Austrian Journal of Statistics
July 2023, Volume 52, 15–45.
http://www.ajs.or.at/
doi:10.17713/ajs.v52i4.1455

A Generalization of LASSO Modeling via Bayesian Interpretation

Gayan Warahena-Liyanage
University of Dayton

Felix Famoye
Central Michigan University

Carl Lee
Central Michigan University

Abstract

The aim of this paper is to introduce a generalized LASSO regression model that is derived using a generalized Laplace (GL) distribution. Five different GL distributions are obtained through the T-$R\{Y\}$ framework with quantile functions of standard uniform, Weibull, log-logistic, logistic, and extreme value distributions. The properties, including quantile function, mode, and Shannon entropy of these GL distributions are derived. A particular case of GL distributions called the beta-Laplace distribution is explored. Some additional components to the constraint in the ordinary LASSO regression model are obtained through the Bayesian interpretation of LASSO with beta-Laplace priors. The geometric interpretations of these additional components are presented. The effects of the parameters from beta-Laplace distribution in the generalized LASSO regression model are also discussed. Two real data sets are analyzed to illustrate the flexibility and usefulness of the generalized LASSO regression model in the process of variable selection with better prediction performance. Consequently, this research study demonstrates that more flexible statistical distributions can be used to enhance LASSO in terms of flexibility in variable selection and shrinkage with better prediction.

Keywords: LASSO regression, beta-Laplace distribution, T-Laplace family, variable selection, prediction.

1. Introduction

Developments in sophisticated data collection techniques have significantly increased the number of potential predictor variables in almost every area of science, entertainment, business, and industry. In the field of science, some applications with many predictive variables include genomics, biomedical imaging, and tumor classifications. Studying customer ratings to recommend or sell new movies and books, analyzing social network profiles to improve the online experience, analyzing sports statistics to help team managers and players to make better decisions are applications with a large number of predictors in the fields of entertainment and business.

With the current COVID-19 pandemic that we are in, it is important to develop more big data

analytics tools to understand pandemic data. Given that more data are freely available for COVID-19, it is imperative to develop big data analytics tools such as the flexible LASSO in this study to better understand virus transmission, risk factors, origins, diagnostics, and other vital data. Table 1 presents some of the significant applications of big data in the COVID-19 pandemic (Haleem, Javaid, Khan, and Vaishya 2020). Applied to reality, the proposed study advances the data analysis techniques in machine learning and data mining.

Table 1: Some of the significant applications of big data in the COVID-19 pandemic

Area of the Applications	Description
(a) Infected cases	Identification of infected cases from the massive amount of data based on medical histories of all patients
(b) Travel history	To analyze the risk and identify people who may have been in contact with the infected patients based on travel history
(c) Symptoms	Identification of suspicious cases based on most significant symptoms
(d) Disease detection	Identification of infected patients at an early stage using the most significant factors
(e) Medical treatments	Rapid development of new medicines and medical equipment that are needed for current and future medical needs via critical data available

When dealing with statistical modeling problems with many predictor variables, our goal is to find a simple model that also has a good predictive ability. The statistical models with fewer predictors are easy to interpret and often lead to a better understanding of the underlying process generating the data. In 1976, British statistician George Box wrote a famous line, "All models are wrong, some are useful." As such, in the model selection process, the approximate best model for the data is selected based on the prediction performance among different models. The traditional model selection methods such as forward and stepwise methods often select too many predictors when size of data is very large due to the standard errors of parameter estimates become very small. As a result, the obtained model based on the data used to train the model has small mean squared error (MSE). However, when it is applied to an independent data, the MSE often is not optimal due to large variance component. Different solutions have been proposed. In general, the resulting 'optimal' model often has higher MSE than the ordinary Least Square model due to the trade off between bias and variance. One such approach is to shrink the values of the regression coefficients smaller or to zero. This is the well known shrinkage or regularization methods.

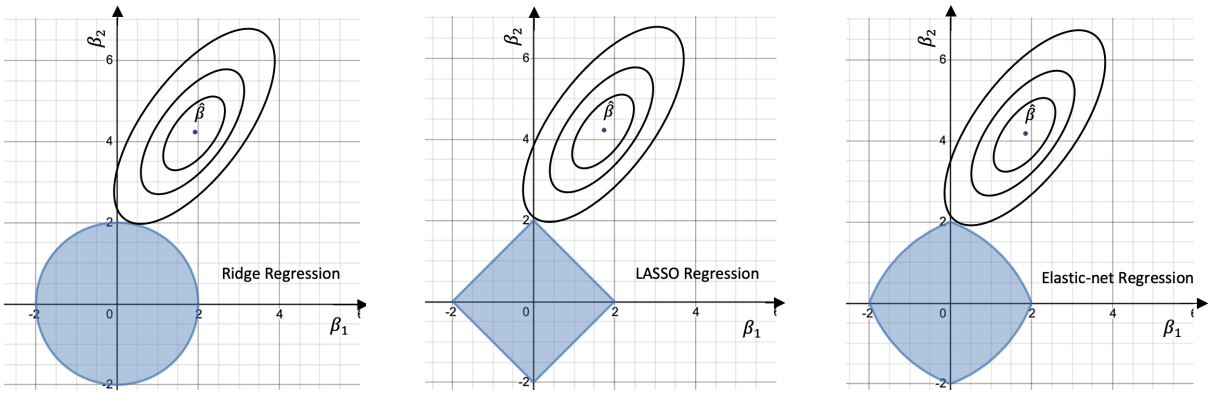
In the last few decades, many shrinkage methods have been proposed (Hastie, Tibshirani, and Wainwright 2015). Some of the methods are ridge regression method (Hoerl and Kennard 1970), Least Absolute Selection and Shrinkage Operator (LASSO) method by Tibshirani (1996), and elastic-net regression method by Zou and Hastie (2005).

Consider the usual linear regression model: given p predictors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p$, our aim is to predict the response Y using a linear model

$$\hat{Y} = \hat{\beta}_0 + \mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_2\hat{\beta}_2 + \cdots + \mathbf{X}_p\hat{\beta}_p. \tag{1}$$

Given a data set of n observations, the vector of estimators $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ is obtained through a model-fitting procedure. In the ordinary least squares (OLS) method, the estimates are obtained by minimizing sum of squared errors (SSE). The ridge regression method obtains

the estimates by minimizing SSE subject to a bound on L_2 norm of the regression coefficients. This constraint leads the regression coefficients to shrink toward zero. This helps to reduce model complexity and multicollinearity. On the other hand, the LASSO method obtains the estimates by minimizing SSE subject to a bound on L_1 norm of the regression coefficients. The L_1 penalty in LASSO can set some regression coefficients to zero. Thus, the LASSO reduces model complexity and multicollinearity and can be used in variable selection. The elastic-net method is a hybrid method that combines the regularization in both LASSO and ridge. Figures 1a, 1b, and 1c represent the contours of the error function and constraint regions of ridge, LASSO, and elastic-net methods, respectively.



(a) (b) (c)

Note: $\hat{\beta}$ is the solution from the OLS method.

Figure 1: Contours of the error function and constraint regions

In this paper, we specifically focus on the LASSO regression method. Even though the LASSO is a promising method on many occasions, it has some drawbacks.

- Due to the nature of the convex optimization problem that LASSO tries to minimize, when $p > n$, LASSO tends to select at most n predictors. This limits the use of LASSO in variable selection.
- When multicollinearity exists in a group of predictors, LASSO tends to pick only one predictor from the group without caring which one to select and ignores others.
- The obtained LASSO model may be over-simplified and results in large bias component due to many parameter estimates are set to zero.

LASSO method is a convex optimization problem, and many optimization methods have been developed for solving LASSO. On the other hand, there is also a Bayesian view of the LASSO estimators. Based on the form of the penalty term in LASSO, the LASSO estimates can be interpreted as posterior mode estimates when regression parameters have independent and identical Laplace priors. Motivated by this, in this paper we introduce a family of generalized Laplace priors that could be useful in defining a generalized LASSO regression model by first developing a family of generalized Laplace distributions.

The paper is organized as follows: In Section 2, we introduce the T-Laplace family of distributions using the T-$R\{Y\}$ framework. We also present some general properties of the proposed generalizations of Laplace distribution. Section 3 explores the beta-Laplace distribution and some of its properties. In Section 4, we derive a generalized LASSO regression model by introducing additional components to the constraint in the ordinary LASSO regression model through the Bayesian interpretation of LASSO with beta-Laplace priors. The geometric interpretations of these additional components and the effects of the parameters from beta-Laplace

distribution in the generalized LASSO regression model are also investigated. Section 5 presents a numerical study and two real data examples to demonstrate the flexibility and usefulness of the generalized LASSO regression model in the process of variable selection with better prediction performance. Finally, Section 6 contains the conclusions and suggestions for further study.

2. T-Laplace family of distributions

Eugene, Lee, and Famoye (2002) introduced the beta-generated family of distributions with CDF $G(x) = \int_0^{F(x)} b(t)dt$ where $b(t)$ is PDF of a beta random variable and $F(x)$ is the CDF of any random variable. Motivated by this idea, Alzaatreh, Lee, and Famoye (2013) introduced T-$X(W)$ family of distributions with the CDF, $G(x) = \int_a^{W(F(x))} r(t)dt$ where $r(t)$ is the PDF of any random variable $T \in [a, b]$, $-\infty \leq a < b \leq \infty$, and $W(F(x))$ is a monotonic and absolutely continuous function. Following that, Aljarrah, Lee, and Famoye (2014) defined T-$X\{Y\}$ by taking $W(F(x))$ to be $Q_Y(F(x))$, the quantile function of any random variable Y. Later, Alzaatreh, Lee, and Famoye (2014) renamed the T-$X\{Y\}$ family as T-$R\{Y\}$ framework by defining a unified notation, which will be used in this article.

Let T, R, and Y be random variables with CDFs $F_T(x) = P(T \leq x)$, $F_R(x) = P(R \leq x)$, and $F_Y(x) = P(Y \leq x)$, respectively. Let the PDFs be denoted by $f_T(x)$, $f_R(x)$, and $f_Y(x)$, respectively. The corresponding quantile functions are $Q_T(p)$, $Q_R(p)$, and $Q_Y(p)$, where the quantile function is defined as $Q_Z(p) = inf\{z : F_Z(z) \geq p\}$, $0 < p < 1$. Assume the random variables $Y \in [c,\ d]$ and $T \in [a,\ b] \subset [c,\ d]$, for $-\infty \leq a < b \leq \infty$ and $\infty \leq c < d \leq \infty$. Using the T-$R\{Y\}$ framework, the CDF and PDF of the random variable X are respectively defined as

$$F_X(x) = \int_a^{Q_Y(F_R(x))} f_T(t)dt = F_T(Q_Y(F_R(x))), \tag{2}$$

$$f_X(x) = f_R(x) \times \frac{f_T(Q_Y(F_R(x)))}{f_Y(Q_Y(F_R(x)))}. \tag{3}$$

One can use $X = Q_R(F_Y(T))$ to generate the random variable X. Since the support of T-$R\{Y\}$ is the same as the support of R, given a random variable R, T-$R\{Y\}$ gives the generalized R distribution for any non-uniform T and Y. So, the T-$R\{Y\}$ framework can be used to generate different families of generalized R distribution.

From equations (2) and (3), the cumulative hazard function and the hazard function of the random variable X can be defined as

$$H_X(x) = -\log(1 - F_T(Q_Y(F_R(x)))), \tag{4}$$

$$h_X(x) = h_R(x) \times \frac{h_T(Q_Y(F_R(x)))}{h_Y(Q_Y(F_R(x)))}. \tag{5}$$

In probability theory and statistics, the Laplace distribution is a continuous probability distribution named after marquis Pierre-Simon Laplace (1749-1824), a French scholar and a polymath. Sometimes the Laplace distribution is also known as the double exponential distribution due to its particular shape. Compared to the Gaussian distribution, the Laplace distribution is also a symmetric distribution, but with moderate tails and a discontinuous first derivative at the mean. The Laplace distribution has various applications in biology, computer science, social studies, physics, finance, and economics. It has also been commonly used over Gaussian distribution in robustness studies.

Let R be a centered Laplace random variable with the scale parameter $\sigma > 0$. The corresponding CDF and PDF of R are defined as

$$F_R(x) = \frac{1}{2} + \frac{1}{2}\,\mathrm{sgn}(x)\left\{1 - e^{-|x|/\sigma}\right\} = \begin{cases} \frac{1}{2}e^{x/\sigma} & \text{if } x < 0, \\ 1 - \frac{1}{2}e^{-x/\sigma} & \text{if } x \geq 0, \end{cases} \tag{6}$$

$$f_R(x) = \frac{1}{2\sigma} e^{-|x|/\sigma} = \frac{1}{2\sigma} \begin{cases} e^{x/\sigma} & \text{if } x < 0, \\ e^{-x/\sigma} & \text{if } x \geq 0, \end{cases} \tag{7}$$

The cumulative hazard function and the hazard function of R are defined as

$$H_R(x) = -\log\left[\frac{1}{2} - \frac{1}{2}\operatorname{sgn}(x)\left\{1 - e^{-x/\sigma}\right\}\right], \tag{8}$$

$$h_R(x) = \begin{cases} \left[\sigma\left\{2\, e^{-x/\sigma} - 1\right\}\right]^{-1} & \text{if } x < 0, \\ \sigma^{-1} & \text{if } x \geq 0. \end{cases} \tag{9}$$

If the random variable R follows the Laplace distribution with CDF, PDF, cumulative hazard function, and hazard function as defined in equations (6), (7), (8), and (9), then equations (2), (3), (4), and (5) give the CDF, PDF, cumulative hazard function, and hazard function of a random variable that follows the T-Laplace$\{Y\}$ distribution.

Applying different random variables T and Y, T-Laplace$\{Y\}$ generates families of generalized Laplace (GL) distributions, and the domain of the resulting distribution is the same as that of Laplace distribution. In Table 2, five different choices of random variable Y and corresponding quantile functions and the domain of T that can be used along with Y are listed.

