Research Article

Robust Bayesian Regularized Estimation Based on t Regression Model

Zean Li¹ and Weihua Zhao²

¹School of Computer Science and Technology, Nantong University, Nantong 226019, China

²School of Science, Nantong University, Nantong 226019, China

Correspondence should be addressed to Weihua Zhao; zhaowhstat@163.com

Received 5 June 2015; Revised 7 September 2015; Accepted 8 September 2015

Academic Editor: Shein-chung Chow

Copyright © 2015 Z. Li and W. Zhao. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The t distribution is a useful extension of the normal distribution, which can be used for statistical modeling of data sets with heavy tails, and provides robust estimation. In this paper, in view of the advantages of Bayesian analysis, we propose a new robust coefficient estimation and variable selection method based on Bayesian adaptive Lasso t regression. A Gibbs sampler is developed based on the Bayesian hierarchical model framework, where we treat the t distribution as a mixture of normal and gamma distributions and put different penalization parameters for different regression coefficients. We also consider the Bayesian t regression with adaptive group Lasso and obtain the Gibbs sampler from the posterior distributions. Both simulation studies and real data example show that our method performs well compared with other existing methods when the error distribution has heavy tails and/or outliers.

1. Introduction

Following the pioneer work of Tibshirani [1], the Lasso (least absolute shrinkage and selection operator) has generated much interest in statistical literatures [2–5]. For the Gaussian linear regression,

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where $\mathbf{x}_i$'s are either fixed or random covariates, $\epsilon_1, \dots, \epsilon_n$, i.i.d. $\sim N(0, \sigma^2)$. The Lasso estimates are viewed as L_1 regularized least squares estimates, which are defined as

$$\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^p |\beta_j|, \quad (2)$$

for $\lambda \geq 0$ and with $n \times 1$ vectors $\mathbf{y} = (y_1, \dots, y_n)^T$, $n \times p$ matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$, and $p \times 1$ vector $\boldsymbol{\beta}$. The key advantage of Lasso lies in its ability to do simultaneous parameter estimation and variable selection and the entire path of Lasso estimates for all values of λ can be efficiently computed

through a modification of the LARS (least angle regression) algorithm of Efron et al. [5].

The Gaussian assumption is not crucial in model (1), but it is useful to make connections to the likelihood framework for regularized t regression in Section 2. The Lasso estimator in (2) is equivalent to minimizing the penalized negative log-likelihood $l(\boldsymbol{\beta}; \mathbf{y})$ as a function of the regression coefficients $\boldsymbol{\beta}$ and using the L_1 -penalty $\|\boldsymbol{\beta}\| = \sum_{j=1}^p |\beta_j|$; equivalence means here that we obtain the same estimator for a potentially different tuning parameter, but the Lasso estimator in (2) does not provide an estimate of the variance parameter σ^2 . In fact, the variance parameter σ^2 plays an important role especially when the error in the regression model has high variance.

On the other hand, the Lasso estimate in (2) based on least square loss may suffer from poor performance when the error distribution has heavy tail which is longer than normal distribution or the error has large variance or contains outliers in the linear regression model, which motivate many researches to consider regularized estimation based on robust method. The most existing robust regularized estimation methods mainly replace the least square loss function in (2)

by some robust loss functions, such as Huber loss, L_1 loss, and quantile loss function. Li and Zhu [6] considered quantile regression with the Lasso penalty and developed its piecewise linear solution path. Wang et al. [7] investigated the least absolute deviance (LAD) estimate with adaptive Lasso penalty (LAD-Lasso) and proved its oracle property. Wu and Liu [8] further discussed the oracle properties of the SCAD (smoothly clipped absolute deviation) and adaptive Lasso regularized quantile regression. Zou and Yuan [9] proposed a new efficient estimation and variable selection procedure which is called composite quantile regression. Chen et al. [10] studied the robust Lasso in the presence of noise with large variance. They considered two editions of robust Lasso, which use the convex combined loss and Huber loss function criterion instead of least square criterion.

Besides the above existing robust regularized method, which, in essence, replace the L_2 loss by other loss criteria, we can also directly consider robust estimation from the standpoint of error distribution in model (1). It is well known that when the error in the linear regression follows a normal distribution, the estimations will be sensitive to outliers, which motivate us to consider using robust error distribution in the linear regression model in (1). The t distribution as an alternative to the normal distribution has frequently been suggested in the literature as a robust extension of the traditional normal models. The t distribution provides a convenient description for regression analysis when the residual term has density with heavy tails and excess kurtosis.

Since the Lasso estimate in (2) can be interpreted as a Bayesian posterior mode estimate when the regression parameters have independent Laplace priors, motivated by this connection, Park and Casella [11] discussed the Bayesian Lasso estimation for linear regression based on Gibbs sampler method. The Bayesian Lasso has several advantages; for example, it can conveniently provide interval estimates, that is, Bayesian credible intervals, that can guide variable selection, the structure of the hierarchical model provides methods for selecting the Lasso parameter, and sometimes its performances outperform non-Bayesian method. Therefore, several authors were subsequently devoted to the variable selection problem based on Bayesian method, such as Park and Casella [11], Hans [12], Li et al. [13], and Kyung et al. [14].

The main goal of this paper is to extend the Bayesian regularized method for t regression model (4), which is a robust edition Bayesian regularized estimation method, and it is expected that the performances will be better than other regularized methods when the error distribution follows a symmetry distribution and there are outliers in regression model.

The rest of the paper is organized as follows. In Section 2, we introduce the t regression with Lasso penalty. In Section 3, we build the framework of Bayesian t regression model with adaptive Lasso penalty and then show the corresponding Bayesian hierarchical model and obtain the Gibbs sample estimation procedure. In Sections 4 and 5, we carry out some simulation studies and real data analysis to illustrate the performance of our new method. Some conclusion and remarks are contained in Section 6. We also discuss the Bayesian t regression with group Lasso penalty and its Gibbs sampler method in the Appendices.

