that the summary is insensitive to the deficiencies in the model for inferences of interest. The probability model for the data, $f(\mathbf{X}|\boldsymbol{\theta})$, induces a likelihood for the parameter based on the robust summary, $g(T(\mathbf{X})|\boldsymbol{\theta})$. After observing data $\mathbf{X} = \mathbf{x}$, we use this likelihood to pass to the posterior

$$\pi^*(\boldsymbol{\theta}|T(\mathbf{x})) \propto g(T(\mathbf{x})|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{2}$$

which is then used for inference. We refer to $f(\mathbf{X}|\boldsymbol{\theta})$ as the 'full likelihood' and to $g(T(\mathbf{X})|\boldsymbol{\theta})$ as the 'restricted likelihood.' Likewise, we refer to (1) as the 'full posterior' and to (2) as the 'restricted posterior.'

We seek to obtain better inference through a wise choice of $T(\mathbf{X})$. When we wish to conduct inference which is insensitive to outliers, classical robust 'M-estimators' are an attractive choice. 'M-estimators', which include the class of MLEs, have been well studied and have many nice properties, including asymptotic normality under standard regularity conditions (Huber, 1974). There exists a rich history of robust estimation, especially in the presence of large outliers. Techniques include various trimming methods, modeling the outliers with thick-tailed distributions or mixture distributions, and employing rules of thumb for cleaning the data. We refer the interested reader to Stigler (1973) for a concise overview of the history of robust estimation during the late 19^{th} and early 20^{th} centuries. There is also a vast history of debate on which robust estimators are the most reliable and whether or not they are even useful. Stigler (1977) conducts a study to compare the performance of several robust estimators on twenty different historical data sets. The suggestion is made that the 10% trimmed mean performs well on a wide variety of data sets and that iterative methods, at the time, were rarely worth the effort needed to calculate them. Modern day computational abilities render the latter argument essentially obsolete. Rocke et al. (1982) compare variance estimates of several robust estimators and conclude that for greatest effectiveness, an analyst should severely trim the data (up to 40-50%) or employ robust 'M-estimators.' Most of these earlier comparisons used data from the physical sciences, such as measurements of the speed of light and the density of earth. In more recent applications, Bollinger and Chandra (2005) argue that results on robust estimators in the physical sciences should not necessarily be applied to the social sciences. In the social sciences, some error processes can be attributed to social processes rather than physical processes. For example, in surveys designed to make inference on wages, responses from employers and employees may not agree. This can tilt the choice of robust estimator. Angrist and Krueger (1999) argue that trimming the data can be beneficial when the extremes obviously do not resemble the true values while Winsorizing the data is useful if the extremes are considered to be exaggerations of the truth. We do not seek to debate the merits of different robust estimation procedures nor do we attempt to make a case for robust estimators of a specific form. Instead, we set out to lay a foundation for using classical robust estimators in the Bayesian framework.

In this article, we concentrate on showing the advantages of the *blended paradigm* in the univariate location and scale setting where the data are contaminated with outliers. We use two forms of $T(\mathbf{X})$: (1) a set of well chosen order statistics, and (2) Huber estimators of location and scale, a class of robust 'M-estimators'. Both of these estimators downweight the effect of outliers. We find that this downweighting carries over to the posterior distribution. We show empirically that our methods produce posteriors that are more concentrated around the target value, reducing the bias and variance of the posterior mean as a location estimator. Additionally, we note that the restricted posterior variance is smaller than the full posterior variance. The rest of the paper is organized as follows: In Section 2 we conduct simulation studies to demonstrate the ability of the *blended paradigm* to reduce bias and variance. In Section 3 we apply the method to a well known data set collected by Simon Newcomb in 1882 in order to estimate the speed of light. This data set contains two outlying observations and offers a good platform to study the properties of our method. Finally, in Section 4 we provide a discussion of the advantages and disadvantages of the method.

2 Proof of Concept–Outliers

In this section we use simulations to study the *blended paradigm* in a univariate setting when the data are contaminated with outliers. In the simulations, data are generated independently from the following mixture distribution

$$h(x|\theta,\sigma^2) = \frac{(1-p)}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right) + \frac{p}{10\sigma}\phi\left(\frac{x-\theta}{10\sigma}\right) \tag{3}$$

where $\phi(\cdot)$ is the pdf of the standard normal distribution. Thus, on average, a proportion of $1-p$ of the data come from a normal distribution with mean θ and variance σ^2. The rest of the data come from a normal distribution with the same mean, but a much larger variance of $100\sigma^2$. For the sake of this study, we set the values of the parameters to be $\boldsymbol{\theta} = (\theta, \sigma^2) = (0, 1)$ and set $p = 0.2$. The parameter θ is the parameter of interest, while σ^2 is a nuisance parameter. Although a nuisance parameter, we shall see that how σ^2 is handled has a major impact on estimation of θ. A smoothed distribution of a typical sample of size $n = 100$ from (3) with the specified parameters is displayed in Figure 1.

