Communications for Statistical Applications and Methods
2022, Vol. 29, No. 6, 641–653

https://doi.org/10.29220/CSAM.2022.29.6.641
Print ISSN 2287-7843 / Online ISSN 2383-4757

Penalized maximum likelihood estimation with symmetric log-concave errors and LASSO penalty

Seo-Young Park[a], Sunyul Kim[b], Byungtae Seo[1,a]

[a]Department of Statistics, Sungkyunkwan University, Korea;
[b]Samsung Electronics Co., LTD, Korea

Abstract

Penalized least squares methods are important tools to simultaneously select variables and estimate parameters in linear regression. The penalized maximum likelihood can also be used for the same purpose assuming that the error distribution falls in a certain parametric family of distributions. However, the use of a certain parametric family can suffer a misspecification problem which undermines the estimation accuracy. To give sufficient flexibility to the error distribution, we propose to use the symmetric log-concave error distribution with LASSO penalty. A feasible algorithm to estimate both nonparametric and parametric components in the proposed model is provided. Some numerical studies are also presented showing that the proposed method produces more efficient estimators than some existing methods with similar variable selection performance.

Keywords: log-concave distribution, variable selection, LASSO, linear programming, CNM

1. Introduction

Linear regression is a simple and insightful tool to investigate the relationship between variables. In the presence of a large number of predictors, however, the model may produce inefficient parameter estimators and suffer multicollinearity problems. In this case, it is desirable to select a smaller subset most related to the response, for a concise and interpretable model. Penalized linear regression methods have gained popularity over the last two decades for this purpose, which adopt a certain penalty function to the objective function of the problem.

Suppose that we have a random sample $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$ where $\boldsymbol{x}_i$ is a p-dimensional covariate and y_i is univariate continuous response variable. A general objective function for the penalized linear regression can be expressed as

$$\sum_{i=1}^{n} g\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right) + n \sum_{j=1}^{p} p_\lambda\left(\beta_j\right),$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of unknown regression coefficients, $g(\cdot)$ is a loss function, and $p_\lambda(\cdot)$ is a penalty function with tuning parameter λ. The penalized regression estimator is then defined as the minimizer of this objective function. A different choice of g and p_λ leads to a different

The work of Byungtae Seo was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A2C1006462).

[1] Corresponding author: Department of Statistics, Sungkyunkwan University, 5-2, Sungkyunkwan street, Jongno-Gu, Seoul 03063, Korea. E-mail: seobt@skku.edu

Published 30 November 2022 / journal homepage: http://csam.or.kr

penalized method. The least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) estimator can be obtained with $p_\lambda(a) = \lambda|a|$. Additional for LASSO, the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), adaptive LASSO (Zou, 2006), and minimax concave penalty (MCP) (Zhang, 2010) have been proposed for p_λ and known to enjoy the oracle property under certain conditions.

For the choice of the loss function, when L_2 loss function is used for $g(\cdot)$, a penalized least squares estimator (PLSE) can be obtained. If the loss function is the Huber's ψ function (Fan and Li, 2001), we can have a penalized robust estimator. The penalized maximum likelihood estimator (PMLE) can be obtained if $g(\cdot)$ is the minus log-density of the error distribution. These methods enable us to jointly estimate regression coefficients and select variables at the same time with the penalty function such as LASSO, SCAD, and MCP. By this means, sacrificing a little bias, we can gain the efficiency of the estimator and improve the overall prediction accuracy.

It is well known that the PLSE and PMLE are identical under the normality of the error. However, the normality of the error is often too strong in practice. For example, it is common that the error distribution has a heavier tail than that of the normal distribution. In this case, the PLSE or the PMLE under the normality can produce biased and inefficient estimators. For the PMLE, it is essential that the error distribution is correctly specified. If not, the PMLE is generally less efficient than PLSE and it could result in a serious bias. To this end, some robust and flexible parametric distributions such as student-t or Laplace distributions are proposed for the error distribution (Lange *et al.*, 1989; Rousseeuw and Leroy, 1987). However, it is still restrictive because those families of distributions cannot cover a wide variety of distributions and this could result in inaccurate parameter estimation. In this paper, to relax such misspecification problems, we propose to use the log-concave error distribution which is a nonparametric family of distributions with a specific shape constraint.