2. t Regression Model and L_1 Norm Regularization

2.1. t Regression Model. The t distribution as an alternative to the normal distribution has frequently been suggested in the literatures; for example, Lange et al. [15] introduced t distributed errors regression models as a robust extension of the traditional normal models; Liu and Rubin [16] discussed the estimations of t distribution by EM algorithm and its extensions; and Lin et al. [17] studied heteroscedasticity diagnostics problem by score test for t linear regression models. And it has often been used in many real applications, especially in finance.

The probability density function of a t distributed variable y with location parameter μ , dispersion parameter σ , and degree of freedom ν , denoted as $t(\mu, \sigma^2, \nu)$, is

$$f(y, \mu, \sigma; \nu) = \frac{\Gamma((1+\nu)/2)}{\sigma(\pi\nu)^{1/2}\Gamma(\nu/2)} \left(1 + \nu^{-1} \left(\frac{y-\mu}{\sigma}\right)^2\right)^{-(1+\nu)/2}, \quad (3)$$

where $\Gamma(\cdot)$ denotes the gamma function, $y, \mu \in \mathbb{R}$, and $\sigma > 0$.

The mean and variance of y are μ ($\nu > 1$) and $\nu\sigma^2/(\nu-2)$ ($\nu > 2$), respectively. When $\nu = \infty$ and 1, the t distribution reduces to the normal and Cauchy distributions, respectively. The univariate linear Student- t regression model can be expressed as

$$y_i = \mathbf{x}_i^T \beta + \varepsilon_i, \quad \varepsilon_i \sim t(0, \sigma^2, \nu), \text{ i.i.d.}, \quad (4)$$

where the $\mathbf{x}_i$'s are known p -vector of covariates and $\beta = (\beta_1, \dots, \beta_p)^T$ are the unknown p -vectors of parameters.

2.2. t Regression with Lasso Penalty. As discussed in Section 1, the L_2 loss in the Lasso model is not robust to heavy-tailed error distribution and/or outliers. This indicates that the Lasso is not an ideal goodness-of-fit measure criterion in the presence of noise with large variance. We consider following regularized t regression, which is defined as

$$\hat{\theta} = \underset{\beta \in \mathbb{R}^p, \sigma^2 > 0}{\operatorname{argmax}} \left\{ \sum_{i=1}^n l_i(\beta, \sigma^2; y_i, \mathbf{x}_i) - \lambda \sum_{j=1}^p |\beta_j| \right\}, \quad (5)$$

where $\theta = (\beta^T, \phi)^T$ and $l_i(\beta, \sigma^2; y_i, \mathbf{x}_i)$ is the log-likelihood function of y_i :

$$\begin{aligned} l_i(\beta, \sigma^2; y_i, \mathbf{x}_i) &= \log \left\{ \frac{\Gamma((1+\nu)/2)}{(\pi\nu)^{1/2}\Gamma(\nu/2)} \right\} - \frac{1}{2} \log \sigma^2 \\ &\quad - \left(\frac{1+\nu}{2} \right) \log \left(1 + \nu^{-1} \frac{(y_i - \mu_i)^2}{\sigma^2} \right). \end{aligned} \quad (6)$$

Comparing (5) with (2), we take the parameter σ^2 into the optimization of the penalized maximum likelihood estimator.

Though we are penalizing only the parameter β , the variance parameter estimate $\hat{\sigma}^2$ is influenced indirectly by the amount of shrinkage λ in (5).

There is main drawback of the estimator in (5) [18]. It is not equivalent under scaling of the response [19]. More precisely, consider the transformation

$$\begin{aligned}\bar{y} &= by, \\ \bar{\beta} &= b\beta, \\ \bar{\sigma} &= b\sigma\end{aligned}\quad (7)$$

$(b > 0),$

which leaves model (2) invariant, but this is not the case for the estimator in (5). We address this drawback by using the penalty term $\lambda\|\beta\|_1/\sigma$, leading to the following estimator:

$$\hat{\theta} = \underset{\beta \in \mathbb{R}^p, \sigma^2 > 0}{\operatorname{argmax}} \left\{ \sum_{i=1}^n l_i(\beta, \sigma^2; y_i, \mathbf{x}_i) - \lambda \sum_{j=1}^p \frac{|\beta_j|}{\sigma} \right\}. \quad (8)$$

Noting the form of the regularized estimation in (8), the Lasso estimates can be interpreted as posterior mode estimates when the regression parameters have independent Laplace (i.e., double-exponential) priors:

$$f(\beta_j | \sigma^2, \lambda) = \frac{\lambda}{2\sqrt{\sigma^2}} \exp \left\{ -\lambda \frac{|\beta_j|}{\sqrt{\sigma^2}} \right\}. \quad (9)$$

Park and Casella [11] argued that conditioning on σ^2 is important because it guarantees a unimodal full posterior.

3. Bayesian Formulation of Robust t Adaptive Lasso

3.1. Bayesian Adaptive Lasso t Regression. It is well known that the ordinary Lasso has a conflict between the consistency for model selection and optimal prediction. As a solution to achieve both estimation and model selection consistency, the adaptive Lasso [20] was proposed and its model selection consistency and asymptotic normality under certain rate of shrinkage parameter λ_n were proved. To extend the idea of the adaptive Lasso [20] to our proposed robust t Lasso regression, we define the robust t adaptive Lasso as

$$\hat{\theta} = \underset{\beta \in \mathbb{R}^p, \phi > 0}{\operatorname{argmax}} \left\{ \sum_{i=1}^n l_i(\beta, \phi; y_i, \mathbf{x}_i) - \sum_{j=1}^p \phi^{1/2} \lambda_j |\beta_j| \right\}, \quad (10)$$

where λ_j 's allow unequal penalties for the coefficients and we define the precision parameter $\phi = \sigma^{-2}$.