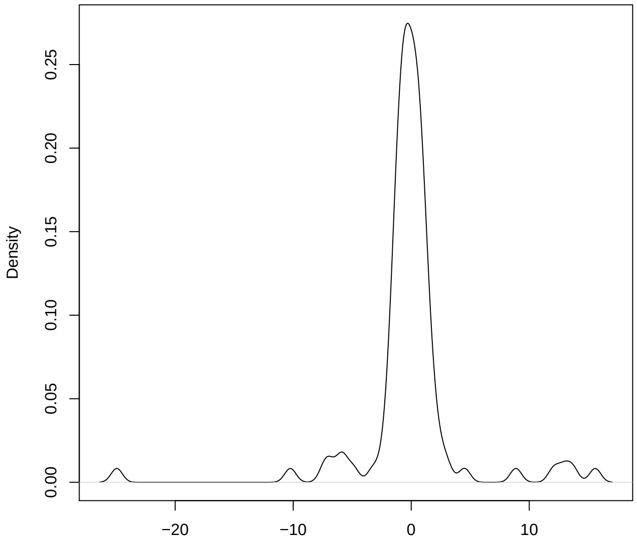


Figure 1: Distribution of a typical sample of size $n = 100$ generated from (3) with $(\theta, \sigma^2) = (0, 1)$ and $p = 0.2$. The samples are characterized by numerous, sizeable outliers.

We (incorrectly) model the 100 independent draws $X_1, \cdots, X_{100}$ from (3) with a normal theory Bayesian model. In particular, the Bayesian model to be used for analysis is

$$\begin{gathered}\sigma^2 \sim IG(\alpha, \beta) \\ \theta \sim N(\mu, \tau^2) \\ X_1, \cdots, X_{100} \overset{\text{iid}}{\sim} N(\theta, \sigma^2)\end{gathered} \tag{4}$$

where $N(a, b)$ represents a normal distribution with mean a and variance b and $IG(c, d)$ represents an inverse gamma distribution with shape and scale parameters c and d, respectively. The hyperparameters $\mu,\ \tau^2,\ \alpha$, and β are set to $5, 4, 5$, and 5, respectively. Notice that the prior for θ is centered around 5 and is off center from the true value of 0.

In this set up, the 'full likelihood,' $f(\mathbf{X}|\boldsymbol{\theta})$, is the product of pdfs of normal densities with mean θ and variance σ^2. The 'restricted likelihood' is induced by the choice of $T(\mathbf{X})$ and the 'full likelihood.'

In Section 2.1 we define $T(\mathbf{X})$ to be a set of order statistics. The restricted likelihood is then the likelihood of this set of order statistics, which is analytically known. Standard MCMC algorithms are used to draw samples from the full and restricted posteriors. In particular, a Metropolis-Hastings algorithm is developed to sample from the restricted posterior. We use a normal proposal, centered at the value of θ at the previous iterate with a standard deviation of 1, for the location parameter θ. For σ^2, we propose from an inverse gamma distribution with shape equal to the prior parameter $\alpha = 5$ and scale $4\sigma^2_{cur}$, where σ^2_{cur} is the value of σ^2 from the previous iterate of the Markov chain. This results in proposal with mean equal to σ^2_{cur}. These proposals result in adequate mixing of the chain. To sample from the full posterior, this algorithm can easily be adjusted and results in good mixing. Another option for the full posterior is to use Gibbs sampling, as the full conditional distributions are available in this case. We compare estimates of the posterior means of the parameters under the full and restricted posteriors over many data sets simulated from (3).

In Section 2.2 we define $T(\mathbf{X})$ to be Huber's robust estimators of location and scale. See Huber (1974) for an overview of these statistics. In this case, the restricted likelihood is not available in closed form and MCMC algorithms are nontrivial. For the purpose of this article, we resort to a direct sampling algorithm to sample from the restricted posterior (2). This algorithm is described in more detail below. We compare the full and restricted posteriors for the data $\mathbf{X} = (X_1, \cdots, X_{100})$ displayed in Figure 1.

2.1 Order Statistics

For the restricted likelihood we choose $T(\mathbf{X}) = (X_{(31)}, \cdots, X_{(70)})$, the middle order statistics from 31 to 70. The motivation behind using these statistics is to down-weight the effect of the outliers. The restricted likelihood follows from the joint pdf of the order statistics:

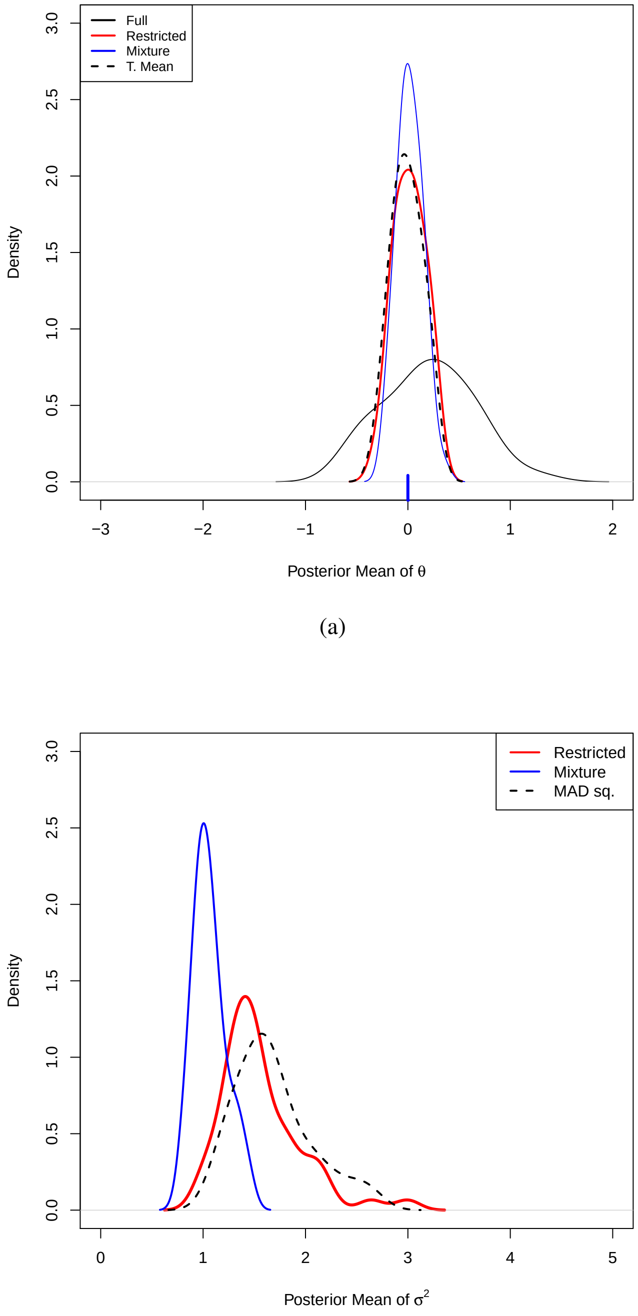
$$g(T(\mathbf{X})|\theta, \sigma^2) = \frac{100!}{30!30!} F(X_{(31)})^{30}[1 - F(X_{(70)})]^{30} \prod_{i=31}^{70} f(X_{(i)}) \tag{5}$$

where $F(\cdot)$ and $f(\cdot)$ are the cdf and the pdf of the normal distribution with mean and variance θ and σ^2, respectively. We develop a simulation to study the distribution of the posterior means of θ and σ^2 under both the full and restricted likelihood models when data come from the mixture distribution (3). The details of the simulation are as follows:

1) simulate a data set of size $n = 100$ from (3). 2) Sample both the full and restricted posteriors, using the Metropolis-Hastings algorithm described above, to obtain chains of length $10,000$ after burn-in. 3) Calculate the sample means for each parameter to obtain estimates of the means under both the full and restricted posteriors. We repeat steps 1-3 of the simulation 50 times to obtain a collection of estimates of the means for each parameter under both the full and restricted posteriors. As a reference, we also estimate the posterior means under the true mixture model (3), with $p = 0.2$ taken to be fixed and known. Figure 2(a) shows the distribution of the 50 estimated posterior means for θ under each model. For comparison to a classical estimate, we also plot the distribution of the 30% trimmed mean. The trimmed mean is the sample mean after 30% of the data is trimmed from each end of the data. The 30% trimming is chosen to align closely with the set of order statistics. When using the restricted likelihood, we can see in Figure 2(a) that we have a reduction in both bias and variance of the posterior mean estimators when compared to the full likelihood estimators. Further, the distribution of the posterior means under the restricted likelihood more closely aligns to that of the posterior means under the true mixture likelihood. Lastly, the distribution of the trimmed means closely resembles the distribution of the means under the restricted posterior. In the analysis, we expect the offset prior distribution to hinder the performance of the estimator. The sample mean and variance of the trimmed means are -0.009 and 0.0225, respectively. Those for the restricted posterior means are 0.016 and 0.0239, suggesting comparable variance and only a slight upward bias induced by the (poor) prior information.