Table 2: Quantile functions of different choices of Y and domains of T

Random variable Y	The quantile function $Q_Y(p)$	Domain of T
(a) Standard uniform	p	$(0,\ 1)$
(b) Weibull	$\lambda(-\log(1-p))^{1/k},\ \lambda,\ k > 0$	$(0,\ \infty)$
(c) Log-logistic	$\alpha[p/(1-p)]^{1/\beta},\ \alpha,\ \beta > 0$	$(0,\ \infty)$
(d) Logistic	$\mu + \lambda\ \log[p/(1-p)],\ \lambda > 0$	$(-\infty,\ \infty)$
(e) Extreme value	$\alpha + \beta \log[-\log(1-p)]$	$(-\infty,\ \infty)$

In the following, we define five families of GL distributions, T-Laplace{uniform}, T-Laplace{Weibull}, T-Laplace{log-logistic}, T-Laplace{logistic}, and T-Laplace{extreme value} using the quantile functions of standard uniform, Weibull, log-logistic, logistic, and extreme value distributions respectively as listed in Table 2 where $F_R(x)$, $f_R(x)$, $H_R(x)$, and $h_R(x)$ are respectively defined as in equations (6), (7), (8), and (9), and $S_R(x) = 1 - F_R(x)$ is the survival function of the random variable R.

i. T-Laplace{uniform} family: The CDF, PDF, and hazard function of T-Laplace{uniform} are respectively given by

$$F_X(x) = F_T(F_R(x)), \tag{10}$$

$$f_X(x) = f_R(x) f_T(F_R(x)), \tag{11}$$

$$h_X(x) = f_R(x)\, h_T(F_R(x)), \tag{12}$$

where $F_T(x)$, $f_T(x)$, and $h_T(x)$ are the CDF, PDF, and hazard function of the random variable T.

ii. T-Laplace{Weibull} family: The CDF, PDF, and hazard function of T-Laplace{Weibull} are respectively given by

$$F_X(x) = F_T\left(\lambda\left[H_R(x)\right]^{1/k}\right), \tag{13}$$

$$f_X(x) = \frac{\lambda f_R(x)\, f_T\left(\lambda\left[H_R(x)\right]^{1/k}\right)\left[H_R(x)\right]^{(1-k)/k}}{k S_R(x)}, \tag{14}$$

$$h_X(x) = (\lambda/k) h_R(x)\, h_T\left(\lambda\left[H_R(x)\right]^{1/k}\right)\left[H_R(x)\right]^{(1-k)/k}.$$

Note that when $\lambda = k = 1$, we get the CDF, PDF, and the hazard function of the T-Laplace{exponential} family.

iii. T-Laplace{log-logistic} family: The CDF, PDF, and hazard function of T-Laplace{log-logistic} are respectively given by

$$F_X(x) = F_T\left(\alpha\left[F_R(x)/S_R(x)\right]^{1/\beta}\right), \tag{15}$$

$$f_X(x) = \frac{\alpha f_R(x)\, f_T\left(\alpha\left[F_R(x)/S_R(x)\right]^{1/\beta}\right)\left[F_R(x)\right]^{(1-\beta)/\beta}}{\beta\left[S_R(x)\right]^{1/\beta}}, \tag{16}$$

$$h_X(x) = \frac{\alpha h_R(x)\, h_T\left(\alpha\left[F_R(x)/S_R(x)\right]^{1/\beta}\right)\left[F_R(x)\right]^{(1-\beta)/\beta}}{\beta\left[S_R(x)\right]^{1/\beta}}.$$

iv. T-Laplace{logistic} family: Using the quantile function (d) in Table 2 with $\mu = 0$, the CDF, PDF, and hazard function of T-Laplace{logistic} are respectively given by

$$F_X(x) = F_T\left(\lambda \log\left(F_R(x)/S_R(x)\right)\right), \tag{17}$$

$$f_X(x) = \frac{\lambda f_R(x)\, f_T\left(\lambda \log\left(F_R(x)/S_R(x)\right)\right)}{F_R(x) S_R(x)}, \tag{18}$$

$$h_X(x) = \frac{\lambda h_R(x)\, h_T\left(\lambda \log\left(F_R(x)/S_R(x)\right)\right)}{F_R(x)}.$$

v. T-Laplace{extreme value} family: Using the quantile function (e) in Table 2 with $\alpha = 0$, the CDF, PDF, and hazard function of T-Laplace{extreme value} are respectively given by

$$F_X(x) = F_T\left(\beta \log\left(H_R(x)\right)\right), \tag{19}$$

$$f_X(x) = \frac{f_R(x)\, f_T\left(\beta \log\left(H_R(x)\right)\right)}{-\log\left(S_R(x)\right) S_R(x)}, \tag{20}$$

$$h_X(x) = \frac{\beta h_R(x)\, h_T\left(\beta \log\left(H_R(x)\right)\right)}{H_R(x)}.$$

2.1. Some properties of the T-Laplace family of distributions

In this section, some of the general properties of the T-Laplace family will be presented.

Lemma 2.1 (**Transformations**). *Given any random variable T with the PDF $f_T(x)$,*

i. the random variable $X = -\sigma \operatorname{sgn}(T - 0.5)\log(1 - 2|T - 0.5|)$ follows the distribution of T-Laplace{uniform} family in Equation (10). An equivalent random variable can be written as

$$X = \begin{cases} \sigma \log(2T) & \text{if } 0 < T < 0.5, \\ -\sigma \log(2 - 2T) & \text{if } 0.5 \le T < 1, \end{cases}$$

ii. the random variable $X = -\sigma \operatorname{sgn}(0.5 - e^{-(T/\lambda)^k})\log(1 - |1 - 2e^{-(T/\lambda)^k}|)$ follows the distribution of T-Laplace{Weibull} family in Equation (13). An equivalent random variable can be written as

$$X = \begin{cases} \sigma[(T/\lambda)^k - \log(2)] & \text{if } 0 < T \le \lambda[-\log(0.5)]^{1/k}, \\ \sigma \log(2 - 2e^{(T/\lambda)^k}) & \text{if } \lambda\left[-\log(0.5)\right]^{1/k} < T < \infty, \end{cases}$$

iii. the random variable $X = -\sigma \operatorname{sgn}\left(0.5(T-1)/(T+1)\right)\log(1-|(T-1)/(T+1)|)$ *follows the distribution of* T*-Laplace*{*log-logistic*} *family in Equation* (15)*. An equivalent random variable can be written as*

$$X = \begin{cases} \sigma \log\left(T/(T+1)\right) & \text{if } 0 < T < 1, \\ -\sigma \log\left(2/(T+1)\right) & \text{if } 1 \leq T < \infty, \end{cases}$$

iv. the random variable $X = -\sigma \operatorname{sgn}\left(0.5(e^T-1)/(e^T+1)\right)\log(1-\left|(e^T-1)/(e^T+1)\right|)$ *follows the distribution of* T*-Laplace*{*logistic*} *family in Equation* (17)*. An equivalent random variable can be written as*

$$X = \begin{cases} \sigma \log\left(2e^T/(e^T+1)\right) & \text{if } -\infty < T < 0, \\ -\sigma \log(2/(e^T+1)) & \text{if } 0 \leq T < \infty, \end{cases}$$

v. the random variable $X = -\sigma \operatorname{sgn}(0.5 - e^{-e^T})\log(1-|2e^{-e^T}-1|)$ *follows the distribution of* T*-Laplace*{*extreme value*} *family in Equation* (19)*. An equivalent random variable can be written as*

$$X = \begin{cases} \sigma \log\left(2 - 2e^{-(e)^T}\right) & \text{if } -\infty < T < \log\left[-\log\left(0.5\right)\right], \\ \sigma(e^T - \log(2)) & \text{if } \log(-\log\left(0.5\right)) \leq T < \infty. \end{cases}$$

Proof. The results follow immediately from the fact that $X = Q_R\left(F_Y\left(T\right)\right)$ can be used to generate the random variable X using the random variable T, where $Q_R\left(p\right) = -\sigma \operatorname{sgn}\left(p-0.5\right)\log\left(1-|2p-1|\right)$, $0 < p < 1$ is the quantile function of the Laplace distribution. So, we can generate a random variable X that follows T-Laplace{uniform} distribution by first simulating the random variable T and then computing $X = -\sigma \operatorname{sgn}\left(T-0.5\right)\log(1-|2T-1|)$. □

Thus, in general, one can compute $E\left(X^r\right)$, the r^{th} non-central moment of the random variable X, by using $E\left(X^r\right) = E([Q_R\left(F_Y\left(T\right)\right)]^r)$. As an example, $E\left(X\right)$ of T-Laplace{uniform} can be computed as

$$\begin{aligned} E\left(X\right) &= E(-\sigma \operatorname{sgn}\left(T-0.5\right)\log\left(1-|2T-1|\right)) \\ &= \begin{cases} \sigma E(\log(2T)) & \text{if } 0 < T < 0.5, \\ -\sigma E(\log(2-2T)) & \text{if } 0.5 \leq T < 1. \end{cases} \\ &= \sigma \int_0^{0.5} \log\left(2T\right) f_T\left(t\right) dt - \sigma \int_{0.5}^{1} \log\left(2-2T\right) f_T\left(t\right) dt, \end{aligned}$$

where $f_T\left(t\right)$ is the PDF of the random variable T.

Lemma 2.2 (**Quantiles**)**.** *Let* $Q_X\left(p\right)$, $0 < p < 1$ *denotes the quantile function of the random variable* X*. Then the quantile functions for the i.* T*-Laplace*{*uniform*}*, ii.* T*-Laplace*{*Weibull*}*, iii.* T*-Laplace*{*log-logistic*}*, iv.* T*-Laplace*{*logistic*}*, and v.* T*-Laplace*{*extreme value*} *distributions are respectively given by*

i. $Q_X\left(p\right) = -\sigma \operatorname{sgn}\left(Q_T\left(p\right)-0.5\right)\log\left(1-|2Q_T\left(p\right)-1|\right)$,

ii. $Q_X\left(p\right) = -\sigma \operatorname{sgn}(0.5 - e^{-(Q_T(p)/\lambda)^k})\log(1-|2e^{(Q_T(p)/\lambda)^k}-1|)$,

iii. $Q_X\left(p\right) = -\sigma \operatorname{sgn}\left(0.5(Q_T\left(p\right)-1)/(Q_T\left(p\right)+1)\right)\log(1-|(Q_T\left(p\right)-1)/(Q_T\left(p\right)+1)|)$,

iv. $Q_X\left(p\right) = -\sigma \operatorname{sgn}\left(0.5(e^{Q_T(p)}-1)/(e^{Q_T(p)}+1)\right)\log(1-\left|(e^{Q_T(p)}-1)/(e^{Q_T(p)}+1)\right|)$,

v. $Q_X(p) = -\sigma \operatorname{sgn}(0.5 - e^{-e^{Q_T(p)}}) \log(1 - |2e^{-e^{Q_T(p)}} - 1|)$,

where $Q_T(p)$, $0 < p < 1$ *is the quantile function of the random variable* T.

Proof. Each result can be shown by solving $F_X(Q_X(p)) = p$ for $Q_X(p)$ where $F_X(\cdot)$ is the CDF defined in equations (10), (13), (15), (17), and (19), respectively. □

Theorem 2.1. *The mode(s) of the* T*-Laplace*$\{Y\}$ *family are the solutions of the equations*

$$x = \sigma \operatorname{sgn}(x) \log\left(\operatorname{sgn}(x)\left(\frac{Q_Y''(F_R(x))}{2Q_Y'(F_R(x))} + \frac{f_T'(Q_Y(F_R(x)))}{2f_T(Q_Y(F_R(x)))} Q_Y'(F_R(x))\right)\right).$$

Proof. First using the fact that $Q_Y(F_Y(x)) = x$, it follows that $Q_Y'(F_R(x)) = 1/f_Y(Q_Y(F_R(x)))$ so that Equation (3) can be written as $f_X(x) = f_R(x) f_T(Q_Y(F_R(x))) Q_Y'(F_R(x))$. The result in Theorem (2.1) can be shown by setting the first derivative of equation $f_X(x) = f_R(x) f_T(Q_Y(F_R(x))) Q_Y'(F_R(x))$ to zero. □

Corollary 2.1. *The mode(s) of the i.* T*-Laplace*$\{$*uniform*$\}$*, ii.* T*-Laplace*$\{$*Weibull*$\}$*, iii.* T*-Laplace*$\{$*log-logistic*$\}$*, iv.* T*-Laplace*$\{$*logistic*$\}$*, and v.* T*-Laplace*$\{$*extreme value*$\}$ *distributions, respectively, are the solutions of the equations*

i. $x = \sigma \operatorname{sgn}(x) \log\left(\operatorname{sgn}(x)\left(\frac{f_T'(F_R(x))}{2f_T(F_R(x))}\right)\right)$,

ii. $x = \sigma \operatorname{sgn}(x) \log\left(\frac{\operatorname{sgn}(x)}{H_R(x)S_R(x)}\left(-H_R(x) + k - 1 + \frac{\lambda\{H_R(x)\}^{1/k} f_T'\left(\lambda\{H_R(x)\}^{1/k}\right)}{k f_T\left(\lambda\{H_R(x)\}^{1/k}\right)}\right)\right)$,

iii. $x = \sigma \operatorname{sgn}(x) \log\left(\frac{\operatorname{sgn}(x)}{2\beta F_R(x)S_R(x)}\left(2\beta F_R(x) - \beta + 1 + \frac{\alpha\{h_R(x)\}^{1/\beta} f_T'\left(\alpha\{h_R(x)\}^{1/\beta}\right)}{f_T\left(\alpha\{h_R(x)\}^{1/\beta}\right)}\right)\right)$,

iv. $x = \sigma \operatorname{sgn}(x) \log\left(\frac{\operatorname{sgn}(x)}{2F_R(x)S_R(x)}\left(2F_R(x) - 1 + \frac{2\lambda f_T'(\lambda \log(h_R(x)))}{f_T(\lambda \log(h_R(x)))}\right)\right)$,

v. $x = \sigma \operatorname{sgn}(x) \log\left(\frac{\operatorname{sgn}(x)}{2H_R(x)S_R(x)}\left(2H_R(x) - 1 + \frac{\beta f_T'(\beta \log(H_R(x)))}{f_T(\beta \log(H_R(x)))}\right)\right)$.