Motivated by the connection with Bayesian estimation and (9), we consider a fully Bayesian analysis using a conditional Laplace prior on β_j :

$$f(\beta_j | \phi, \lambda_j) = \frac{\phi \lambda_j}{2} \exp(-\lambda_j \phi |\beta_j|). \quad (11)$$

For any $a \leq 0$, by the following equality [21]:

$$\frac{a}{2} e^{-a|t|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} \exp\left(-\frac{t^2}{2s}\right) \frac{a^2}{2} \exp\left(-\frac{a^2 s}{2}\right) ds, \quad (12)$$

the Laplace prior (11) on β can be written as

$$\begin{aligned}f(\beta | \phi, \lambda_1, \dots, \lambda_p) &= \prod_{j=1}^p \int_0^\infty \frac{1}{\sqrt{2\pi s_j}} \exp\left(-\frac{\beta_j^2}{2s_j}\right) \\ &\quad \cdot \frac{\phi \lambda_j^2}{2} \exp\left(-\frac{\phi \lambda_j^2 s_j}{2}\right) ds_j.\end{aligned}\quad (13)$$

On the other hand, from Lange et al. [15] and Liu and Rubin [16], $t(\mu, \phi, \nu)$ can be treated as a mixture of normal and gamma distributions; that is, if $u \sim \Gamma(\nu/2, \nu/2)$ and $z | u \sim N(\mu, (\phi u)^{-1})$, then $z \sim t(\mu, \phi, \nu)$. If we further put gamma priors on the parameter λ_j^2 and use the improper prior density on ϕ , then we have the following Bayesian hierarchical model:

$$\begin{aligned}y_i &= \mathbf{x}_i^T \beta + (\phi u_i)^{-1/2} z_i, \\ f(\phi) &\propto \frac{1}{\phi}, \\ f(\mathbf{u}) &\propto \prod_{i=1}^n u_i^{\nu/2-1} \exp\left(-\frac{\nu}{2} u_i\right), \\ f(\mathbf{z}) &\propto \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z_i^2\right), \\ f(\beta, \mathbf{s} | \lambda_1^2, \dots, \lambda_p^2) &\propto \prod_{j=1}^p \frac{1}{\sqrt{2\pi s_j}} \exp\left(-\frac{\beta_j^2}{2s_j}\right) \frac{\phi \lambda_j^2}{2} \exp\left(-\frac{\phi \lambda_j^2 s_j}{2}\right), \\ f(\lambda_1^2, \dots, \lambda_p^2 | a, b) &\propto \prod_{j=1}^p \frac{b^a}{\Gamma(a)} (\lambda_j^2)^{a-1} \exp(-b \lambda_j^2), \\ f(a, b) &\propto \frac{1}{b},\end{aligned}\quad (14)$$

where $\mathbf{u} = (u_1, \dots, u_n)$, $\mathbf{z} = (z_1, \dots, z_n)$, and $\mathbf{s} = (s_1, \dots, s_p)$.

Remark 1. The amount of shrinkage parameters λ_j^2 's depend on the value of hyperparameters a and b . Because larger a and smaller b lead to bigger penalization, it is important to treat a and b as unknown parameters to avoid giving specific values which affect the estimates of the regression coefficients.

Remark 2. Similar to Park and Casella [11], there is also another existing method to obtain the estimation of adaptive Lasso parameters beside sampling from the posterior distribution of λ_j^2 's. We can use the empirical Bayes estimation method to update λ_j^2 's without setting the prior of λ_j^2 's

in the Bayesian hierarchical model. Our little experience shows that the two methods perform well, but the empirical Bayes estimation method can save some computing time; for more details see Park and Casella [11] and Leng et al. [22].

3.2. Gibbs Sampler for All Parameters. Let β_{-i} be the parameter vector β excluding the component β_i , $\mathbf{s}_{-j}$ the vector $\mathbf{s}$ excluding the component s_j , and $\mathbf{u}_{-i}$ the variable excluding the component u_i . Based on the Bayesian hierarchical model (14), the posterior distribution of all parameters can be given by

$$\begin{aligned}
 f(\beta, \phi, \mathbf{u}, \mathbf{s}, \lambda_1^2, \dots, \lambda_p^2, a, b \mid \mathbf{y}, \mathbf{x}) &\propto f(\mathbf{y} \mid \beta, \phi, \mathbf{u}, \mathbf{x}) \\
 &\cdot f(\phi) \prod_{i=1}^n f(u_i) \prod_{j=1}^p f(\beta_j, s_j \mid \lambda_j^2) f(\lambda_j^2 \mid a, b) \\
 &\cdot f(a, b) \propto \phi^{n/2-1} \exp \left\{ -\frac{\phi}{2} \sum_{i=1}^n u_i (y_i - \mathbf{x}_i^T \beta)^2 \right\} \\
 &\cdot \prod_{i=1}^n u_i^{(v-1)/2} \exp \left(-\frac{v}{2} u_i \right) \prod_{j=1}^p \frac{1}{\sqrt{2\pi s_j}} \exp \left(-\frac{\beta_j^2}{2s_j} \right) \\
 &\cdot \frac{\phi \lambda_j^2}{2} \exp \left(-\frac{\phi \lambda_j^2 s_j}{2} \right) \frac{b^a}{\Gamma(a)} (\lambda_j^2)^{a-1} \\
 &\cdot \exp(-b \lambda_j^2) \times b^{-1}.
 \end{aligned} \tag{15}$$

Based on the expression (15), we can obtain the posterior full conditional distribution for all parameters, which can yield a tractable and efficient Gibbs sampler that works as follows.

The full conditional distribution $f(u_i \mid \mathbf{x}, \mathbf{y}, \beta, \phi, \mathbf{u}_{-i}, \mathbf{s}, \lambda_1^2, \dots, \lambda_p^2, a, b)$ is

$$\begin{aligned}
 f(u_i \mid \mathbf{x}, \mathbf{y}, \beta, \phi, \mathbf{u}_{-i}, \mathbf{s}, \lambda_1^2, \dots, \lambda_p^2, a, b) \\
 \propto u_i^{(v+1)/2-1} \exp \left(-\left[\frac{\phi}{2} (y_i - \mathbf{x}_i^T \beta)^2 + \frac{v}{2} \right] u_i \right).
 \end{aligned} \tag{16}$$

Thus, the full conditional distribution of u_i is a gamma distribution.