The rest of this paper is organized as follows. In Section 2, we define the notation and introduce the penalized linear regression models with symmetric log-concave errors. In Section 3, we propose using a smoothed log-concave maximum likelihood estimator (Chen and Samworth, 2013; Kim and Seo, 2018) with a Gaussian kernel for the estimation of the initial regression coefficient. Numerical simulations and real data studies are also conducted to compare the performance of the proposed method with other existing methods in Section 4. We end this paper with some concluding remarks in Section 5.

2. Penalized maximum likelihood method with symmetric log-concave error densities

2.1. Log-concave error distribution in linear regression

The family of log-concave densities is

$$\mathcal{LC} := \left\{ f(t) = e^{\varphi(t)} \mid \varphi(\cdot) \in C, \int e^{\varphi(t)} dt = 1 \right\},$$

where $C := \{\varphi : \mathbb{R} \to (-\infty, \infty) \mid \varphi \text{ is a concave and closed function on } \mathbb{R}\}$. It is convenient to think of log-concave densities as unimodal densities with fairly light tail which is exponentially decaying. The family of log-concave densities includes most of the commonly-used parametric distributions but not limited to uniform, normal, Gamma(r, λ) for $r \geq 1$, Beta(α, β) for $\alpha, \beta \geq 1$, logistic, and Laplace (see. e.g., (Bagnoli and Bergstrom, 2005)).

Many favorable properties of log-concave densities have been studied by (Dümbgen and Rufibach, 2009), (Balabdaoui *et al.*, 2009), and (Dharmadhikari and Joag-Dev, 1988). For example, it is well-

known that convolution, linear transformation, and product measure of log-concave distributions preserve log-concavity. (Dümbgen and Rufibach, 2009) showed that the nonparametric maximum likelihood estimator (NPMLE) of φ uniquely exists and is a piecewise linear concave function whose knots are constrained to the subset of the input data points. In addition, $t_{(1)}$ and $t_{(n)}$ are always treated as knots, and the NPMLE $\hat{\varphi}$ takes the value $-\infty$ on the outside of the effective domains $[t_{(1)}, t_{(n)}]$ where $t_{(1)} \leq \cdots \leq t_{(n)}$ are the order statistics of the sample.

There are several studies that used this log-concave distribution as an error distribution in regression settings. (Dümbgen *et al.*, 2011) showed that the linear regression estimator with the log-concave error distribution is consistent. (Hu *et al.*, 2017) applied the finite mixture of regressions with each component having a log-concave error density and gained robustness by adopting the idea of least trimmed squares. (Kim and Seo, 2018) compared the performance of estimators when the error distribution is assumed as Gaussian scale mixture and log-concave densities based on numerical studies. (Kim and Seo, 2021) proposed a modal linear regression assuming that the error distribution is log-concave.

2.2. Penalized likelihood under symmetric log-concave errors

When we consider the mean regression, our main interest is to represent the expected response as a linear function of covariates. In this case, it is reasonable to assume that the error distribution is symmetric. Because $\mathcal{LC}$ contains both symmetric and asymmetric densities, to properly use the log-concave distribution for the error distribution in the mean regression, we further restrict $\mathcal{LC}$ to the symmetric subfamily. Let us define

$$\mathcal{SC} := \{\varphi(-t) = \varphi(t) \quad \text{for all} \quad t \in \mathbb{R} \mid \varphi(\cdot) \in C\}.$$

Then, the family of symmetric log-concave distributions can be expressed as

$$\mathcal{SLC} := \left\{ f(t) = e^{\varphi(t)} \mid \varphi(\cdot) \in \mathcal{SC}, \int e^{\varphi(t)} dt = 1 \right\}.$$

Now, in the linear regression model,

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n,$$

we assume that the probability density function (pdf) of ϵ is a member of $\mathcal{SLC}$ which contains all log-concave densities symmetric at zero. Then, the log-likelihood function is

$$\ell_n(\boldsymbol{\beta}, \varphi) = \frac{1}{n} \sum_{i=1}^{n} \varphi\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right), \tag{2.1}$$

where $\varphi \in \mathcal{SC}$ and $\int e^{\varphi(t)} dt = 1$.

An equivalent expression of (2.1) without constraint $\int e^{\varphi(t)} dt = 1$ is

$$\ell_n(\boldsymbol{\beta}, \varphi) = \frac{1}{n} \sum_{i=1}^{n} \varphi\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right) - \int e^{\varphi(t)} dt, \tag{2.2}$$

where $\varphi \in \mathcal{SC}$ and $\int e^{\varphi(t)} dt$ is from the normalization constant involved in the density estimation (Silverman, 1982; Dümbgen and Rufibach, 2009). Note that the estimated symmetric log-concave density $\hat{f}$ is just $e^{\hat{\varphi}}$ where $\hat{\varphi}$ is the maximizer of (2.2).

