

Department of Econometrics and Business Statistics

<http://business.monash.edu/econometrics-and-business-statistics/research/publications>

Robust Bayesian Exponentially Tilted Empirical Likelihood Method

Zhichao Liu, Catherine Forbes and Heather Anderson

December 2017

Working Paper 21/17

Robust Bayesian Exponentially Tilted Empirical Likelihood Method

Zhichao Liu, Catherine Forbes, Heather Anderson

December 12, 2017

Abstract

This paper proposes a new Bayesian approach for analysing moment condition models in the situation where the data may be contaminated by outliers. The approach builds upon the foundations developed by Schennach (2005) who proposed the Bayesian exponentially tilted empirical likelihood (BETEL) method, justified by the fact that an empirical likelihood (EL) can be interpreted as the nonparametric limit of a Bayesian procedure when the implied probabilities are obtained from maximizing entropy subject to some given moment constraints. Considering the impact that outliers are thought to have on the estimation of population moments, we develop a new robust BETEL (RBETEL) inferential methodology to deal with this potential problem. We show how the BETEL methods are linked to the recent work of Bissiri *et al.* (2016) who propose a general framework to update prior belief via a loss function. A controlled simulation experiment is conducted to investigate the performance of the RBETEL method. We find that the proposed methodology produces reliable posterior inference for the fundamental relationships that are embedded in the majority of the data, even when outliers are present. The method is also illustrated in an empirical study relating brain weight to body weight using a dataset containing sixty-five different land animal species.

Keywords: Moment condition models, Outliers, Misspecification.

1 Introduction

Traditional parametric Bayesian analysis of data requires the specification of a likelihood function and a prior distribution for the parameters. However, when the data generating process (DGP) is unknown, meaning that the true likelihood function is unavailable, then an alternative, or approximate model is often used in its place. In light of the well known adage of Box and Draper (1987), “All models are wrong, but some are useful”, such models can nevertheless be useful for researchers to understand the fundamental relationships between variables. However, incorrect models may come with a non-negligible risk of leading to grossly misleading inference.

Specifying models through moment restrictions alone is appealing in many situations as a way to reduce the risk of misspecifying a likelihood function. The resulting class of models, commonly referred to as moment condition models, produces inference about parameters from the information supplied by the moment restrictions, thereby circumventing the need for potentially unsuitable distributional assumptions about the sampling density. Methods to estimate and conduct hypothesis tests based on moment condition models have been extensively developed in the Frequentist literature; for example the well-known generalized method of moments (GMM) (e.g. Hansen, 1982, Newey and West, 1987, Hall, 2005 and many others). More recently, empirical likelihood (EL) based methods have become popular (e.g. Qin and Lawless, 1994, Imbens, 1997, Imbens *et al.*, 1998, Newey and Smith, 2004 and Schennach, 2007). Under a Bayesian framework, the estimation of moment condition models requires the formal construction of a proxy likelihood function using information given by the sample moments. See Yin (2009), for example, who proposes the use of the asymptotic distribution of the sample moments as an approximation of the likelihood function – an idea that is closely related to the use of a Laplace approximation as proposed by Chernozhukov and Hong (2003). In related work, and following on from earlier work by Chamberlain and Imbens (2003), Bayesian nonparametric methods are explored and further developed for moment condition models by Bornn *et al.* (2015).

Bayesian EL-based methods build on theoretical work from the Frequentist perspective as initially developed by Owen (1988). As argued by Lazar (2003), the result is a valid substitute for the

true likelihood function in a Bayesian analysis. In addition, Schennach (2005) shows that an exponentially tilted empirical likelihood (ETEL) has a well defined probabilistic interpretation arising from a Bayesian nonparametric procedure, and thus she proposes to obtain the Bayesian ETEL (BETEL) posterior density as being proportional to the product of the prior and the ETEL. Recently Chib *et al.* (2017) propose a method under the BETEL framework to deal with the problem when some moment conditions are misspecified. They introduce additional nuisance parameters in a reformulation of the moment restrictions, resulting in new and valid so-called augmented moment conditions, enabling them to be used in a BETEL setting. Their approach then compares the models as defined by different moment conditions, and selects the model having the largest marginal likelihood, thus promoting inference about the parameters to be made from the moment restrictions consistent with having the most empirical support.

The focus of our work is to deal with the problem when outliers are present in the data while working within a moment condition model setting. Outlier-prone data has been a concern in econometrics and related disciplines for many years (see, for example the reviews given by Stigler, 1973, Meer *et al.*, 1991 and Berger, 1994). Outliers are described as observations that are not generated from the same DGP as the majority, and they have the potential to distort understanding of the fundamental relationship between most of the observations. Robust methods for moment condition models have been proposed under the Frequentist framework, for example Ronchetti and Trojani (2001) and Ortelli and Trojani (2005) who consider robust estimation methods under the GMM setting. Schennach (2005) demonstrates in an example that moment conditions may be selected with robustness in mind, but does not elaborate further. Moment restrictions specified and valid for the majority of observations may in fact be invalid when outliers are present.

This paper contributes to the literature by proposing a novel robust Bayesian method based on the BETEL framework for moment condition models, intended for situations when outliers may be present in the data. Our new method, named the robust BETEL (RBETEL) method, is justified as the distribution that minimises the posterior expected loss function, along the lines of Bissiri *et al.* (2016) and where the relevant loss function arises from an appropriately defined EL ratio (see, Owen, 1990). Parameter values that result in a lower EL ratio correspond to stronger evidence for

an hypothesis associated with the validity of given moment restrictions. The relevant moment conditions are specified in a manner similar to Chib *et al.* (2017), by introducing a nuisance indicator vector that is used to separate the complete set of observations into subsets of outlying and non-outlying data points. The new loss function for the RBETEL setting can be expressed as the EL ratio evaluated using only the non-outlying (or ‘active’) data points given by the indicator vector. We then argue, using the framework of Bissiri *et al.* (2016), that the RBETEL joint posterior is a valid and coherent representation of subjective uncertainty about the minimizer of the expected loss. The robust posterior distribution is seen as the marginal RBETEL posterior for the model parameters, obtained using a Markov chain Monte Carlo (MCMC) approach, and marginalized over the uncertainty regarding the locations of active data points.

We demonstrate the estimation performance of the proposed RBETEL method under controlled simulation settings. The RBETEL method is shown to be robust to outliers in the sense that the posterior mean estimates of the model parameters are close to the (designed) parameter values used to simulate the ‘good’ (i.e. non-outlying) data. In addition, we find that the percentage of the posterior densities whose mass covers cover the designed values is high. The RBETEL method is then used to analyze the relationship between the brain weight and body weight of various land animal species, using a dataset also used by Rousseeuw and Van Zomeren (1990), who apply a Frequentist approach to mitigate potential outliers and examine possible leverage points. A linear model is employed to estimate the relationship between the two variables following Rousseeuw and Van Zomeren (1990). Comparing the estimation results produced by both the BETEL and RBETEL methods, we find that the RBETEL method fits the majority of the observations well and seems to be robust to potential leverage effects. In addition, the posterior mean estimates of the parameters produced by the RBETEL method are similar to the parameter estimates produced by the robust M-estimator of Yohai (1987).

The remainder of the paper is organized as follows. Section 2 provides background information, including a brief review of the original BETEL method of Schennach (2005) and the new framework that updates prior belief via a loss function proposed by Bissiri *et al.* (2016). Section 3 proposes the new RBETEL method, the loss function used for the new method is shown and the posterior

distribution based on the loss function is derived. Details for the computation are also discussed. Simulation experiments are conducted in Section 4. Then in Section 5 the empirical study regarding the relationship between land animals' brain weight and body weight is provided. Section 6 concludes the paper and discusses the future work needed to improve the new method.

2 Background

This section provides essential background for the proposed robust Bayesian exponentially tilted empirical likelihood (RBETEL) method. We first review the standard Bayesian exponentially tilted empirical likelihood (BETEL) framework proposed by Schennach (2005), which is designed to produce inference for a moment condition model. The original BETEL setting does not consider the situation when outliers are present in the data. We build upon the BETEL framework and develop the new RBETEL method in Section 3, which produces Bayesian inference that is robust to outliers that contaminate the observations. In addition, we outline the framework recently proposed by Bissiri *et al.* (2016). The posterior distribution is viewed as an update of prior belief via a loss function under this framework, and it provides an alternative justification for the BETEL posterior distribution. We identify a monotonic relationship between an empirical likelihood ratio and the loss function for the standard BETEL method, and thus a loss function for the new RBETEL method can be obtained in a similar way. The RBETEL posterior distribution can then be justified under the framework of Bissiri *et al.* (2016).

2.1 Bayesian exponentially tilted empirical likelihood

A moment condition model is specified through a set of moment restrictions of the form

$$E^F [g(X; \theta)] = \mathbf{0}, \tag{1}$$

where $g(X; \theta)$ is a given d_g dimensional function of a $d_X \times 1$ random vector X and a $p \times 1$ parameter vector θ . For identification purposes, $d_g \geq p$ must be satisfied. The expectation in (1) is taken with respect to the unknown distribution of X , denoted by F , and $\mathbf{0}$ is a $d_g \times 1$ vector of zeros.

Suppose that a random sample $x_{1:n} = (x_1, \dots, x_n)$ is observed. Following Schennach (2005), if the interior of the convex hull of $\bigcup_{i=1}^n \{g(x_i, \theta)\}$ contains the origin, then the BETEL posterior takes the form

$$\pi_{\text{BETEL}}(\theta|x_{1:n}) \propto \pi(\theta)\widehat{p}(x_{1:n}|\theta), \quad (2)$$

where $\pi(\theta)$ is the assumed prior probability density function (pdf) for θ , and $\widehat{p}(x_{1:n}|\theta)$ is the proxy likelihood function, referred to as the ETEL. The ETEL is given by

$$\widehat{p}(x_{1:n}|\theta) = \prod_{i=1}^n \widehat{w}_i(\theta), \quad (3)$$

where the $\widehat{w}_i(\theta)$, for $i = 1, \dots, n$, are implied probabilities defined by the solution to the constrained minimization problem given by

$$(\widehat{w}_1(\theta), \dots, \widehat{w}_n(\theta)) = \arg \min_{(w_1, \dots, w_n)} \sum_{i=1}^n w_i \ln w_i, \quad (4)$$

subject to

$$\sum_{i=1}^n w_i = 1 \text{ and } \sum_{i=1}^n w_i g(x_i; \theta) = \mathbf{0}. \quad (5)$$

As shown by Schennach (2005), the implied probabilities in (4) correspond to those that minimise the Kullback-Leibler (KL) divergence from the approximating multinomial distribution associated with the probabilities (w_1, \dots, w_n) to the empirical distribution having weights given by $(\frac{1}{n}, \dots, \frac{1}{n})$. Further, these probabilities can be computed conveniently as

$$\widehat{w}_i(\theta) = \frac{\exp(\widehat{\lambda}(\theta)'g(x_i; \theta))}{\sum_{j=1}^n \exp(\widehat{\lambda}(\theta)'g(x_j; \theta))}, \text{ for } i = 1, \dots, n, \quad (6)$$

where the so-called optimal ‘tilting’ parameter, $\widehat{\lambda}(\theta)$, is given by

$$\widehat{\lambda}(\theta) = \arg \min_{\lambda} \sum_{j=1}^n \exp(\lambda'g(x_j; \theta)). \quad (7)$$

Schennach (2005) also shows that the ETEL is equivalent to the limit of a non-parametric Bayesian procedure under a suitably defined non-informative prior. Therefore, as Chib *et al.* (2017) point out, the BETEL method may be used to form the basis of a Bayesian semi-parametric

analysis. In Section 2.2 we justify the BETEL method from a different perspective.

