Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 8, No. 2, June 2020, pp.262-271

Bayesian group Lasso regression for left-censored data

Saja Hussein Aljanabi, Rahim Alhamzawi
Statistics Department, College of Administration and Economics, University of Al-Qadisiyah

ABSTRACT

In this paper, a new approach for model selection in left-censored regression has been presented. Specifically, we proposed a new Bayesian group Lasso for variable selection and coefficient estimation in left-censored data (BGLRLC). A new hierarchical Bayesian modeling for group Lasso has introduced, which motivate us to propose a new Gibbs sampler for sampling the parameters from the posteriors. The performance of the proposed approach is examined through simulation studies and a real data analysis. Results show that the proposed approach performs well in comparison to other existing methods.

Keywords: Left-censored regression, Bayesian group Lasso left-censored regression (BGLRLC), Variable selection (VS)

Corresponding Author:

Saja Hussein Aljanabi,
Statistics Department,
College of Administration and Economics,
University of Al-Qadisiyah, Al-Diwaniyah, Iraq
E-mail: sajaljanabi@yahoo.com

1. Introduction

Statistics is an influential tool for measuring the impact of experimental data and for drawing the accurate conclusions from it. The importance of statistics appears in demonstrating different phenomena by models that are closer to the reality. The latent variable in left censored model contains a large number of observations that are less than a certain value. These data have been widely employed in many fields of science such as economy, epidemiology, chemistry, and geology. The left-censored regression model assumes

$$y_i = \begin{cases} y_i^* & if \quad y_i{}^* > c \\ c & if \quad y_i^* \leq c \end{cases},$$

where y_i is the observed dependent variable, c denotes to the left-censored point, and y_i^* is unobserved dependent variable defined as follows

$$y_i^* = \boldsymbol{x}_i^t \boldsymbol{\beta} + \varepsilon_i \,, \quad i = 1, \dots, n \,, \qquad (1)$$

Here, $\boldsymbol{x}_i^t$ is a $1 \times k$ matrix of covariates, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_G)^{\,t}$, and $\varepsilon_i \sim N(0, \sigma^2)$.

The important problem in the censored regression model is selecting important covariates to increase the accuracy of the model and build a good predictive model. Specifically, as number of variables has been greater in comparison to the sample size or when there is a strong correlation between the covariates, the variance-covariance matrix is singular. Therefore, several methods were introduced to select important variables. For this reason, Akaike, in 1974, presented Akaike Information Criterion (AIC) to select a good predictive model [1]. However, AIC produces the inconsistent model [2]. So, AIC is weak in selecting the optimal model when $n < k$ ([3].

Schwarz presented the Bayesian Information Criterion (BIC) for overcoming problems regarding AIC. Specifically, BIC produces a consistent model. However, the performance of BIC does not work satisfactorily when $k > n$ [4].

The authors of [5] provided an algorithm to deal with the problem of variables selection and to overcome the problem of (AIC) and (BIC). This algorithm is called stochastic search variable selection (SSVS). Although such algorithm produces a good model, it consumes a lot of time and sometimes it does not acquire the correct model.

Recently, regularization approaches were suggested for estimating parameters and selecting variables simultaneously. For example, Ridge regression presented in [6], addressed problem of linear multiplicity by making the variance of the estimations slightly smaller. Later, Tibshirani suggested the Lasso regression by imposing an $L_1 - \text{norm}$ for ordinary least squares (OLS) [7]. Although this is an attractive feature, lasso regression does not select variables well when the explanatory variables are larger than the sample size($k > n$), or when there is a strong correlation between the covariates. To overcome the downsides regarding lasso regression, Zou and Hastie assumed elastic net regression [8], which compromises between the lasso ridge penalty(L_2) as well as penalty (L_1). This method performs well in selecting the important variables and estimating the parameters regarding model with regard to high correlations between variables or even when ($k > n$), as well as group selection. However, it has required high computational cost. Thus, many researchers have proposed ways to solve the problem of group selection and grouping structure of between covariates. Bakin proposed in [9], the development of group lasso as well as the development of group selection methods presented in [10]. Kim et al., presented Lasso group within the generalized linear models [11]. Meier and Bühlmann proposed Lasso group with the logistic regression [12]. The absolute general punishment method provided by [13] is an extension of the Lasso method. Hashem et al., proposed Bayesian quantile regression with group lasso penalty [14].

