Scientific Research Publishing

Advances in Pure Mathematics, 2022, 12, 392-404
https://www.scirp.org/journal/apm
ISSN Online: 2160-0384
ISSN Print: 2160-0368

# $L_{1/2}$ Regularization Based on Bayesian Empirical Likelihood

**Yuan Wang, Wanzhou Ye**

Department of Mathematics, College of Science, Shanghai University, Shanghai, China
Email: ahwy@shu.edu.cn, wzhy@shu.edu.cn

**How to cite this paper:** Wang, Y. and Ye, W.Z. (2022) $L_{1/2}$ Regularization Based on Bayesian Empirical Likelihood. *Advances in Pure Mathematics*, **12**, 392-404.
https://doi.org/10.4236/apm.2022.125029

**Received:** April 8, 2022
**Accepted:** May 28, 2022
**Published:** May 31, 2022

## Abstract

Bayesian empirical likelihood is a semiparametric method that combines parametric priors and nonparametric likelihoods, that is, replacing the parametric likelihood function in Bayes theorem with a nonparametric empirical likelihood function, which can be used without assuming the distribution of the data. It can effectively avoid the problems caused by the wrong setting of the model. In the variable selection based on Bayesian empirical likelihood, the penalty term is introduced into the model in the form of parameter prior. In this paper, we propose a novel variable selection method, $L_{1/2}$ regularization based on Bayesian empirical likelihood. The $L_{1/2}$ penalty is introduced into the model through a scale mixture of uniform representation of generalized Gaussian prior, and the posterior distribution is then sampled using MCMC method. Simulations demonstrate that the proposed method can have better predictive ability when the error violates the zero-mean normality assumption of the standard parameter model, and can perform variable selection.

## Keywords

Bayesian Empirical Likelihood, Generalized Gaussian Prior, $L_{1/2}$ Regularization, MCMC Method

## 1. Introduction

Empirical likelihood is a nonparametric method first proposed by Owen [1] [2] [3], which is an estimation method inspired by maximum likelihood, but does not require assumptions about the distribution of the data. Thus, we can avoid potential problems of model misspecification. Because of the robustness of empirical likelihood, and the fact that it inherits many desirable properties of parametric likelihood, empirical likelihood has been extended to linear models, correlation models, variance models [3], general estimating equations [4], generalized linear models [5] and longitudinal data analysis [6] [7], etc.

On the one hand, the Bayesian method based on empirical likelihood has the advantages of Bayesian inference, and on the other hand, it avoids the risk of incorrect model assumptions, and has received extensive attention from scholars and has developed rapidly. Bayesian empirical likelihood was first proposed by Lazar [8]. It is a semiparametric method that combines parametric priors and nonparametric likelihoods. It not only pays attention to the use of overall information and sample information, but also pays attention to the collection of prior information. After processing, it forms a prior distribution and participates in statistical inference. Lazar [8] replaced the likelihood function in Bayes theorem with an empirical likelihood function, and used Monte Carlo simulation to prove the validity of the obtained posterior distribution. Zhong and Ghosh [9] studied some higher-order properties of Bayesian empirical likelihood. Li, Zhao and Dong [10] applied Bayesian empirical likelihood to linear regression models with censored data. Bedoui and Lazar [11] proposed the Bayesian empirical likelihood for lasso regression and ridge regression. Moon and Bedoui [12] proposed an empirical-likelihood-based Bayesian elastic network model that combines the interpretability and robustness of Bayesian empirical likelihood methods, which can be used for variable selection. In addition, Bayesian empirical likelihood is also extended to quantile structural equation modeling [13], quantile regression [14], etc.

Variable selection under the Bayesian framework, that is, introducing penalty terms into the model in the form of parameter priors. For example, Park and Casella [15] used conditional Laplace prior for complete Bayesian analysis and proposed Bayesian lasso. In addition, Li and Lin [16] proposed Bayesian elastic net using an informative prior. Mallick and Yi [17] proposed a new Bayesian lasso method based on uniform scale mixing of Laplace density. The variable selection based on Bayesian empirical likelihood is to replace the parametric likelihood function in Bayes theorem with a nonparametric likelihood function, which can be studied without making assumptions about the distribution of the data, avoiding problems caused by misspecified models.

This paper is divided into six sections. The first section introduces the research status of empirical likelihood and Bayesian empirical likelihood, and how to select variables based on Bayesian empirical likelihood. Section 2 derives the empirical likelihood function for linear models. Section 3 introduces the basics of Bayesian empirical likelihood. The fourth section is the focus of this paper, where $L_{1/2}$ regularization based on Bayesian empirical likelihood is proposed, and the penalty term is added to the model in the form of a generalized Gaussian prior. Section 5 verifies the effectiveness of the proposed method when the error violates the normality assumption of zero mean of the standard parameter model by simulation. The sixth section is the conclusion of this paper.

## 2. Empirical Likelihood Inference for Linear Models

Suppose we observe a set of data $\left(\boldsymbol{x}_1, y_1\right)\cdots\left(\boldsymbol{x}_n, y_n\right)$, if the relationship between

$\boldsymbol{x}_i$ and $y_i$ is linear, it can be represented by the following mathematical model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_q x_{iq} + \varepsilon_i,\ i = 1,2,\cdots,n, \tag{1}$$

where $\boldsymbol{x}_i = \left(x_{i1}, x_{i2}, \cdots, x_{iq}\right)^{\mathrm{T}}$ is the predictor variable, $y_i$ is the response variable, $\beta_0$ is the unknown intercept, $\beta_j$ is the unknown slope of the explanatory variable $x_{ij}$, and $\varepsilon_i$ is the error. In the standard parametric model, we generally assume that the errors are independent and obey a normal distribution with a mean of zero and a constant variance. But in the empirical likelihood, we relax the distributional assumption of the error, and the error distribution does not necessarily satisfy the normality assumption of zero mean. Next, without loss of generality, assuming that both the predictor and response variables are standardized, then the intercept term $\beta_0$ is equal to zero.