2.2 Updating prior belief via a loss function

In this section, we outline the recent work of Bissiri *et al.* (2016) who propose a framework to update prior beliefs about parameters via a loss function. We then justify the BETEL method as such an updated distribution under this new framework.

Suppose we have a loss function $l(\theta, X)$ which is a function of a parameter θ and a random variable X having an unknown distribution function F . Let θ_0 be defined as the minimizer of the expected loss, i.e.

$$\begin{aligned}\theta_0 &= \arg \min_{\theta} E^F [l(\theta; X)] \\ &= \arg \min_{\theta} \int l(\theta; X) dF.\end{aligned}\tag{8}$$

The objective of Bissiri *et al.* (2016) is to determine the form of a valid and coherent representation of subjective uncertainty in θ_0 , denoted by $\hat{\pi}(\theta|x_{1:n})$, from a prior $\pi(\theta)$ and observations $x_{1:n}$.

Bissiri *et al.* (2016) argue that, although F is unknown, a coherent decision maker should prefer a probability measure $\pi_1(\theta|x_{1:n})$ over $\pi_2(\theta|x_{1:n})$ if the posterior expected loss under $\pi_1(\theta | x_{1:n})$ is strictly less than the posterior expected loss under $\pi_2(\theta | x_{1:n})$, i.e. if

$$\int \int l(\theta; x_{1:n}) dF \pi_1(\theta|x_{1:n}) d\theta < \int \int l(\theta; x_{1:n}) dF \pi_2(\theta|x_{1:n}) d\theta.\tag{9}$$

They also argue that provided the sample is independent of the prior, then a valid and coherent representation of subjective uncertainty in θ_0 arises from minimizing the so-called cumulative (or

additive) loss, given by

$$\begin{aligned}
L(\pi(\theta|x_{1:n}); \pi(\theta), x_{1:n}) &= \int l(\theta; x_{1:n}) \pi(\theta|x_{1:n}) d\theta + \int \pi(\theta|x_{1:n}) \log\left(\frac{\pi(\theta|x_{1:n})}{\pi(\theta)}\right) d\theta \\
&= \int \pi(\theta|x_{1:n}) \log\left(\frac{\pi(\theta|x_{1:n})}{\exp\{-l(\theta; x_{1:n})\} \pi(\theta)}\right) d\theta.
\end{aligned} \tag{10}$$

This cumulative loss represents the expected posterior loss associated with the data $x_{1:n}$ in addition to the expected loss due to the prior $\pi(\theta)$. Bissiri *et al.* (2016) further show that $L(\pi(\theta|x_{1:n}); \pi(\theta), x_{1:n}) \rightarrow \int \int l(\theta; x_{1:n}) dF \pi(\theta|x_{1:n}) d\theta$ as $n \rightarrow \infty$, providing an asymptotic justification for the finite sample minimizer in (10).

The cumulative loss function (10) has the form of the KL divergence between the posterior distribution $\pi(\theta|x_{1:n})$ and another distribution with probability density function (pdf) proportional to $\exp\{-l(\theta; x_{1:n})\} \pi(\theta)$. Therefore, it is straightforward to see that the minimizer of (10) has the form

$$\begin{aligned}
\hat{\pi}(\theta|x_{1:n}) &= \arg \min_{\pi(\theta|x_{1:n})} L(\pi(\theta|x_{1:n}); \pi(\theta), x_{1:n}) \\
&= \frac{\exp\{-l(\theta; x_{1:n})\} \pi(\theta)}{\int \exp\{-l(\theta; x_{1:n})\} \pi(\theta) d\theta}.
\end{aligned} \tag{11}$$

2.2.1 The BETEL loss function

We now define the loss function implied by the BETEL method arising from the framework of Bissiri *et al.* (2016). Given the BETEL posterior in (2), with implied probabilities $\hat{w}_i(\theta)$, for $i = 1, 2, \dots, n$ given in (6), and optimal tilting parameter $\hat{\lambda}(\theta)$ given in (7), it follows directly that the corresponding loss function must be given by

$$l_{\text{BETEL}}(\theta, x_{1:n}) = - \sum_{i=1}^n \log \hat{w}_i(\theta). \tag{12}$$

The BETEL posterior (2) can then be interpreted as the representation of the subjective uncertainty in θ_0 .

Notice that the exponential of the negative loss function is proportional to the EL ratio, $\mathcal{R}(\theta) = \prod_{i=1}^n n\hat{w}_i(\theta)$, since

$$\mathcal{R}(\theta) \propto \prod_{i=1}^n \hat{w}_i(\theta) = \exp(-l_{\text{BETEL}}(\theta; x_{1:n})). \quad (13)$$

Under the Frequentist framework and given an estimate of θ , say $\hat{\theta}$, the EL ratio statistic, $\mathcal{R}(\hat{\theta})$, is commonly used for testing the null hypothesis that the moment restrictions in (1) are valid (see, for example Owen, 1988 and Owen, 1990). The asymptotic properties of the EL ratio using implied probabilities defined by (4) are discussed by Schennach (2007).

A higher EL ratio, or equivalently, a lower $l_{\text{BETEL}}(\theta; x_{1:n})$ in our case, suggests less evidence against the null hypothesis, suggesting that we will be interested in values of θ for which $\mathcal{R}(\theta)$ is high and $l_{\text{BETEL}}(\theta; x_{1:n})$ is low. This intuition provides a guideline for constructing the loss function in the RBETEL method proposed in Section 3.

3 Robust Bayesian exponentially tilted empirical likelihood method

In this section, we develop the proposed RBETEL method. Our approach is to augment the required moment conditions with information about the possible presence of outliers in the dataset. After showing that the standard moment conditions which assume all the observations have the same distribution will be invalid when outliers are present in the data, we modify the moment conditions by introducing an indicator vector into the formulation. We then demonstrate the form of the RBETEL posterior based on these modified moment conditions. The RBETEL posterior distribution is justified under the framework of Bissiri *et al.* (2016) with a loss function related to the empirical likelihood ratio evaluated using some subsets of the data. We detail the computation strategy to sample model parameters and the indicator vector from the RBETEL posterior. Lastly, we discuss the choice of the moment conditions which provide the essential information to separate the good data and outliers.

3.1 A contaminated sample

Given that outliers are thought to be present in the sample, we think of the distribution F from which the data are actually sampled as a mixture of two distinct distributions, denoted by G and B , respectively. In particular, a subset of the complete dataset consisting of K observations are generated from the unknown distribution G , where $\frac{n}{2} < K \leq n$. We refer to G as the ‘good’ distribution, as the values drawn from G are expected to satisfy the stated moment conditions. The remaining $n - K$ observations are assumed to arise from another distribution denoted by B , whose values have the potential to be vastly different from those generated by G . In particular, the moment conditions specified for G will not hold for B . We refer to the distribution B in this context as the ‘bad’ distribution.

The reason for the imposed condition $\frac{n}{2} < K \leq n$ will become clearer in Section 3.3. It relates to the fact that as the RBETEL procedure will use identified subsets of non-outlying datasets in order to construct a suitable likelihood function, it will be important to ensure that at least half of the data come from G .

When the outlying realisations from B appear randomly throughout the sample, with each realization in the complete dataset having an equal probability v of being drawn from G , then the distribution associated with the data, without conditioning on K , is given by the mixture distribution

$$X_i | v \stackrel{iid}{\sim} \begin{cases} G \text{ with prob } v \\ B \text{ with prob } (1 - v), \end{cases} \quad (14)$$

where X_i is a $d_x \times 1$ random variable corresponding to the realized value x_i , for each $i = 1, 2, \dots, n$, and *iid* refers to independent and identically distributed.

The *expected* proportion of outliers, given by $(1 - v)$, may be treated as either fixed or known. By suitable incorporation of a prior distribution for v , uncertainty regarding the expected proportion of outliers may also be accounted for, as we show in Section 3.4. Until then, however, we treat

v as fixed.

Our interest lies in understanding the fundamental relationship between the subset of random variables generated from G alone, and *not* from the contaminated distribution in (14). In particular we want to infer a distribution for the value of the parameter $\theta \in \Theta$ from the *valid* moment conditions

$$E^G [g(X, \theta)] = \mathbf{0}, \quad (15)$$

for some $g(X, \theta)$ satisfying the usual conditions, as detailed in Section 2.1. Note that the expectation in (15) is taken with respect to G , i.e. the distribution that generates the ‘good’ data. We assume that the moment conditions are valid under G , meaning that (15) holds, while at the same time

$$E^B [g(X, \theta)] \neq \mathbf{0}, \quad (16)$$

meaning that the same type of moment restrictions will be invalid under the distribution B that generates the outliers. Therefore, since F is comprised of both G and B , the moment restrictions will also be invalid under the mixture distribution, $F = v * G + (1 - v) * B$, as detailed in (14).

3.2 The RBETEL modified moment conditions

The mixture model for X_i in (14) may be written as the marginal distribution that results from the joint specification of X_i and an auxiliary random variable s_i that indicates whether observation i comes from the ‘good’ or ‘bad’ mixture component, i.e.

$$X_i | s_i \sim \begin{cases} G & \text{if } s_i = 1 \\ B & \text{if } s_i = 0, \end{cases} \quad (17)$$

where the indicators in the vector $s = (s_1, \dots, s_n)$ are, by virtue of the mixture model, *iid* Bernoulli(v) random variables, consistent with $Pr(s_i = 1 | v) = v$.

Using the mixture indicator variable, s_i , the desired moment expression in (15) may be expressed

as

$$E^F [g(X_i, \theta) s_i] = \mathbf{0} \text{ for } i = 1, \dots, n, \quad (18)$$

where we note that the expectation is taken with respect to F rather than G . Now, since

$$E^F [g(X_i, \theta) s_i] = v E^G [g(X_i, \theta) s_i | s_i = 1] + (1 - v) E^B [g(X_i, \theta) s_i | s_i = 0], \quad (19)$$

then $E^F [g(X_i, \theta) s_i] = \mathbf{0}$ only when $E^G [g(X_i, \theta)] = \mathbf{0}$. It is clear that if X_i is an outlier, then the moment condition in (18) will only be valid if the indicator s_i is zero.

3.3 The RBETEL posterior

The RBETEL method produces a posterior distribution for the parameter θ *jointly* with the vector of indicator variables $s = (s_1, s_2, \dots, s_n)$. In order to ensure that the indicators reflect the non-outlying observations (i.e. observations from G rather than from B) we impose the constraint that $K = \sum_{i=1}^n s_i > \frac{n}{2}$ to ensure that the resulting posterior inference about θ is informed by the majority of the data. Thus, given a prior specification for θ (consistent with the distribution G), denoted by $\pi(\theta)$, the RBETEL joint posterior conditional on v is given by

$$\pi_{\text{RBETEL}}(\theta, s | x_{1:n}, v) \propto \pi(\theta) \pi(s | v) \tilde{p}(x_{1:n} | \theta, s) I\left(\sum_{j=1}^n s_j > \frac{n}{2}\right), \quad (20)$$

where the ETEL in this context is denoted by $\tilde{p}(x_{1:n} | \theta, s)$. The indicator function $I_{(\cdot)}$ takes value one if the constraint holds, otherwise it is equal to zero. Although not strictly required, we would anticipate that $\pi(\theta)$ is independent of both s and v , as implied by (20).