Proof. First, we derive formulas for $\frac{Q_Y''(p)}{Q_Y'(p)}$, the ratio between the first and the second derivatives of random variable Y from uniform, Weibull, log-logistic, logistic, and extreme value distributions. Then, the results in equations i-v respectively can be obtained by applying each ratio in Theorem (2.1). □

Theorem 2.2. *The Shannon entropies for the* T*-Laplace*$\{Y\}$ *family are given by*

$$\eta_X = \eta_T + E(\log(f_Y(T))) + \log(2\sigma) + E(|X|)/\sigma, \tag{21}$$

where η_T *is the Shannon entropy of random variable* T.

Proof. First, applying Equation (3) in the definition of the Shannon entropy, $\eta_X = E(-\log[f_X(x)])$, we get $\eta_X = \eta_T + E(\log(f_Y(T))) - E(\log(f_R(X)))$. Then, applying Equation (7), the PDF of centered Laplace random variable R, we get the desired result in Equation (21). □

Corollary 2.2. *The Shannon entropies of the i.* T*-Laplace*$\{$*uniform*$\}$*, ii.* T*-Laplace*$\{$*Weibull*$\}$*, iii.* T*-Laplace*$\{$*log-logistic*$\}$*, iv.* T*-Laplace*$\{$*logistic*$\}$*, and v.* T*-Laplace*$\{$*extreme value*$\}$ *distributions, respectively, are given by*

i. $\eta_X = \eta_T + \log(2\sigma) + E(|X|)/\sigma$,

ii. $\eta_X = \eta_T + \log\left(2\sigma k/\lambda^k\right) + (k-1)E(\log(T)) - E\left(T^k\right)/\lambda^k + E(|X|)/\sigma$,

iii. $\eta_X = \eta_T + \log\left(2\sigma\beta/\alpha^\beta\right) + (\beta-1)E(\log(T)) - 2E\left(\log\left[1+(T/\alpha)^\beta\right]\right) + E(|X|)/\sigma$,

iv. $\eta_X = \eta_T + \log(2\sigma/\lambda) - \mu_T/\lambda - 2E\left(\log\left(1+e^{-T/\lambda}\right)\right) + E(|X|)/\sigma$,

v. $\eta_X = \eta_T + \log(2\sigma/\beta) + \mu_T/\beta - E\left(e^{T/\beta}\right) + E(|X|)/\sigma$,

where μ_T and η_T are the mean and the Shannon entropy for the random variable T.

Proof. The results in i-v can be shown by applying the PDFs, $f_Y(T) = 1$, $(k/\lambda)(T/\lambda)^{k-1}e^{-(T/\lambda)^k}$, $(\alpha/\beta)(T/\alpha)^{\beta-1}/\left(1+(T/\alpha)^\beta\right)^2$, $e^{-T/\lambda}/\lambda\left(1+e^{-T/\lambda}\right)^2$, $e^{T/\beta}e^{-e^{T/\beta}}$ of standard uniform, Weibull, log-logistic, logistic, and extreme value distributions in Equation (21). □

3. Beta-Laplace distribution

In recent years, several GL distributions have been studied through different generalization techniques. Some examples are beta-Laplace distribution (Cordeiro and Lemonte 2011) and Kumaraswamy-Laplace distribution (Aryal and Zhang 2016). In this section, we explore some properties of beta-Laplace distribution using the T-$R\{Y\}$ framework that are not available in the literature.

Let a random variable T follow the beta distribution with parameters a and b. Then the PDF of T is given by $f_T(x) = x^{a-1}(1-x)^{b-1}/B(a,b)$, $a,b>0$, where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the beta function and $\Gamma(\cdot)$ is the gamma function. The CDF of T is the regularized incomplete beta function, $F_T(x) = I_x(a,b) = B_x(a,b)/B(a,b)$, where $B_x(a,b) = \int_0^x u^{a-1}(1-u)^{b-1}du$. From Equation (10), the CDF of the beta-Laplace distribution is defined as

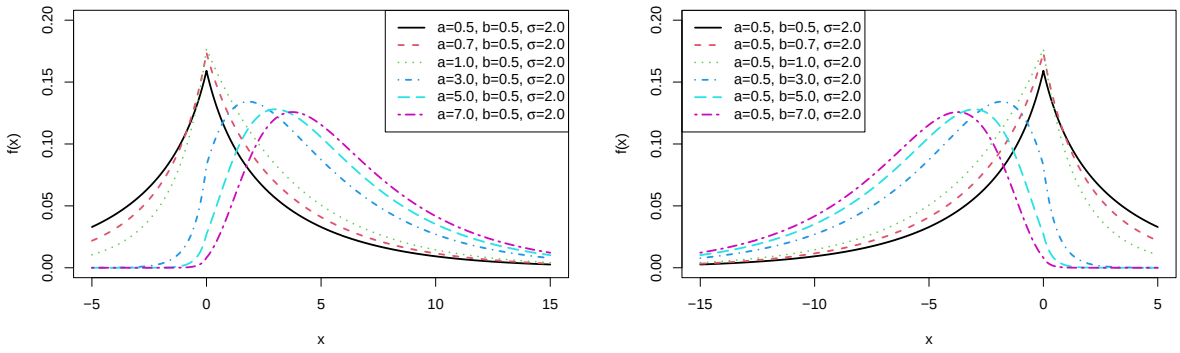
$$F_X(x) = I_{\frac{1}{2}+\frac{1}{2}\operatorname{sgn}(x)\left\{1-e^{-|x|/\sigma}\right\}}(a,\ b) = \begin{cases} I_{\frac{1}{2}e^{x/\sigma}}(a,b) & \text{if } x<0, \\ I_{1-\frac{1}{2}e^{-x/\sigma}}(a,b) & \text{if } x\geq 0. \end{cases} \tag{22}$$

By using Equation (11) , the PDF of the beta-Laplace distribution is given by

$$\begin{aligned} f_X(x) &= \frac{1}{2\sigma B(a,b)}e^{-|x|/\sigma}\left[\frac{1}{2}+\frac{1}{2}\ \operatorname{sgn}(x)\left\{1-e^{-|x|/\sigma}\right\}\right]^{a-1} \\ &\times \left[\frac{1}{2}-\frac{1}{2}\ \operatorname{sgn}(x)\left\{1-e^{-|x|/\sigma}\right\}\right]^{b-1} \\ &= \frac{1}{2^{a+b-1}\sigma B(a,b)}\begin{cases} e^{ax/\sigma}\left(2-e^{x/\sigma}\right)^{b-1} & \text{if } x<0 \\ e^{-bx/\sigma}\left(2-e^{-x/\sigma}\right)^{a-1} & \text{if } x\geq 0. \end{cases} \end{aligned} \tag{23}$$

When $a=b$, the plots are symmetric and when $a=b=1$, the PDF in Equation (23) reduces to the PDF of Laplace distribution in Equation (7). The plots for the PDF of the beta-Laplace distribution for several combinations of parameters σ, a, and b are given in Figure 2. According to the plots in Figure 2a, when $a>b$ the graphs are positively skewed, and the skewness increases as a increases. Also, as a increases, the mode increases when $a>1$.

According to the plots in Figure 2b, when $a<b$ the graphs are negatively skewed, and the skewness increases as b increases. Also, as b increases, the mode decreases when $b>1$. It is



(a) (b)

Figure 2: Plots of the pdf of beta-Laplace distribution

evident that the parameters a and b from beta distribution increase the flexibility in the Laplace distribution.

(Cordeiro and Lemonte 2011) discussed some of the properties of the beta-Laplace distribution using series expansions. Here, we present some additional properties of the beta-Laplace distribution using the general properties of T-Laplace family of distributions in Section 2.

1. **Quantile function**: By using Lemma (2.2), the quantile function of the beta-Laplace distribution can be derived as

$$Q_X(p) = -\sigma \operatorname{sgn}\left(I^{-1}{}_p(a,b) - \frac{1}{2}\right) \log\left(1 - 2\left|I^{-1}{}_p(a,b) - \frac{1}{2}\right|\right),$$

where $I^{-1}{}_p(a,b)$ is the inverse regularized beta function.

2. **Median**: The median of the beta-Laplace distribution is given by

$$M_X = Q_X(0.5) = -\sigma \operatorname{sgn}\left(I^{-1}{}_{0.5}(a,b) - \frac{1}{2}\right) \log\left(1 - 2\left|I^{-1}{}_{0.5}(a,b) - \frac{1}{2}\right|\right).$$

Since there is no general closed-form expression for the median of the beta distribution for arbitrary values of the parameters a and b, we do not have a closed-form expression for the median of the beta-Laplace distribution. Some closed-form expressions of the median for particular values of a and b are obtained in the following.

- Symmetric case: $a = b$
  When $a = b$, we have $I^{-1}{}_{0.5}(a,b) = 0.5$. So, the median, $M_X = Q_X(0.5) = 0$.
- When $a = 1$ and $b > 0$, we have $I^{-1}{}_{0.5}(1,b) = 1 - 2^{-1/b}$. Then,

$$M_X = -\frac{\sigma}{2} \operatorname{sgn}\left(1 - 2^{(b-1)/b}\right) \log\left(1 - \left|1 - 2^{(b-1)/b}\right|\right).$$

- When $a > 0$ and $b = 1$, we have $I^{-1}{}_{0.5}(a,1) = 2^{-1/a}$. Then,

$$M_X = -\frac{\sigma}{2} \operatorname{sgn}\left(2^{(a-1)/a} - 1\right) \log\left(1 - \left|2^{(a-1)/a} - 1\right|\right).$$

A reasonable approximation for the median of the beta distribution when $a, b \geq 1$, is given by (Kerman 2011), $I^{-1}{}_{0.5}(a,b) \approx (a - 1/3)/(a + b - 2/3)$. Using this approximation, an

approximated value for the median of the beta-Laplace distribution for arbitrary values of the parameters σ, a, and b can be found using the formula

$$M_X \approx -\sigma \operatorname{sgn}\left(\frac{a-b}{2(a+b-2/3)}\right) \log\left(1-\left|\frac{a-b}{a+b-2/3}\right|\right).$$

3. **Mode**: By using Theorem (2.1), the mode of the beta-Laplace distribution can be derived as

$$\text{Mode} = \begin{cases} -\sigma \log\left(\frac{a+b-1}{2a}\right) & \text{if } a < b-1, \\ \sigma \log\left(\frac{a+b-1}{2b}\right) & \text{if } a > b+1, \\ 0 & \text{otherwise.} \end{cases}$$

Figure 3 displays the regions of a and b to calculate the mode of beta-Laplace distribution for given values of a and b. When (a, b) is from the red region (i.e. when $a < b-1$), the mode is negative, and when (a, b) is from blue region (i.e. when $a > b+1$), the mode is positive. Also, when (a, b) is from the white region (the band between the two dashed lines), the mode is zero.

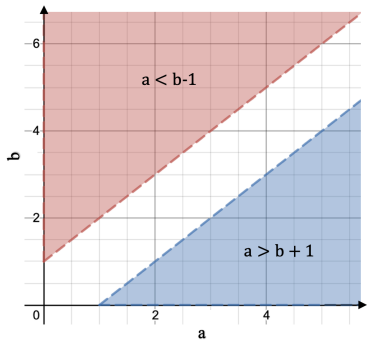


Figure 3: Plot of regions of a and b to calculate mode

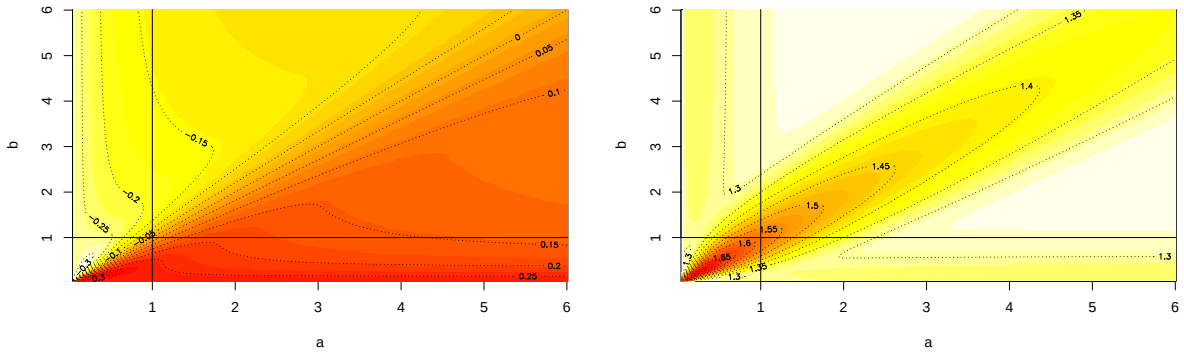
4. **Skewness and Kurtosis**: Here, we present the formulas for skewness and kurtosis of beta-Laplace distribution based on the quantile function of the beta-Laplace distribution. The measure of the Galton's skewness S and the measure of the Moors' kurtosis K are defined based on the quantile functions. Thus, the Galton's skewness (S) and Moors' kurtosis (K) of the beta-Laplace distribution respectively can be found using

$$\begin{aligned} S &= \frac{Q_X(3/4) - 2Q_X(1/2) + Q_X(1/4)}{Q_X(3/4) - Q_X(1/4)}, \\ K &= \frac{Q_X(7/8) - Q_X(5/8) + Q_X(3/8) - Q_X(1/8)}{Q_X(3/4) - Q_X(1/4)}, \end{aligned}$$

where $Q_X(p) = -\sigma \operatorname{sgn}\left({I^{-1}}_p(a,b) - 1/2\right) \log\left(1 - 2\left|{I^{-1}}_p(a,b) - 1/2\right|\right)$. Figure 4 represents the contour plots of skewness (4a) and kurtosis (4b) of beta-Laplace distribution for different combinations of parameters a and b when $\sigma = 1$.