Let

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \cdots & x_{1q} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nq} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix},\ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_q \end{pmatrix},\ \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},\ \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

where $\boldsymbol{X}$ is the design matrix of $n \times q$, $\boldsymbol{\beta}$ is the vector of $q \times 1$, $\boldsymbol{y}$ is the vector of $n \times 1$, $\boldsymbol{\varepsilon}$ is the vector of $n \times 1$. Then the above multiple linear regression model can be expressed as:

$$y_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i,\ i = 1,2,\cdots,n. \tag{2}$$

Also, in linear models, regression coefficients are generally estimated by minimizing the residual sum of squares $\left\|\boldsymbol{y} - \boldsymbol{X\beta}\right\|_2^2$. Using the matrix notation defined above and assuming $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is invertible, the canonical equation is obtained

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\beta} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}.$$

That is, the regression coefficient satisfies the following estimation equation:

$$E\left(\boldsymbol{X}^{\mathrm{T}}\left(\boldsymbol{y} - \boldsymbol{X\beta}\right)\right) = 0.$$

Defining auxiliary variables $Z_i\left(\boldsymbol{\beta}\right) = \boldsymbol{x}_i\left(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)$, the profile empirical likelihood ratio of the regression parameters $\boldsymbol{\beta}$ can be obtained as follows:

$$R\left(\boldsymbol{\beta}\right) = \max_{\omega_i}\left\{\prod_{i=1}^{n} n\omega_i \mid \omega_i \geq 0, \sum_{i=1}^{n}\omega_i = 1, \sum_{i=1}^{n}\omega_i Z_i\left(\boldsymbol{\beta}\right) = \boldsymbol{0}\right\}. \tag{3}$$

Then apply the Lagrange multiplier method to solve the $\omega_i$ that satisfies the formula (3). If you want to find the $\omega_i$ that maximizes $\prod_{i=1}^{n} n\omega_i$, it is equivalent to finding the $\omega_i$ that maximizes $\sum_{i=1}^{n}\log n\omega_i$. Let

$$G = \sum_{i=1}^{n}\log n\omega_i - n\boldsymbol{\eta}^{\mathrm{T}}\sum_{i=1}^{n}\omega_i\boldsymbol{x}_i\left(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right) - \gamma\left(1 - \sum_{i=1}^{n}\omega_i\right), \tag{4}$$

where $\boldsymbol{\eta} = \left(\eta_1, \eta_2, \cdots, \eta_q\right)^{\mathrm{T}}$ and $\gamma$ are Lagrange multipliers. Let the partial derivative of $G$ with respect to $\omega_i$, $\boldsymbol{\eta}$ and $\gamma$ be zero, and the following equations can be obtained:

$$\begin{cases} \dfrac{\partial G}{\partial \omega_i} = 0 \Leftrightarrow \dfrac{n}{n\omega_i} - n\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{x}_i\left(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right) + \gamma = 0 \qquad ① \\ \dfrac{\partial G}{\partial \boldsymbol{\eta}} = 0 \Leftrightarrow -n\sum_{i=1}^{n}\omega_i\boldsymbol{x}_i\left(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right) = \mathbf{0} \qquad ② \\ \dfrac{\partial G}{\partial \gamma} = 0 \Leftrightarrow 1 - \sum_{i=1}^{n}\omega_i = 0 \qquad ③ \end{cases} \tag{5}$$

By multiplying both sides of Equation ① in formula (5) by $\omega_i$ at the same time and summing it up, we can get

$$0 = \sum_{i=1}^{n}\omega_i\frac{\partial G}{\partial \omega_i} = n + \gamma.$$

That is, $\gamma = -n$. Then substitute $\gamma = -n$ into Equation ① in formula (5) to get

$$\omega_i = \frac{1}{n}\cdot\frac{1}{1 + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{x}_i\left(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)}.$$

Then the profile empirical likelihood function of the regression coefficient $\boldsymbol{\beta}$ can be written, which is given by $L_{EL}\left(\boldsymbol{\beta}\right) = \exp\left\{l_{EL}\left(\boldsymbol{\beta}\right)\right\}$, where

$$l_{EL}\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{n}\log\omega_i = -n\log n - \sum_{i=1}^{n}\log\left\{1 + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{x}_i\left(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right\}. \tag{6}$$

Substitute the expression of $\omega_i$ into the ② in formula (5), and the Lagrange multiplier $\boldsymbol{\eta} = \boldsymbol{\eta}\left(\boldsymbol{\beta}\right)$ can be solved by the following equation:

$$\sum_{i=1}^{n}\frac{\boldsymbol{x}_i\left(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)}{1 + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{x}_i\left(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)} = \mathbf{0}. \tag{7}$$

Next, it is proved that under some regular conditions, if $\hat{\boldsymbol{\beta}}$ makes the profile logarithmic likelihood function $l_{EL}\left(\boldsymbol{\beta}\right)$ maximum, then $\hat{\boldsymbol{\beta}}$ converges to the true value $\boldsymbol{\beta}_0$ according to the probability.

**Theorem** (Consistency) Under some regular conditions, if

$$\hat{\boldsymbol{\beta}} = \arg\max l_{EL}\left(\boldsymbol{\beta}\right) = \arg\min\sum_{i=1}^{n}\log\left\{1 + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{x}_i\left(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right\},$$

then $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$.

