Robust Bayesian Regression with Synthetic Posterior

Shintaro Hashimoto\textsuperscript{1} and Shonosuke Sugasawa\textsuperscript{2}

\textsuperscript{1}Department of Mathematics, Graduate School of Science, Hiroshima University
\textsuperscript{2}Center for Spatial Information Science, The University of Tokyo

Abstract

Regression models are fundamental tools in statistics, but they typically suffer from outliers. While several robust methods have been proposed based on frequentist approaches, Bayesian methods would be more preferable in terms of easiness of uncertainty quantification of estimation results. In this article, we propose a robust Bayesian method for regression models by introducing a synthetic posterior distribution based on robust divergence. We also consider introducing shrinkage priors for regression coefficients for variable selection. We provide an efficient posterior computation algorithm using Bayesian bootstrap within Gibbs sampling. We carry out simulation studies to evaluate performance of the proposed method, and apply the proposed method to real datasets.

Key words: Bayesian bootstrap; Bayesian Lasso; Divergence; Gibbs sampling; Linear regression
1 Introduction

The maximum likelihood or least squared methods are basic strategies of the inference for parametric statistical models. However, these methods are strongly affected by outliers, and several alternative robust methods have been developed, e.g. M-estimation, least absolute deviation (LAD) regression and so on. In addition to such problem, we often face with data which have large number of covariates. In this case, the shrinkage estimation of coefficients in regression models such as Lasso (Tibshirani, 1996) is known to be useful, but these methods also suffer from the problem of outliers. To overcome these problems, we may use LAD regression or $t$-distribution as an error distribution together with shrinkage methods, while it is known that such methods dose not work well when the contamination ratio is not small or outliers concentrate in one side of tail region. On the other hand, robust estimation methods based on divergence have been attracted attention because of the automatic rejection of outliers. The pioneering work by Basu et al. (1998) proposed the density power divergence, and they studied robustness in the sense of the influence function. As a related work, Jones et al. (2001) and Fujisawa and Eguchi (2008) considered a slight modified version of the density power divergence, which is now called $\gamma$-divergence, and $\gamma$-divergence is shown to be robust against heavy contamination. Furthermore, Kawashima and Fujisawa (2017) employed $\gamma$-divergence to propose robust and sparse linear regression methods.

In Bayesian framework, we can also construct sparse linear regression models by using shrinkage prior distributions which induce sparsity in regression coefficients. They are roughly divided into two types of priors. The first one is discrete mixture priors known as spike-and-slab priors (e.g. Ishwaran and Rao, 2005; Rockova and George, 2018). Although these priors lead some of coefficients to exactly zero, the complexity of posterior computation is explosively growing with the number of covariates. The second one is use of continuous shrinkage priors such as Laplace (Park and Casella, 2008) and horseshoe priors (Carvalho et al., 2010). The Bayesian linear regression based on the Laplace prior is known as Bayesian lasso. While estimators
based on such priors can not be exactly zero, there exists efficient Markov chain Monte Carlo (MCMC) algorithms for posterior computation. In this article, we adopt the later type of priors for reasons of practical use. The Bayesian approach for sparse regression models is very attractive because it can take uncertainly of estimation into account, i.e., we can easily compute credible intervals whereas frequentist approach is not necessary easy. Also, we can include uncertainty of a tuning parameter in estimation results by averaging over posterior distribution. However, these Bayesian linear regression models also suffer from outliers for the same reasons as frequentist approaches. For a remedy, we may use the Bayesian lasso with heavy-tailed error distributions such as $t$-distribution for robustness, but this strategy is not necessarily a good solution since it is not easy to extend this approach to other regression models such as generalized linear models. Moreover, the use of heavy-tailed distribution cannot necessarily work well under variety of contamination structures.

In this paper, we propose the use of synthetic posterior based on robust divergence combined with shrinkage priors for regression coefficients. Recently, robust Bayesian methods via synthetic posterior have been proposed (e.g. Bissiri et al. 2016; Bhat
tacharya et al. 2019; Miller and Dunson 2019; Nakagawa and Hashimoto 2019), but such methodologies are demonstrated in low-dimensional parametric models to show their good robustness properties through numerical studies. On the other hand, this paper is concerned with regression models with relatively large number of covariates. As discussed later, the main problem using the proposed synthetic posterior is its complicated functional form of model parameters, so that the main bottleneck is difficulty of efficient posterior computation. To solve the problem, we provide an efficient algorithm which generates random samples from the complicated full conditional distribution via optimization of the randomized objective function within Gibbs sampler. This idea is related to Bayesian bootstrap (Rubin 1981), weighted likelihood bootstrap (Newton and Raftery 1994) and nonparametric Bayesian updating (Lyddon et al. 2018). In the optimization step, we use the modified version of Majorization-Minimization algorithm given in Kawashima and Fujisawa (2017). The advantage of the proposed algorithm is that there is no rejection steps, and
performance in terms of mixing and autocorrelation is quite reasonable.

