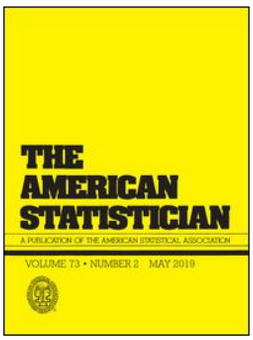

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

A Robustified Posterior for Bayesian Inference on a Large Number of Parallel Effects

J. G. Liao, Arthur Berg & Timothy L. McMurry

To cite this article: J. G. Liao, Arthur Berg & Timothy L. McMurry (2020): A Robustified Posterior for Bayesian Inference on a Large Number of Parallel Effects, The American Statistician, DOI: 10.1080/00031305.2019.1701549

To link to this article: https://doi.org/10.1080/00031305.2019.1701549

THE AMERICAN STATISTICIAN
2020, VOL. 00, NO. 0, 1–7: General
https://doi.org/10.1080/00031305.2019.1701549

A Robustified Posterior for Bayesian Inference on a Large Number of Parallel Effects

J. G. Liao[a], Arthur Berg[a], and Timothy L. McMurry[b]

[a]Division of Biostatistics & Bioinformatics, Pennsylvania State University, Hershey, PA; [b]Division of Biostatistics, University of Virginia, Charlottesville, VA

ABSTRACT

Many modern experiments, such as microarray gene expression and genome-wide association studies, present the problem of estimating a large number of parallel effects. Bayesian inference is a popular approach for analyzing such data by modeling the large number of unknown parameters as random effects from a common prior distribution. However, misspecification of the prior distribution can lead to erroneous estimates of the random effects, especially for the largest and most interesting effects. This article has two aims. First, we propose a robustified posterior distribution for a parametric Bayesian hierarchical model that can substantially reduce the impact of a misspecified prior. Second, we conduct a systematic comparison of the standard parametric posterior, the proposed robustified parametric posterior, and nonparametric Bayesian posterior which uses a Dirichlet process mixture prior. The proposed robustified posterior when combined with a flexible parametric prior can be a superior alternative to nonparametric Bayesian methods.

ARTICLE HISTORY
Received October 2018
Accepted November 2019

KEYWORDS
Large-scale data; Order statistics; Robust inference

1. Introduction

In the past decades, new technologies such as gene microarray and genome-wide association studies have fundamentally changed the landscape of biomedical research. Instead of studying one gene or one single nucleotide polymorphism (SNP) at a time, these technologies allow us to study thousands of genes or SNPs simultaneously. Bayesian approaches have proven effective for analyzing such data by modeling a large number of parallel parameters for individual genes or SNPs as random effects from a common prior. Bayesian methods can improve inference by borrowing information from other genes and by incorporating useful structure such as modeling a large proportion of the genes or SNPs as having no effect on the outcome. This approach automatically adjusts for multiple comparisons and selection bias inherent in the large-scale data setting (Johnstone and Silverman 2004; Efron 2010).

A canonical model for this data structure is

$$y_i = \theta_i + \varepsilon_i, \quad i = 1, \ldots, p, \tag{1}$$

where y_i is the observed measurement, θ_i is the unknown true parameter of interest for the ith gene or SNP, and ε_i is an unobserved random error. For example, θ_i could represent the mean of log expression level for the ith gene probe from a gene expression array with several thousand probes and y_i be the sample average over different samples. In a genome-wide association study, θ_i can be the log odds ratio for the association of disease and the ith SNP and $\hat{\theta}_i$ be an estimator. For many applications, there is often little information to distinguish one θ_i from the others before the data is collected, and in such situations, we can consider the θ_i exchangeable (Gelman et al. 2014, chap. 5). For this article, we shall treat θ_i as random effects drawn from a density π_0, and consider π_0 as a smooth or limiting form of the empirical distribution of the underlying $\theta_1, \ldots, \theta_p$ to be estimated. Letting $\theta = (\theta_1, \ldots, \theta_p)$ and $y = (y_1, \ldots, y_p)$, a suitable posterior distribution that facilitates the inference about θ is

$$f(\theta \mid y, \pi_0) \propto f(y \mid \theta)\pi_0(\theta).$$

We would like to approximate $f(\theta \mid y, \pi_0)$ as closely as possible in our Bayesian inference.

In practice, however, π_0 is usually unknown. A standard way to take advantage of such data structure is through a parametric Bayesian hierarchical model (Gelman et al. 2014, chap. 5) as follows:

$$\begin{aligned} y_i \mid \theta_i &\sim f(y_i \mid \theta_i) \\ \theta_i \mid \eta &\sim f(\theta_i \mid \eta) \\ \eta &\sim f(\eta), \end{aligned} \tag{2}$$

where $f(y_i|\theta_i)$ is defined by the density of error ε_i in (1), $y_1, \ldots, y_p$ are independent given $\theta_1, \ldots, \theta_p$, and $\theta_1, \ldots, \theta_p$ are independent given η. This approach can easily incorporate prior information about the structure of π_0 for improved inference about θ_i. For example, we can choose the working prior $f(\theta_i|\eta)$ to be a Laplace family with scale parameter η if we believe that π_0 is unimodal and long-tailed and a normal family if short-tailed. When the shape of $f(\theta_i|\eta)$ is misspecified and severely deviated from π_0, however, it can lead to inferior inference. For example, excessive shrinkage and therefore bias can occur if a working prior $f(\theta_i \mid \eta)$ has much shorter tails than π_0. This misspecification of working prior can be a serious concern because, while some information about π_0 may be available in a particular application, the information is typically insufficient to

CONTACT J. G. Liao jliao@phs.psu.edu Division of Biostatistics & Bioinformatics, Pennsylvania State University, Hershey, PA 17033.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TAS.

determine how heavy the tails should be, which can substantially influence the extremal effects of θ_i, usually the effects of most practical interest.