Proof: Let $R\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{n}\log\left\{1 + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{x}_i\left(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right\}$, and denote $\boldsymbol{\beta} = \boldsymbol{\beta}_0 + \boldsymbol{u}n^{-1/3}$ for $\boldsymbol{\beta} \in \left\{\boldsymbol{\beta} \middle| \left\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\right\| \le n^{-1/3}\right\}$ where $\left\|\boldsymbol{u}\right\| = 1$. Owen [2] proved that when $\left\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\right\| \le n^{-1/3}$, there is

$$\begin{aligned} \boldsymbol{\eta}\left(\boldsymbol{\beta}\right) &= \left[\frac{1}{n}\sum_{i=1}^{n}Z_i\left(\boldsymbol{\beta}\right)Z_i^{\mathrm{T}}\left(\boldsymbol{\beta}\right)\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n}Z_i\left(\boldsymbol{\beta}\right)\right] + o\left(n^{-1/3}\right) \\ &= O\left(n^{-1/3}\right)\ (a.s.). \end{aligned}$$

Then perform Taylor expansion on $R\left(\boldsymbol{\beta}\right)$, we get:

$$\begin{aligned}
R(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{\beta}) Z_i(\boldsymbol{\beta}) - \frac{1}{2}\sum_{i=1}^{n}\left[\boldsymbol{\eta}^{\mathrm{T}}(\boldsymbol{\beta}) Z_i(\boldsymbol{\beta})\right]^2 + o\left(n^{-1/3}\right)\ (a.s.) \\
&= \frac{n}{2}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i(\boldsymbol{\beta})\right]^{\mathrm{T}} \left[\frac{1}{n}\sum_{i=1}^{n} Z_i(\boldsymbol{\beta}) Z_i^{\mathrm{T}}(\boldsymbol{\beta})\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} Z_i(\boldsymbol{\beta})\right] + o\left(n^{-1/3}\right)\ (a.s.) \\
&= \frac{n}{2}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i(\boldsymbol{\beta}_0) + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial Z_i(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \boldsymbol{u} n^{-1/3}\right]^{T} \left[\frac{1}{n}\sum_{i=1}^{n} Z_i(\boldsymbol{\beta}) Z_i^{\mathrm{T}}(\boldsymbol{\beta})\right]^{-1} \\
&\quad \times \left[\frac{1}{n}\sum_{i=1}^{n} Z_i(\boldsymbol{\beta}_0) + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial Z_i(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \boldsymbol{u} n^{-1/3}\right] + o\left(n^{-1/3}\right)\ (a.s.) \\
&= \frac{n}{2}\left[O\left(n^{-1/2}(\log\log n)^{1/2}\right) + E\left(\frac{\partial Z_i(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right)\boldsymbol{u} n^{-1/3}\right]^{T} \left[E\left(Z_i(\boldsymbol{\beta}_0) Z_i^{\mathrm{T}}(\boldsymbol{\beta}_0)\right)\right]^{-1} \\
&\quad \times \left[O\left(n^{-1/2}(\log\log n)^{1/2}\right) + E\left(\frac{\partial Z_i(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right)\boldsymbol{u} n^{-1/3}\right] + o\left(n^{-1/3}\right)\ (a.s.) \\
&\geq (c-\varepsilon) n^{1/3}\ (a.s.),
\end{aligned}$$

where $c-\varepsilon>0$, $c$ is the smallest eigenvalue of $E\left(\partial Z_i(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}\right)^{\mathrm{T}}\left[E\left(Z_i(\boldsymbol{\beta}_0)Z_i^{\mathrm{T}}(\boldsymbol{\beta}_0)\right)\right]^{-1} E\left(\partial Z_i(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}\right)$. Similarly, it can also be shown that

$$\begin{aligned}
R(\boldsymbol{\beta}_0) &= \frac{n}{2}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i(\boldsymbol{\beta}_0)\right]^{\mathrm{T}} \left[\frac{1}{n}\sum_{i=1}^{n} Z_i(\boldsymbol{\beta}_0) Z_i^{\mathrm{T}}(\boldsymbol{\beta}_0)\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} Z_i(\boldsymbol{\beta}_0)\right] + o(1) \\
&= O(\log\log n)\ (a.s.).
\end{aligned}$$

Since $R(\boldsymbol{\beta})$ is continuous with respect to $\boldsymbol{\beta}$, and $\boldsymbol{\beta}$ is in the sphere $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|$, so $R(\boldsymbol{\beta})$ has a minimum value in the sphere, that is $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$.

## 3. Bayesian Empirical Likelihood

Penalized linear regression and Bayesian linear regression are closely related, and their estimates can be interpreted as Bayesian posterior estimates of parameters under certain priors. For linear models:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{8}$$

Under the assumption that the noise obeys the Gaussian distribution of the regularization framework, from the perspective of probability, the regularized least squares method corresponds to the maximum a posteriori estimate, namely

$$P(\boldsymbol{\beta}|\mathrm{Data}) \propto P(\mathrm{Data}|\boldsymbol{\beta})P(\boldsymbol{\beta}).$$

Then the maximum a posteriori estimate of the parameter $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} P(\boldsymbol{\beta}|\mathrm{Data}) = \arg\max_{\boldsymbol{\beta}} P(\mathrm{Data}|\boldsymbol{\beta})P(\boldsymbol{\beta}),$$

where $P(\boldsymbol{\beta})$ is the prior distribution of the parameter $\boldsymbol{\beta}$. When the parameter $\boldsymbol{\beta}$ obeys the Laplace distribution, the $\mathrm{L}_1$ regularization is derived; when the parameter $\boldsymbol{\beta}$ obeys the Gaussian distribution, the $\mathrm{L}_2$ regularization is derived. From the above, it can be seen that lasso regression and ridge regression are closely related to Bayesian linear models when different priors are placed on the parameters.