For a given indicator vector s we are able to identify the outlier observations, each corresponding to $s_i = 0$, and hence we will want to exclude these observations from the calculation of the implied probability weights. Accordingly, we note that the indicator vector s determines the vector of n empirical probabilities, given by $(\frac{s_1}{\sum_{j=1}^n s_j}, \dots, \frac{s_n}{\sum_{j=1}^n s_j})$. The corresponding RBETEL implied probabilities, given by $(\tilde{w}_1(\theta, s), \dots, \tilde{w}_n(\theta, s))$, are then obtained by minimizing the KL divergence from

the multinomial probabilities $(w_1 s_1, \dots, w_n s_n)$ to the empirical probabilities noted above. Hence

$$(\tilde{w}_1(\theta, s), \dots, \tilde{w}_n(\theta, s)) = \arg \min_{(w_1, \dots, w_n)} \sum_{i=1}^n w_i s_i \ln \left(w_i \sum_{j=1}^n s_j \right), \quad (21)$$

subject to

$$\sum_{i=1}^n w_i s_i = 1 \text{ and } \sum_{i=1}^n w_i s_i g(x_i, \theta) = 0. \quad (22)$$

From (22) it can be seen that the weights for the active observations must sum to one, and the associated weighted sample average must satisfy the required conditions, corresponding to the theoretical specification in (18).

It can be found from (21) that the values of the implied probabilities, $\tilde{w}_i(\theta, s)$, associated with $s_i = 0$, are not uniquely defined. This does not cause any issue in the RBETEL setting, because these implied probabilities corresponding to outliers do not contribute to the value of the ETEL, given by

$$\tilde{p}(x_{1:n} | \theta, s) = \prod_{i=1}^n \left(\tilde{w}_i(\theta, s) \sum_{j=1}^n s_j \right)^{s_i}. \quad (23)$$

It is straightforward to show that the RBETEL implied probabilities corresponding to the unit indicators, i.e. $s_i = 1$, can be represented by

$$\tilde{w}_i(\theta, s) = \frac{\exp(\tilde{\lambda}(\theta, s)' g(x_i, \theta))}{\sum_{j=1}^n s_j \exp(\tilde{\lambda}(\theta, s)' g(x_j, \theta))}, \quad (24)$$

where the optimal tilting parameter is given by

$$\tilde{\lambda}(\theta, s) = \arg \min_{\lambda} \sum_{j=1}^n s_j \exp(\lambda' g(x_j, \theta)). \quad (25)$$

Therefore, although there are actually infinitely many solutions for the RBETEL weights in (24) given s and θ , the collection of the K weights $\{\tilde{w}_i(\theta, s) \text{ for all } i = 1, 2, \dots, n, \text{ where } s_i = 1\}$ is unique if the usual requirements on the functions comprising $g(X, \theta)$ and the parameter space Θ are satisfied.

Now, it is clear that if the indicator vector, s , is known, the RBETEL posterior is the same

as a standard BETEL posterior conditional on a subset of the complete dataset. Hence for any s and θ (given v) the joint RBETEL posterior pdf can be computed up to a normalizing constant, as per (20). However, to extract inference about θ *marginal* of s , we simply marginalise the joint RBETEL posterior over s to obtain

$$\pi_{\text{RBETEL}}(\theta | x_{1:n}, v) \propto \int \pi_{\text{RBETEL}}(\theta, s | x_{1:n}, v) ds. \quad (26)$$

As we show in Section 3.4, computation of the posterior is undertaken using an MCMC approach, and hence this marginalisation will be easy to implement once a sample of (θ, s) draws from the joint posterior is available.

3.4 Posterior computation with MCMC

In this section, we detail an MCMC method for sampling the unknowns θ and s from the RBETEL joint posterior distribution, corresponding to (20). In addition, as uncertainty in the proportion of outliers, v can be accommodated by hierarchically adding a prior distribution for v , in which case the *joint* RBETEL posterior becomes

$$\pi_{\text{RBETEL}}(\theta, s, v | x_{1:n}) \propto \pi(\theta)\pi(v)\pi(s | v)I\left(\sum_{j=1}^n s_j > \frac{n}{2}\right) \prod_{i=1}^n \left(\tilde{w}(\theta, s) \sum_{i=1}^n s_i \right)^{s_i}, \quad (27)$$

with, as before, the marginalisation to $\pi_{\text{RBETEL}}(\theta | x_{1:n})$ being managed via MCMC.

The approach we use here samples values of θ , s and v , iteratively, each from the corresponding full conditional RBETEL posterior. The sampling methods for θ and v , respectively, are reasonably straightforward. Both are undertaken conditionally on s and hence effectively operate as BETEL conditional distributions on an active sub-sample. We detail these two cases first, followed by a description of our approach to sample s .

3.4.1 Sampling θ and v

Only the currently active data subset is needed for sampling θ , as $\pi_{\text{RBETEL}}(\theta | x_{1:n}, v, s) = \pi_{\text{RBETEL}}(\theta | x_{1:n}, s)$. This full conditional distribution is given by

$$\pi_{\text{RBETEL}}(\theta | x_{1:n}, s) \propto \pi(\theta) \prod_{i=1}^n \left(\tilde{w}_i(\theta, s) \sum_{i=1}^n s_i \right)^{s_i}. \quad (28)$$

As the $\tilde{w}_i(\theta, s)$ must be calculated numerically, we suggest sampling from (28) using a random walk Metropolis-Hastings (MH) algorithm (as described, for example, in Chib and Greenberg, 1995).

Ideally, good subjective prior information will be available for the proportion of outliers, v . Noting that the Beta distribution is conjugate for the Binomial distribution, and hence for the independent Bernoulli latent variables contained in s , we suggest the use of a suitably truncated Beta prior for v . In our examples we take the truncation to be $v \in (0.5, 1]$, and use a truncated Beta prior for the parameter v , having hyperparameters α_0 and β_0 and with pdf given by

$$\pi(v) = \frac{v^{\alpha_0-1}(1-v)^{\beta_0-1}}{\int_{0.5}^1 v^{\alpha_0-1}(1-v)^{\beta_0-1} dv} I(v \in (0.5, 1]). \quad (29)$$

We of course acknowledge that other priors may be used, if desired. However, under the truncated Beta prior specified in (29), the full conditional distribution for v has a pdf given by

$$\pi(v | x_{1:n}, \theta, s) \propto v^{\sum_{i=1}^n s_i + \alpha_0 - 1} (1-v)^{n - \sum_{i=1}^n s_i + \beta_0 - 1} I(v \in (0.5, 1]). \quad (30)$$

Note that this full conditional does not depend on either θ or $x_{1:n}$, but corresponds to a kernel of a truncated Beta distribution with parameters given by $\sum_{i=1}^n s_i + \alpha_0$ and $n - \sum_{i=1}^n s_i + \beta_0$, with the value of v restricted to $(0.5, 1]$.

3.4.2 Sampling s

Next, note that the indicators s_i , for $i = 1, \dots, n$, are assumed *a priori* to be *iid* Bernoulli random variables with the probability of success given by v . This implies that the prior distribution for $K = \sum_{j=1}^n s_j$ is *Bernoulli*(v). However, once the data are observed, the conditional joint posterior

distribution for K will no longer be *Bernoulli*(v) and the individual s_j components will no longer be *iid*. In addition, recall that we have imposed the identifying constraint that $K > \frac{n}{2}$ so that we will always include the majority of the observations in the ‘good’, or ‘active’ data group. To overcome these complications, we use an MH approach to sample s . To devise an efficient candidate proposal, however, we do so by sampling a joint draw of the pair (s, K) , accepting with the relevant probability, and then discarding the drawn value of K .

Noting the form of the joint full conditional distribution of s and K is given by

$$\pi_{\text{RBETEL}}(s, K \mid x_{1:n}, \theta, v) \propto v^K (1-v)^{n-K} I_{(K > \frac{n}{2})} I_{(K = \sum_{j=1}^n s_j)} \prod_{i=1}^n (\tilde{w}_i(\theta, s) K)^{s_i}. \quad (31)$$

We denote the joint proposal of (s, K) at the m^{th} iteration of the MCMC procedure by (s^\dagger, K^\dagger) . The approach we take is to first generate K^\dagger , conditionally upon $s^{(m-1)}$, and then generate s^\dagger given both K^\dagger and $s^{(m-1)}$. In particular, only $K^{(m-1)} = \sum_{j=1}^n s_j^{(m-1)}$ is used when generating K^\dagger . Our proposal distribution is therefore given by

$$q_{(s,K)}(s^\dagger, K^\dagger \mid s^{(m-1)}) = q_2(s^\dagger \mid K^\dagger, s^{(m-1)}, \tau) q_1(K^\dagger \mid K^{(m-1)}), \quad (32)$$

where τ is a tuning parameter and is defined below. The proposal distribution (32) is comprised of two parts. The first part, denoted by $q_1(\cdot)$, proposes a new value for the sum of the indicators, K , conditional on its value in the previous MCMC iteration, $K^{(m-1)}$. We use a truncated *Binomial* $\left(n, \frac{c}{n} K^{(m-1)}\right)$ here, with $c = 0.99$ used in this paper, so that K^\dagger will be similar to $K^{(m-1)}$ while ensuring that there is always a positive probability of moving away from $K^{(m-1)} = n$ if that ever occurs. The proposed value K^\dagger must be integers in the interval $(\frac{n}{2}, n]$.

The second part of the proposal involves drawing the candidate indicator vector of s^\dagger , conditionally given the drawn K^\dagger and the last MCMC draw $s^{(m-1)}$. For this second proposal we select the vector s^\dagger such that exactly K^\dagger of its components have $s_i^\dagger = 1$, indicating the set of active observations from the complete dataset in the next iteration of the MCMC. The probability of an observation to stay in the same state, i.e. active with $s^{(m-1)} = 1$ or inactive with $s^{(m-1)} = 0$, as

proportional to some fixed probability, say τ , while the probability of s_i changing to the alternative state given by $1 - \tau$. All this while conditioning on the total number of proposed active observation being fixed at K^\dagger . Accordingly, the pdf corresponding to the second proposal component is

$$q_2\left(s^\dagger \mid K^\dagger, s^{(m-1)}, \tau\right) \propto \prod_{\{i:s_i^{(m-1)}=1\}} \left[\tau^{s_i^\dagger} (1-\tau)^{(1-s_i^\dagger)}\right] \prod_{\{i:s_i^{(m-1)}=0\}} \left[(1-\tau)^{s_i^\dagger} \tau^{(1-s_i^\dagger)}\right] I_{(\sum s_i=K^\dagger)}. \quad (33)$$

At iteration m , the generated indicator vector proposal, s^\dagger , is then accepted, activating a new subset of data, with probability given by

$$R_{s,K} = \min\left(1, \frac{\pi_{\text{RBETEL}}(s^\dagger, K^\dagger \mid x_{1:n}, \theta, v)}{\pi_{\text{RBETEL}}(s^{(m-1)}, K^{(m-1)} \mid x_{1:n}, \theta, v)} \frac{q_{s,K}(s^{(m-1)}, K^{(m-1)} \mid s^\dagger)}{q_{s,K}(s^\dagger, K^\dagger \mid s^{(m-1)})}\right), \quad (34)$$

otherwise, set $(s^{(m)}, K^{(m)}) = (s^{(m-1)}, K^{(m-1)})$. Note that as $K^{(m)}$ may be determined completely from $s^{(m)}$, only $s^{(m)}$ need actually be retained.