The aim of this study, as compared with above studies, is to present new group Lasso for selecting groups regarding the significant variables of left-censored regression. After that, novel Gibbs algorithm for sampling parameters with regard to the variable selection has been conducted. Simulation results as well as real data analysis indicated that the proposed new approach executed excellently in superior results as compared to the present approaches in the literatures.

2. Bayesian group Lasso with left-censored

The Lasso group estimator has been suggested by Yuan and Lin in 2006 to solve the following problem[10]:

$$\widehat{\boldsymbol{\beta}} = argmin(y - X\beta)^t(y - X\beta) + \sum_{g=1}^{G} \lambda_g \|\beta_g\|, \qquad (2)$$

where $y = (y_1, \dots y_n)^t$, $\lambda_g \geq 0$ is the regularization parameters, G is sizes of the groups and $\|\beta_{\boldsymbol{g}}\|$ is the L_1 penalty of $\beta_{\boldsymbol{g}}$. The Lasso group performs well when the structure of the group of variables is known [15]. The attractive feature of this method is the feasibility to get rid of a group of unimportant variables by making their coefficients equal to zero at the same time [11]. This leads to automatic variable selection and estimation of parameters simultaneously. This method has diverse solutions on the level of groups [10]. Lasso group is a generalization or expansion for Lasso as Lasso can be distinctive condition related to the Lasso group. Kyung et al., in 2010, suggested Bayesian group lasso for a linear regression model as treatment of Lasso group for overcoming problems of covariance matrix estimation. Similar to the Bayesian lasso, Kyung et al., proposed a hierarchical representation for the Lasso group imposed on each group a multi-Laplace prior [16]. Based on [10], the Bayesian group lasso for censored data can be written as:

$$\widehat{\boldsymbol{\beta}} = argmin(y^* - X\beta)^t(y^* - X\beta) + \sum_{g=1}^{G} \lambda_g \|\beta_g\| . \qquad (3)$$

Based on [17-21], a key step in any Bayesian analysis is the prior distribution. Based on [22], we assign the following prior distribution for $\boldsymbol{\beta}$ to proceed a Bayesian analysis as follows:

$$\pi(\boldsymbol{\beta}) = C(\lambda)exp\big(-\lambda\|\beta_g\|\big),$$

Here, $C(\lambda) = \frac{\lambda^k}{2^k}$. Thus, we get:

$$\pi(\boldsymbol{\beta}) = \frac{\lambda^k}{2^k} exp\left(-\lambda \left\| \beta_g \right\|\right)$$

$$= \frac{\lambda^{k+1}}{2^k} \int\limits_{\|\boldsymbol{\beta}\|_1 < s} exp\{-\lambda s\} ds$$

Let $V_k(Q) = \frac{2^k s^k}{\Gamma(k+1)}$, $s > 0$

$$= \int\limits_{\|\boldsymbol{\beta}\|_1 < s} \frac{\lambda^{k+1}}{\Gamma(k+1)} s^k exp\{-\lambda s\} ds$$

This study will convert the above-mentioned formula in the following way:

Assuming $r = \lambda\, s \quad \Longrightarrow \quad dr = \lambda\, ds$

$s = \frac{r}{\lambda} \quad \Longrightarrow \quad ds = \frac{dr}{\lambda}$

Then,

$$\frac{\lambda^k}{2^k} e^{-\lambda \|\beta_g\|} = \int\limits_{\|\lambda\beta\| < r} \frac{\lambda^{k+1}}{\Gamma(k+1)} \left(\frac{r}{\lambda}\right)^k \exp(-r)\, \frac{dr}{\lambda}$$

$$= \int_{\|\lambda\beta\| < r} \frac{1}{\Gamma(k+1)}\, r^k\, \exp(-r)\, dr \qquad (4)$$