The Bayesian empirical likelihood is as follows: Let $\boldsymbol{X}=\left(\boldsymbol{x}_1,\cdots,\boldsymbol{x}_q\right)$ be an independent multivariate random variable subject to an unknown distribution $F_{\boldsymbol{\beta}}$, whose unknown distribution $F_{\boldsymbol{\beta}}\in\mathcal{F}_{\boldsymbol{\beta}}$ depends on the parameter $\boldsymbol{\beta}=\left(\beta_1,\cdots,\beta_q\right)^{\mathrm{T}}\in\Omega\in\mathbb{R}^Q$. Assuming that both the predictor and response variables are standardized, then the intercept term is zero. Let the prior of $\boldsymbol{\beta}$ be $\pi\left(\boldsymbol{\beta}\right)$, and when the data distribution is unknown, replace the parameter likelihood function in Bayes theorem with the empirical likelihood function, then the posterior empirical likelihood density is

$$\pi\left(\boldsymbol{\beta}\mid\boldsymbol{X},\boldsymbol{y}\right)=\frac{L_{EL}\left(\boldsymbol{\beta}\right)\pi\left(\boldsymbol{\beta}\right)}{\int_{\Omega}L_{EL}\left(\boldsymbol{\beta}\right)\pi\left(\boldsymbol{\beta}\right)\mathrm{d}\boldsymbol{\beta}}\propto L_{EL}\left(\boldsymbol{\beta}\right)\pi\left(\boldsymbol{\beta}\right). \tag{9}$$

Combining the empirical likelihood inference of multiple linear regression in section 2, we can obtain the posterior inference of the Bayesian empirical likelihood of linear regression as

$$\pi\left(\boldsymbol{\beta}\mid\boldsymbol{X},\boldsymbol{y}\right)\propto\exp\left[\log\left\{\pi\left(\boldsymbol{\beta}\right)\right\}-\sum_{i=1}^{n}\left\{1+\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{x}_i\left(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right\}\right]. \tag{10}$$

## 4. L$_{1/2}$ Regularization Inference Based on Bayesian Empirical Likelihood

### 4.1. Hierarchical Model

Linear regression L$_{1/2}$ regularization penalizes the magnitude of the regression coefficients by imposing an L$_{1/2}$ penalty, that is, it minimizes the penalized residual sum of squares as follows:

$$\min_{\boldsymbol{\beta}}\left(\frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\right\|_2^2+\lambda\left\|\boldsymbol{\beta}\right\|_{1/2}^{1/2}\right). \tag{11}$$

Without loss of generality, we assume that the data is normalized and the intercept term is 0. In formula (11), $\boldsymbol{y}$ is the $n\times1$ vector, $\boldsymbol{\beta}$ is the $q\times1$ vector, $\boldsymbol{X}$ is the $n\times q$ design matrix, $\left\|\boldsymbol{\beta}\right\|_{1/2}^{1/2}=\sum_{j=1}^{q}\left|\beta_j\right|^{1/2}$, and the tuning parameter $\lambda$ controls the degree of penalty. The larger the value of $\lambda$, the larger the shrinkage of the regression parameters.

By observing the form of the penalty term in (11), we find that the regression parameter $\beta_j$ in the L$_{1/2}$ regularization has the form of an independent and identical zero-mean generalized Gaussian prior. The density function expression of the zero-mean generalized Gaussian distribution is:

$$f\left(x\right)=\frac{p}{2\sigma\Gamma\left(1/p\right)}\exp\left(-\frac{\left|x\right|^p}{\sigma^p}\right),$$

where $\Gamma\left(\cdot\right)$ is the gamma function, $\sigma$ is the scale parameter, and $p>0$ is the shape parameter that controls the decay rate of the tail of the distribution. There are two special cases in GGD: when $p=1$, corresponding to the Laplace distribution, and when $p=2$, corresponding to the normal distribution.

Combining the above connections, on the basis of Park and Casella [15], we consider adding a generalized Gaussian prior to the regression parameters $\beta_j$ with mean of 0, shape parameter of $p=1/2$, and scale parameter of $\sqrt{\sigma^2}\lambda^{-2}$. The expression is as follows:

$$\pi\left(\boldsymbol{\beta}\middle|\sigma^2\right)=\prod_{j=1}^{q}\frac{\lambda^2}{2\sqrt{\sigma^2}\Gamma(2+1)}\exp\left\{-\lambda\left(\left|\beta_j\right|/\sqrt{\sigma^2}\right)^{1/2}\right\}. \tag{12}$$

Although most of the existing literatures express the generalized Gaussian distribution as a scale mixture of normal distributions, this representation is not suitable for the Bayesian bridge model of $L_q(0<q<1)$ penalty. Therefore, other representations need to be explored. In this paper, the generalized Gaussian distribution is expressed as a mixture of uniform distribution and gamma distribution, which is:

$$\begin{aligned}&\frac{\lambda^2}{2\sqrt{\sigma^2}\Gamma(2+1)}\exp\left\{-\lambda\left(|x|/\sqrt{\sigma^2}\right)^{1/2}\right\}\\&=\int_{-\sqrt{\sigma^2}u^2<x<\sqrt{\sigma^2}u^2}\frac{1}{2\sqrt{\sigma^2}u^2}\cdot\frac{\lambda^{2+1}}{\Gamma(2+1)}u^{(2+1)-1}\mathrm{e}^{-\lambda u}\mathrm{d}u.\end{aligned}$$