3.5 The RBETEL loss function

We now provide a justification for the RBETEL posterior distribution using the framework of Bisiri *et al.* (2016) that we explained in Section 2.2. After identifying a loss function for the RBETEL method, the joint posterior distribution for (θ, s) taking the form (20) is shown to be a representation of subjective uncertainty in the values of θ and s which minimize the expected RBETEL loss. We note that this derivation is undertaken conditionally on v .

Consistent with the algebraic derivation and intuition provided in Section 2.2.1, the RBETEL loss function is related to an EL ratio defined by

$$\begin{aligned} \tilde{\mathcal{R}}(\theta, s) &= \frac{\prod_{i=1}^n (s_i \tilde{w}_i(\theta, s))^{s_i}}{\prod_{i=1}^n \left(\sum_{j=1}^n s_j\right)^{s_i}} \\ &= \prod_{i=1}^n \left(\tilde{w}_i(\theta, s) \sum_{j=1}^n s_j\right)^{s_i}. \end{aligned} \quad (35)$$

The EL ratio function (35) can be written equivalently as

$$\tilde{\mathcal{R}}(\theta, s) = \prod_{i=1}^n \frac{\tilde{w}_i(\theta, s)}{1/\sum_{j=1}^n s_j} I(s_i = 1), \quad (36)$$

which is the EL ratio evaluated using only the active observations. This EL ratio is higher when evaluated on values of θ and s that are consistent with the desired theoretical moment condition (18). Information provided by the moment conditions regarding the distinguishing characteristics of the good data and outliers enters the RBETEL posterior through this EL ratio.

Now, the RBETEL loss function, $\tilde{l}(\theta, s; x_{1:n})$, is defined following the monotonic relationship (13) found in Section 2.2.1, i.e.

$$\begin{aligned} \tilde{l}(\theta, s; x_{1:n}) &= -\ln \tilde{\mathcal{R}}(\theta, s) \\ &= -\sum_{i=1}^n s_i \ln \left(\tilde{w}_i(\theta, s) \sum_{j=1}^n s_j \right). \end{aligned} \quad (37)$$

Following Bissiri *et al.* (2016), we derive the RBETEL joint posterior distribution, $\pi_{\text{RBETEL}}(\theta, s | x_{1:n}, v)$, with loss function (37) in the appendix. The RBETEL joint posterior distribution is obtained from minimization of a cumulative loss function which is a finite sample version of the expect loss, $E^F [\tilde{l}(\theta, s; x_{1:n})]$.

3.6 Choosing the moment conditions

It is essential to carefully choose the moment restrictions which provide information about the relationship between variables and the distribution of the data. In the literature, where the presence of outliers in the dataset has not been considered, moment conditions are chosen from available model assumptions, for example that the conditional expectation of an error term should equal zero, or that certain products of random variables should, on average, be zero. These are commonly referred to as orthogonality assumptions (see e.g., Hansen, 1982, Wooldridge, 2001 and Stock *et al.*, 2002). Such choices are at the discretion of the analyst, and in general will be problem specific.

Now that outliers are to be accommodated, we assume that the analyst has already in mind

some moment conditions, such as those arising from orthogonality conditions for the ‘good’ data. These conditions are now stated as $E^G[g(\theta, x)] = \mathbf{0}$. However, the RBETEL method also requires that the selected moment condition vector $E^F[g(\theta, x)] \neq \mathbf{0}$ when outliers are present, i.e. when $E^B[g(\theta, x)] \neq \mathbf{0}$. Therefore, it is important to include some additional moment conditions in the specification to ensure this additional need is met. In fact, as discussed in Section 2.2, the key information for identifying outliers comes from the moment conditions that lead to different values of the EL ratio when evaluated using different subsets of observations. Under the RBETEL framework, at least p moment conditions are needed to ensure that the parameter vector θ is identified. Among the restrictions used, at least one moment condition will need to be chosen to fit the characteristics of the good data while simultaneously being misspecified for the outliers. For convenience, we refer to such a moment condition as a ‘*key condition*’. More than one key condition may be included, and in fact more than one will help to improve the identification of the outliers, as these key conditions ensure that the specified set of RBETEL moment conditions are valid only under G , and not under B .

However, since the true DGP is unknown, the key conditions will need to be built from the analyst’s understanding of the characteristics of the good data, as well as how these characteristics might be destroyed by outliers. Again these considerations will be problem specific, however we offer a few suggestions for location and regression problems below.

3.6.1 Suggested key conditions for a location problem

Suppose one is interested in inferring the location, μ , of a population associated with the ‘good’ part of an observed dataset, (x_1, \dots, x_n) , which may be contaminated by outliers. A commonly used moment condition for inferring the location of G is $E^G[x - \mu] = 0$. However, in the RBETEL context, the moment condition given by $E^F[(x - \mu)s] = 0$ alone would not be sufficient because it does not provide information to distinguish observations from G and B .

One approach would be to augment the simple moment condition $E^F[(x - \mu)s] = 0$ above with an additional ‘key condition’ arising from a certain assumption about G . For example, it might be

expected that

$$E^G[(x_i - \mu)^3] = 0,$$

which would lead to the RBETEL additional specification that

$$E^F[(x_i - \mu)^3 s_i] = 0.$$

This key condition suggests that the good data distribution should be symmetric around μ , so that observations that cause substantial skewness in the complete dataset may be considered as outliers.

A second possible approach to the specification of key conditions in the location problem is to consider moment conditions derived from a robust Huber-type location estimator, given by $g(x, \theta) = H(x - \mu)$ where for some fixed $\epsilon_0 > 0$, the $H(\cdot)$ function is given by

$$H(\epsilon) = \begin{cases} 1 & \text{if } \epsilon \geq \epsilon_0, \\ \epsilon/\epsilon_0 & \text{if } |\epsilon| < \epsilon_0, \\ -1 & \text{if } \epsilon \leq -\epsilon_0. \end{cases} \quad (38)$$

A key condition in this case could be formed according to

$$E^F[((x_i - \mu) - H(x_i - \mu)) s_i] = 0.$$

A third alternative would be to incorporate selected statistics known for their (Frequentist) robust properties in to the moment conditions. An example we consider is the key condition given by

$$E^F[(x_i - \mu)^2 s_i] - MAD(x_{1:n})^2 = 0,$$

where $MAD(x_{1:n})$ denotes the median absolute deviation of the observations. This key condition incorporates a robust scale estimate of the data to provide information regarding the expected (squared) distance between good data points and the conditional mean value μ .

4 Simulation experiments

In this section, we demonstrate the performance of the RBETEL method under controlled simulation settings. Section 4.1 examines the RBETEL method to estimate location using the moment conditions, introduced in Section 3.6, with a simulated dataset. Then we generate replicated datasets from the location estimation setting with a fixed proportion of outliers in Section 4.2. We examine the average performance of the RBETEL posterior inference when outliers in the data have different sizes. The standard BETEL method is also applied to estimate the location of the data and the inferential results produced by both methods are compared. In Section 4.3, replicated datasets are considered in a linear regression setting when the data may be contaminated by outliers. Some of these outliers are leverage points which are known to have strong influence on inference in the linear regression context. Under this setting, we consider different numbers of outliers in the generated datasets.

4.1 Location estimation using different key conditions

Here we consider a location estimation example and compare the estimated RBETEL posterior distributions when different key conditions suggested in Section 3.6 are included into the set of moment conditions.

We simulate a dataset $y_{1:n} = (y_1, \dots, y_n)$ from the following data generating process (DGP):

$$y_i = \begin{cases} 1 + e_i, & \text{with probability } 0.95, \\ 6 + e_i, & \text{with probability } 0.05, \end{cases} \quad (39)$$

where the errors, e_i , are independently generated from a Normal distribution with zero mean and variance equal to one, i.e. $e_i \stackrel{iid}{\sim} N(0,1)$, for $i = 1, \dots, n$. One hundred observations are generated (i.e. $n = 100$), with 95% of these having a mean of 1, and these observations are referred to as the good data points. The remaining 5% of the observations are considered to be outliers, and they have a mean of 6. Only a single dataset is used here to present a clear picture of how the entire posterior is impacted by the RBETEL conditions. Simulation experiments that employ replicated

datasets are considered in Sections 4.2 and 4.3.

The simulated dataset is plotted in Figure 1 and five outliers ($O_1 = 6.21$, $O_2 = 4.62$, $O_3 = 6.27$, $O_4 = 6.10$, $O_5 = 6.26$) in the simulated data are highlighted. We can see that the sample mean, which is about 1.17, deviates from the designed location for the good data at 1. On the other hand, the median of the data, which is about 0.95, is very close to the desired value of 1. In addition, the designed DGP produces outliers with expected location much larger than the good data. As a result, the complete dataset seems to have an asymmetric distribution and this differs from the distribution of the good data which is symmetric.

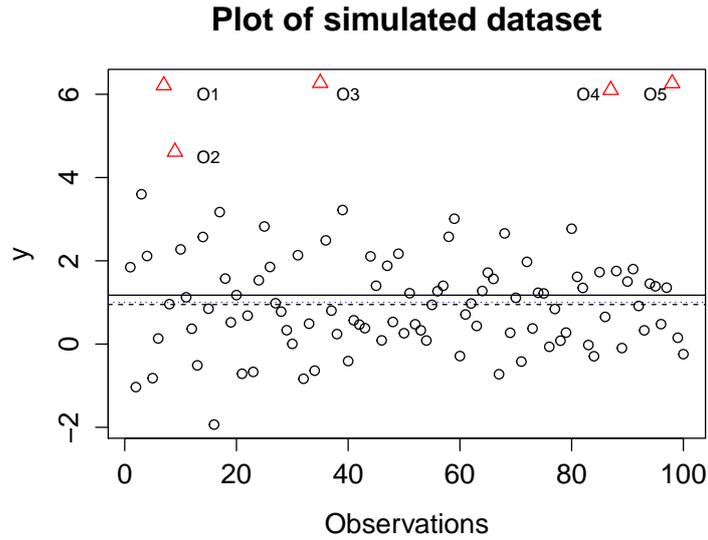


Figure 1: Simulated dataset with 5% of the observations being outliers. The outliers are highlighted with red triangles. The solid black line indicates the sample mean of the data (1.17). The dashed line is the median of the data (0.95). The dotted blue line indicates the value 1 which is the designed location for the good data.

We consider producing posterior distributions for the location of the good data, denoted by μ , using the RBETEL method based on different set of moment conditions. The moment condition $E[(y_i - \mu)s_i]$ is always used, together with some combinations of the following key conditions:

$$C1: E[(y_i - \mu)^3 s_i],$$

$$C2: E[(y_i - \mu) - H(y_i - \mu)] s_i = 0,$$

$$C3: E[(y_i - \mu)^2 s_i] - MAD(y_{1:n})^2 = 0.$$

The function $H(\cdot)$ in $C2$ is given by (38) and the trimming parameter, ϵ_0 , is set to be 1.5. $MAD(y_{1:n})$ in $C3$ denotes the median absolute deviation of the observations.