2.1. Bayesian hierarchical model

This study will provide Bayesian hierarchical model based on hierarchical model reported in [22] as follows:

$$\boldsymbol{y}^* | \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim N_n(X\boldsymbol{\beta}, \sigma^2 \mathrm{I_n})$$

$$y_i = max\{y_i^*, c\}, \quad i = 1, \dots, n$$

$$y_i^* = \boldsymbol{x_i^t}\boldsymbol{\beta} + \varepsilon_i$$

$$\beta_g | r_g \sim Multivariate\ uniform\ for\ g = 1, \dots, G, \qquad (5)$$

$r_1, \dots, r_G | \lambda_1, \dots, \lambda_G \sim \prod_{g=1}^{G} Gamma(k, 1)$,

Here, m_g is the dimension of g,

$$\sigma^2 \sim \frac{b^h}{\Gamma(h)} \sigma^{2^{-h-1}} \exp\left(-\frac{b}{\sigma^2}\right), \qquad b, h > 0$$

$$\lambda_g \sim gamma\,(f, d)$$

2.2. Full conditional distribution

The full conditional distribution is related to y_i^* is as follows:

$$y_i^* | y_i, \boldsymbol{\beta} \sim \begin{cases} \delta(y_i) & if\ y_i > c, \\ N\left(\boldsymbol{x_i'}\boldsymbol{\beta}, \sigma^2\right) & otherwise, \end{cases}$$

where δ can be defined as degenerated distribution. Based on [22], the full conditional posterior distribution regarding ($\boldsymbol{\beta}$) is as follows:

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}^*, \boldsymbol{X}, \lambda) \propto \pi(\boldsymbol{y}^*|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta}|\lambda)$$

$$\propto exp\left\{-\frac{1}{2\sigma^2}(y^* - X\beta)^t(y^* - X\beta)\right\}\prod_{g=1}^{G} I\left\{\|\beta_g\| < \frac{r_g}{\lambda_g}\right\}$$

$$\boldsymbol{\beta}|\boldsymbol{y}^*, \boldsymbol{X}, \boldsymbol{r}, \sigma^2, \lambda_1, \dots, \lambda_G \sim N_k(\widehat{\boldsymbol{\beta}}, \sigma^2(\boldsymbol{X^t X})^{-1})\prod_{g=1}^{G} I\left\{-\frac{r_g}{\lambda_g} < \beta_g < \frac{r_g}{\lambda_g}\right\}, \qquad (6)$$

Here, $\widehat{\boldsymbol{\beta}} = (X^t X)^{-1} X^t y^*$

Thus, full conditional posterior distribution regarding ($\boldsymbol{r}$) is as follows:

$$\pi(\boldsymbol{r}|\boldsymbol{y}^*, \boldsymbol{X}, \boldsymbol{\beta}, \lambda) \propto \pi(\boldsymbol{r}) I\{r_g > \lambda_g \|\beta_g\|\}$$

$$\boldsymbol{r}|\boldsymbol{y}^*, \boldsymbol{X}, \boldsymbol{\beta}, \lambda_1, .., \lambda_G \sim \prod_{g=1}^{G} Exponential(1) I\{r_g > \lambda_g \|\beta_g\|\}, \qquad (7)$$

Similarly, the full conditional posterior distribution regarding $\pi(\sigma^2)$ is as follows:

$$\pi(\sigma^2|\boldsymbol{y}^*, \boldsymbol{X}, \boldsymbol{\beta}) \propto \pi(\boldsymbol{y}^*|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)\pi(\sigma^2)$$

$$\propto (\sigma^2)^{-\frac{n}{2}} exp\left\{-\frac{1}{2\sigma^2}(y^* - X\beta)^t(y^* - X\beta)\right\}\frac{b^h}{\Gamma(h)}\sigma^{2^{-h-1}} \exp\left(-\frac{b}{\sigma^2}\right)$$