Alternatively, nonparametric and semiparametric Bayesian methods (Do, Mueller, and Tang 2005; Bogdan, Ghosh, and Tokdar 2008; Kim, Dahl, and Vannucci 2009; Muralidharan 2010; Martin and Tokdar 2012; Müller and Mitra 2013) can be used that impose minimum structure on π_0. A popular approach is to specify the working prior for θ_i as generated from a Dirichlet process mixture, as described in more detail in Section 4. Although such a working prior generally performs reasonably well for a wide range of π_0, it may not be optimal due to its weaker prior information. Additionally, its focus on flexibility of the model can make it difficult for a statistician to incorporate useful prior information (Carlin and Murray 2013; Hoff 2013; O'Hagan 2013). However, to our knowledge, no systematic comparison of parametric and nonparametric approaches, via simulation study or real datasets, is available in the setting of a large number of effects (p in the order of thousands) to give empirical guidance in real applications.

This article has two aims. First, we propose a simple new method to robustify the posterior distribution of θ_i for parametric hierarchical model (2) by using the asymptotic behavior of order statistics. A unique feature of the proposed method is that it can substantially improve the posterior distribution when $f(\theta_i|\eta)$ is misspecified without affecting the posterior under a correctly specified working prior. Therefore, it can be broadly used. Second, we conduct a systematic performance comparison of the standard parametric posterior, the proposed robustified parametric posterior, and a nonparametric Bayesian posterior which uses a Dirichlet process mixture prior with a normal base distribution. Our study shows that while the nonparametric Bayesian method does provide reasonable performance under different forms of π_0, it can perform poorly when π_0 is severely deviated from normal. The proposed robustified posterior when combined with a flexible parametric prior can be a superior alternative to nonparametric Bayesian methods.

2. The Robustified Posterior

We now present the key result of how to robustify the standard posterior distribution for a general working prior density π given by

$$f(\theta \mid y, \pi) \propto f(y \mid \theta)\pi(\theta). \tag{3}$$

To develop the robustified version denoted by $f_{\text{robust}}(\theta \mid y, \pi)$, we need the following assumption, which is reasonable if y_i is to be a measurement of θ_i.

Assumption 1. We assume the distribution of $y_i \mid \theta_i$ is strictly stochastically increasing in θ_i.

Let Φ_i be the cumulative distribution function of the errors $\varepsilon_i = y_i - \theta_i$ in (1) and let ϕ_i be the corresponding density. We then have $f(y_i \mid \theta_i) = \phi_i(y_i - \theta_i \mid \theta_i)$. Here we allow the distribution of ε_i to depend on θ_i to be more general. Let

$$u_i = \Phi_i(y_i - \theta_i \mid \theta_i). \tag{4}$$

It follows immediately that $u_i \overset{\text{iid}}{\sim} \text{Unif}(0,1)$ when $\varepsilon_i \sim \phi_i(\cdot \mid \theta_i)$ independently for $i = 1, \ldots, p$ as specified in canonical model (1). For a given y_i, $\Phi_i(y_i - \theta_i \mid \theta_i)$ is a strictly decreasing function of θ_i by Assumption 1 and therefore θ_i can be written as a function of u_i: $\theta_i = g_i(u_i)$. We will reformulate the posterior of θ_i, given in (3), in terms of u_i. Then the change-of-variables formula rewrites the posterior (3) as

$$f(u \mid y, \pi) \propto \pi\left(g(u)\right) \phi(y - g(u) \mid g(u)) \left|\frac{d\theta}{du}\right|. \tag{5}$$

where $u = (u_1, \ldots, u_p)$, $g(u) = (g_1(u_1), \ldots, g_p(u_p))$, and $\phi = \phi_1 \times \cdots \times \phi_p$.

An interesting special case occurs when the distribution of ε_i does not depend on θ_i. In this case, $\theta_i = g_i(u_i) = y_i - \Phi_i^{-1}(u_i)$ and $\left|\frac{du}{d\theta}\right| = \prod_{i=1}^{p} \phi_i(y_i - \theta_i)$. Therefore the posterior distribution (5) has a particularly simple form

$$f(u \mid y, \pi) = \pi\left(y - \Phi^{-1}(u)\right). \tag{6}$$

We now return to the general case of (5). Let $\tilde{u} = (u_{[1]}, \ldots, u_{[p]})$ denote the order statistics of $u = (u_1, \ldots, u_p)$. Then a draw u from posterior distribution $f(u \mid y, \pi)$ can be decomposed into two steps:

$$\tilde{u} \sim f(\tilde{u} \mid y, \pi), u \sim f(u \mid \tilde{u}, y, \pi). \tag{7}$$

If the working prior π is misspecified, both distributions in (7) can be distorted from their corresponding distribution under the correctly specified prior π_0. The key idea of our proposed robustified posterior distribution is to replace $f(\tilde{u} \mid y, \pi)$ in (7) with the asymptotic approximation of $f(\tilde{u} \mid y, \pi_0)$, which turns out not to depend on π_0. In the rest of this section, we shall assume that y is generated under model (1) with $\theta_i \sim \pi_0$. In what follows, we show the asymptotic limit of $f(\tilde{u} \mid y, \pi_0)$ is available without knowledge of the correct prior π_0. More specifically, $f(\tilde{u} \mid y, \pi_0)$ converges to a delta distribution on $\{\frac{1}{p+1}, \ldots, \frac{p}{p+1}\}$. The key insight is that when p is large and under the correct π_0, $\tilde{u}$ is well approximated by the quantiles of the uniform distribution on $[0, 1]$; this is the same rationale as justifies the widely used QQ-plot for distribution checking. We formalize this in the following theorem (proof provided in the Appendix).