Before we incorporate LASSO penalty into the log-likelihood function (2.2), without loss of generality, we assume that the responses are centered, so the intercept term is not included in the regression function. Additionally, we also standardize the predictors so that each covariate is centered and has unit variance. Then, the PMLE with the LASSO penalty can be defined as the maximizer of the penalized log-likelihood function

$$\ell_n^*(\boldsymbol{\beta}, \varphi) = \ell_n(\boldsymbol{\beta}, \varphi) - \lambda \|\boldsymbol{\beta}\|_1, \tag{2.3}$$

where $\|\cdot\|_p$ is L_p-norm and $\lambda \geq 0$. Tuning parameter λ controls the complexity of the model. As the tuning parameter increases, it continuously shrinks the coefficients toward 0, and some coefficients shrink to exact 0 with sufficiently large enough λ.

2.3. Estimation

The estimation problem under (2.3) involves the estimation of both $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\varphi \in \mathcal{SC}$. Since there is no explicit way to maximize both parametric and nonparametric components simultaneously, we propose an iterative procedure that alternatively updates $\boldsymbol{\beta}$ and φ in (2.3). The PMLE of $\boldsymbol{\beta}$ and NPMLE of φ can be obtained by iterating the following two procedures until they converge.

First, for fixed $\boldsymbol{\beta}$, NPMLE of φ in (2.3) is equivalent to the NPMLE of φ without the penalty term as in (2.2). Hence, the NPMLE of φ is

$$\hat{\varphi} = \underset{\varphi \in \mathcal{SC}}{\operatorname{argmax}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \varphi\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right) - \int e^{\varphi(t)} dt \right\}. \tag{2.4}$$

Maximization problem (2.4) can be equivalently converted to

$$\hat{\varphi} = \underset{\tilde{\varphi} \in C}{\operatorname{argmax}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \tilde{\varphi}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right) - \int e^{\tilde{\varphi}(t)} dt \right\}, \tag{2.5}$$

where $\tilde{\varphi}(t) = \varphi(t) + \varphi(-t)$. (Dümbgen *et al.*, 2007) introduced an efficient algorithm based on an active set method, and this algorithm is available within R package `logcondens` (Dümbgen and Rufibach, 2011). Although this algorithm is designed to find the NPMLE of φ without symmetric constraint, same algorithm can be used by turning (2.4) into (2.5).

Recently, (Liu and Wang, 2018) proposed a much faster algorithm known as the constrained newton method (CNM) for the log-concave density estimation without the symmetric constraint. From the piecewise linearity of the NPMLE $\hat{\varphi}$, they reparametrized φ in (2.4) where $\varphi \in C$ as

$$\hat{\varphi} = \underset{\alpha, \pi, \delta}{\operatorname{argmax}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\alpha \epsilon_i - \sum_{j=1}^{m} \pi_j \left(\epsilon_i - \delta_j\right)_+ \right) - \int e^{\varphi(t)} dt \right\}, \tag{2.6}$$

where $\alpha \in \mathbb{R}$, $\pi = (\pi_1, \ldots, \pi_m)^\top$, $\delta = (\delta_1, \ldots, \delta_m)^\top$, and $\epsilon_i = y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}$. Notation a_+ means a if $a \geq 0$ and zero otherwise. δ_j's are knots, and $\alpha - \sum_{k=1}^{j} \pi_k$ with a positive π_k is the slope between two knots δ_j and δ_{j+1} for $j = 1, \ldots, m-1$. In CNM, the directional derivative of (2.6) is used to find new knots that increase the likelihood. Changes in the slope at each knot are computed by the quadratically convergent method and the knots with zero changes are removed. To ensure a monotone yet sufficiently large increase of the log-likelihood, line search by Armijo's rule is added for each iteration of the algorithm.