The paper is organized as follows: In Section 2, we introduce the setting of our model and formulate the synthetic posterior distribution. Next, we consider the Bayesian linear regression model under shrinkage priors which are constructed by the scale mixtures of normal. Then a new posterior computation algorithm for synthetic posterior with shrinkage priors is proposed. In Section 3, we provide results of numerical studies and real data analysis. Under several types of contaminations, it is shown that the proposed method outperforms the original Bayesian lasso and Bayesian lasso with the $t$-prior distribution in the sense of the mean squared error, average length of credible intervals and coverage probability. Also, we show mixing properties of the proposed algorithm compared with other methods.

2 Robust Bayesian regression via synthetic posterior

2.1 Settings and synthetic posterior

Suppose we have independent observation $(y_i, x_i)$ for $i = 1, \ldots, n$, where $y_i$ is a continuous response and $x_i = (x_{i1}, \ldots, x_{ip})^t$ is a $p$-dimensional vector of covariates. We consider fitting a regression model $y_i = x_i^t \beta + \varepsilon_i$ with $\varepsilon \sim N(0, \sigma^2)$. Let $\pi(\beta, \sigma^2)$ be prior distribution for the model parameters $\beta$ and $\sigma^2$. Then the standard posterior distribution of the model parameters are given by

$$\pi(\beta, \sigma^2|D) = \pi(\beta, \sigma^2) \exp \left\{ \sum_{i=1}^{n} \log f(y_i; x_i^t \beta, \sigma^2) \right\}, \quad (1)$$

where $f(y_i; x_i^t \beta, \sigma^2)$ is the density function of $N(x_i^t \beta, \sigma^2)$, and $D$ denotes the set of sampled data. When there exist outliers which have large residuals $(y_i - x_i^t \beta)/\sigma$, the posterior distribution $[1]$ is known to be sensitive to such outliers, and it could produce biased or inefficient posterior inference.

To overcome the problem, we propose replacing the log-likelihood function in $[1]$ with robust alternatives. Specifically, we employ $\gamma$-divergence $[\text{Jones et al.}, 2001]$. 


Fujisawa and Eguchi (2008) of the form:

\[ R_\gamma(\beta, \sigma^2) = \frac{n}{\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( \frac{f(y_i; x_i^T \beta, \sigma^2)}{\|f(\cdot; x_i^T \beta, \sigma^2)\|_{\gamma+1}} \right)^\gamma \right\}, \]

where

\[ \|f(\cdot; x_i^T \beta, \sigma^2)\|_{\gamma+1} = \left( \int f(t; x_i^T \beta, \sigma^2)^{1+\gamma} dt \right)^{1/(1+\gamma)}. \]

Note that \( \gamma \) is a tuning parameter that controls the robustness, and \( R_\gamma(\theta) \) reduces to the log-likelihood function as \( \gamma \to 0 \). Straightforward calculation shows that

\[ R(\beta, \sigma^2) = \frac{n}{\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^{n} f(y_i; x_i^T \beta, \sigma^2)^\gamma \right\} + \frac{n\gamma}{2(1+\gamma)} \log \sigma^2 + C, \] (2)

where \( C \) is an irrelevant constant which does not depend on \( \beta \) and \( \sigma^2 \). Based on (2), we define the following synthetic posterior:

\[ \pi_\gamma(\beta, \sigma^2|D) \propto \pi(\beta, \sigma^2) \exp \left\{ R_\gamma(\beta, \sigma^2) \right\} \]

\[ \propto \pi(\beta, \sigma^2) \exp \left[ \frac{n}{\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^{n} f(y_i; x_i^T \beta, \sigma^2)^\gamma \right\} + \frac{n \gamma}{2(1+\gamma)} \log \sigma^2 \right]. \] (3)

Since \( \pi_\gamma(\beta, \sigma^2|D) \) reduces to the standard posterior (1) under \( \gamma \to 0 \), the synthetic posterior (3) can be regarded as a natural extension of the standard posterior (1).