Theorem 1. Let y be generated under model (1) with $\theta_i \sim \pi_0$. Let the order statistics $\tilde{u} = (u_{[1]}, \ldots, u_{[p]})$ be drawn from $f(\tilde{u} \mid y, \pi_0)$. Then

$$\sup \left| u_{[i]} - \frac{i}{p+1} \right| \xrightarrow{p} 0,$$

except on a small subset of y whose probability can be made as small as any $\delta > 0$.

We therefore propose to fix the $\tilde{u}$ in $f(u \mid \tilde{u}, y, \pi)$ in the right-hand side in (7) to be its asymptotic approximation $u_{[i]} = i/(p+1)$ derived under $f(\tilde{u} \mid y, \pi_0)$ and define our robustified posterior as

$$f_{\text{robust}}(u \mid y, \pi) = f\left(u \mid u_{[1]} = \frac{1}{p+1}, \ldots, u_{[p]} = \frac{p}{p+1}, y, \pi\right). \tag{8}$$

In the robust posterior (8), the sample space of u, to be denoted by Γ, consists of the $p!$ permutations of $\{\frac{1}{p+1}, \ldots, \frac{p}{p+1}\}$. This gives the following explicit form of (8)

$$f_{\text{robust}}(u \mid y, \pi) = c(y) f(u \mid y, \pi) \tag{9}$$

on $u \in \Gamma$, where $c^{-1}(y) = \sum_{u \in \Gamma} f(u \mid y, \pi)$. Note that this approach of robustifying the standard posterior through truncation to the discrete space Γ can be applied to the posterior for any general Bayesian model including hierarchical Bayes and empirical Bayes. Using relationship $\theta = g(u)$ as defined at the beginning of this section, we can easily map $f_{\text{robust}}(u \mid y, \pi)$ back to θ as $f_{\text{robust}}(\theta \mid y, \pi)$. In particular, $\theta = g(u) \sim f_{\text{robust}}(\theta \mid y, \pi)$ if $u \sim f_{\text{robust}}(u \mid y, \pi)$.

Based on the discussion above, the robustified posterior (9) is effective over the standard posterior when the misspecified π causes $f(\tilde{u} \mid y, \pi)$ to deviate from $f(\tilde{u} \mid y, \pi_0)$. This can happen when the working prior over-shrinks due to underspecification of the working prior π or shrinks in the wrong direction due to a location shift. On the other hand, it does not improve inference if the misspecified π primarily affects $f(u \mid \tilde{u}, y, \pi)$ in (7), which happens when the working prior is too diffuse and therefore under-shrinks.

For hierarchical model (2), let

$$\pi_h(\theta) = \int f(\theta \mid \eta) f(\eta)\, d\eta \tag{10}$$

be the prior of θ after integrating out η. We can use Theorem 1 to improve the inference of hierarchical model (2) by replacing $f(\theta \mid y, \pi_h)$, as defined in (3), by $f_{\text{robust}}(\theta \mid y, \pi_h)$, which is, however, computationally prohibitive because the mixture density π_h given in (10) is very expensive to evaluate. To get around this computational limitations, note that the standard posterior $f(\theta \mid y, \pi_h)$ can be simulated as the stationary distribution of Gibbs sampler

$$\begin{cases} \theta^{[j]} \sim f(\theta \mid \eta^{[j-1]}, y), \\ \eta^{[j]} \sim f(\eta \mid \theta^{[j]}, y). \end{cases}$$

Now a robustified version of the component $f(\theta \mid \eta, y)$, $f_{\text{robust}}(\theta \mid \eta, y)$, can be constructed as $f_{\text{robust}}(\theta \mid y, \pi)$, where $\pi(\theta) = f(\theta \mid \eta)$, which is much less computationally intensive because $f(\theta \mid \eta)$ is usually a simple parametric distribution. We use it to construct a robustified Gibbs sampler

$$\begin{cases} \theta^{[j]} \sim f_{\text{robust}}(\theta \mid \eta^{[j-1]}, y), \\ \eta^{[j]} \sim f(\eta \mid \theta^{[j]}, y), \end{cases} \tag{11}$$

which can be simulated from rapidly. We shall show in Section 3 that this new Gibbs sampler has a stationary distribution. Our proposed robustified posterior is defined as the stationary distribution of θ for this robustified Gibbs sampler, which can replace the standard posterior $f(\theta \mid y, \pi_h)$ for improved Bayesian inference.

3. A Markov Chain Monte Carlo Algorithm

We now describe a Metropolis algorithm (e.g., Robert and Casella 2009, chap. 6) to sample from $f_{\text{robust}}(u \mid y, \pi)$ in (9), whose sample space Γ consists of $p!$ permutations of $(\frac{1}{p+1}, \ldots, \frac{p}{p+1})$. Due to the enormous number of points in Γ, an effective algorithm must start the chain from a point $u \in \Gamma$ well supported by data y. To accomplish this, note that $q_i = \Phi_i(y_i \mid \theta_i = 0)$ is the p-value of testing $H_0 : \theta = 0$ versus $H_1 : \theta_i < 0$ and $1 - q_i$ is the p-value for testing $H_0 : \theta_i = 0$ versus $H_1 : \theta_i > 0$. In other words, a smaller q_i is stronger evidence for $u_i < 0$ and a larger q_i is stronger evidence for $u_i > 0$. Therefore, we propose to reorder $y_1, \ldots, y_p$ by the value of q_i. To simplify notation, the reordered sequence will still be denoted as $y_1, \ldots, y_p$ but now satisfies

$$q_1 \le q_2 \le \cdots \le q_p.$$

We can then start the Metropolis algorithm from the initial point $u = (\frac{1}{p+1}, \ldots, \frac{p}{p+1})$, which is well supported by the reordered data $y = (y_1, \ldots, y_n)$.