Second, for fixed φ, we maximize (2.3) with respect to $\boldsymbol{\beta}$ with fixed tuning parameter λ. Since the NPMLE $\hat{\varphi}$ is a piecewise linear function whose number of knots is smaller than the number of observations, the sum of piecewise linear concave functions of $\boldsymbol{\beta}$ can be expressed as the minimum of multiple affine functions (Kim and Seo, 2021). We let

$$\hat{\varphi}(t) = \min_{k=1,\ldots,q-1} (a_k t + b_k),$$

where a_k is the slope and b_k is the intercept between k^{th} and $(k+1)^{\text{th}}$ knots whose total number is q. L_1 norm of p-dimensional vector $\boldsymbol{\beta}$ is also a piecewise linear function as $\lambda\|\boldsymbol{\beta}\|_1 = \lambda \sum_{j=1}^{p} |\beta_j| = \lambda \sum_{j=1}^{p} \max(\beta_j, -\beta_j)$ and putting the negative sign turns the maximization problem to the minimization problem. Then the PMLE of $\boldsymbol{\beta}$ with fixed $\hat{\varphi}$ can be calculated as follows:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ \min_{\substack{k^{(i)}=1,\ldots,q-1 \\ (i=1,\ldots,n)}} \sum_{i=1}^{n} \left(-\tilde{a}_{k^{(i)}} \boldsymbol{x}_i^\top \boldsymbol{\beta} + \tilde{a}_{k^{(i)}} y_i + \tilde{b}_{k^{(i)}} \right) + \lambda \sum_{j=1}^{p} \min\left(\beta_j, -\beta_j\right) \right\}, \tag{2.7}$$

where $\tilde{a}_k = a_k/n$ and $\tilde{b}_k = b_k/n$.

We use the linear programming (LP) to solve the maximization problem. The equivalent LP problem for (2.7) is

$$\begin{cases} \underset{\boldsymbol{\beta}}{\operatorname{argmax}} & \mathbf{1}_n^\top \mathbf{u} + \mathbf{1}_p^\top \mathbf{v} \\ \text{subject to} & \begin{cases} \tilde{a}_k(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}) + \tilde{b}_k \geq u_i, & i = 1,\ldots,n \text{ and } k = 1,\ldots,q-1; \\ \lambda\beta_j \geq v_j, & j = 1,\ldots,p; \\ -\lambda\beta_j \geq v_j, & j = 1,\ldots,p, \end{cases} \end{cases} \tag{2.8}$$

where $\mathbf{u} = (u_1,\ldots,u_n)^\top$, $\mathbf{v} = (v_1,\ldots,v_p)^\top$, and $\mathbf{1}_m \in \mathbb{R}^m$ is a vector whose elements are all 1. Suppose $\mathbf{0}_{p\times q}$ is a $p \times q$ matrix and $\mathbf{0}_p$ is a p-dimensional vector whose elements are all 0. Let $\mathbf{a} = (\tilde{a}_1,\ldots,\tilde{a}_{q-1})^\top$ and $\mathbf{b} = (\tilde{b}_1,\ldots,\tilde{b}_{q-1})^\top$ be the normalized slope vector and the normalized intercept vector, respectively. Then, we can reexpress (2.8) as

$$\text{maximize } \mathbf{c}\theta \text{ subject to } Q\theta \preceq \mathbf{w}, \tag{2.9}$$

where $\theta = (\boldsymbol{\beta}^\top, \mathbf{u}^\top, \mathbf{v}^\top)^\top$, $\mathbf{c} = (\mathbf{0}_p^\top, \mathbf{1}_{n+p}^\top)$,

$$Q = \begin{bmatrix} \mathbf{X}\otimes\mathbf{a} & I_{n\times n}\otimes\mathbf{1}_{q-1} & \mathbf{0}_{n(q-1)\times p} \\ -\lambda I_{p\times p} & \mathbf{0}_{p\times n} & I_{p\times p} \\ \lambda I_{p\times p} & \mathbf{0}_{p\times n} & I_{p\times p} \end{bmatrix}, \quad \text{and } \mathbf{w} = \begin{bmatrix} \boldsymbol{y}\otimes\mathbf{a} + \mathbf{1}_n\otimes\mathbf{b} \\ \mathbf{0}_p \\ \mathbf{0}_p \end{bmatrix}.$$

In matrix Q of (2.9), $\otimes$ stands for the Kronecker product, $I_{p\times p}$ is the $p \times p$ identity matrix, $\mathbf{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^\top$, and $\boldsymbol{y} = (y_1,\ldots,y_n)^\top$. By this means, we can turn the estimation problem for $\boldsymbol{\beta}$ into the LP problem as in (2.9). To solve this LP problem, we can use the `R` package `lpSolveAPI` (Konis and Konis, 2020).