The reduced variance in the means of θ under the restricted compared to the full posteriors is related to the parameter σ^2. Under the true mixture distribution (3), $\sigma^2 = 1$. Under the full likelihood (4), σ^2 represents the true variance of the data, 20.8. The estimates for the posterior means of σ^2 under the true likelihood and the restricted likelihood are displayed in Figure 2(b). Also displayed is the distribution of the squared (scaled) median absolute deviation (MAD^2). The median absolute deviation is the median of the absolute deviations from the median. For a consistent estimate of σ under normality, this is scaled by a factor of 1.4826. MAD^2 is then an estimate of σ^2. The estimates under the full likelihood are roughly centered around 20.8 and are much more variable. They range from 6 to 37 with a sample variance of 51 and hence are not shown in the figure. We see that the posterior mean of σ^2 in the restricted setting is more closely tied to its counterpart in the true mixture model than in the full model. Viewed through the lens of the model, σ^2 is apparently smaller under the restricted likelihood than under the full likelihood. A smaller value of σ^2 suggests that the data contain more information about θ, and this translates into a sharper restricted likelihood and a sharper restricted posterior. The increased information moves the posterior toward the the data-based estimate $T(x)$. This helps to explain the reduction in variance of the estimated posterior means for the location parameter. Again, we also see a close relationship between the classical robust estimator MAD^2 and the posterior mean of σ^2 under the restricted model.

We note that using the middle order statistics from 31 to 70 is quite arbitrary and so a natural question to ask is how the trimming value affects the inference? Letting $T(\mathbf{X}) = (X_{(k+1)}, \cdots, X_{(n-k)})$, we repeat the simulations (using the same simulated data) for several trimming values k. The posterior mean estimates of θ for $k = 10, 20$, and 35 are added to Figure 2(a) and displayed in Figure 3(a). We see that the choice of k matters somewhat. On average, the data contain 20% outliers and so we should trim at least 10% from each tail as is suggested by the larger variance of the estimates for $k = 10$. As we trim fewer values from each side, we move toward the traditional Bayesian analysis given



(a)

(b)

Figure 2: Distribution of the 50 estimated posterior means for θ (a) and σ^2 (b) under the full and restricted posterior distributions. The posterior mean estimates when fitting the true likelihood (3) are shown as a reference. The 30% trimmed mean and the MAD^2 (MAD sq.) estimates are also displayed in (a) and (b) respectively. Note that the distribution of the posterior means for σ^2 under the full model is not shown in (b) as it is roughly centered around 20.8 and has a much larger spread than the two posteriors shown.

k	10	20	30	35
$\widehat{R}$	1.03	0.92	0.81	0.75

Table 1: The relative variances of the posterior means. For each k, the table provides the ratio of the sample variance of the 50 full posterior means to the sample variance of the 50 restricted posterior means.

by the full posterior distribution. Trimming more values moves $T(\mathbf{X})$ toward the median.

A second natural question to ask is what is sacrificed if the data are not contaminated with outliers? To explore this question, we re-run the simulations above with the data generated from a single standard normal distribution without outliers. The same model for analysis (4) is used and we compare estimates of the posterior means for θ under the full and restricted likelihoods for several values of k. These estimates appear in Figure 3(b). A measure of relative variation in the estimators ($\widehat{R}$) is provided in Table 1. This table displays the ratio of the sample variance of the 50 full posterior means to the sample variance of the 50 restricted posterior means. As we increase k, the variance of the restricted posterior means increases. This is expected as we are losing information by conditioning on a non-sufficient statistic. However, we are only about 8% less efficient when using $k = 20$. There is a suggestion that using $k = 10$ is not detrimental as the measure of the ratio of variances is above 1. This is reminiscent of a suggestion in Stigler (1977) that the 10% trimmed mean is a reliable estimator. Our result could be a product of simulation variation, since there are only 50 replicates. It may be a product of the simulation conditions, since the data were generated as standard normals.