The full conditional distribution $f(s_j \mid \mathbf{x}, \mathbf{y}, \beta, \phi, \mathbf{s}_{-j}, \mathbf{u}, \lambda_1^2, \dots, \lambda_p^2, a, b)$ is

$$\begin{aligned}
 f(s_j \mid \mathbf{x}, \mathbf{y}, \beta, \phi, \mathbf{s}_{-j}, \mathbf{u}, \lambda_1^2, \dots, \lambda_p^2, a, b) \\
 \propto \frac{1}{\sqrt{s_j}} \exp \left\{ -\frac{1}{2} [\phi \lambda_j^2 s_j + \beta_j^2 s_j^{-1}] \right\}.
 \end{aligned} \tag{17}$$

Thus, the full conditional distribution of s_j is a generalized inverse Gaussian distribution. Now we consider the full conditional distribution of β_j which is given by

$$\begin{aligned}
 f(\beta_j \mid \mathbf{x}, \mathbf{y}, \beta_{-j}, \phi, \mathbf{s}, \mathbf{u}, \lambda_1^2, \dots, \lambda_p^2, a, b) \\
 \propto \exp \left\{ -\frac{1}{2} \left[\left(\phi \sum_{i=1}^n u_i x_{ij}^2 + \frac{1}{s_j} \right) \beta_j^2 \right. \right. \\
 \left. \left. - 2\phi \sum_{i=1}^n u_i x_{ij} \tilde{y}_{ij} \beta_j \right] \right\},
 \end{aligned} \tag{18}$$

where $\tilde{y}_{ij} = y_i - \sum_{k=1, k \neq j} x_{ik} \beta_k$. Let $\tilde{\sigma}_j^{-2} = \phi \sum_{i=1}^n u_i x_{ij}^2 + 1/s_j$ and $\tilde{\mu}_j = \tilde{\sigma}_j^2 \phi \sum_{i=1}^n u_i x_{ij} \tilde{y}_{ij}$. Then the full conditional distribution of β_j is just the normal distribution $N(\tilde{\mu}_j, \tilde{\sigma}_j^2)$. The full conditional distribution of ϕ is

$$\begin{aligned}
 f(\phi \mid \mathbf{x}, \mathbf{y}, \beta, \mathbf{s}, \mathbf{u}, \lambda_1^2, \dots, \lambda_p^2, a, b) &\propto \phi^{n/2+p-1} \\
 &\cdot \exp \left\{ -\phi \left[\frac{\sum_{i=1}^n u_i (y_i - \mathbf{x}_i^T \beta)^2 + \sum_{j=1}^p \lambda_j^2 s_j}{2} \right] \right\}.
 \end{aligned} \tag{19}$$

So the full conditional distribution of ϕ is a gamma distribution.

The full conditional distribution of λ_j^2 is

$$\begin{aligned}
 f(\lambda_j^2 \mid \mathbf{x}, \mathbf{y}, \beta, \phi, \mathbf{s}, \mathbf{u}, \lambda_{k,k=1,\dots,p,k \neq j}^2, a, b) \\
 \propto (\lambda_j^2)^a \exp \left(-\left[\frac{\phi s_j}{2} + b \right] \lambda_j^2 \right).
 \end{aligned} \tag{20}$$

That is, the full conditional distribution of λ_j^2 is a gamma distribution. At last, the full conditional distributions of a and b are

$$\begin{aligned}
 f(a \mid \mathbf{x}, \mathbf{y}, \beta, \phi, \mathbf{s}, \mathbf{u}, \lambda_1^2, \dots, \lambda_p^2, b) \\
 \propto \left(\frac{b^a}{\Gamma(a)} \right)^p \left(\prod_{j=1}^p \lambda_j^2 \right)^a, \\
 f(b \mid \mathbf{x}, \mathbf{y}, \beta, \phi, \mathbf{s}, \mathbf{u}, \lambda_1^2, \dots, \lambda_p^2, a) \\
 \propto b^{ap-1} \exp \left(-b \sum_{j=1}^p \lambda_j^2 \right).
 \end{aligned} \tag{21}$$

Obviously, the full conditional distribution of b is just a gamma distribution. However, the full conditional posterior distribution of c does not have a closed form. Fortunately, we note that the full conditional posterior distribution of c is a log-concave function, so the adaptive rejection sampling algorithm [23] can be used to sample from this distribution. The adaptive rejection sampling algorithm is an efficient sampling method from any univariate probability density function which is log-concave.

Remark 3. The Lasso was originally designed for variable selection for the posterior mode of β_j 's can be exactly zeros due to the nature of the laplace prior in (11). However, the posterior draws of β_j 's cannot be exact zeros since β_j 's are drawn from the continuous posterior distribution. A post hoc thresholding rule may overcome this difficulty. Alternatively, kyung et al. [24] recommended using the credible interval on the posterior mean to guide variable selection.

Remark 4. For variable selection, sometimes, several explanatory variables may be represented by a group of derived input variables. In this case the selection of important variables corresponds to the group of variables. The group Lasso penalty [13, 25] takes the group structure into account and can do variable selection at the group level. For the Bayesian t regression model with adaptive group Lasso penalty, we give the details in the Appendices.

4. Simulation Studies

The main goal of this section is to assess the performance of our method (BAT.L) through some simulations with comparison to several Bayesian and non-Bayesian methods. The Bayesian methods include the Bayesian Lasso (BLS.L) [11] and Bayesian adaptive Lasso (BALS.L) [22]. The non-Bayesian methods include Lasso (LASSO) [1] and adaptive Lasso (ALASSO) [20]. The data in the simulation studies are generated by

$$y_i = \mathbf{x}_i^T \beta + \epsilon_i, \quad i = 1, \dots, n, \quad (22)$$

$\mathbf{x}_i \sim N(0, \Sigma)$, where the covariance matrix Σ is an AR(1) matrix; that is, $\Sigma_{jk} = \rho^{|j-k|}$ with correlation $\rho = 0.5$, and the error distributions of ϵ_i considered here are the following six situations:

The first choice is a standard normal distribution, $N(0, 1)$.