2.2 Posterior computation

In what follows, we consider the standard prior for \((\beta, \sigma^2)\), that is, normal prior for \( \beta \) and inverse gamma prior for \( \sigma^2 \), independently, given by

\[ \pi(\beta, \sigma^2) \propto \exp \left( -\frac{1}{2} \beta^T S_\beta^{-1} \beta \right) (\sigma^2)^{-a/2-1} \exp \left( -\frac{a}{2\sigma^2} \right), \]

where \( S_\beta \) and \( a \) are hyperparameters. Since the synthetic posterior distribution (3) is not a familiar form, the posterior computation to generate random samples of \((\beta, \sigma^2)\) is not straightforward. We may use crude approaches such as MH algorithm, but the acceptance probabilities might be very small even when the dimension of covariates \( x_i \)
is moderate. To avoid such undesirable situation, we adopt approximated sampling strategy using Bayesian bootstrap [Rubin 1981] or nonparametric Bayesian updating [Lyddon et al. 2018], which generate posterior samples as the minimizer of the following weighted objective function:

\[
L_w(\beta, \sigma^2) = \frac{-n}{\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^{n} w_i f(y_i; x_i^t \beta, \sigma^2) \gamma \right\} + \frac{1}{2} \beta^t S_{\beta}^{-1} \beta \\
+ \left(1 + \frac{a}{2} - \frac{n \gamma}{2(1 + \gamma)} \right) \log \sigma^2 + \frac{a}{\sigma^2},
\]

(4)

where \((w_1, \ldots, w_n) \sim n \cdot \text{Dirichlet}(1, \ldots, 1)\), so that \(\sum_{i=1}^{n} w_i = n\). The minimization of (4) can be efficiently carried by Majorization-Minimization (MM) algorithm [Hunter and Lange 2004] obtained by slight modification of one given in [Kawashima and Fujisawa 2017]. From Jensen’s inequality, with current values \((\beta^{(k)}, \sigma^2_{(k)})\) of the model parameters, the upper bound of the objective function (4) can be obtained as follows:

\[
L_w(\beta, \sigma^2) \\
= -\sum_{i=1}^{n} s_i^{(k)} \log f(y_i; x_i^t \beta, \sigma^2) + \frac{1}{2} \beta^t S_{\beta}^{-1} \beta + \frac{a}{2} - \frac{n \gamma}{2(1 + \gamma)} \log \sigma^2 + \frac{a}{\sigma^2} + C,
\]

(5)

where \(C\) is an irrelevant constant and \(s_i^{(k)}\) is a new weight in the \(k\)th iteration, defined as

\[
s_i^{(k)} = \frac{w_i f(y_i; x_i^t \beta^{(k)}, \sigma^2_{(k)})^\gamma}{\sum_{j=1}^{n} w_j f(y_j; x_j^t \beta^{(k)}, \sigma^2_{(k)})^\gamma},
\]

(6)

noting that \(\sum_{i=1}^{n} s_i^{(k)} = n\). Since the upper bound given in (5) is a familiar function of \(\beta\) and \(\sigma^2\), the minimizer can be easily obtained and the updating process is given
by

$$\beta^{(k+1)} = \left\{ \frac{1}{\sigma^2(k)} \sum_{i=1}^{n} s_i^{(k)} x_i x_i^t + S^{-1}_\beta \right\}^{-1} \frac{1}{\sigma^2(k)} \sum_{i=1}^{n} s_i^{(k)} x_i y_i,$$

$$\sigma^2(k+1) = \left( 2 + n + \frac{n}{1 + \gamma} \right)^{-1} \left\{ a + \sum_{i=1}^{n} s_i^{(k)} (y_i - x_i^t \beta^{(k+1)})^2 \right\}.$$

(7)

Therefore, the MM-algorithm to get the minimizer of (4) repeats calculation of the weight (10) and updating parameter values via (7) until convergence. Since all the steps are obtained in analytical forms, the MM-algorithm is very efficient and easy to implement.

For generating $B$ random samples from synthetic posterior distribution (3), we generate $B$ samples of $w_i$’s and solve minimization problems of (4) for $B$ times via the MM-algorithm. The resulting samples can be approximately regarded as posterior samples of (3), and the approximation error would be negligible when $n$ is large (e.g. Newton and Raftery, 1994; Lyddon et al., 2018; Newton et al., 2018).

2.3 Incorporating shrinkage priors

When the dimension of $x_i$ is moderate or large, it is desirable to select a subset of $x_i$ that are associated with $y_i$, which corresponds to shrinkage estimation of $\beta$. Here we rewrite the regression model to explicitly express an intercept term as $y_i = \alpha + x_i^t \beta + \epsilon_i$.