Then, without assuming the distribution form of the data, the empirical likelihood function is used to replace the parameter likelihood function, and the Bayesian hierarchical model can be expressed as:

$$\begin{aligned}&L_{EL}(\boldsymbol{\beta})\sim\exp\left\{-\sum_{i=1}^{n}\log\left[1+\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{x}_i\left(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right]\right\},\\&\boldsymbol{\beta}|\boldsymbol{u},\sigma^2\sim\prod_{j=1}^{q}\mathrm{Uniform}\left(-\sqrt{\sigma^2}u_j^2,\sqrt{\sigma^2}u_j^2\right),\\&\boldsymbol{u}|\lambda\sim\prod_{j=1}^{q}\mathrm{Gamma}(2+1,\lambda),\\&\sigma^2\sim\pi\left(\sigma^2\right)\mathrm{d}\sigma^2.\end{aligned} \tag{13}$$

In the above hierarchical model, we choose $\pi\left(\sigma^2\right)=IG(a,b)$. Assuming that the priors of different parameters are independent, then the joint posterior density can be expressed as:

$$\pi\left(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\lambda\middle|\boldsymbol{y},\boldsymbol{X}\right)\propto L_{EL}(\boldsymbol{\beta})\pi\left(\boldsymbol{\beta}\middle|\boldsymbol{u},\sigma^2\right)\pi(\boldsymbol{u}|\lambda)\pi(\lambda)\pi\left(\sigma^2\right)\mathrm{d}\sigma^2. \tag{14}$$

Given $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{u}$, $\lambda$ and $\sigma^2$, the full conditional distribution of $\boldsymbol{\beta}$ is:

$$\begin{aligned}&\pi\left(\boldsymbol{\beta}\middle|\boldsymbol{y},\boldsymbol{X},\boldsymbol{u},\lambda,\sigma^2\right)\propto L_{EL}(\boldsymbol{\beta})\pi\left(\boldsymbol{\beta}\middle|\boldsymbol{u},\sigma^2\right)\\&\qquad\propto\exp\left\{-\sum_{i=1}^{n}\log\left[1+\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{x}_i\left(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)\right]\right\}\prod_{j=1}^{q}I\left\{\left|\beta_j\right|<\sqrt{\sigma^2}u_j^2\right\}.\end{aligned} \tag{15}$$

From the expression of the full conditional distribution of $\boldsymbol{\beta}$, we know that its full conditional distribution has no closed form.

Similarly, given the conditions of $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$, $\lambda$ and $\sigma^2$, the full conditional distribution of $\boldsymbol{u}$ is:

$$\pi\left(\boldsymbol{u}\middle|\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta},\lambda,\sigma^2\right)\propto\pi\left(\boldsymbol{\beta}\middle|\boldsymbol{u},\sigma^2\right)\pi(\boldsymbol{u}|\lambda)\propto\prod_{j=1}^{q}\mathrm{e}^{-\lambda u_j}I\left\{u_j>\left(\frac{\left|\beta_j\right|}{\sqrt{\sigma^2}}\right)^{1/2}\right\}. \tag{16}$$

Analogously, given the conditions of $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$, $\boldsymbol{u}$ and $\lambda$, the full conditional distribution of $\sigma^2$ is:

$$\pi\left(\sigma^2\middle|\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{u},\lambda\right)\propto\pi\left(\boldsymbol{\beta}\middle|\boldsymbol{u},\sigma^2\right)\pi\left(\sigma^2\right)\mathrm{d}\sigma^2$$
$$\propto\left(\frac{1}{\sigma^2}\right)^{(q-1)/2+a+1}\exp\left(-\frac{b}{\sigma^2}\right)I\left\{\sigma^2>\max_j\frac{\beta_j^2}{u_j^4}\right\}. \tag{17}$$

In the expression of the prior distribution, we find that tuning parameter $\lambda$ is introduced into the model in the form of hyperparameters that play a role in controlling the accuracy of the prior distribution. The larger the value of $\lambda$, the more concentrated the prior distribution is at mean 0; the smaller the value of $\lambda$, the more scattered the prior distribution is at mean 0. In this paper, we specify a gamma prior Gamma(c, d) for the penalty parameter $\lambda$.

In model (13), when the latent variable $u_j$ is marginalized and the generalized Gaussian prior is directly used, the full conditional distribution of $\lambda$ given $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$, $\boldsymbol{u}$ and $\sigma^2$ is:

$$\pi\left(\lambda\middle|\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta},\boldsymbol{u},\sigma^2\right)\propto\lambda^{(c+2q)-1}\exp\left\{-\lambda\left(d+\sum_{j=1}^{q}\left|\beta_j\right|^{1/2}\right)\right\}. \tag{18}$$

### 4.2. The Framework of the Algorithm

Regarding $\boldsymbol{u}$, $\sigma^2$ and $\lambda$ in the Bayesian hierarchical model, this paper uses the Gibbs algorithm to sample.

1) The full conditional distribution of $u_j$ is the left-truncated exponential distribution $\exp(\lambda)I\left\{u_j>\left(\left|\beta_j\right|/\sqrt{\sigma^2}\right)^{1/2}\right\}$, and two-step sampling is considered. First generate $u_j^*$ from the exponential distribution $\exp(\lambda)$, and then let $u_j=u_j^*+\left(\left|\beta_j\right|/\sqrt{\sigma^2}\right)^{1/2}$.

2) The full conditional distribution of $\sigma^2$ is the left-truncated inverse gamma distribution, and two-step sampling is considered. First generate $\sigma^{2*}$ from the right-truncated gamma distribution $\mathrm{Gamma}\left((q-1)/2+a,b\right)I\left\{\sigma^{2*}<\max_j\left(1/\left(\beta_j^2/u_j^4\right)\right)\right\}$, then let $\sigma^2=1/\sigma^{2*}$.

3) The full conditional distribution of $\lambda$ is the gamma distribution, and $\lambda$ is generated directly from the gamma distribution $\mathrm{Gamma}\left(c+2q,d+\sum_{j=1}^{q}\left|\beta_j\right|^{1/2}\right)$.