Based on the contour plots of Galton's skewness (S) in Figure 4a, the following properties are observed:

- Case (i) If $a = b$, then $S = 0$ and beta-Laplace distribution is symmetric.
- Case (ii) $a < 1$ and $b < 1$: If $a < b$, then $S < 0$ and beta-Laplace distribution is negatively skewed. If $a > b$, $S > 0$ and beta-Laplace distribution is positively skewed.
- Case (iii) $a > 1$ and $b > 1$: If $a < b$, then $S < 0$ and beta-Laplace distribution is negatively skewed. If $a > b$, $S > 0$ and beta-Laplace distribution is positively skewed.



(a) (b)

Figure 4: Galton's skewness(S) and Moors' kurtosis (K) of beta-Laplace distribution

- Case (iv) $a < 1$ and $b \geq 1$: $S < 0$ and beta-Laplace distribution is negatively skewed.
- Case (v) $a \geq 1$ and $b < 1$: $S > 0$ and beta-Laplace distribution is positively skewed.

Based on the contour plots of Moors' kurtosis (K) in Figure 4b, Moors' kurtosis (K) is a decreasing function of a and b.

5. **Shannon entropy**: By using Corollary (2.2) and the fact that $\eta_T = \log(B(a,b)) - (a-1)\psi(a) - (b-1)\psi(b) + (a+b-2)\psi(a+b)$, where $\psi(\cdot)$ is the digamma function, the Shannon entropy of beta-Laplace distribution can be obtained as

$$\eta_X = \log(B(a,b)) - (a-1)\psi(a) - (b-1)\psi(b) + (a+b-2)\psi(a+b) + \log(2\sigma) + E\left(|X|\right)/\sigma.$$

4. A generalized LASSO regression model

We introduce a generalized family of LASSO regression models using the T-Laplace{uniform} family in Sub-section 4.2 and a generalized LASSO model using beta-Laplace distribution in Sub-section 4.3. First, we present an overview of the LASSO regression model.

4.1. LASSO regression model

The Least Absolute Selection and Shrinkage Operator (LASSO) method is a powerful penalized regression method that was first formulated by Tibshirani (1996). Given a data set of n observations, let $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ denotes the vector of responses, and $\mathbf{X}$ be an $n \times p$ matrix with $x_i \in \mathbb{R}^p$ where p is number of predictors in the model. Then, LASSO finds the solution $(\hat{\beta}_0, \hat{\boldsymbol{\beta}})$ to the optimization problem

$$\begin{aligned} \underset{\beta_0, \boldsymbol{\beta}}{\text{minimize}} \quad & \frac{1}{2n} \|\mathbf{y} - \beta_0 \mathbf{1} - \mathbf{X}\boldsymbol{\beta}\|_2^{\,2} \\ \text{subject to} \quad & \|\boldsymbol{\beta}\|_1 \leq t, \end{aligned} \tag{24}$$

where $\mathbf{1}$ is the vector of n ones, and $\|\cdot\|_1$ and $\|\cdot\|_2$ denote L_1 and L_2 norms, respectively. The bound t in the constraint limits the sum of the parameter estimates' absolute values and controls how well the data can be fitted. The value of t can be obtained through a cross-validation process. Suppose that we standardize each column of the matrix of predictors $\mathbf{X}$ so that each predictor is centered with a zero-mean and unit variance. This allows us to ignore the units of

the predictors. We also assume that the vector of responses, $\mathbf{y}$ is also centered with a zero-mean. With these centering conditions, we can omit the intercept β_0 from the LASSO optimization. Once the optimal solution $\hat{\boldsymbol{\beta}}$ is found for the centered data, it is possible to recover the optimal solution for the uncentered data: $\hat{\boldsymbol{\beta}}$ is same and $\hat{\beta}_0 = \bar{\mathbf{y}} - \hat{\boldsymbol{\beta}}\bar{\mathbf{X}}$, where $\bar{\mathbf{y}}$ is the mean of the uncentered response vector $\mathbf{y}$ and $\bar{\mathbf{X}}$ is the vector of means of the uncentered columns in $\mathbf{X}$. Because of that we omit the intercept $\hat{\beta}_0$ from the models for the remainder of this paper.

An equivalent form of the optimization problem in (24), the so called Lagrangian form, is given as

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2{}^2 + \lambda \|\boldsymbol{\beta}\|_1 , \tag{25}$$

where $\lambda \geq 0$ is the regularization parameter that controls the strength of the penalty $\|\boldsymbol{\beta}\|_1$. The value of λ can be determined through a cross-validation process. The regularization parameter λ and the bound t have a reverse relationship. As the value of t becomes infinity, the LASSO becomes the OLS, and λ becomes zero. On the other hand, as t becomes zero, all the parameter estimates become zero, and λ becomes infinity.

LASSO has a convex objective function and a convex constraint. So, it is a convex optimization problem, and many sophisticated optimization methods have been developed to solve it. On the other hand, Tibshirani (1996) identified a Bayesian view of LASSO estimators. According to the author LASSO estimates can be interpreted as posterior mode estimates when regression parameters have independent and identical Laplace priors. Park and Casella (2008) implemented the first explicit Bayesian approach for LASSO and presented a model of the form

$$\begin{aligned} \mathbf{y} \mid \boldsymbol{\beta}, \sigma &\sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_{n\times n}) \\ \boldsymbol{\beta} \mid \lambda, \sigma &\sim \prod_{j=1}^{p} \frac{\lambda}{2\sigma} e^{-\frac{\lambda}{\sigma}|\beta_j|}, \end{aligned}$$

using the independent and identical Laplace prior for each β_j. Then, the negative log of full conditional posterior distribution for $\boldsymbol{\beta} \mid \mathbf{y}, \lambda, \sigma$ is proportional to

$$\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2{}^2 + \frac{\lambda}{\sigma} \|\boldsymbol{\beta}\|_1 , \tag{26}$$

where an additive constant that is independent of $\boldsymbol{\beta}$ is dropped. Given any fixed values of λ and σ, the posterior mode in Equation (26) coincides with LASSO estimate with the regularization parameter $\sigma\lambda$.

4.2. A generalization of LASSO regression model using T-Laplace family

Motivated by the Bayesian view of the LASSO estimators in Sub-section 4.1, we introduce some additional components to the constraint in the ordinary LASSO regression model using the T-Laplace{uniform} family.

Park and Casella (2008) assumed that the priors of regression parameters $\boldsymbol{\beta}$ are independent and identical Laplace distributions. Given the distribution of the random variable T, we assume that the priors of $\boldsymbol{\beta}$ are independent and identical T-Laplace{uniform} distributions with the CDF and PDF given in equations (10) and (11), respectively. For example, if T is a beta random variable, then we assume that the priors of $\boldsymbol{\beta}$ are independent and identical beta-Laplace{uniform} distributions. So, we can write a general class of prior distributions for $\boldsymbol{\beta}$ using the PDF of T-Laplace{uniform} in Equation (11) with $R \sim$ Laplace(0,σ_1) as

$$\pi(\boldsymbol{\beta}|\ \sigma_1) \quad = \quad \prod_{j=1}^{p} f_R(\beta_j) f_T(F_R(\beta_j))$$

$$
\begin{aligned}
&= \prod_{j=1}^{p} \frac{1}{2\sigma_1} e^{-|\beta_j|/\sigma_1} f_T\left(\frac{1}{2} + \frac{1}{2}\operatorname{sgn}(\beta_j)\left\{1 - e^{-|\beta_j|/\sigma_1}\right\}\right) \\
&= \frac{1}{(2\sigma_1)^p} e^{-\sum_{j=1}^{p}|\beta_j|/\sigma_1} \prod_{j=1}^{p} e^{\log\left\{f_T\left(\frac{1}{2}+\frac{1}{2}\operatorname{sgn}(\beta_j)\left\{1-e^{-|\beta_j|/\sigma_1}\right\}\right)\right\}} \\
\text{Note that } \pi(\boldsymbol{\beta}|\ \sigma_1) &\propto e^{-\|\boldsymbol{\beta}\|_1/\sigma_1 + \sum_{j=1}^{p}\log\left\{f_T\left(\frac{1}{2}+\frac{1}{2}\operatorname{sgn}(\beta_j)\left\{1-e^{-|\beta_j|/\sigma_1}\right\}\right)\right\}}. \qquad (27)
\end{aligned}
$$

As in Park and Casella (2008), we also assume that $\mathbf{y} \mid \boldsymbol{\beta}, {\sigma_2}^2 \sim N(\mathbf{X}\boldsymbol{\beta}, {\sigma_2}^2\mathbf{I}_{n\times n})$. Here $\mathbf{y}$ is mean centered and $\mathbf{X}$ is standardized so that each predictor is centered with a zero-mean and unit variance. Then, we can write the likelihood function as

$$
\begin{aligned}
L_n(\boldsymbol{\beta}|\ \sigma_2) &= \prod_{i=1}^{n} p(y_i \mid \boldsymbol{\beta}, {\sigma_2}^2) \\
&= \frac{1}{\sqrt{(2\pi)^p \det({\sigma_2}^2\mathbf{I})}} e^{-\frac{1}{2{\sigma_2}^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\intercal}\mathbf{I}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})} \\
\text{Note that } L_n(\boldsymbol{\beta}|\ \sigma_2) &\propto e^{-\frac{1}{2{\sigma_2}^2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^{\,2}}. \qquad (28)
\end{aligned}
$$

By the Bayes' theorem, the full conditional of $\boldsymbol{\beta} \mid \mathbf{y}, \sigma_1, \sigma_2$ can be obtained using

$$
p(\boldsymbol{\beta} \mid \mathbf{y}, \sigma_1, \sigma_2) \propto L_n(\boldsymbol{\beta}|\ \sigma_2)\pi(\boldsymbol{\beta}|\ \sigma_1). \qquad (29)
$$

By substituting equations (27) and (28) in Equation (29), we get

$$
p(\boldsymbol{\beta} \mid \mathbf{y}, \sigma_1, \sigma_2) \propto e^{-\frac{1}{2{\sigma_2}^2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^{\,2} - \|\boldsymbol{\beta}\|_1/\sigma_1 + \sum_{j=1}^{p}\log\left\{f_T\left(\frac{1}{2}+\frac{1}{2}\operatorname{sgn}(\beta_j)\left\{1-e^{-|\beta_j|/\sigma_1}\right\}\right)\right\}}. \qquad (30)
$$

From Equation (30), we can write the negative log of full conditional posterior for $\boldsymbol{\beta} \mid \mathbf{y}, \sigma_1, \sigma_2$ as

$$
\frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^{\,2} + \frac{{\sigma_2}^2}{\sigma_1}\left[\|\boldsymbol{\beta}\|_1 - \sigma_1\sum_{j=1}^{p}\log\left\{f_T\left(\frac{1}{2}+\frac{1}{2}\operatorname{sgn}(\beta_j)\left\{1-e^{-|\beta_j|/\sigma_1}\right\}\right)\right\}\right], \qquad (31)
$$

where an additive constant and a multiplicative constant that are independent of $\boldsymbol{\beta}$ are dropped. Based on Equation (31), we define the following optimization problem:

$$
\begin{aligned}
&\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^{\,2} \\
&\text{subject to} \quad \|\boldsymbol{\beta}\|_1 - \sigma_1\sum_{j=1}^{p}\log\left\{f_T\left(\frac{1}{2}+\frac{1}{2}\operatorname{sgn}(\beta_j)\left\{1-e^{-|\beta_j|/\sigma_1}\right\}\right)\right\} \le t,
\end{aligned} \qquad (32)
$$

The optimization problem in (32) is a generalized LASSO regression problem defined by using the T-Laplace{uniform} family. If the random variable T is from the standard uniform distribution, then this generalized LASSO becomes the ordinary LASSO in (24). It is clear that the generalized LASSO introduces some additional components to the constraint in the ordinary LASSO. The objective function of the generalized LASSO in (32) is convex while the convexity of the constraint depends on the distribution of the random variable T. The Lagrangian form of the optimization problem in (32) is given by

$$
\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^{\,2} + \lambda\|\boldsymbol{\beta}\|_1 - \lambda\sigma_1\sum_{j=1}^{p}\log\left\{f_T\left(\frac{1}{2}+\frac{1}{2}\operatorname{sgn}(\beta_j)\left\{1-e^{-|\beta_j|/\sigma_1}\right\}\right)\right\}, \qquad (33)
$$

where $\lambda \geq 0$.