We consider normal prior for $\alpha$, that is, $\alpha \sim N(0, S_\alpha)$ with known $S_\alpha > 0$. For coefficients $\beta$, we introduce shrinkage priors expressed as scale mixture of normals given by

$$\pi(\beta) = \prod_{k=1}^{p} \int_{0}^{\infty} \phi(\beta_k; 0, u_k) g(u_k; \lambda) du_k,$$

(8)

where $g(\cdot; \lambda)$ is a mixing distribution which may depends on some tuning parameter $\lambda$. Many existing shrinkage priors are included in this class by appropriately choosing the mixing distribution $g(\cdot)$. Among many others, we consider two priors for $u_i$: exponential distribution which results in Laplace prior of $\beta$ known as Bayesian Lasso (Park and Casella, 2008), and half-Cauchy distribution for $\sqrt{u_k}$ which results in horseshoe prior for $\beta$ (Carvalho et al., 2010). The detailed settings of these two
priors are given as follows:

(Laplace) \( u_k | \lambda \sim \text{Exp}(\lambda^2/2), \lambda^2 \sim \text{Ga}(c_1, c_2) \)

(Horseshoe) \( u_k | \xi_k, \lambda \sim \text{IG}(1/2, \lambda/\xi_k), \xi_k \sim \text{IG}(1/2, 1), \lambda \sim \text{Ga}(c_1, c_2) \),

where \( c_1 \) and \( c_2 \) are fixed hyperparameters, \( \text{Ga}(a, b) \) denotes the gamma distribution with shape parameter \( a \) and rate parameter \( b \), and \( \text{IG}(a, b) \) is the inverse gamma distribution with shape parameter \( a \) and scale parameter \( b \). Note that \( \xi_k \) in the horseshoe prior is an additional latent variable to make posterior computation easier, and the density of \( u_k | \lambda \) is proportional to \( u_k^{-1/2}(u_k + \lambda)^{-1} \), so that \( \sqrt{u_k} | \lambda \) follows half-Cauchy distribution.

Using the mixture representation (8), the full conditional distribution of \((\alpha, \beta, \sigma^2)\) given the other parameters including \( u_k \)'s is given by

\[
\exp\left(-\frac{\alpha^2}{2S_\alpha} - \frac{1}{2} \beta^T U^{-1} \beta \right) (\sigma^2)^{-a/2-1} \exp \left(-\frac{a}{2\sigma^2} \right) \\
\times \exp \left[ \frac{n}{\gamma} \log \left\{ \frac{1}{n} \sum_{i=1}^n f(y_i; \alpha + x_i^T \beta, \sigma^2)^\gamma \right\} + \frac{n\gamma}{2(1 + \gamma)} \log \sigma^2 \right],
\]

where \( U = \text{diag}(u_1, \ldots, u_p) \). Similarly to the previous section, we can generate approximate posterior samples of \((\alpha, \beta, \sigma^2)\) by minimizing the weighted objective function obtained by replacing \( f(y_i; \alpha + x_i^T \beta, \sigma^2)^\gamma \) with \( w_i f(y_i; \alpha + x_i^T \beta, \sigma^2)^\gamma \) in the above expression, where \( w_i \) is defined in the same way in the previous section. Then, the objective function can be minimized by a similar MM-algorithm given in the previous section. Under given regression coefficients \( \beta \), the full conditional distribution of latent variables \( u_1, \ldots, u_p \) and hyperparameter \( \lambda \) are the same as the case with the standard linear regression, so that we can use the existing Gibbs sampling methods.

We summarize our MCMC algorithm under the two shrinkage priors in what follows.

**MCMC algorithm under shrinkage prior**

- (Sampling from \( \alpha, \beta \) and \( \sigma^2 \)) Generate \( w_i = n \cdot \text{Dir}(1, \ldots, 1) \), and set initial values \( \alpha_0, \beta_0 \) and \( \sigma_0^2 \). Repeat the following procedures until convergence:
– Compute the following weight:

\[ s^{(k)}_i = \frac{w_i f(y_i; \alpha^{(k)} + x^t_i \beta^{(k)}, \sigma^2_{(k)})}{\sum_{j=1}^n w_j f(y_j; \alpha^{(k)} + x^t_j \beta^{(k)}, \sigma^2_{(k)})}, \quad i = 1, \ldots, n. \]  

(10)