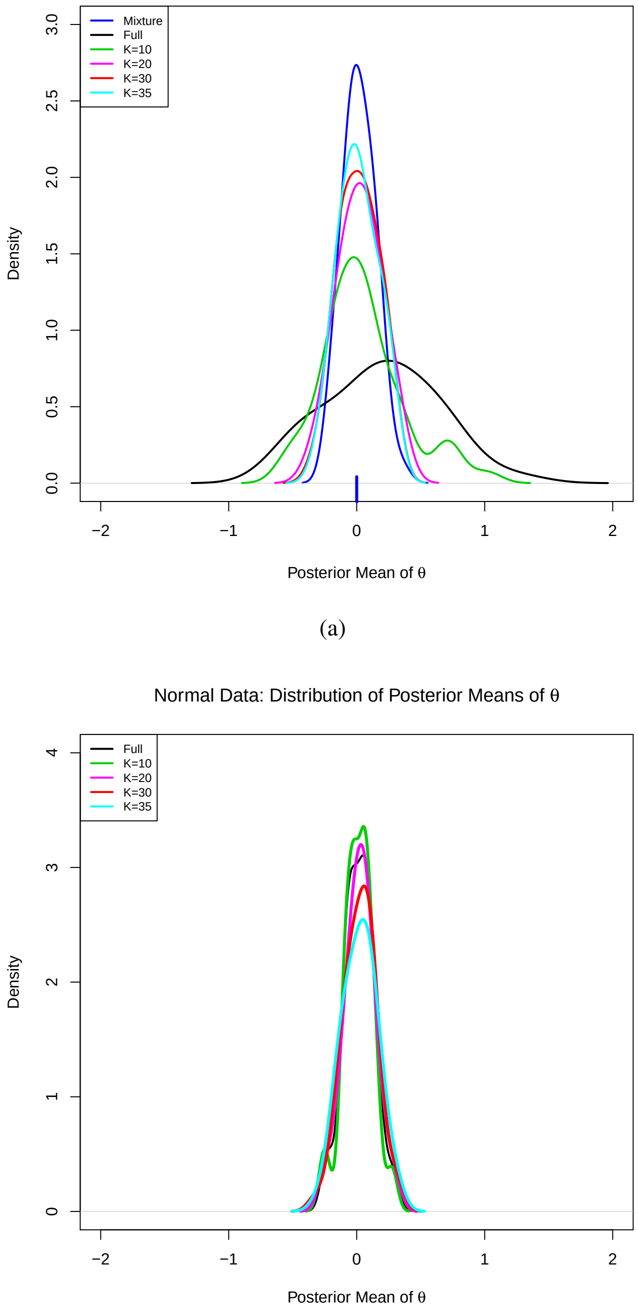
2.2 Huber Statistics

Next, we define $T(\mathbf{X})$ to be Huber's simultaneous estimators of location and scale. For observed data $\mathbf{x} = (x_1, \cdots, x_n)'$, let $T(\mathbf{x}) = (\hat{\theta}, \hat{\sigma})$. These estimators are defined as the simultaneous solutions to

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \hat{\theta}}{\hat{\sigma}}\right) = 0 \tag{6}$$

$$\sum_{i=1}^{n} \chi\left(\frac{x_i - \hat{\theta}}{\hat{\sigma}}\right) = 0. \tag{7}$$

Here, $\psi(x) = \max[-c, \min(c, x)]$ is the derivative of Huber's loss function. For simplicity we set $c = 1.345$ as this is the value where 95% relative efficiency is obtained when the data are normally distributed. $\chi(x) = \psi^2(x) - E[\psi^2(Z)]$ is Huber's 'Proposal 2' (Huber, 1974) where the expectation is taken with respect to the standard normal distribution. The restricted likelihood, $g(T(\mathbf{X})|\theta, \sigma^2)$, is intractable in this case. However, for the observed $\mathbf{x}$, we can sample from $\pi^*(\boldsymbol{\theta}|T(\mathbf{x}))$ using a direct sampling technique. For each point on a fine grid of (θ, σ^2) values, we sample $T(\mathbf{x}_1), \cdots, T(\mathbf{x}_N) \sim g(T(\mathbf{X})|\theta, \sigma^2)$ for a large value of N. With this sample, we then estimate $g(T(\mathbf{x})|\theta, \sigma^2)$ using kernel density estimation. Finally, we use each of these estimates on the fine grid of (θ, σ^2) values to obtain an



(a)

Normal Data: Distribution of Posterior Means of θ

(b)

Figure 3: (a) Distribution of the posterior mean for θ under the full and restricted posteriors for trimming values $k = 10, 20, 30$, and 35. The simulated data come from the mixture distribution (3). (b) Same as in (a), only the simulated data come from a single standard normal distribution.

estimate of the restricted posterior. This technique works well for a low-dimensional parameter $\boldsymbol{\theta}$, but it breaks down for high-dimensional $\boldsymbol{\theta}$ due to limitations in density estimation and computation.