The following Metropolis algorithm is easy to implement and also efficient based on our experience. Let $u = (u_1, \ldots, u_p) \in \Gamma$ be the current value of the Markov chain. For $2 \le i \le p$, swap the elements u_{i-1} and u_i in u and call the resulting vector u^c as the candidate for the next value of the chain. Then accept u^c as the next value with probability

$$\min\left(1, \frac{f(u^c \mid y, \pi)}{f(u \mid y, \pi)}\right) = \min\left(1, \prod_{j=i-1}^{i} \frac{f(u_j^c \mid y_j, \pi)}{f(u_j \mid y_j, \pi)}\right).$$

The second equality follows from the fact that the standard posterior $f(u \mid y, \pi)$ has the form of $\prod_{i=1}^{p} f(u_i \mid y_i, \pi)$ and u^c and u are the same except for the two swapped elements. Perform the above update in the order of $i = 2, \ldots, p$ and keep repeating this operation. The Markov chain then converges to the desired robust posterior $f_{\text{robust}}(u \mid y, \pi)$. We implemented this algorithm using Rcpp package for R (Eddelbuettel et al. 2011) in about 30 lines of code.

We can now easily implement robustified Gibbs sampler (11). Drawing from $f(\eta \mid \theta^{[j]}, y)$ is usually straightforward. The Metropolis–Hasting algorithm in the previous paragraphs can be used to sample from $f_{\text{robust}}(u \mid \eta^{[j-1]}, y)$ and therefore $f_{\text{robust}}(\theta \mid \eta^{[j-1]}, y)$. Note that u has a finite sample space of Γ and the chain is irreducible so long as $f(u \mid \eta, y) > 0$ for all $u \in \Gamma$ under any given η. It follows that this robustified Gibbs sampler (11) has a unique stationary distribution, whose density can be written down explicitly from the two conditional distributions in the Gibbs sampler by the Hammersley-Clifford theorem (Besag 1974).

4. Simulation Study

Our simulation study compares the performance of 11 Bayesian parametric and nonparametric inference methods. Our study is conducted as follows.

Step 1: For $p = 1000$ and $p = 2000$, generate $y_i = \theta_i + \varepsilon_i$ (for $i = 1, \ldots, p$), where $\theta_i \overset{\text{iid}}{\sim} \pi_0$ and $\varepsilon_i \overset{\text{iid}}{\sim} N(0, 1)$.

Step 2: Reorder data $y_1, \ldots, y_p$ by the values of $q_1, \ldots, q_p$ as described in Section 3. Also arrange the generated $\theta_1, \ldots, \theta_p$ in the corresponding order.

Step 3: For dataset $y_1, \ldots, y_p$ from Step 2, generate 5000 draws from the posterior distribution of θ_i, $i = 1, \ldots, p$, for each of the 11 Bayesian methods described below. Compute the mean of these 5000 draws as estimator $\hat{\theta}_i$ and compute $\hat{\theta}_i - \theta_i$ as estimation errors, $i = 1, \ldots, p$. In addition, the standard deviation of the 5000 posterior draws is computed as an estimator of the standard deviation of the posterior distribution of θ_i.

Three forms of π_0 are used in this simulation study. The first form, π_0^N, is normal distribution $N(0, 2^2)$, which serves as an example of a light-tailed distribution. The second form, π_0^t, is the scaled t-distribution with five degrees of freedom and a standard deviation of two, which represents a heavy-tailed distribution. The third form, π_0^h, is given by

$$\pi_0^h = 0.9\pi_0^{N,\text{trunc}} + 0.1\pi_0^{t,\text{trunc}},$$

where $\pi_0^{N,\text{trunc}}$ is π_0^N truncated to interval $[-3.290, 3.290]$ and $\pi_0^{t,\text{trunc}}$ is π_0^t truncated to $(-\infty, -3.290) \cup (3.290, \infty)$. This hybrid distribution has the form of π_0^N in the middle and the form of π_0^t in the tails.

We now describe the 11 estimation methods compared in this simulation.

Method 1a (Laplace). Standard posterior for hierarchical model (2) with Laplace working prior: $\varepsilon_i \overset{\text{iid}}{\sim} N(0, 1)$, $\theta_i \mid \eta_1 \sim \text{Laplace}(0, \eta_1)$, and $\log(\eta_1) \sim \text{Unif}(-0.69, 1.61)$, where η_1 is the scale parameter of the Laplace distribution, $-0.69 \approx \log(0.5)$ and $1.61 \approx \log(5)$.

Method 1b (R Laplace). Robustified posterior of Method 1a.

Method 2a (Normal). Standard posterior for hierarchical model (2) with normal working prior: $\theta_i \mid \eta_2 \sim \mathcal{N}(0, \eta_2^2)$ and $\log(\eta_2) \sim \text{Unif}(-0.69, 1.61)$.

Method 2b (R Normal). Robustified posterior of Method 2a.

Method 3a (Mixture). Standard posterior for hierarchical model (2) with a mixture working prior:

$$\theta_1, \ldots, \theta_p \mid \lambda, \eta_1, \eta_2 \sim \lambda \prod_{i=1}^{p} \text{Laplace}(\theta_i \mid 0, \eta_1) + (1 - \lambda) \times \prod_{i=1}^{p} \mathcal{N}(\theta_i \mid 0, \eta_2^2),$$

where $\lambda \sim \text{Bernoulli}(1/2)$, $\log(\eta_2) \sim \text{Unif}(-0.69, 1.61)$ and $\eta_1 = \eta_2/\sqrt{2}$. With this choice of η_1, the two mixture components Laplace$(0, \eta_1)$ and N$(0, \eta_2)$ have equal variance.

Method 3b (R Mixture). Robustified posterior of Method 3a.