$$\sigma^2|\boldsymbol{y}^*, \boldsymbol{X}, \boldsymbol{\beta} \sim Inverse\ Gamma\left(\frac{n}{2} + h,\ \frac{1}{2}(y^* - X\beta)^t(y^* - X\beta) + b\right) \qquad (8)$$

$$\pi(\lambda|\boldsymbol{\beta}) \propto \pi(\boldsymbol{\beta}|\lambda)\ \pi(\lambda)$$

$$\propto {\lambda_g}^{m_g} {\lambda_g}^{f-1} exp(-\lambda d) \prod_{g=1}^{G}\left\{\lambda_g < \frac{r_g}{\|\beta_g\|}\right\}$$

$$\propto {\lambda_g}^{(m_g+f)-1} exp(-\lambda d) \prod_{g=1}^{G}\left\{\lambda_g < \frac{r_g}{\|\beta_g\|}\right\}$$

$$\lambda_g \sim gamma\left((m_g + f), d\right) \prod_{g=1}^{G}\left\{\lambda_g < \frac{r_g}{\|\beta_g\|}\right\} \qquad (9)$$

3. Computation

The conditional posterior distributions of each parameter have a standard model. This enables us to Gibbs sampling with regard to the model's parameters. The full common distributions can be found by applying the MCMC algorithm as follows:

1- Updating y_i^*

$$y_i^*|y_i,\boldsymbol{\beta}\sim\begin{cases}\delta(y_i) & if\ y_i>c\\ N(\boldsymbol{x_i'\beta},\sigma^2) & otherwise,\end{cases}$$

2- Updating $(\boldsymbol{\beta})$ from multivariate normal distribution with mean $(X^tX)^{-1}X^ty^*$ and the variance $(\sigma^2(X^tX)^{-1})$.

3- Updating (r_g) from exponential distribution $(\lambda_g)I\{r>\|\lambda_g\beta_g\|\}$ as follows:

$$Updating\ r_g^*\sim Exponential(1).$$

$$r_g=r_g^*+\|\lambda_g\beta_g\|$$

4- Updating (σ^2) from the Inverse Gamma with shape parameter $\frac{n}{2}+h$ in addition to rate parameter $(\frac{(y^*-X\beta)^t(y^*-X\beta)}{2}+b)$.

5-Updating (λ_g) from Gamma distribution, the shape (m_g+f) as well as rate (d)

4. Simulation studies

In this section, we have demonstrated the performance of the proposed method for Bayesian group lasso regression for left censored data (BGLRLC) by simulations. This method is compared with the standard left-censored regression (SLCR), Bayesian left censored regression (BLCR) and Bayesian Lasso left censored regression (BLLCR). These four methods are assessed based on the median of mean absolute deviations over 250 simulations. The convergence of the BLLCR algorithm is checked by trace plots using the coda package in R. We also display the histograms of the posterior samples of BLLCR algorithm using the psych package in R.

4.1. Simulation 1

This simulation study considers prediction accuracy with 5 variables simulated from standard multivariate normal distribution. We simulate 100 observations from the model:

$$y_i=\max\{y_i^*,0\}$$

where $y_i^*=\boldsymbol{x}_i^t\boldsymbol{\beta}+\boldsymbol{\varepsilon}_i$ and $\boldsymbol{x}_i^t$ represents a vector of 6 covariates and $\boldsymbol{\varepsilon}_i$ is simulated from standard normal distribution. The true regression coefficients, including the intercept term, are $\beta=((0,\ 2,\ 0),\ (0,\ 0,\ 0))$ which divided in two groups $(0,\ 2,\ 0)\ and\ \ (0,\ 0,\ 0)$. Of the 100 observations, fifty of them are censored. Therefore, the censoring ratio is 50%.

Table 1 explains the MMAD results for simulation 1. We can observe that the BGLRLC has smallest MMAD and SD, which confirms that our proposed method performs better than the other methods in terms of MMAD and SD.