– Update the parameter values:

\[
\begin{align*}
\alpha^{(k+1)} &= \left( \frac{n}{\sigma^2_{(k)}} + \frac{1}{\lambda} \right)^{-1} \frac{1}{\sigma^2_{(k)}} \sum_{i=1}^n s^{(k)}_i (y_i - x^t_i \beta^{(k)}), \\
\beta^{(k+1)} &= \left( \frac{1}{\sigma^2_{(k)}} \sum_{i=1}^n s^{(k)}_i x_i x^t_i + U^{-1} \right)^{-1} \frac{1}{\sigma^2_{(k)}} \sum_{i=1}^n s^{(k)}_i x_i (y_i - \alpha^{(k+1)}), \\
\sigma^2_{(k+1)} &= \left( 2 + a + \frac{n}{1 + \gamma} \right)^{-1} \left\{ a + \sum_{i=1}^n s^{(k)}_i (y_i - \alpha^{(k+1)} - x^t_i \beta^{(k+1)})^2 \right\}.
\end{align*}
\]

We adopt the final values as sampled values from the full conditional distribution.

• (Sampling from \(u_1, \ldots, u_p\) and \(\lambda\))

– (Laplace prior) The full conditional distribution of \(1/u_k\) is inverse-Gaussian with parameters \(\mu = \sqrt{\lambda/\beta^2_k}\) and \(\delta = \lambda\) in the parametrization of the inverse-Gaussian density given by

\[
f(x) = \frac{\delta}{2\pi} x^{-3/2} \exp \left\{ -\frac{\delta(x - \mu)^2}{2\mu^2x} \right\}, \quad x > 0.
\]

The full conditional distribution of \(\lambda^2\) is \(\text{Ga}(c_1 + p, c_2 + \sum_{k=1}^p u_k/2)\).

– (Horseshoe prior) The full conditional distribution of \(u_k, \xi_k\) and \(\lambda\) are \(\text{IG}(1, \lambda/\xi_k + \beta^2_k/2)\), \(\text{IG}(1, 1 + \lambda/u_k)\) and \(\text{Ga}(c_1 + p_2, c_2 + \sum_{k=1}^p u_k^{-1} \xi_k^{-1})\), respectively.

Note that the sampling scheme in the main parameter \(\beta\) does not use the previous sampled values of \(\beta\) and there is no rejection steps, thereby autocorrelation of \(\beta\) generated from the above MCMC algorithm is expected to be very small, which will be demonstrated in our numerical studies in Section 3.
3 Numerical studies

3.1 Simulation study

We here evaluate the performance of the proposed methods through simulation studies. We first compare the point and interval estimation performance of the proposed methods with those of some existing methods. To this end, we consider the following regression model with \( n = 100 \) and \( p = 20 \):

\[
y_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n,
\]

where \( \alpha = 0.5, \beta_1 = \beta_4 = 0.5 \) and \( \beta_7 = \beta_{10} = \beta_{13} = 2 \) and the other \( \beta_k \)'s were set to 0. The covariates \( x_i = (x_{i1}, \ldots, x_{ip}) \) were generated from a multivariate normal distribution \( N_p(0, \Sigma) \) with \( \Sigma = (\rho^{i-j})_{1 \leq i,j \leq p} \) with \( \rho = 0.2 \). For the error term \( \varepsilon_i \), we adopted contaminated structure given by \( \varepsilon_i \sim (1 - \omega_i)N(0, 1) + \omega_i f_c \), where \( f_c \) is a contamination distribution and \( \omega_i \) is contamination probability which might be heterogeneous over samples. We considered two settings \( f_c \), that is, (I) \( f_c \sim N(0, 10^2) \) and (II) \( f_c \sim N(10, 1) \), noting that the contamination distribution has very large variance in scenario (I) while the contamination distribution tends to produce large values in scenario (II). Regarding the contamination probability, we considered the following scenarios:

(Homo) \( \omega_i = \omega \in \{0, 0.05, 0.10, 0.15, 0.20\} \)

(Hetero) \( \omega_i = \delta \times \text{logistic}(-3.3 + x_{i10}), \quad \delta \in \{0, 1, 2, 3, 4\} \).

Note that the contamination probability in the first setting is constant over the samples, which is refereed to homogeneous contamination. On the other hand, in the second setting, the probability depends on covariates and is different over the samples, so that it is refereed to heterogenous contamination.