The BETEL method is also applied to this simulated dataset and inference is based on a single moment condition $E[y - \mu] = 0$.

We produce 30,000 MCMC draws, discarding the first 10,000 draws and keeping the subsequent 20,000 draws for posterior inference. We assign a flat Normal prior with mean equal to zero and variance equal to 100 to the location parameter μ , i.e. $\mu \sim N(0, 100)$. The priors for the indicator vector, s , and the probability parameter, v , are the same as those discussed in Section 3.4.1. The hyperparameters for $\pi(v)$, given in (29), are set to $\alpha_0 = 50$ and $\beta_0 = 5$.

The inferential results produced by the standard BETEL and RBETEL methods are summarized in Table 1 and Table 2, respectively. We report posterior means (Post.Mean), posterior standard deviations (Post.SD) and time-series standard errors (TS.SE) for the location parameter, μ , produced by both methods. For the RBETEL method, we also report the expected proportion of good data, v . In addition, we report the probabilities of each outlier being used in the MCMC iterations; each of these probabilities can be interpreted as the probability of the corresponding observation being a good data point.

The inferential results show that the RBETEL method produces robust inference for the location parameter μ based on moment conditions that include any combination of key conditions $C1$, $C2$ and $C3$. The posterior means are all close to 1 which is the designed location of the good dataset. On the other hand, the posterior mean estimate for parameter μ given by the standard BETEL method is 1.2148, which is near the sample mean of the complete dataset. The posterior standard deviations and the time-series standard errors produced by the RBETEL method with different moment conditions are also similar.

Figure 2 shows the kernel estimates of the posterior densities for μ produced by standard BETEL method, and the RBETEL methods based on different moment conditions. The posterior densities produced by the RBETEL method all center near the designed value 1, while the BETEL posterior does not cover the designed value.

The estimated probabilities of the outliers, O_1 , O_2 , O_3 , O_4 and O_5 , being treated as good data (shown in Table 2) are small in all the cases, while the corresponding probabilities for good data points are all larger than 90%. We find that these probabilities are smaller when condition C_1 , which suggests that the good data have a symmetric distribution, is included into the set of moment conditions. This suggests that a key condition that provides correct information about the distribution of the good data seems to improve the performance of the RBETEL method.

	Post.Mean	Post.SD	TS.SE
μ	1.2148	0.0485	0.0025

Table 1: BETEL posterior summaries for the location parameter given the simulated dataset.

		Estimates			$Pr(S_i = 1)$				
		Post.Mean	Post.SD	TS.SE	O_1	O_2	O_3	O_4	O_5
$C1$	μ	0.95	0.12	0.0047	0.01	0.10	0.01	0.02	0.02
	v	0.90	0.03	0.0004					
$C2$	μ	0.98	0.12	0.0051	0.09	0.14	0.09	0.09	0.08
	v	0.90	0.03	0.0003					
$C3$	μ	0.96	0.12	0.0045	0.19	0.23	0.17	0.18	0.17
	v	0.91	0.03	0.0003					
$C1\&C2$	μ	0.97	0.12	0.0049	0.05	0.16	0.04	0.05	0.05
	v	0.90	0.03	0.0003					
$C1\&C3$	μ	0.94	0.13	0.0051	0.03	0.11	0.03	0.03	0.03
	v	0.90	0.03	0.0004					
$C2\&C3$	μ	0.96	0.12	0.0047	0.10	0.14	0.10	0.09	0.09
	v	0.90	0.03	0.0003					
$C1\&C2$ & $C3$	μ	0.97	0.13	0.0057	0.07	0.19	0.06	0.07	0.06
	v	0.90	0.03	0.0003					

Table 2: RBETEL posterior summary results based on different moment conditions. The left panel reports posterior mean, posterior standard deviation and time-series standard error for parameters μ and v . The right panel shows the probability of each outlier being included in the sub-samples for analysis.

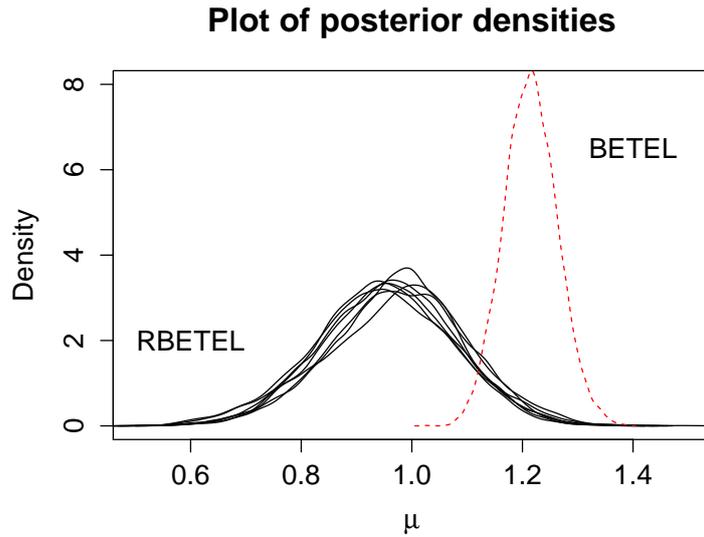


Figure 2: Posterior densities of μ produced by the standard BETEL method (red dashed curve), and the RBETEL methods (black solid curve) based on different moment conditions.

4.2 Inference regarding location under different outlier sizes

In this section, we examine the performance of the RBETEL inference when outliers in the dataset have different sizes. The RBETEL inferential results are also compared with the estimates produced by the standard BETEL method, using replicated datasets.

The DGP we consider is given by

$$y_i = \begin{cases} \mu_0 + e_i, & \text{with probability } 0.95, \\ \xi_0 + e_i, & \text{with probability } 0.05, \end{cases}$$

where μ_0 is the mean for 95% of the observations which are good data, and the other 5% of observations are treated as outliers and they have a mean equal to ξ_0 . The errors are independently generated from a Normal distribution with zero mean and variance equal to one, i.e. $e_i \stackrel{iid}{\sim} N(0, 1)$, for $i = 1, \dots, n$.

We simulate 100 datasets and each consists of $n = 1000$ observations. The mean for the good data is set to be $\mu_0 = 1$ for all the simulated datasets. We consider cases when the mean of the outliers, ξ_0 , is equal to 2, 4 and 6, which correspond to small, medium and large outlier sizes, respectively.

We use all three key conditions examined in the previous section as moment conditions for RBETEL approach and the moment conditions are given by

$$\begin{aligned} E[(y_i - \mu)s_i] &= 0 \\ E[(y_i - \mu)^3 s_i] &= 0 \quad \text{for } i = 1, \dots, n, \\ E[((y_i - \mu) - H(y_i - \mu))s_i] &= 0, \\ E[(y_i - \mu)^2 s_i] - MAD(y_{1:n})^2 &= 0. \end{aligned} \tag{40}$$

It is clear that the data characteristics in (40) are not suitable for the complete dataset when outliers are present. This implies that when applying the standard BETEL method, $E^F[y - \mu] = 0$

is the only moment condition we can use to estimate μ .

We produce 30,000 MCMC draws throughout the simulation experiments, discarding the first 10,000 draws and keeping the subsequent 20,000 draws for posterior inference. We assign a flat Normal prior with a mean equal to zero and a variance equal to 100 to the location parameter μ , i.e. $\mu \sim N(0, 100)$. The priors for the indicator vector, s , and the probability parameter, v , are the same as those discussed in Section 3.4.1. The hyperparameters for $\pi(v)$, given by (29), are set to $\alpha_0 = 500$ and $\beta_0 = 50$.

The inferential results are summarized in Table 3. We report the average posterior means (Av.Post.Mean), average posterior standard deviations (Av.Post.SD) and average time series standard errors (Av.TS.SE). In addition, we look at the proportion of coverage (P.O.C) which is the proportion of the 95% posterior credible intervals (C.I.), out of 100 simulated datasets, that cover the designed location of the good data, i.e. $\mu_0 = 1$.

We can see that the results produced by the standard BETEL method are influenced by outliers in the data. The BETEL posterior means for datasets with small, medium and large outliers are given by 1.0531, 1.1545 and 1.2505, respectively. These posterior mean estimates are all near the sample means associated with the corresponding complete datasets. When the size of outliers is small, 92% of the standard BETEL posteriors still cover the designed location of the good data. The influence of outliers is not so obvious, although on average the posterior means for μ indeed deviate from the designed value of $\mu_0 = 1$. However, it becomes more apparent that the BETEL method is non-robust with respect to outliers when the size of the outliers gets large. The BETEL posterior mean for the location μ deviates further from the designed value when the average size of the outliers is increased. When medium and large sized outliers are present in the data, none of the 95% C.I.s cover the designed mean.

The newly proposed RBETEL method performs well regardless of the size of the outliers, in the sense that the average posterior means are near the designed location for the good data $\mu_0 = 1$. All of the 95% credible intervals cover the desired value in all the cases. The average RBETEL

posterior means for datasets with small, medium and large sized outliers are given by 1.0086, 1.0095 and 1.0071, respectively. In fact, we have not found any evidence suggesting that the performance of RBETEL posterior mean estimator is affected the by average size of the outliers.

We show boxplots of the posterior means produced by both methods for simulated data with different sized outliers in Figure 3. We can see that the posterior means produced by the RBETEL method are concentrated around the designed location of good data for all cases. On the other hand, the posterior means given by standard BETEL method deviate further from the mean of the good data as the average size of outliers increases.

We also find that the posterior standard deviations produced by the RBETEL method are larger than the BETEL posterior standard deviations, which suggests that the RBETEL posteriors are on average more defused than the posteriors produced by the BETEL method under this simulation setting. This is mainly because the RBETEL posterior for μ is obtained by marginalizing over the posterior distributions conditional on subsets of the observations, with the proportion of outliers being unknown. It is also possible that this uncertainty may be offset by sharper posterior inference provided by the uncontaminated ‘good’ data. However, this possibility does not appear to have occurred here. We will return to this point again in Section 4.3.

BETEL				
Size	Av.Post.Mean	Av.Post.SD	Av.TS.SE	P.O.C
Small	1.0531	0.0653	0.0016	0.92
Median	1.1545	0.0663	0.0015	0
Large	1.2505	0.0671	0.0016	0

RBETEL				
Size	Av.Post.Mean	Av.Post.SD	Av.TS.SE	P.O.C
Small	1.0086	0.1352	0.0058	1
Median	1.0095	0.1373	0.0051	1
Large	1.0071	0.1377	0.0055	1

Table 3: Estimation results for μ using BETEL and RBETEL methods. The reported values are average posterior mean, average posterior standard deviation, average time series standard error and the proportion of 95% C.I.s that cover the designed value $\mu_0 = 1$.