The second choice is a normal distribution with variance 9, $N(0, 3^2)$.

The third choice is a t distribution with degree of freedom $\nu = 3$, $t(3)$.

The fourth choice is Laplace distribution, $Lp(0, 1)$.

The last two choices are the mixture of two normal distributions: $0.8N(0, 1) + 0.2N(0, 6^2)$ and the mixture of two Laplace distributions: $0.9Lp(0, 1) + 0.1Lp(0, 5)$.

For each simulation study and each choice of the error distribution, we run 50 simulations. In each simulation, we generate a training set with 20 observations, a validation set with 20 observations, and a testing set with 200 observations. The validation set is used to choose the penalty parameters in LASSO and ALASSO. After the penalty parameters are selected, we combine the training set and validation set together to estimate β . Since BAT.L, BLS.L, and BALS.L do not need the validation set since they estimate the penalty automatically, so we combine the training set and validation

set together for estimation. The testing set is used to evaluate the performance for these methods. Within each simulation study, we consider three different settings for β .

Simulation 1: $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$.

Simulation 2: $\beta = (5, 0, 0, 0, 0, 0, 0, 0)^T$.

Simulation 3: if $j = 5, 10, 15, 20$, $\beta_j = 5$; else, $\beta_j = 0$, $j \leq 20$.

Simulations 1, 2, and 3 are corresponding to the sparse case, very sparse case, and sparse recovery problem in the predictors, respectively.

Note that we intentionally choose error distributions that are different from the t distribution and Laplace distribution to see how the Bayesian t regression adaptive Lasso estimation depends on the error assumption. Our simulation results show that, in terms of parameter estimation accuracy, the Bayesian t adaptive Lasso method still performs well even when this error distribution assumption is violated.

Unless otherwise specified, the freedom degree $\nu = 4$ is used in all simulations for BAT.L. Since the true model is known, we can compute the median of mean absolute deviations (MMAD), that is, median $(1/200 \sum_{i=1}^{200} |\mathbf{x}_i^T \hat{\beta} - \mathbf{x}_i^T \beta^{\text{true}}|)$, where the median of mean absolute deviations is calculated based on 50 simulations. We present the results for simulation in Table 1. The parentheses in these tables are the standard deviations of the MMAD obtained by 500 bootstrap resamplings of the 50 mean absolute deviations.

Simulations show that, in terms of the MMAD, our new regularized method performs better than the other four methods in general, especially for the nonnormal error distributions.

5. Land Rent Data

In this section, we demonstrate the performance of BAT.L together with other methods by a real data set.

Weisberg [26] reported a data set about land rent data. The data was collected by Douglas Tiffany to study the variation in rent paid in 1977 for agricultural land planted with alfalfa. The variables are average rent per acre planted to alfalfa (y), average rent paid for all tillable land (x_1), density of dairy cows (number per square mile) (x_2), and proportion of farmland used as pasture (x_3) and $x_4 = 1$ if liming is required to grow alfalfa; $x_4 = 0$, otherwise. The unit of analysis is a county in Minnesota; the 67 counties with appreciable rented farmland are included. The response variable alfalfa, which has mean 42.1661 and standard deviation 22.5866, is a high protein crop that is suitable feed for dairy cows. It is thought that rent for land planted with alfalfa relative to rent for other agricultural purposes would be higher in areas with a high density of dairy cows and rents would be lower in counties where liming is required, since that would mean additional expense.

We first center all the variables so that the intercept is not considered and the predictors are standardized and then use above five regularized methods to fit the whole data set. For the three Bayesian methods, we use the Bayesian credible

TABLE 1: Simulation results of MMAD for different distribution errors[†].

Method	$N(0, 1)$	$N(0, 3)$	$t(3)$	$Lp(0, 1)$	Normal mixture	Laplace mixture
Simulation 1						
BAT.L	0.2321 (0.1211)	1.0274 (0.4014)	0.3012 (0.1287)	0.2721 (0.2201)	0.3734 (0.4105)	0.4168 (0.3710)
BLS.L	0.3613 (0.1068)	0.9694 (0.3339)	0.4980 (0.2126)	0.4958 (0.1556)	0.8623 (0.3160)	0.8498 (0.5587)
BALS.L	0.2147 (0.1145)	0.9433 (0.4071)	0.3517 (0.2010)	0.3349 (0.1770)	0.7209 (0.3799)	0.6407 (0.6589)
LASSO	0.3221 (0.1311)	0.9755 (0.4875)	0.4396 (0.2771)	0.4306 (0.1635)	0.8073 (0.3775)	0.7993 (0.5934)
ALASSO	0.2219 (0.1165)	0.9609 (0.4882)	0.3504 (0.2628)	0.3265 (0.1962)	0.8352 (0.4046)	0.6840 (0.6998)
Simulation 2						
BAT.L	0.1362 (0.0658)	0.3932 (0.2607)	0.2188 (0.1658)	0.1745 (0.0906)	0.2587 (0.2985)	0.2330 (0.1060)
BLS.L	0.3081 (0.0877)	0.6778 (0.2811)	0.4881 (0.2249)	0.4709 (0.1606)	0.7045 (0.2552)	0.7356 (0.2784)
BALS.L	0.1308 (0.1100)	0.4518 (0.3194)	0.2479 (0.2349)	0.2015 (0.1759)	0.3700 (0.2961)	0.4355 (0.2568)
LASSO	0.1910 (0.1089)	0.6835 (0.3459)	0.3506 (0.2245)	0.3112 (0.1935)	0.5540 (0.3064)	0.4977 (0.3626)
ALASSO	0.0932 (0.1219)	0.4152 (0.3220)	0.2187 (0.2509)	0.1508 (0.2113)	0.2429 (0.3322)	0.3974 (0.3141)
Simulation 3						
BAT.L	0.3253 (0.1049)	0.8987 (0.4659)	0.3108 (0.1538)	0.3610 (0.3654)	0.3915 (0.1939)	0.4196 (0.0112)
BLS.L	0.6694 (0.1443)	1.7382 (0.3742)	0.8894 (0.3500)	0.9042 (0.2057)	1.6136 (0.6524)	1.3718 (0.0775)
BALS.L	0.3183 (0.1374)	0.9895 (0.3686)	0.5150 (0.2234)	0.4676 (0.1962)	0.9364 (0.5522)	0.6973 (0.0660)
LASSO	0.4469 (0.1449)	1.3959 (0.4807)	0.7231 (0.3199)	0.6462 (0.2530)	1.2599 (0.6710)	1.1223 (0.1315)
ALASSO	0.2803 (0.1463)	0.9267 (0.4087)	0.4601 (0.2306)	0.4513 (0.2290)	0.9204 (0.8101)	0.6814 (0.0987)