4.3. A generalized LASSO regression model using beta-Laplace distribution

Let T be a random variable from beta distribution with the PDF

$$f_T(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, \ a,b>0, \tag{34}$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the beta function and $\Gamma(\cdot)$ is the gamma function. By applying the PDF in Equation (34) in the general class of priors in (27), the corresponding prior distribution generated from the beta-Laplace distribution for $\boldsymbol{\beta}$ is given by

$$\pi(\boldsymbol{\beta}|\, a,b,\sigma_1) \propto e^{-\|\boldsymbol{\beta}\|_1/\sigma_1+(a-1)\sum_{j=1}^{p}\log\left\{\frac{1}{2}+\frac{1}{2}\operatorname{sgn}(\beta_j)\left\{1-e^{-|\beta_j|/\sigma_1}\right\}\right\}+(b-1)\sum_{j=1}^{p}\log\left\{\frac{1}{2}-\frac{1}{2}\operatorname{sgn}(\beta_j)\left\{1-e^{-|\beta_j|/\sigma_1}\right\}\right\}} \tag{35}$$

By applying the prior distribution for $\boldsymbol{\beta}$ in Equation (35) to Equation (29) and following the same approach used to define the optimization problem in (32), a generalized LASSO regression problem derived from the beta-Laplace distribution is defined by

$$\begin{aligned} \underset{\boldsymbol{\beta}}{\text{minimize}} \quad & \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2{}^2 \\ \text{subject to} \quad & \|\boldsymbol{\beta}\|_1 - \sigma_1(a-1)\sum_{j=1}^{p}\log\left\{\frac{1}{2}+\frac{1}{2}\operatorname{sgn}(\beta_j)\left\{1-e^{-|\beta_j|/\sigma_1}\right\}\right\} \\ & \quad -\sigma_1(b-1)\sum_{j=1}^{p}\log\left\{\frac{1}{2}-\frac{1}{2}\operatorname{sgn}(\beta_j)\left\{1-e^{-|\beta_j|/\sigma_1}\right\}\right\} \leq t, \end{aligned} \tag{36}$$

It is noticed that there are two additional constraint components besides the L_1 norm. These two components are

$$C_1 = \left[-\sigma_1(a-1)\sum_{j=1}^{p}\log\left\{\frac{1}{2}+\frac{1}{2}\operatorname{sgn}(\beta_j)\left\{1-e^{-|\beta_j|/\sigma_1}\right\}\right\}\right],$$

which is based on the CDF of the random variable R, and

$$C_2 = \left[-\sigma_1(b-1)\sum_{j=1}^{p}\log\left\{\frac{1}{2}-\frac{1}{2}\operatorname{sgn}(\beta_j)\left\{1-e^{-|\beta_j|/\sigma_1}\right\}\right\}\right],$$

which is based on the survival function of R.
We can write the Lagrangian form of the optimization problem in (36) as

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \frac{1}{2n}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2{}^2 + \lambda\left(\|\boldsymbol{\beta}\|_1 + C_1 + C_2\right), \tag{37}$$

where $\lambda \geq 0$. The values of λ, and hyper parameters a, b, and σ_1 can be determined through a cross-validation process. When $a = b = 1$, we have the ordinary LASSO regression model in (25).

Based on the components L_1 norm, C_1, and C_2, we define the following six methods to build the generalized LASSO regression model in (36):

- Method 1: Use C_2 only in the constraint in (36).
- Method 2: Use C_1 only in the constraint in (36).

- Method 3: Use C_1 and C_2 only in the constraint in (36).
- Method 4: Use L_1 norm and C_2 only in the constraint in (36).
- Method 5: Use L_1 norm and C_1 only in the constraint in (36).
- Method 6: Use L_1 norm, C_1 and C_2 in the constraint in (36).

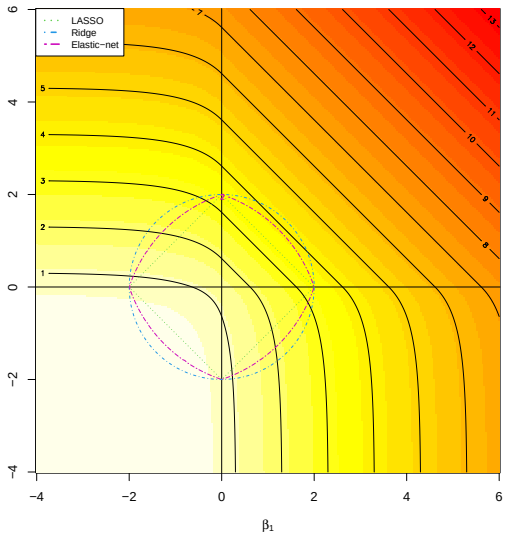
Some insights of the geometric interpretation

Figure 5 represents the constraint regions of each method. The levels of the constraint of each method are presented by black solid contour lines. Also, by looking at C_1 and C_2, and taking the limits as β_j goes to 0, $-\infty$, and $+\infty$ we get the following:

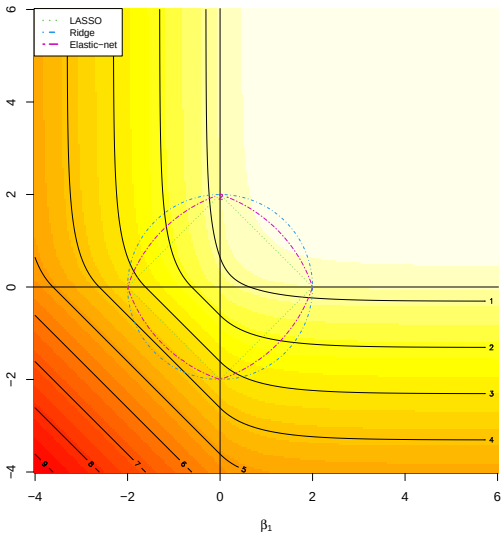
$$\begin{aligned}
\lim_{\beta_j \to 0} C_1 &= \lim_{\beta_j \to 0} C_2 \quad = -\log(0.5) \\
\lim_{\beta_j \to -\infty} C_1 &= \lim_{\beta_j \to +\infty} C_2 \to -\log(0) \\
\lim_{\beta_j \to +\infty} C_1 &= \lim_{\beta_j \to -\infty} C_2 = -\log(1) = 0
\end{aligned}$$

Some insights from these limits and Figure 5:

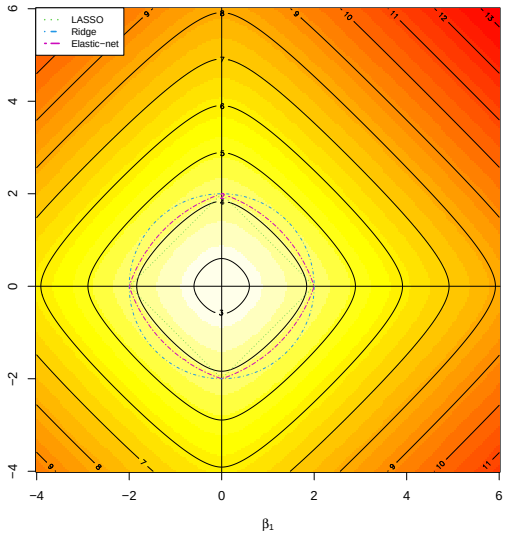
- When $\boldsymbol{\beta}$ goes to zero, both C_1 and C_2 become a constant $-\log(0.5)$. The result should be similar to ordinary LASSO estimates.
- When $\boldsymbol{\beta}$ goes to positive infinity for C_1 or $\boldsymbol{\beta}$ goes to negative infinity for C_2, both C_1 and C_2 go to zero. When $\boldsymbol{\beta}$ goes to negative infinity for C_1, or $\boldsymbol{\beta}$ goes to positive infinity for C_2, both C_1 and C_2 go to positive infinity. This indicates these estimates will shrink during minimization process. This is where the shrinkage occurs. That is, C_1 shrinks very negative estimate to be less negative (pull it back to the zero direction); C_2 shrinks very positive estimate to be less positive (pull it back to the zero direction).
- In methods 3, 4, 5, and 6, the constraint regions have sharp corners, edges and curved contours. The sharp corners and edges encourage variable selection while the curved contours encourage strongly correlated variables to share coefficients. Compared to the constraint region of ordinary LASSO model in (24), the constraint regions of the generalized LASSO model in (36) in methods 4 and 5 are asymmetric.



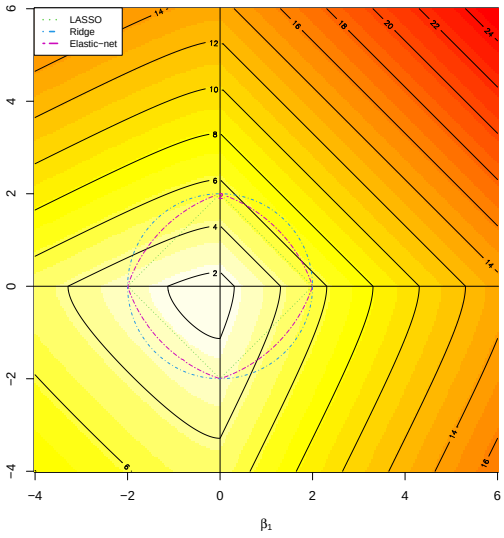
(a) Method 1: C_2 with $b = 2, \sigma_1 = 1$



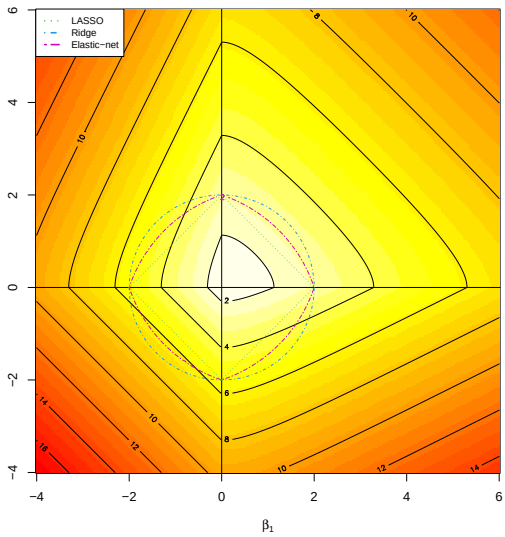
(b) Method 2: C_1 with $a = 2, \sigma_1 = 1$



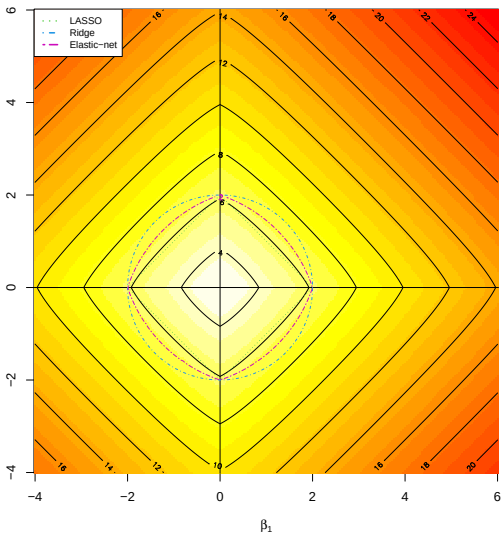
(c) Method 3: $C_1 + C_2$ with $a = b = 2, \sigma_1 = 1$



(d) Method 4: L_1norm $+ C_2$ with $b = 2, \sigma_1 = 1$



(e) Method 5: L_1norm $+ C_1$ with $a = 2, \sigma_1 = 1$

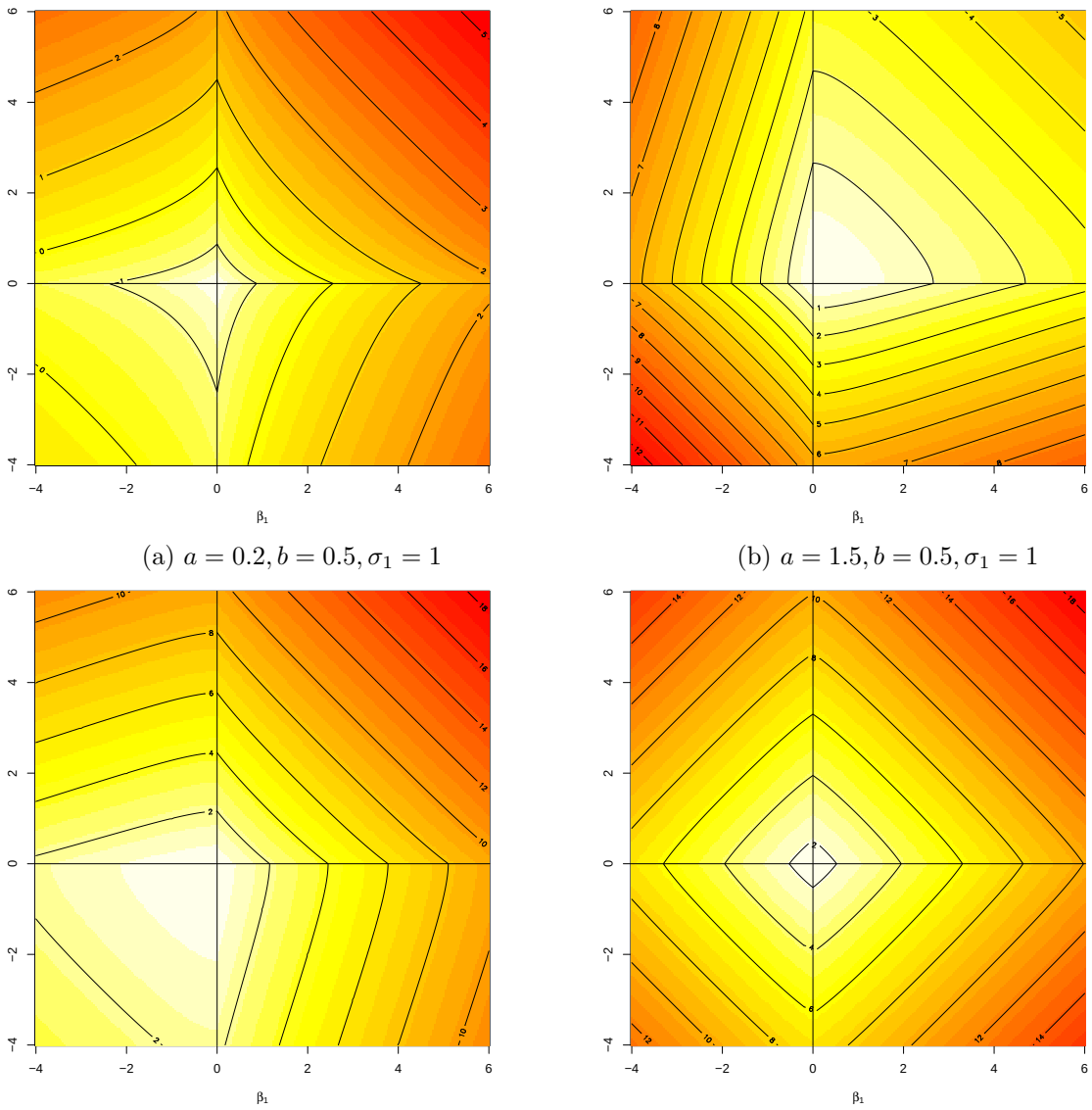


(f) Method 6: L_1norm$+C_1+C_2$ with $a = b = 2, \sigma_1 = 1$

Figure 5: Contours of the constraint in the generalized LASSO model

Note: The black solid contour lines represents the levels of the constraint of each method.