Table 1. MMADs for Simulation 1

Method	MMAD	SD
BGLRLC	0.4012	0.0811
SLCR	0.6644	0.1576
BLCR	0.8614	0.0824
BLLCR	0.4146	0.1415

Table 2 explains the parameter estimates for simulation 1. We can see from Table 2 that the BGLRLC gives very closed results to the true values compared to the other methods.

Table 2. Parameter estimations for simulation 1

Method	β_0	β_1	β_2	β_3	β_4	β_5
β (True)	0.00000	2.00000	0.00000	0.00000	0.00000	0.00000
BGLRLC	0.00000	1.77659	0.00000	0.1875	0.00000	-0.15947
SLCR	-0.16671	1.91309	0.01213	0.10076	-0.12246	-0.33888
BLCR	-0.42281	2.06575	0.00501	0.09376	-0.10687	-0.35883

The trace plots in Figure 1 show that the samples of the BGLRLC method traverse the posterior space very fast**.**

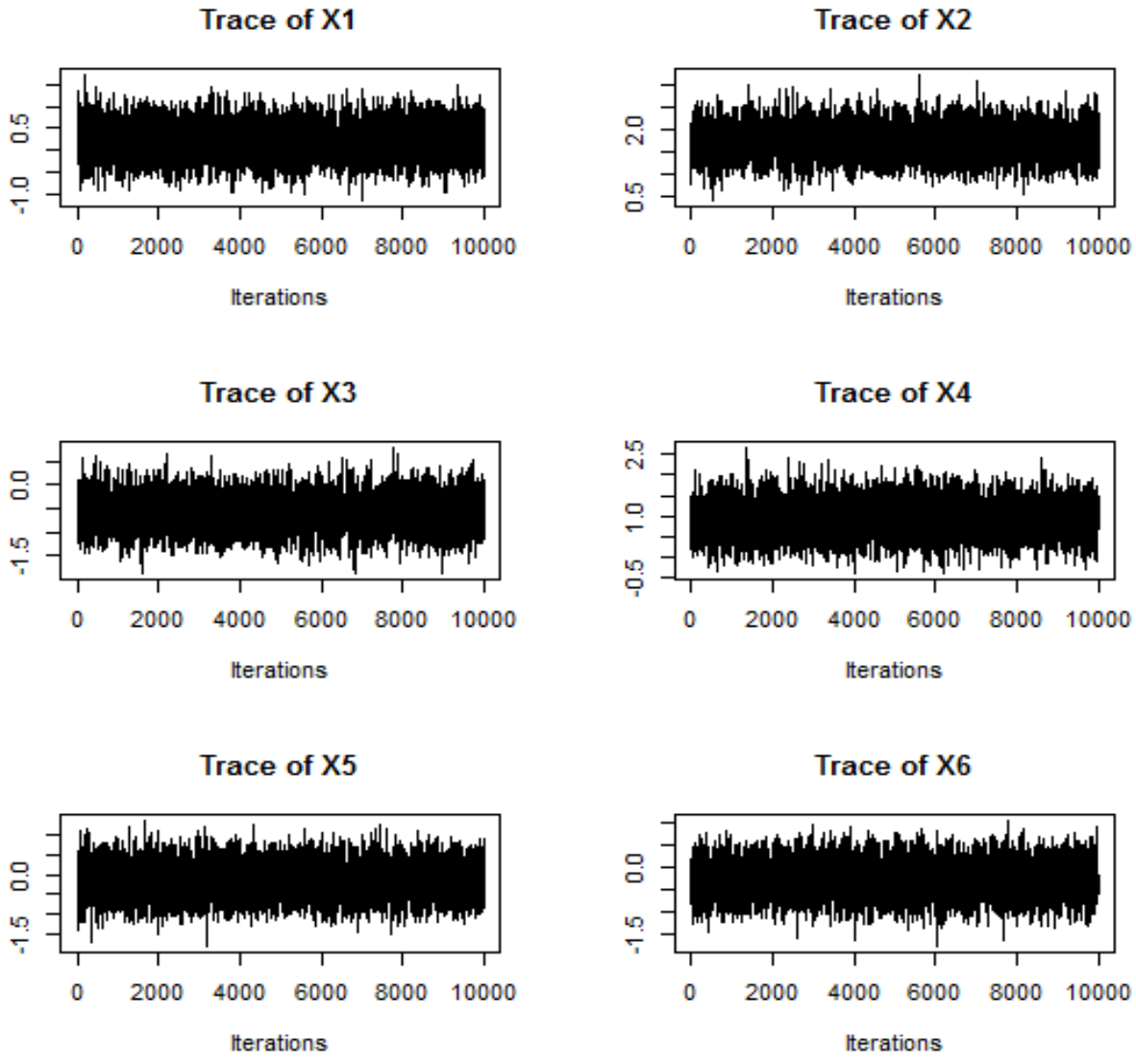

Figure 1. Trace plots for the variables in Simulation 1

The posterior histograms of the BGLRLC method in Figure 2 show that the conditional posteriors, indeed, the desired stationary truncated univariate normal distributions.