Regarding the regression parameter $\boldsymbol{\beta}$, since its full conditional distribution has no closed form, this paper considers sampling using the tailored M-H algorithm adopted by Chib [18] and Bedoui [11]. Among them, the candidate generation density in the M-H algorithm is a multivariate t distribution, its location parameter is the mode of the logarithmic empirical likelihood function for the linear model, and the dispersion matrix is the inverse of the negative Hessian matrix of the logarithmic empirical likelihood function evaluated at this mode.

## 5. Simulation

In this section, simulation experiments are performed to verify the effectiveness of $L_{1/2}$ regularization based on Bayesian empirical likelihood (BEL). We generate data from the following multiple linear regression models:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y}$ is the $n\times 1$ response variable, $\boldsymbol{X}$ is a $n\times 8$ design matrix, $\boldsymbol{\varepsilon}$ is the $n\times 1$ error vector and $n$ is the sample size. The data for the design matrix $\boldsymbol{X}$ comes from a multivariate Gaussian distribution with a mean of zero and a covariance matrix of $\boldsymbol{\Sigma} = 0.2^{|i-j|}, i, j \in \{1,2,\cdots,8\}$. The regression coefficients $\boldsymbol{\beta} = (3,1.5,0,0,2,0,0,0)^{\mathrm{T}}$ are a $8\times 1$ regression vector.

In standard parametric models, it is generally assumed that the errors follow a normal distribution with zero mean. However, in empirical likelihood, there is no need to make assumptions about the error distribution, which can avoid making false assumptions about the error distribution and make the model more robust.

We assume the error violates the zero-mean normality assumption of the standard parametric model, $\varepsilon_i$ is independent and identically distributed from a normal distribution with mean −3 and variance $3^2$. Under this model, we generate training datasets with three different sample sizes ( $n$ = 50, 100, 200). And produce a test set of the same size. Furthermore, the Bayesian empirical likelihood-based $L_{1/2}$ regularization method (BEL) proposed in this paper is compared with the Bayesian bridge regression model for scale mixture of normal based on generalized Gaussian density (BBR.N) proposed by Polson [19], the Bayesian bridge regression model for scale mixtures of triangular based on generalized Gaussian density (BBR.T) proposed by Polson [19], and Bayesian lasso model (BLASSO) proposed by Park and Casella [15]. Among them, the exponent of the regularization term of BBR.N and BBR.T is selected as $\alpha = 0.5$, corresponds to the $L_{1/2}$ penalty using the parametric likelihood function.

For the hyperparameters in the hierarchical model, we choose $a$ = 10, $b$ = 0.1, $c$ = 2 and $d$ = 2 to conduct numerical simulations. And generate 50 sets of training data sets, that is, repeat the experiment 50 times, fit the model on the training data set, iterate 15,000 times for each experiment, discard the first 5000 times, and calculate the mean of the regression coefficients of the last 10,000 times as the estimated value. Then calculate its performance on the test dataset.

The evaluation indicators are the mean square error (MSE) and the mean absolute deviation (MAE) on the test set, and the calculation expressions are as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{19}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{20}$$

In order to exclude the influence of possible extreme values, this paper uses the median of these 50 data to evaluate the performance of the four methods, namely the median of mean square error (MMSE) and the median of mean absolute deviation (MMAE).

Table 1 shows the values of the median of mean squared error and the median of mean absolute error of the four methods at three different sample sizes. As can be seen from Table 1, when the error distribution violates the normality assumption of zero mean of standard parametric model, especially when the sample size is small ( $n$ = 50, 100), the BEL method outperforms the other three methods. And with the increase of sample size, the values of MMSE and MMAE of the four methods all showed a downward trend.

Figure 1 shows the boxplots of the values of MSE computed on the test set for the four evaluation methods at three different sample sizes. It can also be seen from the figure that when the sample size is small ( $n$ = 50, 100), the BEL method is significantly better than the other three methods. And when the sample size is 200, the BEL method is slightly better than the other three methods. In general, it can be seen that the BEL method performs better in small samples when the error violates the zero-mean normality assumption.

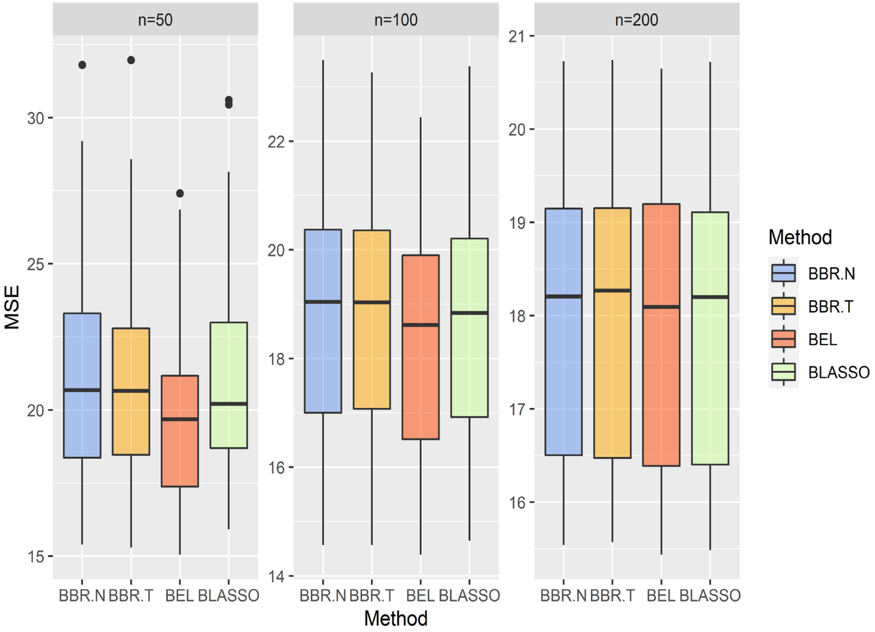

Figure 1. Boxplots of the values of MSE for the four methods.