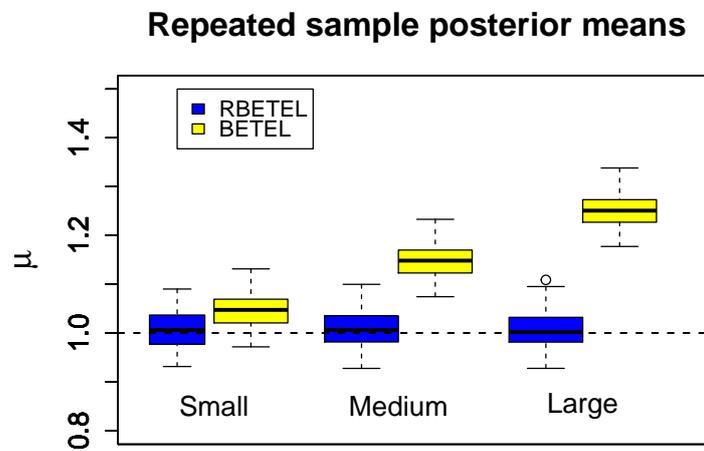


Figure 3: Boxplots of the posterior means produced by the RBETEL and standard BETEL methods. The boxplot on the left (blue)/right (yellow) of each pair corresponds to the posterior means given by the RBETEL/BETEL method. The horizontal line indicates the designed location for the good data, $\mu_0 = 1$.

4.3 Linear regression with different proportions of outliers

In this section we detail the results of a separate simulation experiment where we consider inferring a linear relationship between variables when outliers may be present in the data. We examine the situations when different proportions of observations are treated as outliers. A few outliers in the simulated data are designed to be leverage points and these outliers are designed so that they have strong influence on non-robust inference. This way the difference between RBETEL inference and that of the standard BETEL method can be seen more easily.

We generate the independent variable, x_i , from a Normal distribution with a mean equal to zero and variance equal to 5, i.e.

$$x_i \stackrel{iid}{\sim} N(0,5) \text{ for } i = 1, \dots, n.$$

After sorting the independent variables in an ascending order, the dependent variables, $y_{1:n} = (y_1, \dots, y_n)$, are simulated by firstly generating

$$y_i^\Delta = \begin{cases} \delta_0^* + \delta_1^* x_i + e_i & \text{with probability } v^*, \\ \delta_0^* + \delta_1^* x_i + 3e_i, & \text{with probability } 1 - v^*, \end{cases} \quad (41)$$

and then setting

$$y_i = \begin{cases} y_i^\Delta & \text{for } i = 1, \dots, n-3, \\ y_i^\Delta - 10, & \text{for } i = n-2, n-1, n. \end{cases} \quad (42)$$

The errors e_i are generated independently from a Normal distribution with mean equal to zero and standard deviation equal to one, i.e. $e_i \stackrel{iid}{\sim} N(0,1)$, for $i = 1, \dots, n$. The expected proportion of observations that are good data is given by v^* . The other observations are generated with error standard deviation equal to 3, and they are considered to be outliers. The parameters in the DGP (41) with superscript ‘star’ denote the designed values, and we set $\delta_0^* = 2$ and $\delta_1^* = 1$. We generate $n = 1000$ observations for each simulated dataset, and reduce the last three observations by 10 units so that they are leverage points.

We plot a simulated dataset with $v^* = 0.95$ in Figure 4. We can see that there are a few outliers in this data, and the leverage points on the right of the plot cause the ordinary least square (OLS) estimation of the regression line to have a smaller slope comparing to the designed value. A Frequentist robust regression inferential method proposed by Yohai (1987) is also employed to estimate the regression line, and it can recover the designed regression line very well in this case. The regression line estimated by the robust method basically overlaps with the designed regression line in the plot.

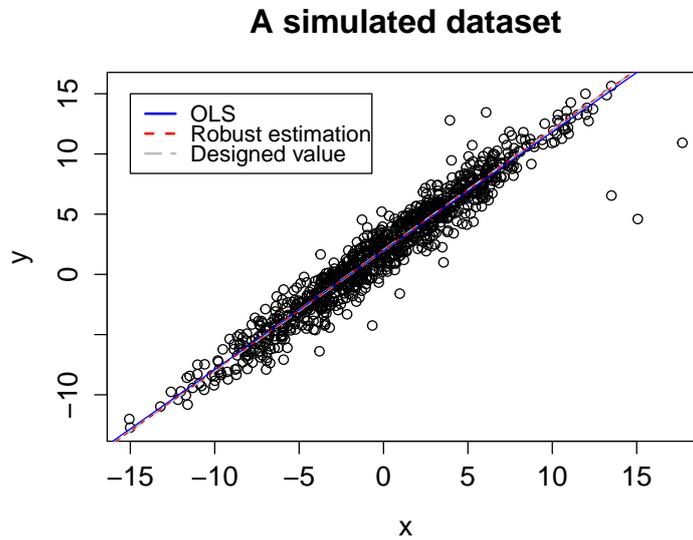


Figure 4: A simulated dataset with 5% outliers and some leverage points. The solid blue line is estimated by OLS. The dashed red line is obtained from robust regression. The dotted black line is the designed regression line.

Commonly used moment conditions for estimating the linear regression model, when outliers are not considered, are given by

$$\begin{aligned}
 E[y_i - \delta_0 - \delta_1 x_i] &= 0 \\
 E[(y_i - \delta_0 - \delta_1 x_i)x_i] &= 0.
 \end{aligned}
 \tag{43}$$

We apply the standard BETEL method to estimate the linear regression relationship of the simulated data based on the moment conditions (43).

We also include the basic moment conditions (43) for the RBETEL inferential method, and these moment conditions are reformulated as

$$\begin{aligned} E[(y_i - \delta_0 - \delta_1 x_i) s_i] &= 0, \\ E[(y_i - \delta_0 - \delta_1 x_i) x_i s_i] &= 0. \end{aligned} \tag{44}$$

In addition, we add some key conditions constructed following the suggestions in Section 3.6.

The first key condition is based on the symmetric distribution assumption for the error term, and it is given by

$$E[(y_i - \delta_0 - \delta_1 x_i)^3 s_i] = 0. \tag{45}$$

This key condition may be violated when a subset of data contain outliers, in particular, when the outliers are asymmetrically distributed around the designed linear regression line.

Another set of key conditions is built incorporating a Huber-type location estimator for the error term. These key conditions are given by

$$\begin{aligned} E[((y_i - \delta_0 - \delta_1 x_i) - H(y_i - \delta_0 - \delta_1 x_i)) s_i] &= 0, \\ E[((y_i - \delta_0 - \delta_1 x_i) x_i - H(y_i - \delta_0 - \delta_1 x_i) x_i) s_i] &= 0, \end{aligned} \tag{46}$$

where Huber's function $H(\cdot)$ is given in (38) and we set the trimming parameter ϵ_0 to be 1.5.

We construct the last RBETEL key condition for linear regression by incorporating a robust scale for the error term. We notice that the distinguishing characteristics of the outliers and the good data in the linear regression case is that an outlier is further away from the robust regression line than a good data point. A robust scale estimator for the errors can be seen as a summary statistic of the data that is related to the squared vertical distance of the good data to a regression line. We employ the robust regression methodology proposed by Yohai (1987) and denote the robust scale estimate for the error term by $T(x_{1:n}, y_{1:n})$, and then the key condition is given by

$$E\left[\left((y_i - \delta_0 - \delta_1 x_i)^2 - T(x_{1:n}, y_{1:n})\right) s_i\right] = 0. \tag{47}$$

We examine the cases when v^* equals 1, 0.98, 0.95 and 0.92 in our experiments. 100 replications of datasets are simulated from the designed DGP for each v^* . For both methods considered here, we produce and discard the first 20,000 MCMC draws and keep the subsequent 30,000 draws for inference. A flat Normal prior with a mean equal to zero and a variance equal to 100 is assigned to the parameters δ_0 and δ_1 . When applying the RBETEL method, the priors for s and v are specified according to Section 3.4. The hyperparameters for $\pi(v)$ are set to $\alpha_0 = 500$ and $\beta_0 = 50$.

The Bayesian inferential results for parameters δ_0 and δ_1 are summarized in Table 4. Both BETEL and RBETEL methods estimate the parameters accurately when the datasets do not contain outliers. The average posterior means are close to the desired values and all the posterior 95% C.I.s cover the designed values for the parameters. When outliers are present in the data, we can see that the posterior means produced by the standard BETEL method deviate from the designed values and the proportions of 95% C.I.s that cover the designed parameter values drop. In particular, the percentage of coverage drops to around 35% for the slope parameter, δ_1 , because of the effect of leverage points. On the other hand, the RBETEL method shows robustness with respect to outliers. The RBETEL average posterior means for the parameters are close to the designed values that generate the good data and the percentages of posterior 95% C.I.s covering the desired values are all higher than 97% regardless of the proportion of outliers in the data.

Figure 5 shows the boxplots of the posterior means produced by the BETEL and RBETEL methods for the linear regression coefficients δ_0 and δ_1 . In all cases, the posterior means produced by the RBETEL method for both parameters are concentrated around the designed parameter values. However, the posterior means produced by the standard BETEL method are not centered around the designed values when outliers are present. In particular, the BETEL posterior means for the slope coefficient δ_1 are always below the designed value due to the effect of leverage points. In this setting, we also note that the average posterior standard deviation under the RBETEL method is actually slightly lower when compared against that from the BETEL method, highlighting the fact that trimming out contaminants from the dataset may actually sharpen the resulting posterior.

BETEL									
v^*	δ_0				δ_1				P.O.C
	Av.Post.Mean	Av.Post.SD	Av.TS.SE	P.O.C	Av.Post.Mean	Av.Post.SD	Av.TS.SE	P.O.C	
1	2.00245	0.03185	0.00195	1	0.99975	0.00633	0.00014	1	
0.98	1.96481	0.03827	0.00289	0.82	0.97791	0.01265	0.00049	0.34	
0.95	1.96470	0.04100	0.00324	0.82	0.97712	0.01281	0.00048	0.35	
0.92	1.96703	0.04429	0.00375	0.78	0.97742	0.01298	0.00049	0.34	

RBETEL									
v^*	δ_0				δ_1				P.O.C
	Av.Post.Mean	Av.Post.SD	Av.TS.SE	P.O.C	Av.Post.Mean	Av.Post.SD	Av.TS.SE	P.O.C	
1	2.00227	0.02015	0.00285	1	0.99973	0.01043	0.00012	1	
0.98	1.99816	0.03117	0.00257	0.97	0.99815	0.01241	0.00049	0.98	
0.95	1.99939	0.03827	0.00388	0.98	0.99921	0.00853	0.00029	0.99	
0.92	2.00163	0.03539	0.00331	0.98	0.99941	0.00750	0.00025	0.98	

Table 4: Posterior summaries for δ_0 and δ_1 using traditional BETEL and RBETEL methods in the linear regression setting. The reported values are average posterior mean (Av.Post.Mean), average posterior standard deviation (Av.Post.SD), average time series standard error (Av.TS.SE) and proportion of coverage (P.O.C) of the designed values $\delta_0^* = 2$ and $\delta_1^* = 1$.