For the simulated dataset, we applied the proposed robust methods with Laplace and horseshoe priors, denoted by RBL and RHS, respectively, as well as the standard
Moreover, we also applied the regression model with error term following $t$-distribution with 3 degrees of freedom, denoted by $tBL$, as a possible competitor of robust Bayesian methods. The tuning parameter $\gamma$ in the proposed method is set to $\gamma = 0.2$. In applying all the methods, we generated 2000 posterior samples after discarding the first 500 samples as burn-in. For point estimates of $\beta_k$’s, we computed posterior median of each element of $\beta_k$’s, and their performance is evaluated via mean squared error (MSE) defined as $p^{-1} \sum_{k=1}^p (\hat{\beta}_k - \beta_k)^2$. We also computed $95\%$ credible intervals of $\beta_k$’s, and calculated average lengths (AL) and coverage probability (CP) defined as $p^{-1} \sum_{k=1}^p |CI_k|$ and $p^{-1} \sum_{k=1}^p I(\beta_k \in CI_k)$, respectively. These values were averaged over 300 replications of simulating datasets.

In Figures 1, we presented the results of logarithm of MSE (log-MSE) with error bars corresponding three times estimated Monte Carlo errors. When there is no outliers, the standard BL method performs quite well while the proposed two robust methods (RBL, RHS) perform almost in the same way. On the other hand, the performance of BL gets worse as the ratio of being outliers increases, compared with the other robust methods. Comparing the proposed methods with $tBL$, it is observed that they perform similarly when the contaminated error distribution is symmetric and has large variance as in Scenario (I), possibly because $t$-distribution can be effectively adapted to such structure. However, when the contamination distribution has different mean as in Scenario (II), the performance of $tBL$ gets worse than the proposed robust methods as the contamination increases. Comparing the proposed two robust methods, RHS using horseshoe prior is slightly better than RBL using Laplace prior in all the scenarios, which could be reasonable since horseshoe prior is known to have better performance than Laplace prior as shrinkage priors for regression coefficients. Regarding interval estimation, the results for AL and CP are given in Figure 2 and Table 1, respectively. From Table 1 CPs seem reasonable in all the methods while $tBL$ slightly show over-coverage properties. On the other hand, Figure 2 reveals a similar trend to one observed in log-MSE, so that the credible intervals of BL are shown to be very inefficient when there exist outliers. Also it should be pointed out that credible intervals of $tBL$ is inefficient compared with the proposed
methods even under Scenario (I). Comparing the proposed two robust methods, RHS provides slightly more efficient interval estimation than RBL, which is consistent with the results of log-MSE in Figure 1.

We next checked mixing properties of the proposed MCMC algorithm (denoted by RBL-proposal) compared with standard methods. We consider the same setting as above simulations. In particular, we show only results in the case of (II)-Homo. Note that we can obtain similar results in other cases, i.e., (I)-Homo, (I)-Hetero and (II)-Hetero. We set $\omega = 0.20$. In addition to the case of $p = 20$, we also consider the case of $p = 40$, where we set the same true non-zero regression coefficients as the case of $p = 20$. For comparison, we employed (non-robust) Bayesian lasso (BL) and robust Bayesian lasso with Langevin algorithm (e.g. Welling and Teh, 2011) with the step size 0.01 (denoted by RBL-Langevin). Note that the Langevin algorithm is applied to the full conditional distribution of $\beta$, so that this algorithm has only difference from the proposed algorithm in sampling from the full conditional distribution of $\beta$. Figures 3 and 4 show mixing and autocorrelation results for one-shot posterior simulation of $\beta_1$, $\beta_2$ and $\beta_{10}$. As is well-known, since the ordinal BL have the efficient Gibbs sampling algorithm, we can find that mixing and autocorrelation of them are not bad. However, sample paths of BL are affected by outliers. The second row shows results of Langevin algorithm for our proposed model. From Figures 3 and 4, mixings are not good and autocorrelations are very high. The bottom of Figure 3 and 4 show results based on the proposed algorithm. In this case, sample paths are not affected by outliers and it seems that there is no autocorrelation among posterior samples because posterior samples of regression coefficients are obtained by solutions of the optimization problem for each iterations. Hence, the proposed method is robust against outliers and enables us to generate posterior samples without being worried about the problem of autocorrelation. Furthermore, our algorithm does not depend on any tuning parameters (for example, we have to set some tuning parameters in Langevin or Metropolis algorithms). Figures 5 and 6 show the case of $p = 40$. Similar to the case of $p = 20$, our algorithm performs well, while the use of Langevin algorithm may cause more higher autocorrelation. We tried other possible choices of
tuning parameters in Langevin algorithm, but the results did not change much.