3. Choice of initial values

The estimation procedure introduced in Section 2.3 could find a local maximizer as the likelihood has multiple local modes. Hence, we need a careful initial estimator, especially for $\boldsymbol{\beta}$. (Kim and Seo,

2021) used a differential evolution algorithm to obtain an initial estimator near the global maximum of (2.2) but this algorithm is time-consuming and does not guarantee the global convergence.

As an alternative to find a suitable initial value, we propose using a smoothed log-concave maximum likelihood estimator (MLE). The smoothed version of the log-concave MLE can be constructed using the convolution of the estimated log-concave distribution and a Gaussian distribution, which preserves the log-concavity shape constraint (Prékopa and András, 1971). (Chen and Samworth, 2013) introduced the smoothed log-concave MLE as

$$\tilde{f}_{\hat{\varphi}}(t) = \int e^{\hat{\varphi}(\tau)} K_h(t-\tau)\, d\tau,$$

where $K_h(\cdot)$ is the Gaussian kernel and $h > 0$ is a smoothing parameter called bandwidth. The log-likelihood function with the smoothed log-concave estimator is then defined as

$$\tilde{\ell}_n(\boldsymbol{\beta}, \hat{\varphi}) = \sum_{i=1}^{n} \log \tilde{f}_{\hat{\varphi}}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right). \tag{3.1}$$

Now, because $\tilde{\ell}_n(\boldsymbol{\beta}, \hat{\varphi})$ is a smoother function of $\boldsymbol{\beta}$ than $\ell_n(\boldsymbol{\beta}, \hat{\varphi})$, we can stably estimate $\boldsymbol{\beta}$.

For fixed $\boldsymbol{\beta}$, we can find NPMLE of φ using CNM algorithm as in Section 2.3. With a given NPMLE $\hat{\varphi}$ which is a piecewise linear function, smoothed version of estimated log-concave distribution can be written as follows:

$$\begin{aligned}
\tilde{f}_{\hat{\varphi}}(t) &= \int e^{\hat{\varphi}(\tau)} \frac{1}{\sqrt{2\pi h^2}} e^{-\frac{1}{2h^2}(t-\tau)^2} d\tau, \\
&= \sum_{k=1}^{q-1} \int_{\delta_k}^{\delta_{k+1}} \exp(b_k + a_k \tau) \frac{1}{\sqrt{2\pi h^2}} e^{-\frac{1}{2h^2}(t-\tau)^2} d\tau, \\
&= \sum_{k=1}^{q-1} \exp\left(a_k t + \frac{h^2}{2} a_k^2 + b_k\right) \left(\Phi\left(\frac{\delta_{k+1} - t - h^2 a_k}{h}\right) - \Phi\left(\frac{\delta_k - t - h^2 a_k}{h}\right)\right),
\end{aligned}$$

where δ_k's are knots, and Φ is the CDF of the standard normal density (Kim and Seo, 2018).

For fixed φ and given h, we can update $\boldsymbol{\beta}$ through a quadratically convergent method (Liu and Wang, 2018). As smoothed log-concave estimator $\tilde{f}_{\hat{\varphi}}(\cdot)$ is twice-differential anywhere, we can compute

$$\begin{aligned}
\frac{\partial \tilde{\ell}_n(\boldsymbol{\beta}, \hat{\varphi})}{\partial \boldsymbol{\beta}} &= -\sum_{i=1}^{n} \boldsymbol{x}_i \frac{\tilde{f}'_{\hat{\varphi}}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)}{\tilde{f}_{\hat{\varphi}}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)} = \mathbf{S} \text{ and} \\
\frac{\partial^2 \tilde{\ell}_n(\boldsymbol{\beta}, \hat{\varphi})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} &= \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \frac{\tilde{f}''_{\hat{\varphi}}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right) \tilde{f}_{\hat{\varphi}}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right) - \tilde{f}'_{\hat{\varphi}}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2}{\tilde{f}_{\hat{\varphi}}\left(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\right)^2} = \mathbf{H},
\end{aligned}$$

where $\tilde{f}'_{\hat{\varphi}}(t) = d/dt \tilde{f}_{\hat{\varphi}}(t)$ and $\tilde{f}''_{\hat{\varphi}}(t) = d^2/dt^2 \tilde{f}_{\hat{\varphi}}(t)$. If we let $\boldsymbol{\beta}^*$ be an update of the vector $\boldsymbol{\beta}$ and the maximization problem of $\tilde{\ell}_n(\boldsymbol{\beta}, \hat{\varphi})$ in (3.1) can be replaced approximately as

$$\min_{\boldsymbol{\beta}^*} \left\| \mathbf{R}\boldsymbol{\beta}^* - \mathbf{R}\boldsymbol{\beta} - \mathbf{R}^{-T}\mathbf{S} \right\|^2,$$