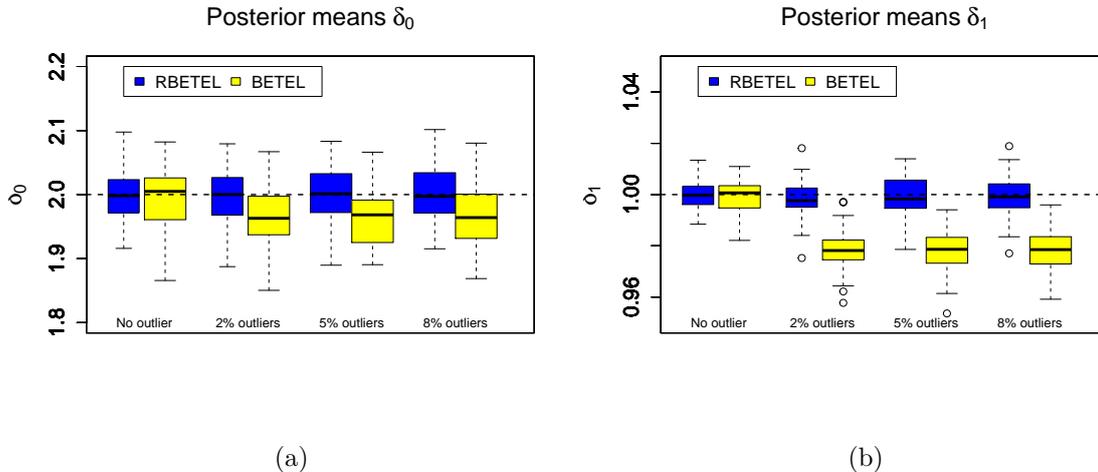


Figure 5: Boxplots of the posterior means for the parameters δ_0 and δ_1 . The boxplot on the left (blue)/right (yellow) of each pair corresponds to the posterior means given by the RBETEL/BETEL method. The horizontal line indicates the designed parameter values for the good data.

5 Empirical example

We apply the RBETEL method to estimate the relationship between log average brain weight (y) and log average body weight (x) for a dataset relating to sixty-five species of land animals. The same example is employed in the work of Rousseeuw and Van Zomeren (1990), who develop a Frequentist approach to unmask multivariate outliers and leverage points. Following Rousseeuw and Van Zomeren (1990), a linear regression model is employed to estimate the relationship, i.e.

$$y_i = \delta_0 + \delta_1 x_i + \epsilon_i \text{ for } i = 1, \dots, 65, \quad (48)$$

where the errors, ϵ_i , are assumed to be iid.

A scatter plot of the data is given in Figure 6, showing a clear linear relationship between the two variables for most of the observations. We use OLS estimation and robust M-estimation methods to produce a preliminary analysis of the linear relationship and estimation results are summarized in Table 5. There are three obvious outliers at the right hand side of the plot and these outliers are leverage points. A regression line produced by robust M-estimation, with parameter estimates given by $\hat{\delta}_0^M = 2.12$ and $\hat{\delta}_1^M = 0.75$, seems to fit the majority of the data well. On the other hand,

non-robust OLS estimation produces a regression line with an apparent lower slope due to the effect of the outliers, and the parameter estimates in this case are given by $\hat{\delta}_0^{OLS} = 2.17$ and $\hat{\delta}_1^{OLS} = 0.59$.

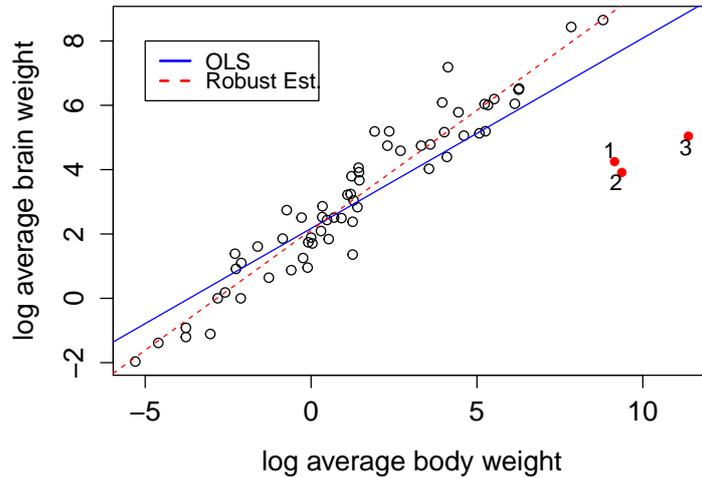


Figure 6: Plot of log average brain weights against log average body weights of sixty-five land animal species. The solid blue line is estimated by OLS. The dashed red line is obtained from a Frequentist robust M-estimation method.

OLS				
Parameter	Estimate	Std.Error	t-statistic	p-value
δ_0	2.1717	0.1620	13.40	$< 2e - 16$
δ_1	0.5915	0.0411	14.37	$< 2e - 16$

Robust estimation				
Parameter	Estimate	Std.Error	t-statistic	p-value
δ_0	2.1175	0.0981	21.59	$< 2e - 16$
δ_1	0.7460	0.0249	29.94	$< 2e - 16$

Table 5: Estimation results produced by OLS method and robust M-estimation method.

We employ the same moment conditions used in the simulation experiment discussed in Section 4.3 for standard BETEL and RBETEL methods. For comparison, we also infer the linear relationship using a parametric Bayesian method based on model (48), and assuming a Student-t

distribution for the errors with degree of freedom ν , i.e. $\epsilon_i \stackrel{iid}{\sim} St(\nu)$. We assign a gamma prior to the degree of freedom parameter, ν , with shape equal to 10 and rate equal to 5. We give a flat Normal prior with mean zero and variance equal to 100 to the parameters δ_0 and δ_1 . For RBETEL estimation, the prior distribution of the indicator vector, s , conditional on the probability parameter, v , is shown in Section 4.3.4.1. The prior of the probability parameter, v , given by (29) is a truncated Beta distribution and we set the hyperparameters to be $\alpha_0 = 30$ and $\beta_0 = 15$. We produce 50,000 MCMC draws from the posterior distribution, and use the first 20,000 draws to warm-up the chains and the subsequent 30,000 draws to make inference.

The inferential results are summarized in Table 6, and Figure 7 provides the plots of the marginal posterior densities for the parameters δ_0 and δ_1 , respectively. We can see that the RBETEL posterior mean estimates for the parameters, $\hat{\delta}_0^{RBETEL} = 2.1482$ and $\hat{\delta}_1^{RBETEL} = 0.7512$, are similar to the robust Frequentist estimates produced by M estimation. The posterior mean estimates produced by the standard BETEL method are given by $\hat{\delta}_0^{BETEL} = 2.2016$ and $\hat{\delta}_1^{BETEL} = 0.6055$, which are similar to those estimated by the non-robust OLS method. We find that the 95% credible intervals for δ_1 produced by the two methods considered here have little overlap. This suggests that inference about the relationship between the variables estimated by these two methods are significantly different. The parametric Bayesian approach assuming a Student-t distribution for the error term shows some extent of robustness with respect to outliers. The posterior mean estimates produced by this parametric approach for δ_0 and δ_1 are 2.1114 and 0.7191, respectively, and they are both between the corresponding estimates produced by the standard BETEL and RBETEL methods. Notably, however, the RBETEL posteriors here are much sharper than are the posteriors that result from the parametric and the BETEL methodologies.

The estimated regression lines with parameter values given by the posterior means are added to the data plot in Figure 8. We can see that the regression line estimated by the RBETEL method fits the majority of the observations well. The red points marked by number 1, 2 and 3 in Figure 8, are the outliers identified by the RBETEL method, while the corresponding estimated probabilities that these points are ‘good’, i.e. $Pr(s_i = 1)$, are 0.053, 0.032 and 0.001, respectively. All other

observations have estimated probabilities that they are ‘good’ that are greater than 0.8. Compared to the robust estimate produced by the RBETEL method, the regression line produced by the traditional BETEL method has an obviously lower slope due to the effect of the leverage points. The estimated regression line produced by the parametric Bayesian approach with Student-t errors is quite similar to the one produced by RBETEL method, but it has a lower slope. Presumably, the influence of outliers has been reduced by using the fat-tailed Student-t distribution, but it is unable to capture the asymmetry of the data distribution due to the presence of leverage points.

BETEL				
Parameter	Post.Mean	95% C.I.	Post.SD	T.S.S.E
δ_0	2.2016	(1.9893, 2.4433)	0.1171	0.0120
δ_1	0.6055	(0.5137, 0.7041)	0.0496	0.0057

RBETEL				
Parameter	Post.Mean	95% C.I.	Post.SD	T.S.S.E
δ_0	2.1482	(1.9389, 2.3606)	0.1098	0.0081
δ_1	0.7512	(0.7029, 0.8033)	0.0256	0.0009

Parametric Bayesian				St error
Parameter	Post.Mean	95% C.I.	Post.SD	T.S.S.E
δ_0	2.1114	(1.7590, 2.4521)	0.1676	0.0223
δ_1	0.7191	(0.6360, 0.8014)	0.0221	0.0022

Table 6: Summary of posterior inferential results given by the BETEL method, RBETEL method and Bayesian parametric estimation with Student-t errors.

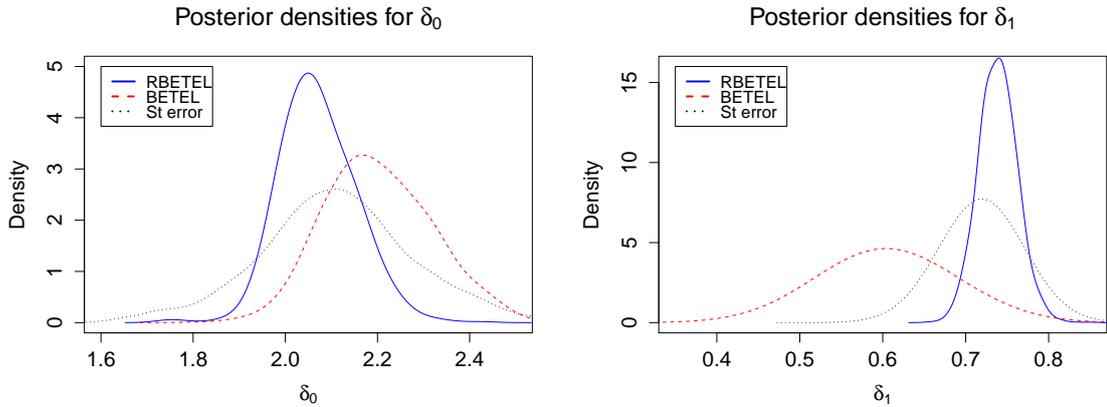


Figure 7: Plot of posterior densities obtained using the BETEL method, RBETEL method and a traditional Bayesian parametric regression method under an assumption of Student-t errors.

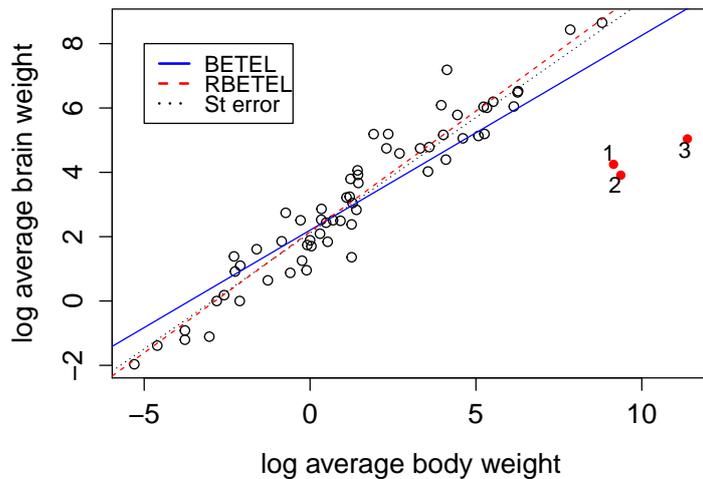


Figure 8: Regression lines obtained by the BETEL, RBETEL methods and traditional Bayesian parametric approach with Student-t errors. The red points 1, 2 and 3 are outliers, as identified by the RBETEL method.