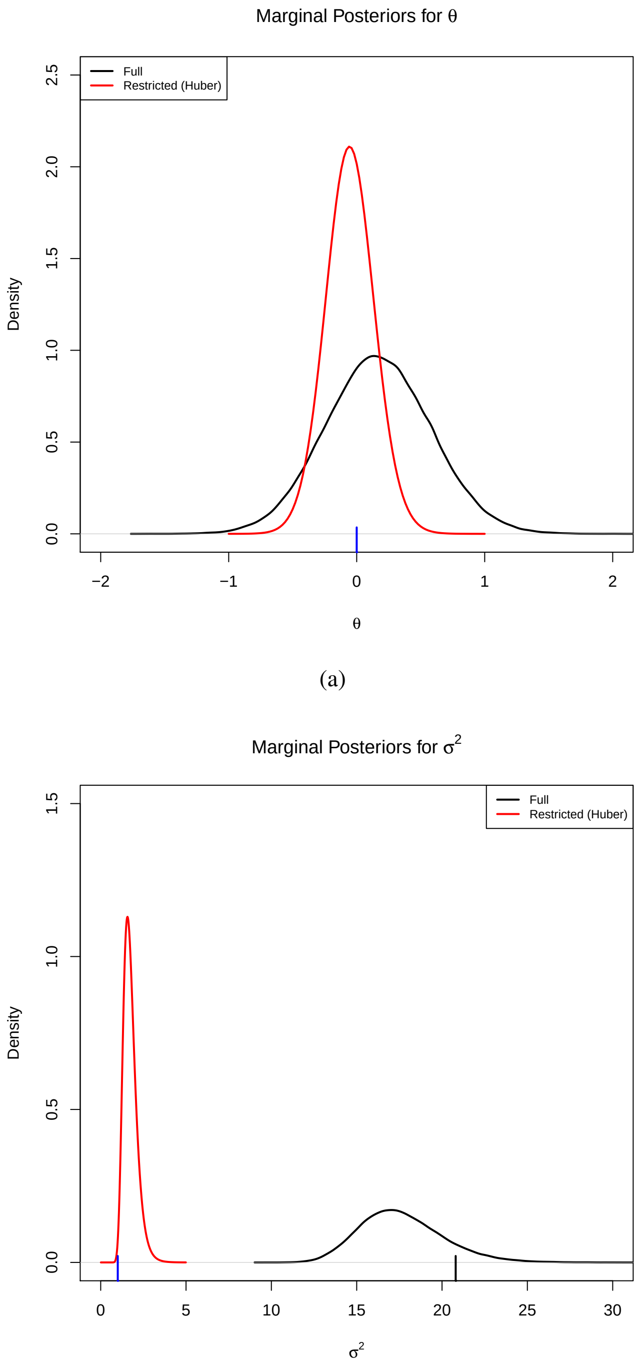
We compare the full posterior and the restricted posterior conditional on the data set displayed in Figure 1. The marginal posteriors for θ and σ^2 under the two likelihoods are shown in Figures 4(a) and 4(b) respectively. We see that the restricted posterior distribution for θ is more centered and tighter around the true parameter value. Again, this reduction in variance can be attributed to the impact of σ^2 on the posterior for θ.

3 Data Analysis–Estimating the Speed of Light

Our simulation studies show the potential the *blended paradigm* has for obtaining more concentrated posterior distributions and for reducing the bias and variance of posterior mean estimates. Next, we apply the *blended paradigm* to a very well known data set from an experiment conducted by Simon Newcomb in 1882 to make inference about the speed of light. Details of this experiment can be found in Stigler (1977). A smoothed version of the 66 data points is shown in Figure 5. The given values times 10^{-3} plus 24.8 are Newcomb's measurements in millionths of seconds for light to traverse a distance of approximately 7.443 km. One advantage of using this data set is that there is a modern consensus on the true value of the speed of light that did not exist when the experiment was conducted. The true value on this scale is now considered to be 33.02 and is marked on the axis in Figure 5. This corresponds to a time of 24.833 millionths of a second for light to travel 7.443 km which, subject to rounding error, equates to the now accepted $299,792.458$ km/s. The most notable aspect of the data are the two outliers on the low end, and the data have previously been used to assess the quality of robust estimators. Stigler (1977) computes several robust estimates including the median, various trimmed means, Huber's location estimator, and Tukey's biweight location estimator and suggests that the 10% trimmed mean is reliable. The 10% trimmed mean is 27.43 and the average of the ten estimates considered is reported as 27.3. Chan and Rhodin (1980) report an estimate of 27.6 based on optimally chosen sample quantiles. Huber's estimate of location is 27.39 with an approximate 95% confidence interval of $(26.12, 28.66)$

We apply the *blended paradigm* method and model the data with the usual normal theory model (4). Thus, the parameter θ can be interpreted as the speed of light in units corresponding to Newcomb's scale (i.e., the deviation from $24,800$ nanoseconds it takes for light to traverse the known distance). The variation in the data can be attributed to measurement error, and σ^2 governs the magnitude of this error.

The prior distribution for θ is chosen to reflect the knowledge about the speed of light that could have been available at the time of Newcomb's experiment. In particular, we make use of several estimates of the speed of light made from the years 1726 to 1879. In km/s, the estimates are $301,000$, $313,000$, $298,000$, and $299,910$ and were made by James Bradley, Armand Fizeau, Leon Foucault, and Albert Michelson, respectively (Froome and Essen, 1969). These values convert to -70, -1018.12, 178.96, and 19.88 on Newcomb's scale. We set the prior parameter $\mu = -222.32$, the mean of these values, and $\tau = 540.43$, the standard deviation of these values. This results in a relatively non-informative prior, reflecting the uncertainty of the true value at the time of the experiment. We take an ad hoc approach to setting the prior distribution for σ^2. The sample variance of the data without the outliers is 25.8 and with the outliers is 115.4. It seems reasonable to take the prior mean



(a)

(b)

Figure 4: Full (black) and restricted (red) posteriors for θ (a) and σ^2 (b). The restricted likelihood is that which is induced by Huber's robust estimators for location and scale. The posteriors are conditioned on the data set that appears in Figure 1.

of σ^2 to be 50 and thus we set $\alpha = 5$ and $\beta = 50(5-1) = 200$.