The following 4 estimation methods are nonparametric Bayes that uses a Dirichlet process (DP) mixture prior with a normal base and a gamma distribution prior on the concentration parameter α. Specifically, we consider the model:

$$\begin{aligned}
&\theta_i \mid G, \sigma \sim \int \mathcal{N}(\mu, \sigma^2)\, dG(\mu)\\
&\sigma^2 \sim \text{Inv-Gamma}(1, 5)\\
&G \mid G_0, \alpha \sim \text{DP}(\alpha, G_0), \text{where}\\
&\quad G_0 = N(0, \sigma_b^2)\\
&\quad \sigma_b^2 \sim \text{Inv-Gamma}(1, 2)\\
&\quad \alpha \sim \text{Gamma}(\texttt{shape}, \texttt{scale}).
\end{aligned}$$

The prior Inv-Gamma$(1, 5)$ for σ^2 have 5% and 95% quantiles of 0.054 and 7.90, respectively. The prior Inv-Gamma$(1, 2)$ for σ_b^2 are chosen to be consistent with the three true underlying models π_0^N, π_0^t and π_0^h above.

We will refer to the above Dirichlet process model as DP(`shape`, `scale`), where `shape` and `scale` are the two parameters in the Gamma prior on concentration parameter α. It is well known that the prior placed on α can substantially affect the G drawn from the Dirichlet process, see the recent article by Canale and Prünster (2017) and Shi et al. (2019). Therefore, we conduct a sensitivity analysis and provide results under four sets of `shape` and `scale` parameters as considered in Escobar and West (1995) and Shi et al. (2019).

Method 4a (DP1). DP$(1, 1)$; that is, $\alpha \sim \text{Gamma}(1, 1)$ in the nonparametric Bayes with Dirichlet process mixture prior. The 5% and 95% quantiles of this Gamma distribution are 0.025 and 3.69, respectively.

Method 4b (DP2). DP$(2, 4)$. The 5% and 95% quantiles of Gamma$(2, 4)$ are 0.97 and 22.29, respectively.

Method 4c (DP3). DP$(0.1, 10)$. The 5% and 95% quantiles of Gamma$(0.1, 10)$ are 0 and 9.78, respectively.

Method 4d (DP4). DP$(10, 0.1)$. The 5% and 95% quantiles of Gamma$(10, 0.1)$ are 0.48 and 1.71, respectively.

Our last estimation method is:

Method 5 (Flat). Standard posterior under flat prior $\theta_i \sim N(0, 1000^2)$, which leads to little shrinkage. The mean of posterior distribution $\theta \mid y_i$ is practically y_i itself, which also closely matches the maximum likelihood estimator $\hat{\theta}_i = y_i$.

For $p = 1000$, Figure 1 presents the distribution of $\hat{\theta}_1 - \theta_1$ and $\theta_p - \hat{\theta}_p$ over 100 replications under each of π_0^N, π_0^t, π_0^h, and for each of the 11 estimation methods using boxplots. Figure 2 presents the same results for $p = 2000$. Note that we can combine $\hat{\theta}_i - \theta_i$ and $\theta_{p+1-i} - \hat{\theta}_{p+1-i}$ into one boxplot to save space because they have the same distribution due to the symmetries of π_0 and the working priors. Concentration of the distribution around 0 in a boxplot represents a good estimation method while concentration below and above 0 represent under-shrinkage and over-shrinkage, respectively. Note also that $\theta_1, \ldots, \theta_p$ have been reordered in Step 2 above by the $q_1, \ldots, q_p$ values. Therefore, θ_1 and θ_{1000} are the two most extremal effects usually of the greatest practical interest.

We now summarize the performance of the 11 estimation methods as presented in Figures 1 and 2. Method 1a (Laplace) under-shrinks considerably under π_0^N as expected because the working Laplace prior has much heavier tails. It works well under heavier-tailed π_0^t and π_0^h. Method 1b (R Laplace) offers no improvement. Method 2a (Normal) works very well under π_0^N but severely over-shrinks under π_0^t and π_0^h. Method 2b (R Normal) substantially improves over Method 2a under π_0^t and π_0^h but still performs poorly due to the restrictive short tails in the working prior. Method 3a (Mixture), with its flexible mixture working prior, works well under both π_0^N and π_0^t but not so well under π_0^h. Method 3b (R Mixture) offers considerable improvement over Method 3a under π_0^h for $p = 2000$. All the four DP methods perform well under π_0^N but less so under π_0^t and π_0^h particularly for DP$(0.1, 10)$ and DP$(10, 0.1)$. This could

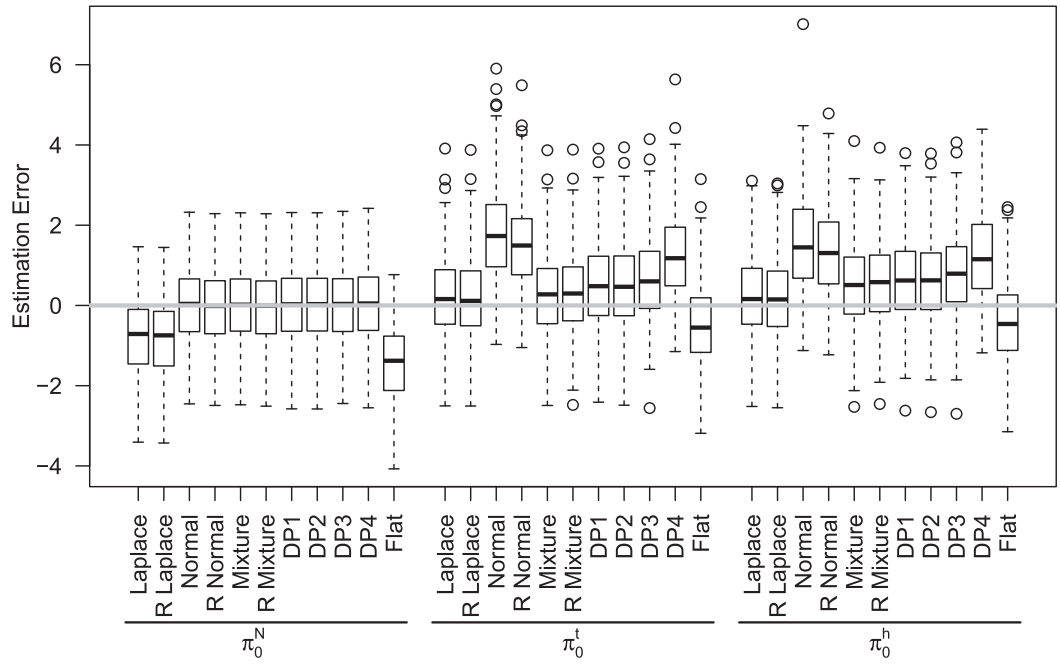

Figure 1. $p = 1000$. Boxplots of the estimation errors $\hat{\theta}_1 - \theta_1$ and $\theta_{1000} - \hat{\theta}_{1000}$ (combined in one graph for having the same distribution), the two most extreme random effects, over 100 replications.