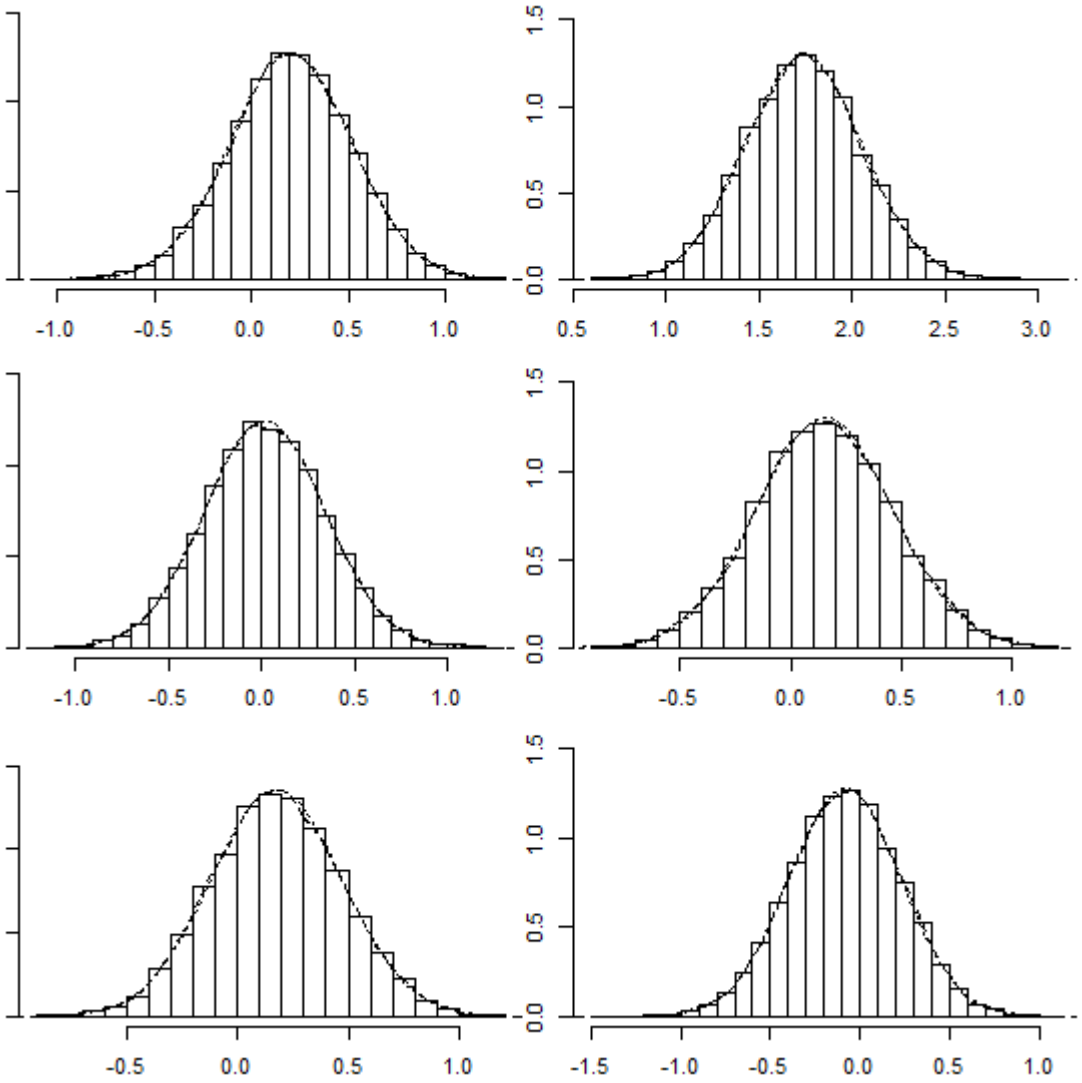

Figure 2. Posterior histograms for simulation 1

4.2. Simulation 2

This simulation study is similar to simulation 1 except that we set the following:
$\beta = ((1,\ 2,\ 0),\ (0,\ 0,\ 0),\ (1,\ 1,\ 0),\ (0,\ 0,\ 0),\ (1,\ 1,\ 0))$. This setup allows us to illustrate the performance of the BGLRLC method in the case with group structures in the covariates.
Table 3 explains the MMAD results for simulation 1. We can see that the BGLRLC has smallest MMAD and SD compared to the other methods in the comparison.

Table 3. MMADs for Simulation 2

Method	MMAD	SD
BGLRLC	1.00315	0.11384
SLCR	1.04711	0.14893
BLCR	1.01459	0.45348
BLLCR	1.00395	0.19435

5. Real data

We illustrate the proposed method with the data of the absence of university students. This dataset has 200 observations on 19 variables. The response variable is the number of absence days, while the other eighteen variables are covariates as follows:

y : The number of absence days.
x_1: Gender.
x_2: Age.
x_3: Social status.
x_4: The distance between the student's residence and the university in (km).
x_5: Student bears a number of family responsibilities.
x_6: Family problems that occur within the family.
x_7: Student health status (chronic diseases).
x_8: The death of a family member or relatives.
x_9: Inability to get up early.
x_{10}: Stay up late on social media, late at night.