The objective function and the constraint of ordinary LASSO problem in Sub-section 4.1 are both convex. The objective function of the generalized LASSO problem in (36) is also convex. However, the convexity of the constraint in (36) depends on the values of a and b. Figure 6 represents the contours of the constraint region in Method 6 at different values of a and b. When $a < 1$ or $b < 1$, the constraint region becomes non-convex. On the other hand, when $a \geq 1$ and $b \geq 1$, the constraint region becomes convex and hence the generalized LASSO problem in (36) is a convex optimization problem.



(a) $a = 0.2, b = 0.5, \sigma_1 = 1$

(b) $a = 1.5, b = 0.5, \sigma_1 = 1$

(c) $a = 0.5, b = 1.5, \sigma_1 = 1$

(d) $a = 1.5, b = 1.5, \sigma_1 = 1$

Figure 6: Contours of the constraint in method 6 at different values of a and b

Note: The black solid contour lines represents the levels of the constraint.

5. Applications

In this section, we present a numerical study to demonstrate the effects of the convexity of the constraint in (36) in the process of variable selection. We will also analyze two real data sets to illustrate the flexibility and usefulness of the generalized LASSO regression model in (37) in the process of variable selection with better prediction performance.

5.1. A numerical study

Zheng, Maleki, Weng, Wang, and Long (2017) studied the performance between ordinary LASSO with L_p regularized least squares models when $p \in [0, 1)$. When $p < 1$, the constraint region becomes non-convex and hence promotes sparsity. Moreover, $p < 1$ leads to more accurate solutions compared to the ordinary LASSO since L_p norm models sparsity better. Authors also discussed some iterative algorithms to obtain local minima of non-convex optimization problem that occurs when $p \in [0, 1)$. Furthermore, the convergence of the iterative algorithms to the global minima depends on the initialization. In this sub-section, we investigate the effects of parameters a and b of the optimization problem in (37) on the sparsity of the model.

We synthesize a data set using the approach by Song and Liang (2015). The data set contains $n = 200$ observations with $p = 15$ predictors. All predictors are generated from multivariate normal distribution $N(0, I_n)$. The random errors are generated from $N(0, \sigma^2 I_n)$ with $\sigma = 1.5$. The response variable is generated using the first eight predictors where the coefficients are given by (2.63 , 2.28, -1.43, 2.16, 1.73, 1.06, -1.7, -2.43) and the random errors. The first eight predictors are the true predictors of the model. The rest seven predictors are generated in a way to have high correlations with the response variable by first randomly generating 1000 predictors from $N(0, I_n)$ and then choosing the top seven predictors that are highly correlated with the response variable. We label these seven predictors as false predictors of the model.

To this data set, we apply the Method 6 with $\sigma_1 = 1.0$ and several combinations of a and b. An optimization algorithm based on the optim() function in R statistical software is implemented to solve the generalized LASSO problem in (37) and the regularization parameter, λ^* is selected through a 10-fold cross-validation process over a grid of values for λ.

Table 3: Number of false predictors captured by the model

a\b	**0.1**	**0.3**	**0.5**	**0.7**	**1.0**	**1.3**	**1.5**	**1.7**	**2.0**	**2.3**	**2.5**	**2.7**	**3.0**	**3.3**	**3.5**	**3.7**	**4.0**
0.1	6	4	4	5	3	3	2	1	1	1	1	1	1	1	1	0	0
0.3	4	4	4	4	4	3	2	1	1	1	1	1	1	1	1	0	0
0.5	2	2	2	2	3	3	1	1	1	1	1	1	1	1	1	1	0
0.7	1	1	2	2	2	2	1	1	1	1	1	1	1	1	1	1	0
1.0	1	1	1	1	1	2	2	1	1	0	1	1	1	1	1	1	0
1.3	1	1	1	1	1	1	0	0	1	0	0	0	1	1	1	1	0
1.5	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	0
1.7	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0
2.0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
2.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3.5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4.0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Note: We decide a false predictor is captured when the absolute value of the regression coefficient estimate is less than 0.001.

Table 3 presents the number of false predictors captured by Method 6 for several combinations of a and b. We decide a false predictor is captured when the absolute value of the regression coefficient estimate is less than 0.001. When $a < 1$ or $b < 1$, the generalized LASSO problem becomes non-convex and has the ability to capture more false predictors. When $a = b = 1$, we have the ordinary LASSO and it captures only one out of seven false predictors. Compared to the ordinary LASSO, when $a = b = 0.1$, the generalized LASSO captures six out of seven false predictors. Based on the number of false predictors captured in Table 3, it is evident that the values of a and b have an impact on the sparsity of the generalized LASSO model in (37).

When the values of a or b less than one, we have a sparse model and as the values of a and b increase the model becomes more dense.

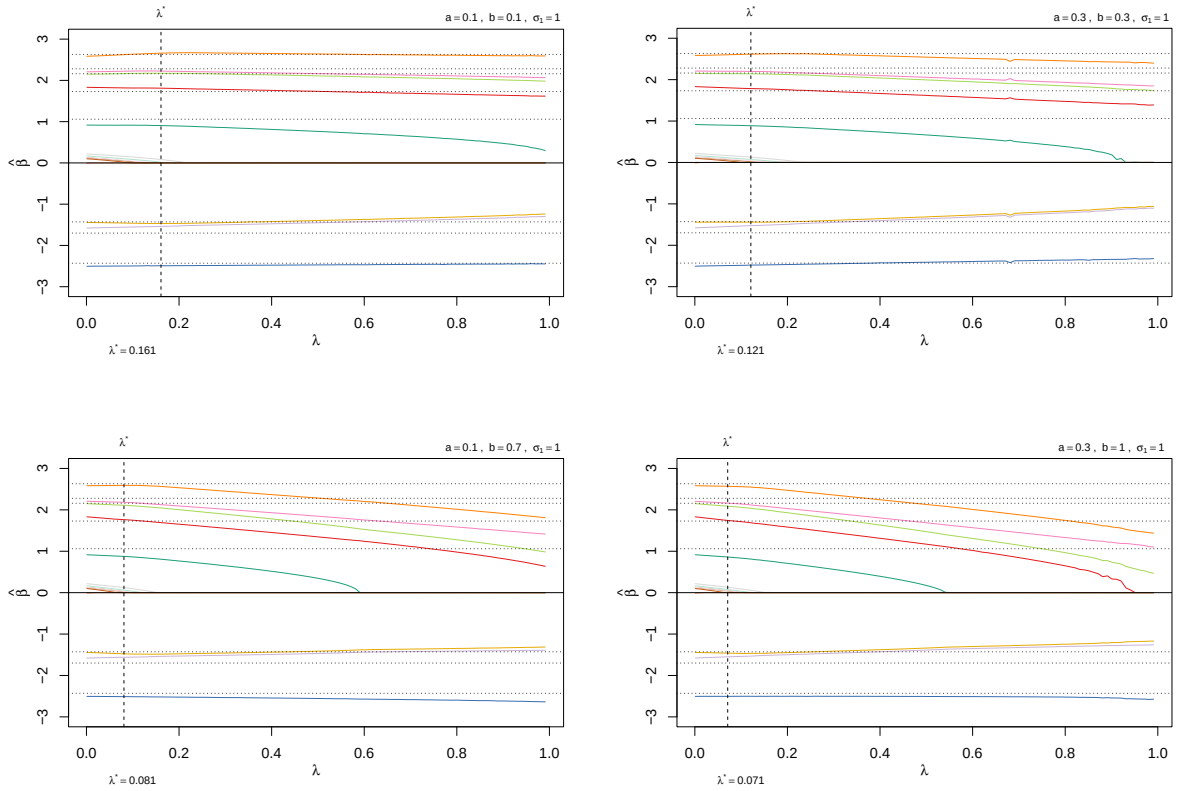


Figure 7: Regularization paths of method 6 for several combinations of a and b

Figure 7 shows the regularization paths of Method 6 for several combinations of a and b. The dotted horizontal lines indicate the true values of the eight true predictors while the colored lines represent the paths of the regression coefficient estimates. The vertical dashed line indicates the value of the regularization parameter (λ^*) which gives the smallest cross-validation error. As the value of λ increases the estimates of the regression coefficients go to zero and hence the model becomes sparse. When $a = b = 1$, the ordinary LASSO regression model selects a value for λ^* which captures only one false predictor. However, by changing the values of a and b it is possible to build a model that selects a value for λ^* that can capture more false predictors in the synthetic data set. Thus, compared to the ordinary LASSO, the generalized LASSO regression model in (37) has much more flexibility in the variable selection.

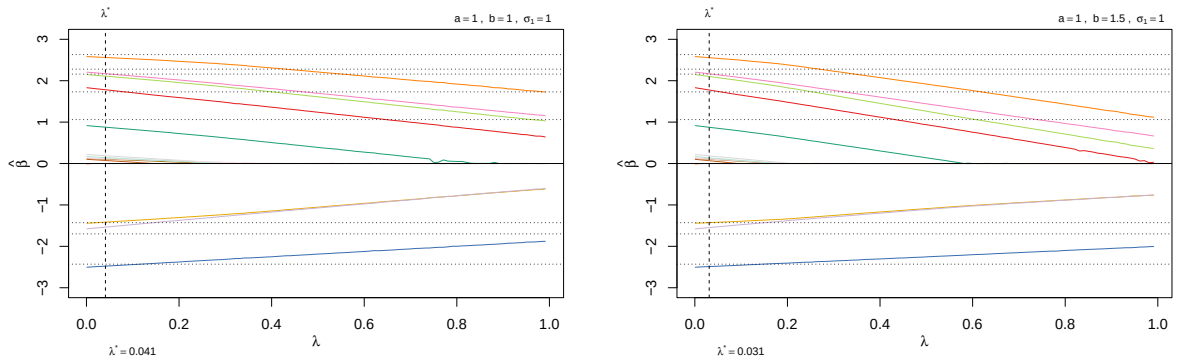


Figure 7: Regularization paths of method 6 for several combinations of a and b

5.2. Diabetes data

In this Sub-section, we analyze Diabetes data set (Efron, Hastie, Johnstone, and Tibshirani 2004) to illustrate the flexibility and usefulness of the generalized LASSO regression model in (37) in the process of variable selection with better prediction performance.

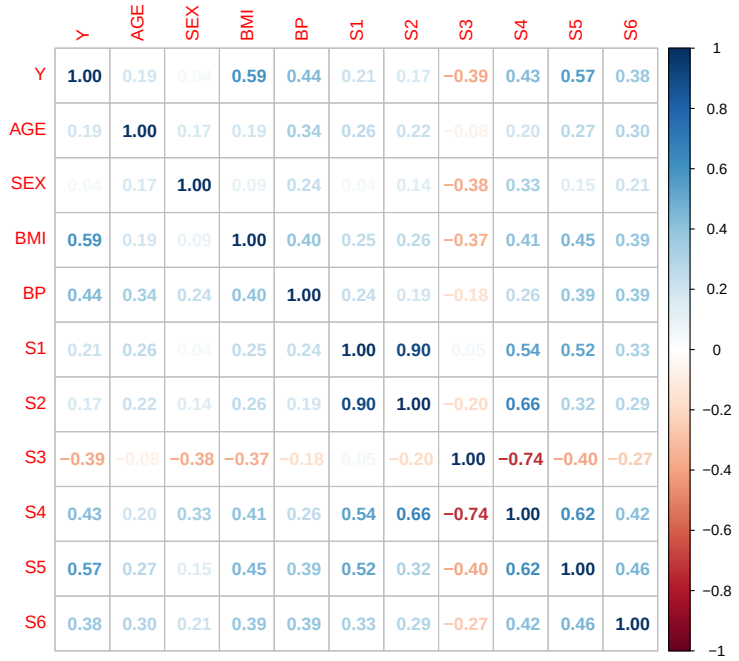


Figure 8: Correlation plot for diabetes data

The data set consists of 442 observations with ten baseline predictors: AGE, SEX, BMI, BP, and Serum measurements (S1–S6) and one response variable (Y) that measures the disease progression one year after baseline. Figure 8 displays the correlation plot for the Diabetes data set. Based on the correlation plot, it can be observed that predictors S1 and S2 are strongly correlated while predictors S3 and S4 are also strongly correlated. First, we randomly split the data set into two: training set (75%) and testing set (25%). Then we apply OLS, LASSO, ridge, elastic-net, and methods 1–6 to the training set. Ideally, the values of a, b, and σ_1 in methods

1–6 should be determined through cross-validation. However, in this example, we set $a = b = 2$ and $\sigma_1 = 1$. The regularization parameters λ of methods 1–6 are determined through a 10-fold cross-validation process.