Table 1. Values of MMSE and MMAE for the four methods.

| | | MMSE(MMAE) | | | |
|---|---|---|---|---|---|
| Error distribution | sample size | BEL | BBR.N | BBR.T | BLASSO |
| $N(-3,3^2)$ | $n = 50$ | **19.69**<br>(3.57) | 20.68<br>(3.68) | 20.66<br>(3.70) | 20.21<br>(3.73) |
| | $n = 100$ | **18.62**<br>(3.56) | 19.04<br>(3.58) | 19.04<br>(3.59) | 18.84<br>(3.58) |
| | $n = 200$ | **18.09**<br>(3.50) | 18.21<br>(3.51) | 18.27<br>(3.52) | 18.20<br>(3.51) |

Figure 2 shows the boxplots of 50 MAEs calculated on the test set by the four evaluation methods by repeating 50 experiments under each sample size of simulation experiment. When the number of observations is 50, 100, the MMAE of the BEL method significantly smaller than the other three methods. When the number of observations is 200, the MMAE of the BEL method is slightly smaller than the other three methods.

Table 2 shows the number of times each component of the regression coefficients is excluded using the scaled neighborhood criterion proposed by Li and Lin [16] on 50 training datasets with three different sample sizes. It can be seen from Table 2 that the four methods can better play the role of identifying important variables and unimportant variables, that is, they can play the role of variable selection. When the number of observations is 50 and 100, the BEL method can more accurately identify non-zero variables.

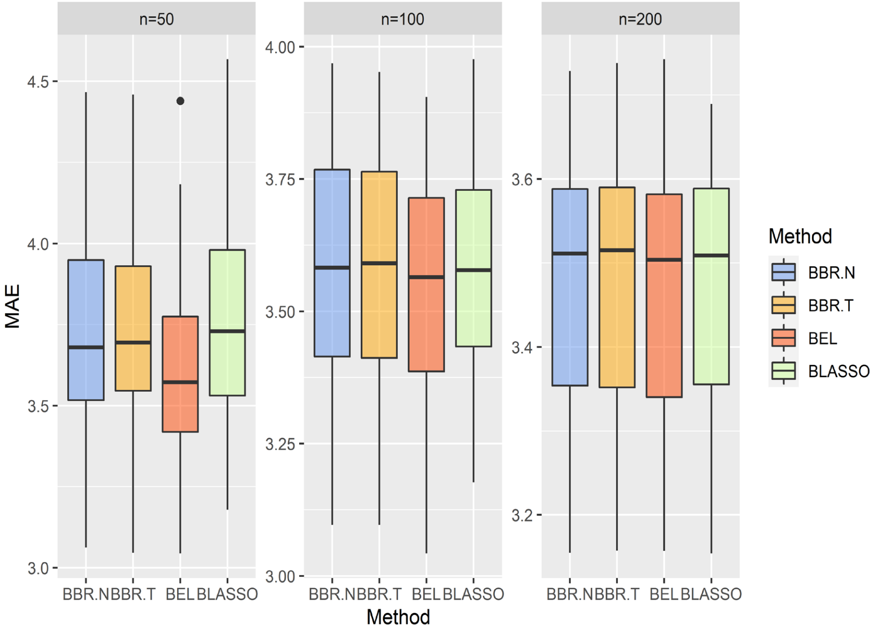

Figure 2. Boxplots of the values of MAE for the four methods.

Table 2. The number of times the regression component was removed based on 50 repetitions of the simulation.

| size | Method | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ | $\beta_5$ | $\beta_6$ | $\beta_7$ | $\beta_8$ |
|---|---|---|---|---|---|---|---|---|---|
| $n=50$ | BEL | 0 | 0 | 36 | 32 | 0 | 31 | 31 | 33 |
| | BBR.N | 0 | 4 | 38 | 35 | 0 | 37 | 37 | 36 |
| | BBR.T | 0 | 4 | 38 | 34 | 0 | 35 | 38 | 36 |
| | BLASSO | 0 | 3 | 40 | 35 | 0 | 36 | 45 | 41 |
| $n=100$ | BEL | 0 | 0 | 34 | 37 | 0 | 34 | 32 | 35 |
| | BBR.N | 0 | 2 | 40 | 38 | 0 | 37 | 33 | 38 |
| | BBR.T | 0 | 2 | 38 | 38 | 0 | 37 | 36 | 37 |
| | BLASSO | 0 | 1 | 43 | 41 | 0 | 40 | 38 | 39 |

Continued

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $n = 200$ | BEL | 0 | 0 | 34 | 34 | 0 | 32 | 38 | 33 |
| | BBR.N | 0 | 0 | 40 | 32 | 0 | 34 | 31 | 36 |
| | BBR.T | 0 | 0 | 39 | 34 | 0 | 32 | 30 | 37 |
| | BLASSO | 0 | 0 | 39 | 39 | 0 | 40 | 41 | 36 |

## 6. Conclusions

This paper proposes a new method for variable selection, which is $L_{1/2}$ regularization based on Bayesian empirical likelihood. This method introduces the $L_{1/2}$ penalty into the model in the form of generalized Gaussian prior. Replace the parametric likelihood function in Bayes theorem with a nonparametric likelihood function, and derive the posterior distribution through the Bayesian hierarchical model, then use MCMC method to sample from the posterior distribution. Simulations demonstrate that the proposed method BEL outperforms BBR.N, BBR.T and BLASSO when the errors violate the zero-mean normality assumption for standard parametric models. Especially when the sample size is small, the prediction accuracy of the BEL method is better. In addition, the proposed method can perform variable selection well.