To create the restricted likelihoods, we use similar forms of $T(\mathbf{X})$ as we did in the simulations. Namely, a set of order statistics and Huber's robust estimators. We change the set of order statistics slightly. Since the outliers occur on the low end of the data we take $T(\mathbf{X}) = (X_{(k+1)}, \cdots X_{(66)})$ for $k = 2, 4$, and 8. In other words, we are only trimming the bottom values. The posterior distributions for θ and σ^2 are displayed in Figures 6(a) and 6(b) respectively. The true value of 33.02 and the sample mean, $\bar{x} = 26.2$, are marked on the axis of Figure 6(a) as a reference. In this display, we first notice that the marginal posteriors for θ under each model are being pulled down by the outliers. However, the posterior distributions for the restricted models are tighter and less biased (they are shifted closer to the truth). Using the full likelihood, we obtain a posterior which is centered very near $\bar{x}$. The posterior distributions for θ using the restricted likelihoods of the order statistics are similar for the varying values of k. The means for $k = 2, 4$, and 8 are $27.33, 27.43$, and 27.52, respectively, and their increasing order makes intuitive sense. They also compare reasonably with the 10% trimmed mean reported above. Using the restricted likelihood of Huber's statistics results in a posterior that is only slightly more variable than under the order statistics. The mean of this posterior is 27.39, which essentially matches the original Huber estimate. The similarity with the order statistics is important to note because to apply the *blended paradigm* in the regression setting, it is likely that we would have to resort to the use of robust regression estimators, such as Huber's statistics. Also, the similarity of the posterior means with the classical estimates reported above is not surprising considering the relatively non-informative prior for θ. Lastly, the marginal posteriors for σ^2 in Figure 6(b) show a similar result as the simulations above. The concentration of the restricted posteriors around smaller values than in the full posterior explains the tighter posteriors for θ.

4 Discussion

In this article, we have introduced the *blended paradigm* which seeks to combine classical robust inference and the Bayesian framework. We have shown the potential for the method to result in more concentrated posteriors in situations where we do not believe that the likelihood fully captures the data-generating process. Applying the method to Newcomb's data set, we were able to make inference which compared well to other, previously considered, classical robust estimators. While there may be better ways to make inference with this particular data set, we believe that the *blended paradigm* introduces a philosophically important idea. Mainly, the ability to posit a model we believe to be close to the truth and to us carefully chosen insufficient statistics to make Bayesian inference that is insensitive to deviations from the posited model.

One criticism of our method is that there is a loss of information as the result of conditioning on non-sufficient statistics. The full data set is indeed sufficient under the posited model, but this becomes less important if we do not fully believe the model in the first place. We seek to preserve a core part of the likelihood while dropping a portion that is sensitive to plausible departures from the model used for analysis.

Disadvantages, mostly computational, do exist. In the regression setting, robust regression estimators, such as Huber's estimators, seem like a logical choice when constructing the restricted likelihood. These likelihoods are analytically intractable and the direct sampling technique is only feasible when the number of regression parameters is modest. Thus, a current research question is how to develop efficient MCMC algorithms to sample from