Table 4: Regression coefficient estimates

	Model									
Predictors	OLS	LASSO	ridge	elastic-net	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6
AGE	-0.65090	.	-0.30745	.	-0.50873	0.39017	-0.20833	-0.00906	0.23128	-0.00064
SEX	-11.34000	-9.65074	-9.90095	-9.63684	-10.73777	-5.62116	-10.26569	-9.47742	-4.51198	-8.02667
BMI	26.06000	26.28249	25.11432	26.11002	25.86784	27.99046	26.39907	25.99478	28.05809	25.93001
BP	18.57000	17.37213	17.20682	17.30652	17.79112	16.42076	17.89192	17.03153	15.77341	16.14491
S1	-39.71000	-9.85397	-6.776626	-9.54005	-9.58100	-8.72247	-12.95845	-7.72758	-7.04683	-3.46306
S2	22.58000	.	-2.389684	.	-0.44406	0.22379	0.86373	-0.18306	0.01433	-0.86722
S3	9.81200	-4.60136	-6.864659	-5.03250	-6.34466	0.24464	-2.02052	-7.33515	0.00014	-10.41868
S4	17.44000	10.8883	10.34761	10.48153	10.24595	12.61183	13.88346	7.82349	11.18310	3.39779
S5	30.31000	19.4899	17.76703	19.34952	19.41769	19.23549	20.17135	19.17917	18.86554	18.36158
S6	3.29400	2.74109	3.850925	2.87404	2.95917	2.67535	3.01802	2.57170	2.40973	2.18371

Table 5: Mean squared errors

	Model									
	OLS	LASSO	ridge	elastic-net	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6
training MSE	1367.15	1377.15	1379.80	1377.55	1376.66	1394.13	1374.33	1380.32	1403.25	1390.95
testing MSE	1679.90	1665.21	1668.70	1664.74	1662.40	1699.46	1673.50	1657.08	1704.83	1653.40

Table 6 and Table 7 present the regression coefficient estimates of each model and mean squared errors (MSEs) from the training and testing data sets, respectively. It appears that if we use only one component from L_1 norm, C_1, and C_2 in the constraint of the generalized LASSO model in (36), then Method 1 (using C_2 only) seems to provide the best test MSE. Method 6 (L_1 norm + C_1 + C_2) gives the best MSE from the test data. Based on the MSEs from test data, Method 4 (L_1 norm + C_2) does well too. Observing the estimates in Table 6, they seem to show the patterns that we observed under the geometric interpretation in Sub-section 4.3. That is, Method 1 (using C_2 only) shrinks positive estimates and takes care of multicollinearity between S1 and S2. Method 2 (using C_1 only) takes care of multicollinearity between S3 and S4. Compared to all other models, Method 5 (L_1 norm + C_1) takes care the multicollinearity between S1 and S2 as well as the multicollinearity between S3 and S4. These observations seem to indicate adding the components C_1 and/or C_2 to L_1 norm is a good approach and improves the flexibility and usefulness of the ordinary LASSO regression model in (24) in the process of variable selection with better prediction performance.

5.3. Stress reactions to COVID-19 data

To illustrate the flexibility and applicability of the generalized LASSO regression model in (37) in the selection of variables with better prediction performance, in this Sub-section we analyze a set of data collected from the study by Flesia, Monaro, Mazza, Fietta, Colicino, Segatto, and Roma (2020).

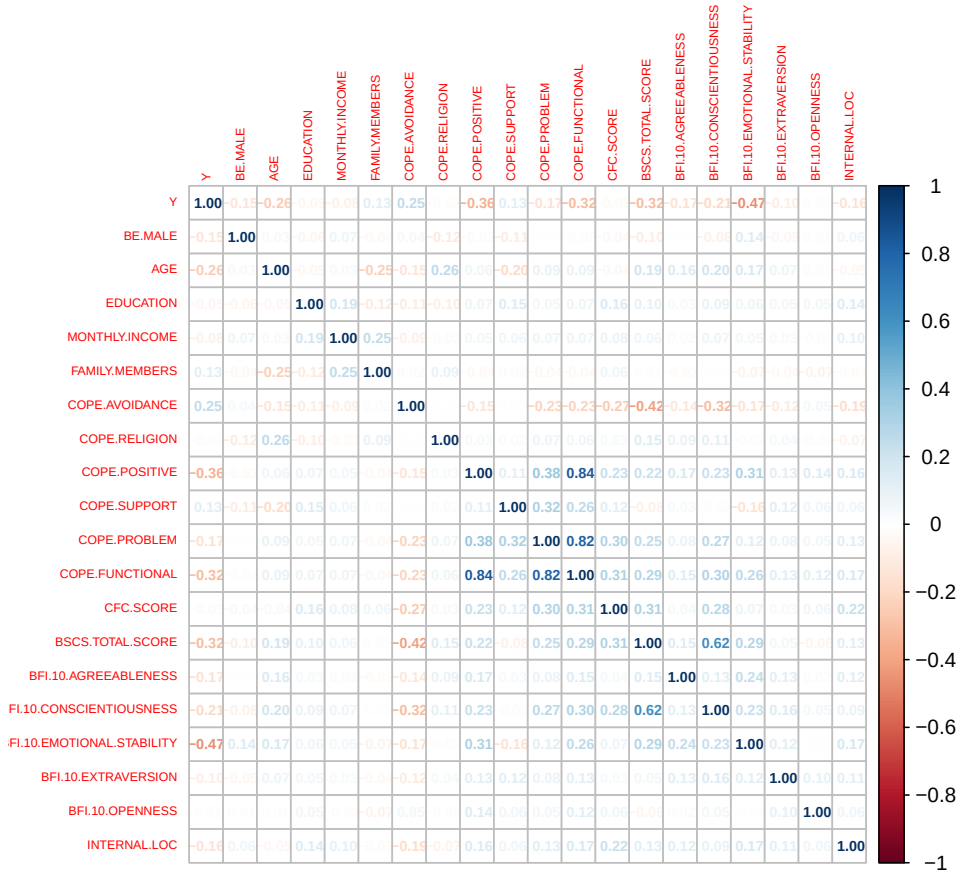


Figure 9: Correlation plot for stress reactions to COVID-19 data

Data consists of responses from 2053 participants to an online survey conducted during the period of 20-31 March 2020. For illustration, we analyzed all 2053 observations and used 19 baseline predictors: gender, age, education, monthly income, number of family members, six measurements under the Coping Orientations to the Problems Experienced (COPE-NVI-25), total score in the Consideration of Future Consequences (CFC) scale, total score in the Brief Self-Control Scale (BSCS), five measurements under the Big Five Inventory (BFI), internal subscale of the short version of the Locus of Control Scale (LOC) and one response variable (Y) that measures perceived stress scale. For more details on the data, see Flesia *et al.* (2020). The correlation plot for the stress reactions to COVID-19 data is shown in Figure 9. The correlation plot reveals a strong correlation between COPE.FUNCTIONAL and COPE.PROBLEM, as well as a strong correlation between COPE.FUNCTIONAL and COPE.POSITIVE. The data set is split into two randomly: a training set (75%) and a testing set (25%), then we apply OLS, LASSO, ridge, elastic-net, and methods 1–6 to the training set. Ideally, the values of a, b, and σ_1 in methods 1–6 should be determined through cross-validation. In this example, however, we set $a = b = 0.5$ and $\sigma_1 = 1$. The regularization parameters λ of methods 1–6 are determined through a 10-fold cross-validation process.

Table 6: Regression coefficient estimates

	Model									
Predictors	OLS	LASSO	ridge	elastic-net	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6
BE.MALE	-1.05800	-1.00608	-0.99070	-1.00355	-0.98626	-0.99025	-0.87181	-0.92883	-0.96164	-0.95158
AGE	-0.70440	-0.67147	-0.68867	-0.67248	-0.76853	-0.76999	-0.36731	-0.69769	-0.73210	-0.72017
EDUCATION	-0.08088	-0.03980	-0.08459	-0.04435	-0.14538	-0.13091	0.09139	-0.05464	-0.09356	-0.07572
MONTHLY.INCOME	-0.37550	-0.32281	-0.34136	-0.32412	-0.35788	-0.36838	-0.42789	-0.32722	0.33396	-0.33448
FAMILY.MEMBERS	0.56800	0.53346	0.53821	0.53339	0.45530	0.48759	0.74620	0.50416	0.46335	0.48749
COPE.AVOIDANCE	0.69610	0.68606	0.68281	0.68491	0.62834	0.67613	1.27520	0.71677	0.65458	0.69173
COPE.RELIGION	0.09329	0.04628	0.08802	0.05006	0.05406	0.07203	0.11380	0.05618	0.02689	0.04287
COPE.POSITIVE	-7285.0	-1.20656	-1.04890	-1.18947	-1.31089	-35.91800	-0.62456	-1.28699	-1.26597	-1.28864
COPE.SUPPORT	0.37410	0.31937	0.38896	0.32772	0.41603	0.42494	0.54051	0.40679	0.38322	0.39648
COPE.PROBLEM	-6875.0	.	0.06598	.	-0.09396	-32.76500	0.23478	-0.00017	-0.00077	-0.00008
COPE.FUNCTIONAL	11850.0	-0.33625	-0.57998	-0.35780	-0.30735	55.93778	-1.18579	-0.34233	-0.36635	-0.34817
CFC.SCORE	0.60290	0.53050	0.55026	0.53218	0.53211	0.56226	0.64942	0.52104	0.50478	0.52147
BSCS.TOTAL.SCORE	-1.30700	-1.18796	-1.18722	-1.18606	-1.17455	-1.21993	-0.97928	-1.09914	-1.11522	-1.12477
BFI.10.AGREEABLENESS	0.03009	.	-0.00259	.	-0.00148	0.02437	0.20413	0.00080	0.00001	0.00040
BFI.10.CONSCIENTIOUSNESS	0.32870	0.19075	0.23365	0.19395	0.30167	0.35555	0.34535	0.24552	0.23111	0.25407
BFI.10.EMOTIONAL.STABILITY	-1.87500	-1.88275	-1.80123	-1.87408	-1.89271	-1.90533	-1.99410	-1.89668	-1.90699	-1.90313
BFI.10.EXTRAVERSION	-0.14380	-0.07922	-0.13526	-0.08479	-0.16589	-0.16371	0.16359	-0.07831	-0.10700	-0.09496
BFI.10.OPENNESS	0.21840	0.18775	0.22866	0.19137	0.15991	0.17489	0.26855	0.16205	0.13740	0.15146
INTERNAL.LOC	-0.25130	-0.23174	-0.26676	-0.23520	-0.26316	-0.25826	0.13212	-0.20112	-0.22836	-0.21551

Table 7: Mean squared errors

	Model									
	OLS	LASSO	ridge	elastic-net	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6
train MSE	15.28393	15.32990	15.32473	15.32885	15.32040	15.31433	15.67544	15.32629	15.32664	15.32229
test MSE	15.73405	15.70487	15.70954	15.70674	15.62415	15.59866	16.53035	15.62504	15.62954	15.61503

The regression coefficient estimates and mean square errors (MSEs) from the training and testing data sets are presented in Table 6 and Table 7. Using only one of the components from L_1 norm, C_1, and C_2 in the constraint of the generalized LASSO model in (36), Method 2 (using C_1 only) appears to provide the best test MSE. Based on the estimates in Table 6, it can be seen that Method 1 (using C_2 only), Method 4 (L_1 norm + C_2), Method 5 (L_1 norm + C_1) and Method 6 (L_1 norm + $C_1 + C_2$) adequately handle the multicollinearity between COPE.FUNCTIONAL and COPE.POSITIVE, as well as the multicollinearity between COPE.FUNCTIONAL and COPE.PROBLEM with better prediction ability. Based on the coefficient estimates, Method 3 ($C_1 + C_2$ only) also accounts for multicollinearity between COPE.FUNCTIONAL and COPE.POSITIVE, as well as between COPE.FUNCTIONAL and COPE.PROBLEM. As a result, adding C_1 and/or C_2 to L_1 norm may be a good approach that significantly enhances the flexible and applicability of the ordinary LASSO regression model in (24) during the variable selection process with better prediction performance.

6. Concluding remarks

In this article, we develop a generalized LASSO regression model from a generalized Laplace distribution through the Bayesian interpretation of LASSO. A family of generalized Laplace distributions is introduced and studied using the T-$R\{Y\}$ framework by Aljarrah *et al.* (2014). Five different generalized Laplace families are obtained using quantile functions of standard uniform, Weibull, log-logistic, logistic, and extreme value distributions. Various general properties of the new families including quantile function, mode, and Shannon entropy are derived. A particular case of T-Laplace{uniform} family called the beta-Laplace distribution is explored.

Some additional components to the constraint in the ordinary Lasso regression model are obtained through the Bayesian interpretation of LASSO with beta-Laplace priors. The geometric interpretations of these additional components are presented. Using a numerical study, the effects of the parameters from beta-Laplace distribution in the generalized LASSO regression model are discussed. Two real data sets are analyzed to illustrate the flexibility and usefulness of the generalized LASSO regression model in the process of variable selection with better prediction performance. The comparison with other existing shrinkage methods indicates adding

the additional components to the constraint in ordinary LASSO regression model improves the flexibility and applicability of LASSO in variable selection with better prediction performance.

From the standpoint of practical applications, we think it will be an interesting study to use Bayesian techniques to estimate regression parameters in the generalized LASSO regression model by assuming that the regression parameters have independent and identical generalized Laplace priors. Here, we will refer to works by Tibshirani (1996) and Park and Casella (2008). Although shrinking the regression parameters or setting some coefficients to zero can sometimes improve the prediction accuracy, this will introduce some bias but lower the variance. We think it would also be interesting to study the generalized LASSO regression model's performance in terms of bias and variance, and a simulation study can be conducted to compare its bias with other existing methods. In a future paper, we will continue investigating this aspect, and we hope that our study will serve as a reference for future research in this area.

Acknowledgements

The authors wish to thank the anonymous reviewers and the Editor for their careful readings and for their comments that greatly improved the manuscript.

References

Aljarrah MA, Lee C, Famoye F (2014). "On Generating T-X Family of Distributions Using Quantile Functions." *Journal of Statistical Distributions and Applications*, **1**(1), 1–17. doi:10.1186/2195-5832-1-2.

Alzaatreh A, Lee C, Famoye F (2013). "A New Method for Generating Families of Continuous Distributions." *Metron*, **71**(1), 63–79. doi:10.1007/s40300-013-0007-y.

Alzaatreh A, Lee C, Famoye F (2014). "T-normal Family of Distributions: A New Approach to Generalize the Normal Distribution." *Journal of Statistical Distributions and Applications*, **1**(1), 1–18. doi:10.1186/2195-5832-1-16.

Aryal G, Zhang Q (2016). "Characterizations of Kumaraswamy Laplace Distribution with Applications." *Stochastics and Quality Control*, **31**(2), 1–18. doi:10.1515/eqc-2016-0009.

Cordeiro GM, Lemonte AJ (2011). "The Beta Laplace Distribution." *Statistics and Probability Letters*, **81**(8), 973–982. doi:10.1016/j.spl.2011.01.017.