[†] In the parentheses are standard deviations of the MMADs obtained 500 bootstrap resampling. The bold numbers correspond to the smallest MMAD in each category.

TABLE 2: MSE of five different regularized methods for land rent data.

Method	BAT.L	LASSO	ALASSO	BLS.L	BALS.L
MSE	94.0764	114.1111	104.3413	109.9957	104.8580

interval to guide variable selection. We found that the five different methods all select x_1 and x_2 as important variables. Furthermore, in order to evaluate the performance of BAT.L together with other methods, we randomly select 40 samples as train data set and the remaining 27 samples as test data set; the mean square errors for test data set are reported in the Table 2 for five different regularized methods.

As we can see from Table 2, the performance of BAT.L outperforms other variable selection procedures for this data set.

6. Conclusion Remarks

In this paper, we studied the regularized t regression model based on Bayesian method. This new method extended the Bayesian Lasso [11] through replacing the least square loss function by the log-likelihood function of t distribution and the Lasso penalty by adaptive Lasso penalty. Bayesian hierarchical models are developed and Gibbs samplers are derived for our new method. Both the simulation studies and the real data analysis show that BAT.L performs better than other existing Bayesian and non-Bayesian methods when the error distribution has heavy tails and/or outliers. We also discussed the Bayesian t regression model with adaptive group Lasso penalty.

There is room to improve our methods. In this paper, we treat the dispersion parameter σ in t regression model as constant. In fact, the dispersion parameter σ may be different for different observations, which need us to consider the problem of regularized t regression with varying dispersion model; that is,

$$\begin{aligned}
 y_i &= \mathbf{x}_i^T \beta + \varepsilon_i \\
 \log \sigma_i &= \mathbf{z}_i^T \gamma \\
 \varepsilon_i &\sim t(0, \sigma_i^2, \nu)
 \end{aligned} \tag{23}$$

$i = 1, \dots, n,$

where $\mathbf{z}_i$'s are covariates, which constitute, in general, although not necessary, a subset of $\mathbf{x}_i$'s. For the varying dispersion t regression model, it is important to consider how to simultaneously do variable selection for both regression coefficients β and γ based on Bayesian method or non-Bayesian method. Furthermore, it also be interesting to consider the problem of variable selection for varying dispersion t regression models with high dimensional covariates. Research in this aspect is ongoing.

Appendices

A. Bayesian t Regression with Adaptive Group Lasso Penalty

In the Appendices, we consider regression problems where some explanatory variables may be represented by a group of derived input variables. In this case, the selection of

important variables corresponds to the group of variables. The group Lasso penalty [13, 25] takes the group structure into account and can do variable selection at the group level.

Suppose that the predictors are grouped into G groups and β_g is the coefficient vector of the g th group predictors $\mathbf{x}_{ig}$. Then $\beta = (\beta_1^T, \dots, \beta_G^T)^T$ and $\mathbf{x}_i = (\mathbf{x}_{i1}^T, \dots, \mathbf{x}_{iG}^T)^T$. Let d_g be the dimension of the vector β_g and let $\mathbf{K}_g$ be a known $d_g \times d_g$ positive definite matrix ($g = 1, \dots, G$). Define $\|\beta_g\|_{\mathbf{K}_g} = (\beta_g^T \mathbf{K}_g \beta_g)^{1/2}$. We consider the following adaptive group Lasso t regression:

$$\hat{\theta} = \underset{\beta \in \mathbb{R}^p, \phi > 0}{\operatorname{argmax}} \left\{ \sum_{i=1}^n l_i(\beta, \phi; y_i, \mathbf{x}_i) - \phi^{1/2} \sum_{g=1}^G \lambda_g \|\beta_g\|_{\mathbf{K}_g} \right\}. \quad (\text{A.1})$$

Pick the prior of β_g as the Laplace prior:

$$\begin{aligned} f(\beta_g | \phi, \lambda_g) \\ = C_{d_g} \sqrt{\det(\mathbf{K}_g)} \left(\sqrt{\phi \lambda_g} \right)^{d_g} \exp \left(-\sqrt{\phi \lambda_g} \|\beta_g\|_{\mathbf{K}_g} \right), \end{aligned} \quad (\text{A.2})$$

where $C_{d_g} = 2^{-(d_g+1)/2} (2\pi)^{-(d_g-1)/2} / \Gamma((d_g + 1)/2)$. By the equality (12), we have

$$\begin{aligned} f(\beta_g | \phi, \lambda_g) &= \int_0^\infty \frac{1}{\sqrt{\det(2\pi s_g \mathbf{K}_g^{-1})}} \exp \left\{ -\frac{1}{2} \right. \\ &\quad \cdot \beta_g^T (s_g \mathbf{K}_g^{-1})^{-1} \beta_g \left. \right\} \frac{(\phi \lambda_g^2 / 2)^{(d_g+1)/2}}{\Gamma((d_g + 1)/2)} s_g^{(d_g-1)/2} \\ &\quad \cdot \exp \left(-\frac{\phi \lambda_g^2}{2} s_g \right) ds_g. \end{aligned} \quad (\text{A.3})$$