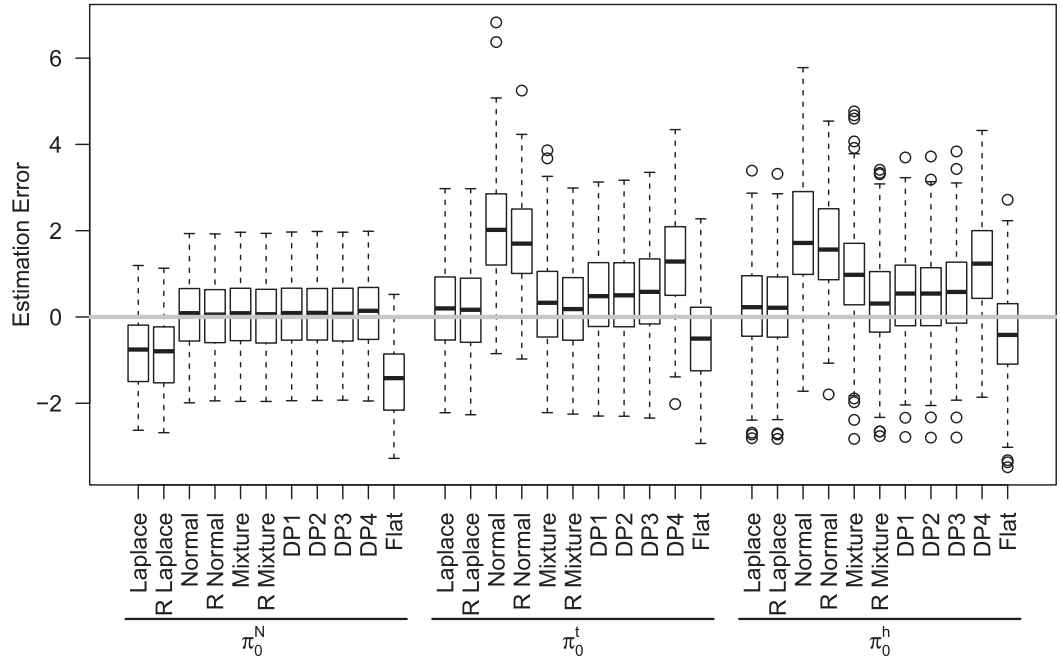

Figure 2. $p = 2000$. Boxplots of the estimation errors $\hat{\theta}_1 - \theta_1$ and $\theta_{2000} - \hat{\theta}_{2000}$ (combined in one graph for having the same distribution), the two most extreme random effects, over 100 replications.

be due to the normal base G_0. Finally, Method 5 (flat) performs poorly due to a lack of any shrinkage.

There are a few cases in Figures 1 and 2 in which the robustified posterior provides little improvement over the standard posterior. This can happen when the working prior is already optimal or close to optimal, such as Method 1a under π_0^t and Method 2a under π_0^N. It can also happen when the working prior has longer tails than π_0 as noted in Section 2, such as Method 1a under π_0^N.

Tables 1 and 2 give, for $p = 1000$ and $p = 2000$, respectively, the mean square error of each estimation method as the average of $(\hat{\theta}_i - \theta_i)^2$ and $(\theta_{p+1-i} - \hat{\theta}_{p+1-i})^2$ on over 100 replications for $i = 1, 2, 3$. The performance ranking of the 11 methods summarized above for $i = 1$ in Figures 1 and 2 still holds for $i = 2, 3$ but the difference between different methods is smaller. Method 3b (R Mixture) is the clear winner in terms of overall performance.

We have therefore shown that the mean of the robustified posterior is a better estimator of the underlying θ_i than the

Table 1. $p = 1000$, mean square error of $\hat{\theta}_i$ and $\hat{\theta}_{p+1-i}$ (combined for having the same underlying value) for the 11 methods under $\pi_0^N, \pi_0^t, \pi_0^h$.

i	π_0	Lapl	R Lapl	Norm	R Norm	Mixt	R Mixt	DP1	DP2	DP3	DP4	Flat
1	π_0^N	1.36	1.43	0.77	0.77	0.77	0.77	0.76	0.76	0.77	0.77	2.83
	π_0^t	1.08	1.07	4.55	3.36	1.16	1.31	1.38	1.38	1.61	2.79	1.27
	π_0^h	1.04	1.03	3.87	3.02	1.42	1.46	1.51	1.52	1.81	2.74	1.17
2	π_0^N	1.28	1.34	0.72	0.73	0.71	0.72	0.72	0.72	0.72	0.72	2.71
	π_0^t	1.44	1.44	2.66	2.38	1.45	1.49	1.52	1.53	1.59	1.74	1.92
	π_0^h	1.09	1.09	2.13	1.93	1.23	1.27	1.22	1.23	1.40	1.44	1.47
3	π_0^N	1.23	1.28	0.81	0.82	0.81	0.82	0.81	0.81	0.81	0.80	2.51
	π_0^t	1.19	1.18	2.01	1.86	1.17	1.23	1.21	1.22	1.30	1.27	1.67
	π_0^h	1.08	1.07	1.69	1.60	1.22	1.24	1.19	1.20	1.33	1.29	1.51

Table 2. $p = 2000$, mean square error of $\hat{\theta}_i$ and $\hat{\theta}_{p+1-i}$ (combined for having the same underlying value) for the 11 methods under $\pi_0^N, \pi_0^t, \pi_0^h$.