6 Conclusion and future work

This paper develops a robust Bayesian exponentially tilted empirical likelihood method which produces robust inference with respect to outliers for models based on moment conditions. The new RBETEL method is built upon the BETEL framework proposed by Schennach (2005). We point

out that the empirical likelihood function in the BETEL framework is closely related to the EL ratio which is a nonparametric statistic for testing the validity of the moment conditions. Therefore, we reinterpret the BETEL method under the framework proposed by Bissiri *et al.* (2016), where a loss function is constructed based on the EL ratio. The resulting posterior for the parameters can be shown as the representation of the subjective uncertainty in the minimizer of the expected loss. Also inspired by the work of Chib *et al.* (2017), we propose to reformulate the moment conditions in the RBETEL approach by introducing an indicator vector which separates the complete dataset into good data and outliers. The RBETEL loss function is constructed as the EL ratio evaluated using a subset of the observations given by the indicator vector. Careful construction of key conditions which ensure that the RBETEL moment conditions are valid for the good data but invalid when outliers are present, ensures that the RBETEL approach is able to identify outliers from the dataset, and therefore produce robust inference about the parameters of interest.

We conduct simulation experiments to assess the performance of the RBETEL method under simple location and linear regression settings. The RBETEL method produces accurate posterior distributions for the model parameters when outliers are present, in the sense that all marginal posteriors are centered over the relevant designed parameter values. This is not always true for the traditional Bayesian method, particularly when large outliers are present. In addition, when the data does not contain outliers, the RBETEL method seems to perform as well as the original BETEL method. We also illustrate the RBETEL method under an empirical setting where the relationship between log average brain weight and log average body weight of sixty-five land animal species is of interest. The empirical data contain outliers which act as leverage points. The RBETEL method is able to produce robust inference regarding the relationship between the two variables and identify influential outliers in the data.

This paper only considers the RBETEL method under simple linear settings, but we expect that this method can be applied to a wide range of problems. Of course, this method is built for moment condition models, and some well known moment based problems that we will consider in the future include instrumental variable regression and stochastic volatility settings.

Appendix: derivation of the RBETEL posterior with given loss function

In this appendix, we derive the joint RBETEL posterior for the parameters θ and s conditional on given proportion of outliers, v . Given the loss function (37), we aim to find a representation of subjective uncertainty in the parameter set (θ, s) that minimizes the expected loss given by

$$E^F [\tilde{l}(\theta, s; x_{1:n})] = \int \tilde{l}(\theta, s; x_{1:n}) dF,$$

where F is the unknown distribution that generates the data, given by (14).

Following the discussion in Section 2.2, we want to find the probability measures $\tilde{\pi}(\theta | s, x_{1:n})$ and $\tilde{\pi}(s | v, x_{1:n})$ which are the minimizers of

$$\sum_s \left[\left(\int \int \tilde{l}(\theta, s; x_{1:n}) dF \pi(\theta | s, x_{1:n}) d\theta \right) \pi(s | v, x_{1:n}) \right]. \quad (49)$$

This can be done by minimizing the finite sample version of (49) recursively subject to probability measures $\pi(\theta | s, x_{1:n})$ and $\pi(s | v, x_{1:n})$. Note that the joint probability measure $\tilde{\pi}$ over the pair (θ, s) denotes the target minimizer of the loss function (49).

First, we define a cumulative loss function

$$\begin{aligned} L_1(\pi(\theta | s, x_{1:n}); \pi(\theta), x_{1:n}) &= \\ & \int \tilde{l}(\theta, s; x_{1:n}) \pi(\theta | s, x_{1:n}) d\theta + \int \pi(\theta | s, x_{1:n}) \log \left(\frac{\pi(\theta | s, x_{1:n})}{\pi(\theta)} \right) d\theta \\ &= \int \pi(\theta | s, x_{1:n}) \log \left(\frac{\pi(\theta | s, x_{1:n})}{\exp\{-\tilde{l}(\theta, s; x_{1:n})\} \pi(\theta)} \right) d\theta. \end{aligned} \quad (50)$$

According to Bissiri *et al.* (2016), this loss function is the Bayesian finite sample version of

$$\int \int \tilde{l}(\theta, s; x_{1:n}) dF \pi(\theta | s, x_{1:n}) d\theta.$$

The cumulative loss function (50) takes the form of KL divergence, so it is straightforward to find

that the minimizer of (50) is given by

$$\begin{aligned}\tilde{\pi}(\theta | s, x_{1:n}) &= \arg \min_{\pi(\theta | s, x_{1:n})} L_1(\pi(\theta | s, x_{1:n}); \pi(\theta), x_{1:n}) \\ &= \frac{\exp\{-\tilde{l}(\theta, s; x_{1:n})\} \pi(\theta)}{\int \exp\{-\tilde{l}(\theta, s; x_{1:n})\} \pi(\theta) d\theta}.\end{aligned}\tag{51}$$

Substituting (51) into (50), we have the expression for the minimum of L_1 which is given by

$$\begin{aligned}\min L_1(\pi(\theta | s, x_{1:n}); \pi(\theta), x_{1:n}) &= L_1(\tilde{\pi}(\theta | s, x_{1:n}); \pi(\theta), x_{1:n}) \\ &= -\log \int \exp\{-\tilde{l}(\theta, s; x_{1:n})\} \pi(\theta) d\theta.\end{aligned}\tag{52}$$

Then the second cumulative loss function is corresponding to the expected loss

$$\sum_s \left[\min \left(\int \int \tilde{l}(\theta, s; x_{1:n}) \pi(\theta | s, x_{1:n}) d\theta dF \right) \pi(s | v, x_{1:n}) \right],$$

it is given by

$$\begin{aligned}L_2(\pi(s | v, x_{1:n}); \pi(s | v), x_{1:n}) &= \\ &= \sum_s \pi(s | v, x_{1:n}) \min L_1(\pi(\theta | s, x_{1:n}); \pi(\theta), x_{1:n}) + \sum_s \pi(s | v, x_{1:n}) \log \left(\frac{\pi(s | v, x_{1:n})}{\pi(s | v)} \right) \\ &= \sum_s \pi(s | v, x_{1:n}) \log \left(\frac{\pi(s | v, x_{1:n})}{\exp\{-\min L_1(\pi(\theta | s, x_{1:n}); \pi(\theta), x_{1:n})\} \pi(s | v)} \right) \\ &= \sum_s \pi(s | v, x_{1:n}) \log \left(\frac{\pi(s | v, x_{1:n})}{\int \exp\{-\tilde{l}(\theta, s; x_{1:n})\} \pi(\theta) d\theta \pi(s | v)} \right).\end{aligned}\tag{53}$$

It is straightforward to find that the minimizer of (53) is given by

$$\begin{aligned}\tilde{\pi}(s | v, x_{1:n}) &= \arg \min_{\pi(s | v, x_{1:n})} L_2(\pi(s | v, x_{1:n}); \pi(s | v), x_{1:n}) \\ &= \frac{\int \exp\{-\tilde{l}(\theta, s; x_{1:n})\} \pi(\theta) d\theta \pi(s | v)}{\sum_s \int \exp\{-\tilde{l}(\theta, s; x_{1:n})\} \pi(\theta) d\theta \pi(s | v)},\end{aligned}\tag{54}$$

and then the minimum of L_2 is given by

$$\begin{aligned} \min L_2(\pi(s | v, x_{1:n}); \pi(s | v), x_{1:n}) &= L_2(\tilde{\pi}(s | v, x_{1:n}); \pi(s | v), x_{1:n}) \\ &= -\log \sum_s \int \exp\{-\tilde{l}(\theta, s; x_{1:n})\} \pi(\theta) d\theta \pi(s | v). \end{aligned} \quad (55)$$

Finally, the desired joint RBETEL posterior is given by the product of (51) and (54), i.e.

$$\begin{aligned} \pi_{\text{RBETEL}}(\theta, s | v, x_{1:n}) &= \tilde{\pi}(\theta | s, x_{1:n}) \tilde{\pi}(s | v, x_{1:n}) \\ &\propto \pi(\theta) \pi(s | v) \exp\{-\tilde{l}(\theta, s; x_{1:n})\}. \end{aligned} \quad (56)$$

References

- Berger, J. O. (1994). An overview of robust Bayesian analysis. *Test*, 3(1):5–124.
- Bissiri, P. G., Holmes, C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.
- Bornn, L., Shephard, N., and Solgi, R. (2015). Moment conditions and Bayesian nonparametrics. *arXiv preprint arXiv:1507.08645*.
- Box, G. and Draper, N. (1987). *Empirical Model Building and Response Surfaces*. John Wiley and Sons.
- Chamberlain, G. and Imbens, G. W. (2003). Nonparametric applications of Bayesian inference. *Journal of Business & Economic Statistics*, 21(1):12–18.
- Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics*, 115(2):293–346.
- Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335.
- Chib, S., Shin, M., and Simoni, A. (2017). Bayesian estimation and comparison of moment condition models. *arXiv:1606.02931*.
- Hall, A. R. (2005). *Generalized method of moments*. Oxford University Press.
- Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50:1029–1054.
- Imbens, G. W. (1997). One-step estimation for over-identified generalized method of moments models. *Review of Economic Studies*, 64:359–383.
- Imbens, G. W., Spady, R. H., and Johnson, P. (1998). Information theoretic approach to inference in moment condition models. *Econometrica*, 66:333–357.
- Lazar, N. A. (2003). Bayesian empirical likelihood. *Biometrika*, 90:319–326.

- Meer, P., Mintz, D., Rosenfeld, A., and Kim, D. Y. (1991). Robust regression methods for computer vision: A review. *International Journal of Computer Vision*, 6(1):59–70.
- Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, 72(1):219–255.
- Newey, W. K. and West, K. D. (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review*, pages 777–787.
- Ortelli, C. and Trojani, F. (2005). Robust efficient method of moments. *Journal of Econometrics*, 128(1):69–97.
- Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, pages 90–120.
- Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75:237–249.
- Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22:300–325.
- Ronchetti, E. and Trojani, F. (2001). Robust inference with GMM estimators. *Journal of Econometrics*, 101:37–69.
- Rousseeuw, P. J. and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639.
- Schennach, S. M. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika*, 92:31–46.
- Schennach, S. M. (2007). Point estimation with exponentially tilted empirical likelihood. *The Annals of Statistics*, 2:634–672.
- Stigler, S. M. (1973). Simon Newcomb, Percy Daniell, and the history of robust estimation 1885–1920. *Journal of the American Statistical Association*, 68(344):872–879.
- Stock, J. H., Wright, J. H., and Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4):518–529.

- Wooldridge, J. M. (2001). Applications of generalized method of moments estimation. *The Journal of Economic Perspectives*, 15(4):87–100.
- Yin, G. (2009). Bayesian generalized method of moments. *Bayesian Analysis*, 4(2):191–207.
- Yohai, V. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 15:642–656.