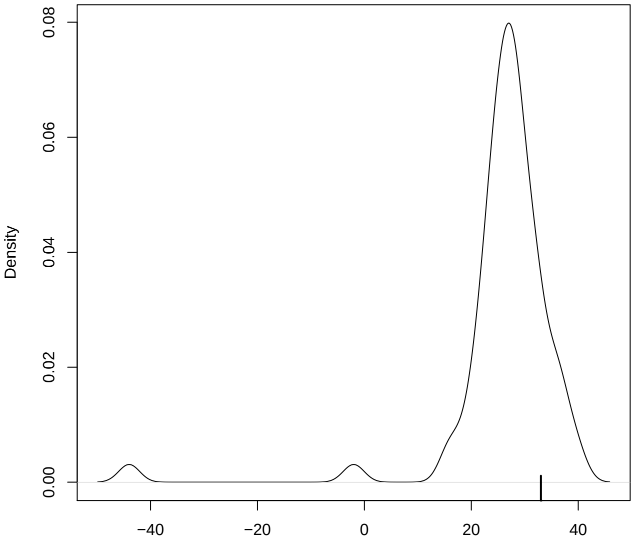
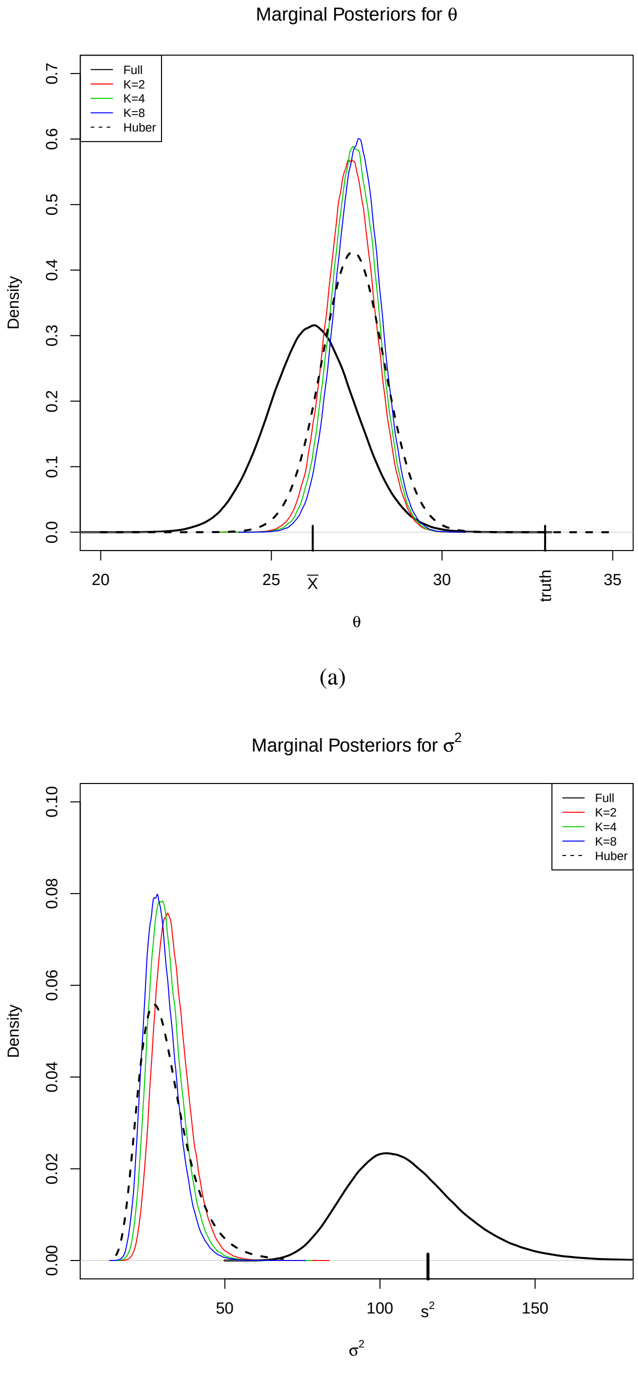


Figure 5: Data collected by Simon Newcomb in an experiment studying the speed of light. The data are given as deviations from 24.8 millionths of seconds for light to travel a known distance. For example, a recorded value of 29 corresponds to a measured time of $.029 + 24.8 = 24.829$ millionths of a second ($.024829$ nanoseconds) for light to travel the known distance. The currently accepted true value of 33.02 is marked on the axis.

the restricted posterior, $\pi^*(\boldsymbol{\theta}|T(\mathbf{X}))$. In addition to traditional MCMC techniques, the Approximate Bayesian Computation (ABC) techniques developed by Beaumont et al. (2002), Marjoram et al. (2003), Sisson et al. (2007) and others, show promise. ABC can be applied in situations where sampling from an intractable likelihood is relatively easy. Samples are obtained from a distribution 'close' to $\pi^*(\boldsymbol{\theta}|T(\mathbf{X}))$ and used to conduct inference. The philosophy of ABC is focused on choice of a 'good' summary of the data, with 'good' taken to be a statistic that is nearly sufficient. This stands in contrast to this work, where we deliberately select an insufficient statistic for conditioning in hope of providing stable and robust inference for the parameters of greatest interest.



(a)

(b)

Figure 6: Posterior distributions for θ (a) and σ^2 (b) under each model for Newcomb's data. The lines labeled with values of k represent restricted posteriors using order statistics with trimming value k. The dashed line represents the restricted posterior using Huber's estimators. The sample variance s^2 is marked in (b) as a reference.

References

Angrist, J. D. and Krueger, A. B. (1999). Empirical strategies in labor economics. In Ashenfelter, O. and Card, D., editors, *Handbook of Labor Economics*, volume 3 of *Handbook of Labor Economics*, chapter 23, pages 1277–1366. Elsevier.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, 162:2025–2035.

Bollinger, C. R. and Chandra, A. (2005). Iatrogenic specification error: A cautionary tale of cleaning data. *Journal of Labor Economics*, 23(2):pp. 235–257.

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

Froome, K. and Essen, L. (1969). *The Velocity of Light and Radio Waves*. Academic Press.

Huber, P. (1974). *Robust Statistics*. Wiley, New York.

Justel, A. and Peña, D. (1996). Gibbs sampling will fail in outlier problems with strong masking. *Journal of Computational and Graphical Statistics*, 5(2):pp. 176–189.

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(26):pp. 15324–15328.

Rocke, D. M., Downs, G. W., and Rocke, A. J. (1982). Are robust estimators really necessary? *Technometrics*, 24(2):pp. 95–101.

Sisson, S. A., Fan, Y., and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 104(6):pp. 1760–1765.

Stigler, S. M. (1973). Simon Newcomb, Percy Daniell, and the history of robust estimation 1885-1920. *Journal of the American Statistical Association*, 68(344):pp. 872–879.

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