	π_0	Lapl	R Lapl	Norm	R Norm	Mixt	R Mixt	DP1	DP2	DP3	DP4	Flat
1	π_0^N	1.43	1.49	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.76	2.99
	π_0^t	1.12	1.12	5.81	4.15	1.40	1.14	1.47	1.48	1.61	3.07	1.30
	π_0^h	1.22	1.21	5.38	4.06	2.82	1.42	1.46	1.45	1.57	3.00	1.33
2	π_0^N	1.57	1.63	0.88	0.89	0.88	0.89	0.88	0.88	0.88	0.88	3.12
	π_0^t	1.02	1.02	3.02	2.55	1.13	1.02	1.17	1.19	1.20	1.35	1.46
	π_0^h	1.11	1.11	2.67	2.42	1.74	1.19	1.20	1.20	1.27	1.42	1.45
3	π_0^N	1.35	1.40	0.76	0.76	0.75	0.76	0.75	0.76	0.76	0.76	2.83
	π_0^t	1.13	1.13	2.58	2.32	1.23	1.13	1.15	1.16	1.17	1.21	1.61
	π_0^h	1.04	1.04	2.14	2.01	1.49	1.13	1.10	1.12	1.17	1.17	1.39

mean of the standard posterior. It is natural to wonder about the impact of the robustification on the standard deviation of the posterior distribution. For $i = 1$, $p = 2000$, and under π_0^N, the average of the standard deviation of posterior draws (computed at Step 3 above) over 100 replicates is 0.9995 for Laplace, 0.9868 for R Laplace, 0.8947 for Normal, 0.8530 for R Normal, 0.8952 for Mixture and 0.8548 for R Mixture. Their respective values are 0.9998, 0.9897, 0.8980, 0.7065, 1.0004, and 0.9848 under π_0^t and 0.9996, 0.9868, 0.9117, 0.7524, 0.9644, and 0.9668 under π_0^h. Similar values are obtained for $i = 2, 3$ and also for $i = 1, 2, 3$ and $p = 1000$. Therefore, the robustified posterior distribution has similar or reduced variation compared to the standard posterior distribution.

The standard posterior estimates in Methods 1a, 2a, and 3a are computed using RStan (Carpenter et al. 2016). The robustified posterior estimates for Methods 1b, 2b, and 3b are obtained by our own R code. Function `DPMmeta` in the DPpackage (Jara et al. 2011) is used for the nonparametric Bayesian estimation in Methods 4a, 4b, 4c, and 4d. The reproducibility of the reported result has been confirmed by some independent further replications. Our complete R code for this simulation study is available at *http://sites.google.com/site/jiangangliao*.

5. Discussion

This article proposes a robustified posterior for improving inference on a large number of parallel effects. By providing significant protection against misspecified priors, our method encourages the use and specification of genuinely informative priors instead of defaulting to a weak and ineffective prior. For example, Method 3a (R Mixture) in Section 4 can be an excellent choice if we believe that the tails of π_0 are between a short-tailed normal and a long-tailed t-distribution. Other approaches to enhance the robustness of Bayesian inference have been proposed in different contexts and models. For example, Lazar (2003) replaces the likelihood function in the Bayesian posterior by an empirical likelihood, which achieves improved robustness by reduced specification in the likelihood. Also, Hoff (2007) proposed to replace the likelihood of the complete data by the likelihood of the rank of the data to remove nuisance parameters in a semiparametric copula estimation. The robustified posterior in this article is specifically developed for estimating a large number of parallel effects. By using asymptotic behavior of order statistics and the unique structure of parallel effects, our method has the distinctive advantage of improving robustness with little or no loss of inferential efficiency even when the working prior is correctly specified.

Finally, we have previously proposed a rank-based robustified posterior in which the posterior of θ_i is computed conditioned on the rank of y_i among $y_1, \ldots, y_p$ instead of the value of y_i itself (Liao, McMurry, and Berg 2014). The rank-based posterior has similar properties as the robustified posterior in this article but works well only when error ε_i have similar variation across $i = 1, \ldots, p$. In contrast, the robustified posterior in this article only requires the error distribution in (1) to be continuous.

Appendix: Proof of Theorem 1

Proof. Formally, first consider the marginal distribution of order statistics $\tilde{u}$:

$$f(\tilde{u} \mid \pi_0) = \int f(\tilde{u} \mid y, \pi_0) f(y \mid \pi_0) dy,$$

where $f(y \mid \pi_0) = \int f(y \mid \theta) \pi_0(\theta) d\theta$ is the marginal distribution of y. It follows from (4) that $f(\tilde{u} \mid \pi_0)$ is the joint distribution of the order statistics from uniform $[0, 1]$ (see, e.g., Shao 1999, p. 72). Let $\tilde{u}$ be a draw from $f(\tilde{u} \mid \pi_0)$. The Glivenko–Cantelli theorem and the Berry–Esseen theorem state that, as $p \to \infty$, the empirical distribution of $\tilde{u}$ converges to the function $F(x) = x$ uniformly on $x \in [0, 1)$. Recent refinements to these theorems (Fresen 2011, Lemma 2) are able to characterize the behavior of the order statistics $u_{[i]}$ directly:

$$\sup_{1 \le i \le p} \left| u_{[i]} - \frac{i}{p+1} \right| \to 0 \tag{A.1}$$

in probability.