Efron B, Hastie T, Johnstone I, Tibshirani R (2004). "Least Angle Regression." *Annals of Statistics*, **32**(2), 407–499. doi:10.1214/009053604000000067.

Eugene N, Lee C, Famoye F (2002). "Beta-normal Distribution and Its Applications." *Communications in Statistics-Theory and methods*, **31**(4), 497–512. doi:10.1081/STA-120003130.

Flesia L, Monaro M, Mazza C, Fietta V, Colicino E, Segatto B, Roma P (2020). "Predicting Perceived Stress Related to the Covid-19 Outbreak through Stable Psychological Traits and Machine Learning Models." *Journal of clinical medicine*, **9**(10), 3350. doi:10.3390/jcm9103350.

Haleem A, Javaid M, Khan IH, Vaishya R (2020). "Significant Applications of Big Data in COVID-19 Pandemic." *Indian Journal of Orthopaedics*, **54**(4), 526–528. doi:10.1007/s43465-020-00129-z.

Hastie T, Tibshirani R, Wainwright M (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC. ISBN 978-1498712163.

Hoerl AE, Kennard RW (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics*, **12**(1), 55–67. doi:10.1080/00401706.1970.10488634.

Kerman J (2011). "A Closed-form Approximation for the Median of the Beta Distribution." *arXiv preprint arXiv:1111.0433*. doi:10.48550/arXiv.1111.0433.

Park T, Casella G (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, **103**(482), 681–686. doi:10.1198/016214508000000337.

Song Q, Liang F (2015). "High-dimensional Variable Selection with Reciprocal L_1-regularization." *Journal of the American Statistical Association*, **110**(512), 1607–1620. doi:10.1080/01621459.2014.984812.

Tibshirani R (1996). "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society: Series B*, **58**(1), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x.

Zheng L, Maleki A, Weng H, Wang X, Long T (2017). "Does L_p-minimization Outperform L_1-minimization?" *IEEE Transactions on Information Theory*, **63**(11), 6896–6935. doi:10.1109/TIT.2017.2717585.

Zou H, Hastie T (2005). "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society: Series B (statistical methodology)*, **67**(2), 301–320. doi:10.1111/j.1467-9868.2005.00503.x.

Appendix

Throughout this study, we used R statistical software. The R codes for the generalized LASSO regression model with beta-Laplace priors are included in this appendix.

```
#Define CDF of Laplace distribution
L_cdf=function(x,mu,sig){
  y=0.5+0.5*sign(x-mu)*(1-exp(-abs(x-mu)/sig))
  return(y)
}

#Define PDF of Laplace distribution
L_pdf=function(x,mu,sig){
  y=0.5*exp(-abs(x-mu)/sig)/sig
  return(y)
}

#Define CDF of beta-Laplace distribution
#a,b>0
BL_cdf=function(a,b,mu,sig,x){
  pbeta(L_cdf(x,mu,sig),a,b)
}

#Define PDF of beta-Laplace distribution
BL_pdf=function(a,b,mu,sig,x){
  (L_cdf(x,mu,sig)^(a-1))*((1-L_cdf(x,mu,sig))^(b-1))
        *L_pdf(x,mu,sig)/beta(a,b)
}

#Plots of PDF of beta-Laplace distribution
x=seq(-15,5,by=0.001)
para=c(0.5,0.5,0.5,0.5,0.5,0.5)
parb=c(0.5,0.7,1,3,5,7)
```

```
parsig=c(2,2,2,2,2,2)
y1=BL_pdf(a=para[1],b=parb[1],mu=0,sig=parsig[1],x)
plot(x,y1,col=1,type="l",lty=1,lwd=2,xlab="x",ylab="f(x)",
ylim=c(0,0.2))
for (i in 2:length(para)) {
  y1=BL_pdf(a=para[i],b=parb[i],mu=0,sig=parsig[i],x)
  lines(x,y1,col=i,lty=i,lwd=2)
}

legend("topleft",c(
  expression(paste(a,'=0.5, ',b,'=0.5, ',sigma,'=2.0')),
  expression(paste(a,'=0.5, ',b,'=0.7, ',sigma,'=2.0')),
  expression(paste(a,'=0.5, ',b,'=1.0, ',sigma,'=2.0')),
  expression(paste(a,'=0.5, ',b,'=3.0, ',sigma,'=2.0')),
  expression(paste(a,'=0.5, ',b,'=5.0, ',sigma,'=2.0')),
  expression(paste(a,'=0.5, ',b,'=7.0, ',sigma,'=2.0'))),
       col=c(1,2,3,4,5,6),lty=c(1,2,3,4,5,6),
       lwd=c(2,2,2,2,2,2))

#Analyzing a data set
install.packages("corrplot")
library(corrplot)

#Reading the data file
diabetes<-read.table(file.choose(), header = TRUE)
X<-as.matrix(diabetes[,1:10])
Y<-diabetes$Y

#Make Y is centered and X is standardized
MakeStandardized<-function(Y,X){
  Y.mean=mean(Y)
  Y=Y-Y.mean
  X.mean=apply(X ,2, mean)
  X.sd=apply(X,2,sd)
  for(i in 1:dim(X)[2]){
    for(j in 1:dim(X)[1]){
      X[j,i]<-(X[j,i]-X.mean[i])/X.sd[i]
    }
  }
  return(list(Y,X))
}

#Correlation plot
M<-round(cor(cbind(Y,X)),2)
corrplot(M, method="number")

#Splitting data into train and test
colnames(X)<-NULL
proportion_split = 0.75
train = sample(1:nrow(X), round(nrow(X)* proportion_split))
#train=sample (1: nrow(X), nrow(X)/2)
test=(- train )
Y.train=Y[train]
Y.train.mean=mean(Y.train)
Y.test=Y[test]-mean(Y.train)

X.train=X[train,]
X.train.mean=apply(X.train ,2, mean)
```

```
X.train.sd=apply(X.train,2,sd)
X.test=X[test,]
for(i in 1:dim(X.test)[2]){
  for(j in 1:dim(X.test)[1]){
    X.test[j,i]<-(X.test[j,i]-X.train.mean[i])/X.train.sd[i]
  }
}

A<-MakeStandardized(Y.train,X.train)
Y.train<-A[[1]]
X.train<-A[[2]]

#Lasso regression using the glmnet package
library(glmnet)

cv.lasso <- cv.glmnet(X.train,Y.train, alpha = 1,
standardize=FALSE) #k-flod crossvalidation to get lambda
cv.lasso$lambda.min #gives smallest crossvalidation error
#running the lasso with alpha=1 with
 cv.lasso$lambda.min as the tunning parameter
model <- glmnet(X.train, Y.train, alpha = 1,
lambda = cv.lasso$lambda.min,standardize=FALSE)
# Display regression coefficients
(coefs<-coef(model))

#train MSE
(obj<-0.5*crossprod((X.train %*% coefs)- Y.train)/length(Y.train))

#test MSE
(obj<-0.5*crossprod((X.test %*% coefs)- Y.test)/length(Y.test))

#Ridge regression using the glmnet package

cv.ridge <- cv.glmnet(X.train,Y.train, alpha = 0,standardize=FALSE) #k-fold
 crossvalidation to get lambda
cv.ridge$lambda.min #gives smallest crossvalidation error
#running the lasso with alpha=1 with
cv.lasso$lambda.min as the tunning parameter
model <- glmnet(X.train, Y.train,alpha = 0,lambda = cv.ridge$lambda.min,
standardize=FALSE)
# Display regression coefficients
(coefs<-coef(model))

#train MSE
(obj<-0.5*crossprod((X.train %*% coefs)- Y.train)/length(Y.train))

#test MSE
(obj<-0.5*crossprod((X.test %*% coefs)- Y.test)/length(Y.test))

#Elastic-net regression using the glmnet package
cv.elastic <- cv.glmnet(X.train,Y.train,
 alpha = 0.5,standardize=FALSE) #k-flod
 crossvalidation to get lambda
cv.elastic$lambda.min #gives smallest
 crossvalidation error
model <- glmnet(X.train, Y.train,
 alpha = 0.5,lambda = cv.elastic$lambda.min,standardize=FALSE)
# Display regression coefficients
```

```
(coefs<-coef(model))

#train MSE
(obj<-0.5*crossprod((X.train %*% coefs)- Y.train)/length(Y.train))

#test MSE
(obj<-0.5*crossprod((X.test %*% coefs)- Y.test)/length(Y.test))

#Define the generalized LASSO regression model with beta-Laplace priors
minimize.lasso2 <- function(par, X, y, lambda){
  activate=4
  sig1=1
  a=2
  b=2
  penalty=0
  rss <- 0.5*crossprod((X %*% par) - y)/length(y)#Residual sum of squares
  ones<-rep(1,times=length(par))
  constraint1<-(a-1)*log(0.5+0.5*sign(par)
*(1-exp(-abs(par)/sig1)))#cdf of laplace
  constraint2<-(b-1)*log(0.5-0.5*sign(par)
*(1-exp(-abs(par)/sig1)))#survival of Laplace
  if(activate==0){
    penalty<-lambda * (abs(par) %*% ones)/sig1 #Original penalty term from lasso
  }
  if(activate==1){
    penalty <- lambda * (constraint2 %*% ones) #Method 1
  }
  if(activate==2){
    penalty <- lambda * (constraint1 %*% ones) #Method 2
  }
  if(activate==3){
    penalty <- lambda * (constraint1 %*% ones+constraint2
%*% ones) #Method 3
  }
  if(activate==4){
    penalty <- lambda * ((abs(par) %*% ones)/sig1+constraint2
%*% ones)#Method 4
  }
  if(activate==5){
    penalty <- lambda * ((abs(par) %*% ones)/sig1+constraint1
%*% ones) #Method 5
  }
  if(activate==6){
    penalty <- lambda * ((abs(par) %*% ones)/sig1+constraint1
%*% ones+constraint2 %*% ones) #Method 6
  }
  return(rss + penalty)
}

# Method 1: Survival function of Laplace
 distribution as the only constraint.
# Method 2: CDF of Laplace distribution
as the only constraint.
# Method 3: Survival function and CDF of
Laplace distribution as the only constraints.
# Method 4: L1 norm and Survival function of
 Laplace as constraints.
# Method 5: L1 norm and CDF of Laplace
```

```
as constraints.
# Method 6: L1 norm, Survival function,
 and CDF of Laplace as constraints

#used a grid search for a tunning
parameter: 10-fold crossvalidation
n=100
p=dim(X.train)[2]
grid = cv.lasso$lambda.min^seq (3,-3, length = n)

k=10
folds=sample (1:k,nrow(X.train),replace =TRUE)
cv.errors =matrix (NA ,k,n, dimnames
 =list(NULL , paste (1:n) ))
op.converge=matrix (NA ,k,n, dimnames
 =list(NULL , paste (1:n) ))
#par.list=matrix (NA ,n,p, dimnames
=list(NULL , paste (1:p) ))

for(j in 1:k){
  for(i in 1:n) {
    Y.train.cross<-Y.train[folds !=j]
    Y.train.mean.cross=mean(Y.train.cross)
    Y.test.cross=Y.train[folds ==j]-
Y.train.mean.cross

    X.train.cross=X.train[folds !=j,]
    X.train.mean.cross=apply(X.train.cross ,2, mean)
    X.train.sd.cross=apply(X.train.cross,2,sd)
    X.test.cross=X.train[folds ==j,]
    for(r in 1:dim(X.test.cross)[2]){
      for(s in 1:dim(X.test.cross)[1]){
        X.test.cross[s,r]<-(X.test.cross[s,r]-
X.train.mean.cross[r])/X.train.sd.cross[r]
      }
    }
    A<-MakeStandardized(Y.train.cross,X.train.cross)
    #options(show.error.messages = FALSE)
    err<-try(
      {
        op.result <- optim(rep(0, p),
 fn = minimize.lasso2, method = 'Nelder-Mead',
 X =A[[2]] , y =A[[1]] , lambda =grid[i],
control=list(maxit=10000) )
      }
    )
    if(class(err)!="try-error"){
      op.cf<-op.result$par
      # for(t in 1:p){
      #   par.list[i,t]<-op.cf[t]
      # }
      op.converge[j,i]=op.result$convergence
      if(op.result$convergence!=0){
        cv.errors [j,i]=NA
      }
      else{
        cv.errors [j,i]=crossprod(
(X.test.cross %*% op.cf) - Y.test.cross)/length(Y.test.cross)
```

```
      }
    }
    else{
      print("sorry")
      i=i+1
    }
  }
}

#op.converge
max(op.converge)

#cv.errors
mean.cv.errors =apply(cv.errors ,2, mean,na.rm=TRUE)
#mean.cv.errors
min(mean.cv.errors,na.rm = TRUE)
(ind=which(mean.cv.errors==min
(mean.cv.errors,na.rm = TRUE),TRUE))
grid[ind]

op.result <- optim(rep(0, p),
fn = minimize.lasso2, method= 'BFGS',
 control=list(maxit=10000), X = X.train,
 y = Y.train, lambda=grid[ind])
op.result$convergence
(op.cf <- op.result$par)
op.result$value

#train MSE
(obj<-0.5*crossprod((X.train %*% op.cf)- Y.train)/length(Y.train))

#test MSE
(obj<-0.5*crossprod((X.test %*% op.cf)- Y.test)/length(Y.test))
```


Affiliation:

Gayan Warahena-Liyanage
Department of Mathematics
University of Dayton
Dayton, OH, 45469, USA
E-mail: gwarahenaliyanage1@udayton.edu

Felix Famoye
Department of Statistics, Actuarial and Data Sciences
Central Michigan University
Mt.Pleasant, MI, 48859, USA

Carl Lee
Department of Statistics, Actuarial and Data Sciences
Central Michigan University
Mt.Pleasant, MI, 48859, USA

Austrian Journal of Statistics
published by the Austrian Society of Statistics
http://www.ajs.or.at/
http://www.osg.or.at/

Volume 52
July 2023

Submitted: 2021-12-16
Accepted: 2022-07-26