x_{11}: The psychological state of the student.
x_{12}: Suffering from insomnia at night.
x_{13}: Preparing for exams (revision).
x_{14}: There is no limitation by management and teachers in accounting for absent students.
x_{15}: Excessive demand for homework and failure to perform it.
x_{16}: Bad weather (rain, hail,).
x_{17}: The security situation in the student's surrounding environment.
x_{18}: Feeling satisfied with his affiliation with the university or his major.

Table 4. Posterior mean for parameter estimates of real data example

Variables	BGLRLC	BLLCR	SLCR	BLCR
Intercept	7.291	6.991	7.105	7.192
X_1	0	0	-0.407	-0.417
X_2	0.219	0.181	-0.024	-0.026
X_3	0	0	0.322	0.461
X_4	0	0.034	0.022	0.022
X_5	0	0	-0.078	-0.077
X_6	0	0	-0.188	-0.187
X_7	0	0	-0.681	-0.658
X_8	0.392	0.387	0.398	0.388
X_9	0	0	0.088	0.079
X_{10}	0	0	0.053	0.059
X_{11}	0	0	0.937	0.896
X_{12}	0	0	-0.216	-0.215
X_{13}	0	0	-0.32	-0.309
X_{14}	0	0	0.226	0.22
X_{15}	0	0	0.242	0.228
X_{16}	0	0	0.194	0.198
X_{17}	0	0	-0.104	-0.094
X_{18}	0	0	-0.221	-0.219

Table 4 summarized the results of the real data example. To evaluate the methods, the DIC was computed for the four methods of BGLRLC, BLLCR, SLCR, and BLCR. The values were 1101.2219, 1273.448, 1192.844 and 1153.2781, respectively. The results of DIC show that the BGLRLC has better performance than the other methods.