If we further put gamma priors on the parameter λ_g^2 , then we have the following Bayesian hierarchical model:

$$\begin{aligned} y_i &= \mathbf{x}_i^T \beta + (\phi u_i)^{-1/2} z_i, \\ f(\phi) &\propto \frac{1}{\phi}, \\ f(u_i) &\propto u_i^{v/2-1} \exp \left(-\frac{v}{2} u_i \right), \\ f(z_i) &\propto \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{1}{2} z_i^2 \right), \\ f(\beta_g | s_g) & \end{aligned}$$

$$\begin{aligned} &\propto \frac{1}{\sqrt{\det(2\pi s_g \mathbf{K}_g^{-1})}} \exp \left\{ -\frac{1}{2} \beta_g^T (s_g \mathbf{K}_g^{-1})^{-1} \beta_g \right\}, \\ f(s_g | \lambda_g^2) & \\ &\propto \frac{(\phi \lambda_g^2 / 2)^{(d_g+1)/2}}{\Gamma((d_g + 1)/2)} s_g^{(d_g-1)/2} \exp \left(-\frac{\phi \lambda_g^2}{2} s_g \right), \\ f(\lambda_g^2 | a, b) &\propto \frac{b^a}{\Gamma(a)} (\lambda_g^2)^{a-1} \exp(-b \lambda_g^2), \\ f(a, b) &\propto \frac{1}{b}. \end{aligned} \quad (\text{A.4})$$

B. The Gibbs Sampler for Bayesian Adaptive Group Lasso t Regression

Based on the Bayesian hierarchical model (A.4), the posterior distribution of all parameters can be given by

$$\begin{aligned} f(\beta, \phi, \mathbf{u}, \mathbf{s}, \lambda_1^2, \dots, \lambda_G^2, a, b | \mathbf{y}, \mathbf{x}) &\propto f(\mathbf{y} | \beta, \phi, \mathbf{u}, \mathbf{x}) \\ &\cdot f(\phi) \prod_{i=1}^n f(u_i) \prod_{g=1}^G f(\beta_g, s_g | \lambda_g^2) f(\lambda_g^2 | a, b) \\ &\cdot f(a, b) \\ &\propto \phi^{n/2-1} \exp \left\{ -\frac{\phi}{2} \sum_{i=1}^n u_i (\mathbf{y}_i - \mathbf{x}_i^T \beta)^2 \right\} \prod_{i=1}^n u_i^{(v-1)/2} \\ &\cdot \exp \left(-\frac{v}{2} u_i \right) \prod_{g=1}^G s_g^{-d_g/2} \exp \left\{ -\frac{1}{2} \beta_g^T (s_g \mathbf{K}_g^{-1})^{-1} \beta_g \right\} \\ &\cdot \left(\frac{\phi \lambda_g^2}{2} \right)^{(d_g+1)/2} s_g^{(d_g-1)/2} \exp \left(-\frac{\phi \lambda_g^2}{2} s_g \right) \\ &\cdot \prod_{g=1}^G \frac{b^a}{\Gamma(a)} (\lambda_g^2)^{a-1} \exp(-b \lambda_g^2) \times b^{-1}. \end{aligned} \quad (\text{B.1})$$

By the expression (B.1), we can obtain the posterior full conditional distribution for all parameters, which can yield a tractable and efficient Gibbs sampler. It is easy to see that the full conditional distribution of u_i is the same as in the adaptive Lasso t regression.

The full conditional distribution $f(s_g | \mathbf{x}, \mathbf{y}, \beta, \phi, \mathbf{s}_{-g}, \mathbf{u}, \lambda_1^2, \dots, \lambda_g^2, a, b)$ is

$$\begin{aligned} f(s_g | \mathbf{x}, \mathbf{y}, \beta, \phi, \mathbf{s}_{-g}, \mathbf{u}, \lambda_1^2, \dots, \lambda_g^2, a, b) \\ \propto \frac{1}{\sqrt{s_g}} \exp \left\{ -\frac{1}{2} [\phi \lambda_g^2 s_g + \beta_g^T \mathbf{K}_g \beta_g s_g^{-1}] \right\}. \end{aligned} \quad (\text{B.2})$$

Thus, the full conditional distribution of s_g is a generalized inverse Gaussian distribution.

The full conditional distribution of β_g is given by

$$f(\beta_g | \mathbf{x}, \mathbf{y}, \beta_{-g}, \phi, \mathbf{s}, \mathbf{u}, \lambda_1^2, \dots, \lambda_g^2, a, b) \propto \exp \left\{ -\frac{1}{2} \left[\beta_g^T \left(\phi \sum_{i=1}^n u_i \mathbf{x}_{ig} \mathbf{x}_{ig}^T + s_g^{-1} \mathbf{K}_g \right) \beta_g - 2\phi \sum_{i=1}^n u_i \tilde{y}_{ig} \mathbf{x}_{ig}^T \right] \right\}, \quad (\text{B.3})$$

where $\tilde{y}_{ig} = y_i - \sum_{k=1, k \neq g}^G \mathbf{x}_{ik} \beta_k$.

Let $\tilde{\Sigma}_g^{-1} = \phi \sum_{i=1}^n u_i \mathbf{x}_{ig} \mathbf{x}_{ig}^T + s_g^{-1} \mathbf{K}_g$ and $\tilde{\mu}_j = \phi \tilde{\Sigma}_g \sum_{i=1}^n u_i \tilde{y}_{ig} \mathbf{x}_{ig}^T$. Then the full conditional distribution of β_g is just the normal distribution $N(\tilde{\mu}_g, \tilde{\Sigma}_g)$.