Subsequent research may consider Bayesian empirical likelihood inference combining $L_{1/2}$ penalty and $L_2$ penalty, which is a flexible penalty method. Consider adding a spike-and-slab prior to the parameters, the expression is as follows:

$$\pi\left(\boldsymbol{\beta}\middle|\boldsymbol{\delta}\right)=\prod_{j=1}^{q}\left\{\left(1-\delta_j\right)\psi\left(\beta_j;\lambda_1,\sigma_1^2\right)+\delta_j\varphi\left(\beta_j;\lambda_2,\sigma_2^2\right)\right\}. \tag{21}$$

where

$$\psi\left(\beta_j;\lambda_1,\sigma_1^2\right)=\lambda_1^2\Big/\left[2\sqrt{\sigma_1^2}\,\Gamma\left(2+1\right)\right]\times\exp\left\{-\lambda_1\left(\left|\beta_j\right|\Big/\sqrt{\sigma_1^2}\right)^{1/2}\right\},$$

$$\varphi\left(\beta_j;\lambda_2,\sigma_2^2\right)=\sqrt{\lambda_2/2\pi\sigma_2^2}\times\exp\left\{-\lambda_2\beta_j^2/2\sigma_2^2\right\},$$

and $\delta_j\in\{0,1\}$. When $\delta_j=1$, it indicates that the *j*th predictor is more important and should be kept. When $\delta_j=0$, it indicates that the *j*th predictor is not important and should be removed from the model. Compared with applying a single prior distribution to the parameters, this mixed prior can well combine the advantages of variable selection and sparse recovery.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

[1] Owen, A.B. (1988) Empirical Likelihood Ratio Confidence Intervals for a Single Functional. *Biometrika*, **75**, 237-249. https://doi.org/10.1093/biomet/75.2.237

[2] Owen, A.B. (1990) Empirical Likelihood Ratio Confidence Regions. *The Annals of*

*Statistics*, **18**, 90-120. https://doi.org/10.1214/aos/1176347494

[3] Owen, A.B. (1991) Empirical Likelihood for Linear Models. *The Annals of Statistics*, **19**, 1725-1747. https://doi.org/10.1214/aos/1176348368

[4] Qin, J. and Lawless, J. (1994) Empirical Likelihood and General Estimating Equations. *The Annals of Statistics*, **22**, 300-325. https://doi.org/10.1214/aos/1176325370

[5] Kolaczyk, E.D. (1994) Empirical Likelihood for Generalized Linear Models. *Statistica Sinica*, **4**, 199-218.

[6] Nadarajah, T., Variyath, A. and Loredo-Osti, J. (2020) Empirical Likelihood Based Longitudinal Data Analysis. *Open Journal of Statistics*, **10**, 611-639. https://doi.org/10.4236/ojs.2020.104037

[7] Huang, T., Fan, Y. and Sun, Z. (2019) Robust Element-Wise Empirical Likelihood Estimation Method for Longitudinal Data. *Journal of Applied Mathematics and Physics*, **7**, 1408-1420. https://doi.org/10.4236/jamp.2019.76094

[8] Lazar, N.A. (2003) Bayesian Empirical Likelihood. *Biometrika*, **90**, 319-326. https://doi.org/10.1093/biomet/90.2.319

[9] Zhong, X. and Ghosh, M. (2016) Higher-Order Properties of Bayesian Empirical Likelihood. *Electronic Journal of Statistics*, **10**, 3011-3044. https://doi.org/10.1214/16-EJS1201

[10] Li, C.J., Zhao, H.M. and Dong, X.G. (2019) Bayesian Empirical Likelihood and Variable Selection for Censored Linear Model with Applications to Acute Myelogenous Leukemia Data. *International Journal of Biomathematics*, **12**, 799-813. https://doi.org/10.1142/S1793524519500505

[11] Bedoui, A. and Lazar, N.A. (2020) Bayesian Empirical Likelihood for Ridge and Lasso Regressions. *Computational Statistics & Data Analysis*, **145**, 106917. https://doi.org/10.1016/j.csda.2020.106917

[12] Moon, C. and Bedoui, A. (2020) Bayesian Elastic Net Based on Empirical Likelihood. arXiv: 2006.10258. https://doi.org/10.48550/arXiv.2006.10258

[13] Zhang, Y. and Tang, N. (2017) Bayesian Empirical Likelihood Estimation of Quantile Structural Equation Models. *Journal of Systems Science & Complexity*, **30**, 122-138. https://doi.org/10.1007/s11424-017-6254-x

[14] Yang, Y. and He, X. (2012) Bayesian Empirical Likelihood for Quantile Regression. *The Annals of Statistics*, **40**, 1102-1131. https://doi.org/10.1214/12-AOS1005

[15] Park, T. and Casella, G. (2008) The Bayesian Lasso. *Journal of the American Statistical Association*, **103**, 681-686. https://doi.org/10.1198/016214508000000337

[16] Li, Q. and Lin, N. (2010) The Bayesian Elastic Net. *Bayesian Analysis*, **5**, 151-170. https://doi.org/10.1214/10-BA506

[17] Mallick, H. and Yi, N. (2014) A New Bayesian Lasso. *Statistics and Its Interface*, **7**, 571-582. https://doi.org/10.4310/SII.2014.v7.n4.a12

[18] Chib, S., Shin, M. and Simoni, A. (2018) Bayesian Estimation and Comparison of Moment Condition Models. *Journal of the American Statistical Association*, **113**, 1656-1668. https://doi.org/10.1080/01621459.2017.1358172

[19] Polson, N.G., Scott, J.G. and Windle, J. (2014) The Bayesian Bridge. *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*), **76**, 713-733. https://doi.org/10.1111/rssb.12042