![Graphs showing log-MSE against tuning parameters for Homo and Hetero scenarios.](image)

Figure 1: Average values of log-MSE with ±3 Monte Carlo error bands based on 300 replications in four simulation scenarios.

### 3.2 Real data examples

We compare results of the proposed methods with that of the non-robust Bayesian Lasso through applications to two real datasets. We use two well-known datasets, Boston Housing data ([Harrison and Rubinfeld, 1978](https://www.jstor.org/stable/2921792)) and Diabetes data from [Efron et al., 2004](https://www.journalofroyalsocietyinterface.com/content/jrsi/1/1/169), which have \((n, p) = (506, 13)\) and \((n, p) = (442, 10)\), respectively. In our applications, we standardized all the covariates, so that they have 0 mean and 1 standard deviation. We used the log-transformed response values in the Boston
Figure 2: Average values of average length of 95% credible intervals with ±3 Monte Carlo error bands based on 300 replications in four simulation scenarios.

Housing data and original response values in the Diabetes data. For the datasets, we applied the proposed robust Bayesian methods with Laplace prior (RBL) and horseshoe prior (RHS). We set $\gamma = 0.2$ in both methods. For comparison, we also applied the standard non-robust Bayesian Lasso (BL). Based on 4000 posterior samples after discarding 1000 posterior samples, we computed posterior medians as well as 95% credible intervals of regression coefficients, which are shown in Figure 7. It is observed that the two robust methods, RBL and RHS produce similar results while the standard BL provides quite different results than the others in some coefficients. In particular, in the Diabetes dataset, the two robust methods detected 5th and 7th
| \( \omega \) | (I)-Homo | (II)-Homo |
|---|---|---|
| 0 | 95.9 | 95.0 | 93.7 | 93.3 | 95.9 | 95.0 | 93.7 | 93.3 |
| 0.05 | 96.7 | 96.5 | 94.1 | 93.9 | 96.2 | 96.9 | 94.1 | 93.7 |
| 0.10 | 96.6 | 97.7 | 94.5 | 94.5 | 96.3 | 98.6 | 94.4 | 94.0 |
| 0.15 | 96.2 | 98.3 | 95.6 | 95.5 | 96.3 | 99.1 | 94.8 | 94.5 |
| 0.20 | 96.7 | 99.0 | 96.0 | 95.9 | 96.5 | 98.7 | 95.3 | 94.8 |

| \( \delta \) | (I)-Hetero | (II)-Hetero |
|---|---|---|
| 0 | 95.9 | 95.0 | 93.7 | 93.3 | 95.9 | 95.0 | 93.7 | 93.3 |
| 1 | 96.4 | 96.5 | 94.3 | 94.2 | 96.1 | 97.1 | 93.8 | 93.8 |
| 2 | 96.1 | 97.7 | 94.5 | 94.3 | 94.9 | 98.6 | 94.1 | 93.5 |
| 3 | 96.4 | 98.2 | 95.6 | 95.3 | 94.1 | 98.8 | 94.5 | 94.3 |
| 4 | 96.4 | 98.4 | 96.2 | 96.1 | 93.4 | 96.7 | 95.5 | 94.8 |

variables as significant ones based on their credible intervals while the credible intervals of the BL method contains 0, which shows that the robust methods may be able to detect significant variables that the non-robust method cannot. A similar phenomena is observed in the 3rd covariate in the Boston Housing data. Comparing two figures, the degree of difference of the results between the two robust methods and the non-robust method in the Diabetes data is smaller than that in the Boston Housing data, for example, the posterior medians among the three methods are almost identical in the Diabetes data. This might show that the ratio of outliers in the Diabetes dataset is smaller than that of the Boston Housing data.

### 4 Conclusions and discussion

We proposed a new robust Bayesian regression method by using synthetic posterior based on \( \gamma \)-divergence. Using a technique of nonparametric Bayesian updating that optimizes a weighted objective function within Gibbs sampling, we developed an efficient posterior computation algorithm for obtained posterior samples of regression coefficients under shrinkage priors. The numerical performance of the proposed
method compared with existing methods are investigated through simulation and real data examples.

Although our presentation is focused on a linear regression in this paper, the proposed method can be conceptionally extended to other models such as generalized linear models (GLM). However, under general GLM, corresponding objective func-
Figure 4: Autocorrelation of posterior samples for three methods (BL, RBL-Langevin and RBL-proposal) in (II)-Homo case with $\omega = 0.20$ and $p = 20$.