The full conditional distribution of λ_g^2 is

$$f(\lambda_g^2 | \mathbf{x}, \mathbf{y}, \beta, \phi, \mathbf{s}, \mathbf{u}, \lambda_{k,k=1,\dots,G,k \neq g}^2, a, b) \propto (\lambda_g^2)^{(d_g+1)/2+a-1} \exp \left(- \left[\frac{\phi s_g}{2} + b \right] \lambda_g^2 \right). \quad (\text{B.4})$$

That is, the full conditional distribution of λ_g^2 is a gamma distribution.

The full conditional distribution of ϕ is

$$f(\phi | \mathbf{x}, \mathbf{y}, \beta, \phi, \mathbf{s}, \mathbf{u}, \lambda_1^2, \dots, \lambda_g^2, a, b) \propto \phi^{n/2+G(d_g+1)/2-1} \cdot \exp \left\{ -\phi \left[\frac{\sum_{i=1}^n u_i (y_i - \mathbf{x}_i^T \beta)^2 + \sum_{g=1}^G \lambda_g^2 s_g}{2} \right] \right\}. \quad (\text{B.5})$$

At last, the full conditional distributions of a and b are

$$f(a | \mathbf{x}, \mathbf{y}, \beta, \phi, \mathbf{s}, \mathbf{u}, \lambda_1^2, \dots, \lambda_p^2, b) \propto \left(\frac{b^a}{\Gamma(a)} \right)^G \left(\prod_{g=1}^G \lambda_g^2 \right)^a, \quad (\text{B.6})$$

$$f(b | \mathbf{x}, \mathbf{y}, \beta, \phi, \mathbf{s}, \mathbf{u}, \lambda_1^2, \dots, \lambda_g^2, a) \propto b^{aG-1} \exp \left(-b \sum_{g=1}^G \lambda_g^2 \right).$$

Again, the full conditional distribution of b is gamma distribution and we can use the adaptive rejection sampling algorithm (Gilks, 1992) to obtain the sampler of a .

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The research was supported in part by the Research Project of Social Science and Humanity Fund of the Ministry of Education (14YJC910007).

References

- [1] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society, Series B: Methodological*, vol. 58, no. 1, pp. 267–288, 1996.
- [2] W. J. Fu, "Penalized regressions: the bridge versus the LASSO," *Journal of Computational and Graphical Statistics*, vol. 7, no. 3, pp. 397–416, 1998.
- [3] K. Knight and W. Fu, "Asymptotics for lasso-type estimators," *The Annals of Statistics*, vol. 28, no. 5, pp. 1356–1378, 2000.
- [4] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.
- [5] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.
- [6] Y. Li and J. Zhu, " L_1 -norm quantile regression," *Journal of Computational and Graphical Statistics*, vol. 17, no. 1, pp. 163–185, 2008.
- [7] H. Wang, G. Li, and G. Jiang, "Robust regression shrinkage and consistent variable selection through the LAD-Lasso," *Journal of Business and Economic Statistics*, vol. 25, no. 3, pp. 347–355, 2007.
- [8] Y. Wu and Y. Liu, "Variable selection in quantile regression," *Statistica Sinica*, vol. 19, no. 2, pp. 801–817, 2009.
- [9] H. Zou and M. Yuan, "Composite quantile regression and the oracle model selection theory," *The Annals of Statistics*, vol. 36, no. 3, pp. 1108–1126, 2008.
- [10] X. Chen, Z. J. Wang, and M. J. McKeown, "Asymptotic analysis of robust LASSOs in the presence of noise with large variance," *IEEE Transactions on Information Theory*, vol. 56, no. 10, pp. 5131–5149, 2010.
- [11] T. Park and G. Casella, "The Bayesian lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008.
- [12] C. Hans, "Bayesian lasso regression," *Biometrika*, vol. 96, no. 4, pp. 835–845, 2009.
- [13] Q. Li, R. Xi, and N. Lin, "Bayesian regularized quantile regression," *Bayesian Analysis*, vol. 5, no. 3, pp. 533–556, 2010.
- [14] M. Kyung, J. Gill, M. Ghosh, and G. Casella, "Penalized regression, standard errors, and Bayesian lassos," *Bayesian Analysis*, vol. 5, no. 2, pp. 369–411, 2010.
- [15] K. L. Lange, R. J. Little, and J. M. Taylor, "Robust statistical modeling using the t distribution," *Journal of the American Statistical Association*, vol. 84, no. 408, pp. 881–896, 1989.
- [16] C. Liu and D. Rubin, "ML estimation of the t distribution using EM and its extensions, ECM and ECME," *Statistica Sinica*, vol. 5, no. 1, pp. 19–39, 1995.
- [17] J.-G. Lin, L.-X. Zhu, and F.-C. Xie, "Heteroscedasticity diagnostics for t linear regression models," *Metrika*, vol. 70, no. 1, pp. 59–77, 2009.
- [18] N. Städler, P. Bühlmann, and S. Van de Geer, " L_1 -penalization for mixture regression models," *TEST*, vol. 19, no. 2, pp. 209–256, 2010.
- [19] E. Lehmann, *Theory of Point Estimation*, Wadsworth and Brooks/Cole, Pacific Grove, Calif, USA, 1983.
- [20] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.
- [21] D. F. Andrews and C. L. Mallows, "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society Series B: Methodological*, vol. 36, pp. 99–102, 1974.

- [22] C. Leng, M. Tran, and D. Nott, “Bayesian adaptive lasso,” <http://arxiv.org/abs/1009.2300>.
- [23] W. R. Gilks and P. Wild, “Adaptive rejection sampling for Gibbs sampling,” *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 41, no. 2, pp. 337–348, 1992.
- [24] M. Kyung, J. Gill, M. Ghosh, and G. Casellax, “Penalized regression, standard errors, and Bayesian lassos,” *Bayesian Analysis*, vol. 5, no. 2, pp. 369–411, 2010.
- [25] M. Yuan and Y. Lin, “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society Series B*, vol. 68, no. 1, pp. 49–67, 2006.
- [26] S. Weisberg, *Applied Linear Regression*, Wiley, New York, NY, USA, 1985.