6. Conclusions

This paper introduced a new procedure for selecting the left-censored regression model using the SUM as the prior distribution. Additionally, we introduced a new model for the Bayesian hierarchy of the Lasso group, which motivates us to suggest a new Gibbs sampling tool for sampling parameters. Simulation results and analyses of the absence of university students have shown that our procedure performed in a higher performance as compared to other procedures in the literature.

References

[1] H. Akaike, "A new look at the statistical model identification", In *Selected Papers of Hirotugu Akaike*, Springer, New York, NY, pp. 215-222, 1974.

[2] R. Nishii, "Asymptotic properties of criteria for selection of variables in multiple regression", *The Annals of Statistics*, pp.758-765, 1984.

[3] J. Dziak, R. Li, and L. Collins, "Critical Review and Comparison of Variable Selection Procedures for Linear Regression", State College, PA: Pennsylvania State University, 2005.

[4] G. Schwarz, Estimating the dimension of a model, *The Annals of Statistics*, vol. 6, pp.461–464, 1978.

[5] E. I. George, and R. E. McCulloch, "Variable selection via Gibbs sampling", *Journal of the American Statistical Association*, *88*(423), 881-889, 1993.

[6] A. E. Hoerl, and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems", *Technometrics*, vol.12, no.1, pp.55-67, 1970.

[7] R. Tibshirani, "Regression shrinkage and selection via the lasso", *Journal of the Royal Statistical Society: Series B (Methodological)*, vol.58, no.1, 267-288, 1996.

[8] H. Zou, and T. Hastie, "Regularization and variable selection via the elastic net", *Journal of the royal statistical society: series B (statistical methodology)*, vol.*67*, no.2, pp.301-320, 2005.

[9] S. Bakin, "Adaptive regression and model selection in data mining problems", PhD Thesis, Australian National University, 1999.

[10] M. Yuan, and Y. Lin, "Model selection and estimation in regression with grouped variables", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol.68, no.1, 49-67, 2006.

[11] Y. Kim, J. Kim, and Y. Kim, "Blockwise sparse regression", *Statistica Sinica*, vol16, no.2, p.375, 2006.

[12] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol.70, no.1, pp.53-71, 2008.

[13] P. Zhao, G. Rocha, and B. Yu, "The composite absolute penalties family for grouped and hierarchical variable selection", *The Annals of Statistics*, vol.37, no.6A, 3468-3497, 2009.

[14] H. Hashem, V. Vinciotti, R. Alhamzawi, and K. Yu, "Quantile regression with group lasso for classification", *Advances in Data Analysis and Classification*, vol.10, no.3, pp.375-390, 2016.

[15] J. Huang, and T. Zhang, "The benefit of group sparsity", *The Annals of Statistics*, vol.38, no.4, pp.1978-2004.

[16] 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.

[17] R. Alhamzawi, and K. Yu, " Power prior elicitation in Bayesian quantile regression", Journal of Probability and Statistics, vol.2011, 2011.

[18] R. Alhamzawi, and H. T. M. Ali, "The Bayesian adaptive lasso regression", *Mathematical biosciences*, vol.303, pp.75-82, 2018.

[19] R. Alhamzawi, and H. T. M. Ali, "The Bayesian elastic net regression", *Communications in Statistics-Simulation and Computation*, vol.47, no.4, pp.1168-1178, 2018.
[20] R. Alhamzawi, K. Yu, V. Vinciotti, and A. Tucker, "Prior elicitation for mixed quantile regression with an allometric model", *Environmetrics*, vol.22, no.7, pp.911-920, 2011.
[21] Z. Y. Algamal, R. Alhamzawi, and H. T. M. Ali, "Gene selection for microarray gene expression classification using Bayesian Lasso quantile regression", *Computers in biology and medicine*, vol.97, pp.145-152, 2018.
[22] H. Mallick, and N. Yi, "Bayesian group bridge for bi-level variable selection", *Computational statistics & data analysis*, vol.110, pp.115-133, 2017.