Now we show the asymptotics in (A.1), derived under the marginal distribution $f(\tilde{u} \mid \pi_0)$, can be extended to the conditional distribution $f(\tilde{u} \mid y, \pi_0)$. It follows from Equation (A.1) that, for any given $\delta_1 > 0$ and as $p \to \infty$, we have

$$\text{pr}\left(\sup \left| u_{[i]} - \frac{i}{p+1} \right| > \delta_1 \right) \to 0.$$

Now for any y, define

$$D(y) \equiv \text{pr}\left(\sup \left| u_{[i]} - \frac{i}{p+1} \right| > \delta_1 \mid y \right),$$

where the right side is the conditional probability given y. It follows that

$$\text{pr}\left(\sup \left| u_{[i]} - \frac{i}{p+1} \right| > \delta_1 \right) = E_y\left(D(y) \right),$$

where the expectation on the right is with respect to $y \sim f(y \mid \pi_0)$. Since $D(y) \geq 0$ for every y and $E_y(D(y)) \to 0$, we have, for any $\delta_2 > 0$,

$$\text{pr}\left(D(y) > \delta_2\right) \to 0$$

as $p \to \infty$, where the probability is evaluated with respect to $y \sim f(y \mid \pi_0)$. In other words, except on a small set of y whose probability goes to 0, we have

$$\text{pr}\left(\sup \left| u_{[i]} - \frac{i}{p+1} \right| > \delta_1 \mid y \right) \leq \delta_2,$$

for every y when p is sufficiently large. □

Acknowledgments

The authors are deeply grateful to the two anonymous reviewers, an associate editor, and editor Dr. Jeske for their expert guidance in the revision of the article.

References

Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems," *Journal of the Royal Statistical Society*, Series B, 36, 192–236. [3]

Bogdan, M., Ghosh, J. K., and Tokdar, S. T. (2008), "A Comparison of the Benjamini-Hochberg Procedure With Some Bayesian Rules for Multiple Testing," in *Beyond Parametrics in Interdisciplinary Research: Festschrift in Honor of Professor Pranab K. Sen*, eds. N. Balakrishnan, Edsel A. Peña and Mervyn J. Silvapulle, Beachwood, OH: Institute of Mathematical Statistics, pp. 211–230. [2]

Canale, A., and Prünster, I. (2017), "Robustifying Bayesian Nonparametric Mixtures for Count Data," *Biometrics*, 73, 174–184. [4]

Carlin, B. P., and Murray, T. A. (2013), Comment on "Bayesian Nonparametric Inference–Why and How," by Müller and Mitra, *Bayesian Analysis*, 8, 303–310. [2]

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2016), "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*, 20, 1–37. [6]

Do, K., Mueller, P., and Tang, F. (2005), "A Nonparametric Bayesian Mixture Model for Gene Expression," *Journal of the Royal Statistical Society*, Series C, 54, 1–18. [2]

Eddelbuettel, D., François, R., Allaire, J., Ushey, K., Kou, Q., Russel, N., Chambers, J., and Bates, D. (2011), "Rcpp: Seamless R and C++ Integration," *Journal of Statistical Software*, 40, 1–18. [3]

Efron, B. (2010), *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction* (Vol. 1), Cambridge, UK: Cambridge University Press. [1]

Escobar, M. D., and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588. [4]

Fresen, D. (2011), "Simultaneous Concentration of Order Statistics," Technical Report, arXiv no. 1102.1128. [6]

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2014), *Bayesian Data Analysis* (Vol. 3), New York: Taylor & Francis. [1]

Hoff, P. D. (2007), "Extending the Rank Likelihood for Semiparametric Copula Estimation," *The Annals of Applied Statistics*, 1, 265–283. [6]

——— (2013), Comment on "Bayesian Nonparametric Inference–Why and How" by Müller and Mitra, *Bayesian Analysis*, 8, 311–318. [2]

Jara, A., Hanson, T. E., Quintana, F. A., Müller, P., and Rosner, G. L. (2011), "DPpackage: Bayesian Semi- and Nonparametric Modeling in R," *Journal of Statistical Software*, 40, 1–30. [6]

Johnstone, I. M., and Silverman, B. W. (2004), "Needles and Straw in Haystacks: Empirical Bayes Estimates of Possibly Sparse Sequences," *The Annals of Statistics*, 32, 1594–1649. [1]

Kim, S., Dahl, D. B., and Vannucci, M. (2009), "Spiked Dirichlet Process Prior for Bayesian Multiple Hypothesis Testing in Random Effects Models," *Bayesian Analysis*, 4, 707–732. [2]

Lazar, N. A. (2003), "Bayesian Empirical Likelihood," *Biometrika*, 90, 319–326. [6]

Liao, J., McMurry, T., and Berg, A. (2014), "Prior Robust Empirical Bayes Inference for Large-Scale Data by Conditioning on Rank With Application to Microarray Data," *Biostatistics*, 15, 60–73. [6]

Martin, R., and Tokdar, S. T. (2012), "A Nonparametric Empirical Bayes Framework for Large-Scale Multiple Testing," *Biostatistics*, 13, 427–439. [2]

Müller, P., and Mitra, R. (2013), "Bayesian Nonparametric Inference—Why and How" (with discussion), *Bayesian Analysis*, 8, 269–302. [2]

Muralidharan, O. (2010), "An Empirical Bayes Mixture Method for Effect Size and False Discovery Rate Estimation," *The Annals of Applied Statistics*, 4, 422–438. [2]

O'Hagan, A. (2013), "Comment on Article by Müller and Mitra," *Bayesian Analysis*, 8, 319–322. [2]

Robert, C., and Casella, G. (2009), *Introducing Monte Carlo Methods With R*, New York: Springer Science & Business Media. [3]

Shao, J. (1999), *Mathematical Statistics*, New York: Springer. [6]

Shi, Y., Martens, M., Banerjee, A., and Laud, P. (2019), "Low Information Omnibus (LIO) Priors for Dirichlet Process Mixture Models," *Bayesian Analysis*, 14, 677–702. [4]