Tions from $\gamma$-divergence is not tractable since it includes integrals or infinite sums (Kawashima and Fujisawa 2019), thereby the posterior computation would not be feasible even if we use a similar techniques used in this paper. Only logistic regression for binary outcomes could be a tractable regression model in which $\gamma$-divergence is obtained in an analytical form, and outliers in logistic regression is typically related
Figure 5: Trace plot of posterior samples for three methods (BL, RBL-Langevin and RBL-proposal) in (II)-Homo case with $\omega = 0.20$ and $p = 40$. Red line corresponds to the true value.

to a mislabeling problem (e.g. Hung et al., 2018). The detailed investigation including developing an efficient posterior computation algorithm could be valuable.

Finally, while $\gamma$ controls the degree of robustness, it has been fixed throughout this paper. In our numerical studies, we fixed $\gamma = 0.2$ and revealed that the simulation results are quite supportive since the purposed method performs significantly better
Figure 6: Autocorrelation of posterior samples for three methods (BL, RBL-Langevin and RBL-proposal) in (II)-Homo case with $\omega = 0.20$ and $p = 40$. than the other methods as well as the performance even under the case without any outliers is comparable with the non-robust method. However, it would be interesting to consider a data-dependent choice of $\gamma$. 

19
Figure 7: 95% credible intervals with posterior medians (×) of regression coefficients based on the standard Bayesian Lasso (BL), the proposed robust Bayesian methods with Laplace prior (RBL) and horseshoe prior (RHS), applied to Boston Housing data and Diabetes data.

Acknowledgement

This work is partially supported by Japan Society for Promotion of Science (KAKENHI) grant numbers 18K12757 and 17K14233.

References

Basu, A., I. R. Harris, N. L. Hjort, and M. C. Jones (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* 85, 549–559.

Bhattacharya, A., D. Pati, and Y. Yang (2019). Bayesian fractional posteriors. *The Annals of Statistics* 47, 39–66.

Bissiri, P. G., C. Holmes, and S. G. Walker (2016). A general framework for updating...
belief distributions. *Journal of the Royal Statistical Society: Series B* 78, 1103–1130.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika* 97, 465–480.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics* 32, 407–499.

Fujisawa, H. and S. Eguchi (2008). Robust parameter estimation with a small bias against heavy contamination. *Journal of Multivariate Analysis* 99, 2053–2081.

Harrison, D. and D. L. Rubinfeld (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management* 5, 81–102.

Hung, H., Z. Jou, and S. Huang (2018). Robust mislabel logistic regression without modeling mislabel probabilities. *Biometrics* 74, 145–154.

Hunter, D. R. and K. Lange (2004). A tutorial on mm algorithms. *The American Statistician* 58, 30–37.

Ishwaran, H. and J. S. Rao (2005). Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics* 33, 730–773.

Jones, M. C., N. L. Hjort, I. R. Harris, and A. Basu (2001). A comparison of related density-based minimum divergence estimators. *Biometrika* 88, 865–873.

Kawashima, T. and H. Fujisawa (2017). Robust and sparse regression via gamma-divergence. *Entropy* 19, 608.

Kawashima, T. and H. Fujisawa (2019). Robust and sparse regression in glm by stochastic optimization. *Japanese Journal of Statistics and Data Science to appear.*

Lyddon, S., S. G. Walker, and C. Holmes (2018). Nonparametric learning from bayesian models with randomized objective functions. *32nd Conference on Neural Information Processing Systems (NIPS 2018).*
Miller, J. W. and D. B. Dunson (2019). Robust bayesian inference via coarsening. *Journal of the American Statistical Association* 114, 1113–1125.

Nakagawa, T. and S. Hashimoto (2019). Robust bayesian inference via $\gamma$-divergence. *Communications in Statistics: Theory and Methods*, to appear.

Newton, M. A., N. G. Polson, and J. Xu (2018). Weighted bayesian bootstrap for scalable bayes. *arXiv*.

Newton, M. A. and A. E. Raftery (1994). Approximate bayesian inference with weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B* 56, 3–48.

Park, T. and G. Casella (2008). The bayesian lasso. *Journal of the American Statistical Association* 103, 681–686.

Rockova, V. and E. I. George (2018). The spike-and-slab lasso. *Journal of the American Statistical Association* 113, 431–444.

Rubin, D. B. (1981). The bayesian bootstrap. *The Annals of Statistics* 9, 130–134.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* 58, 267–288.

Welling, M. and Y. W. Teh (2011). Bayesian learning via stochastic gradient langevin dynamics. *International Conference on Machine Learning (ICML 2011).*