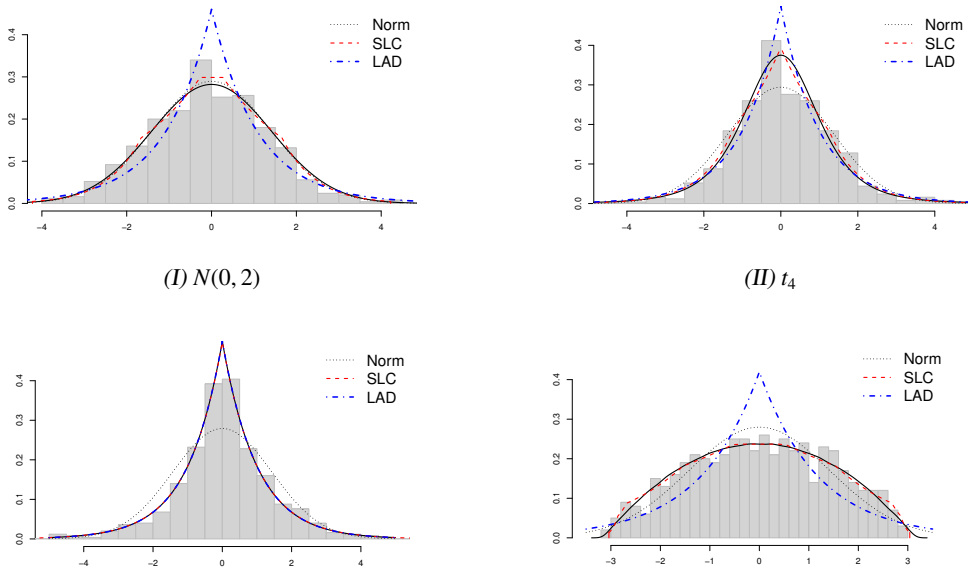

Figure 1: *Estimated error densities based on one simulated sample of size* $n = 500$ *for (I)–(IV).*

where $\|\cdot\|$ denotes the L_2-norm and $\mathbf{R}^{-T} = (\mathbf{R}^\top)^{-1}$. $\mathbf{R}$ is calculated from the spectral decomposition of $\mathbf{H}$ as $\mathbf{H} = -\mathbf{P}\Lambda\mathbf{P}^\top = -\mathbf{R}^\top\mathbf{R}$. Hence, $\mathbf{R} = \Lambda^{1/2}\mathbf{P}^\top$ and the Moore-Penrose generalized inverse $\mathbf{P}\Lambda^{-1/2}$ is used in place of $\mathbf{R}^{-1}$. This problem can be solved by the function `lsi` in R package `lsei` (Wang *et al.*, 2020).

To choose the bandwidth h which determines the degree of smoothing, (Chen and Samworth, 2013) proposed $\hat{h}^2 = \hat{\Sigma} - \tilde{\Sigma}$ where $\hat{\Sigma} = (n-1)^{-1}\sum_{i=1}^n (t_i - \bar{t})^2$ and $\tilde{\Sigma} = \int (t-\bar{t})^2 e^{\hat{\varphi}(t)} dt$. This choice makes the sample variance and estimated variance from $\tilde{f}_{\hat{\varphi}}$ identical. For our purpose, however, this choice produces too small h^2 to produce a sufficiently smooth likelihood. Empirically, we find that the result is not that sensitive to initial estimators by multiplying constant 3 to $\hat{h}^2$.

4. Numerical examples

4.1. Simulation studies

We conduct simulation studies to evaluate the finite-sample performance of the proposed estimator (SLC) comparing with other existing methods such as penalized least squares (NORM) (Tibshirani, 1996) and penalized least absolute deviation (LAD) (Wang *et al.*, 2007) regression with LASSO as the penalty function. NORM can be obtained through R package `glmnet` (Friedman *et al.*, 2010), and LAD can be implemented through R package rqPen (Sherwood and Maidman, 2020).

Selection of the tuning parameter λ is of paramount importance in penalized estimation. When λ is too small, there will be too many variables included in the final model and the variance of estimators would increase. Conversely, the selected model is too simple and has a large model bias in case of selecting too large λ. So it is crucial to keep the balance in selecting the value of λ. Cross-validation is a very common method, but it is time-consuming. We apply the Bayesian information criterion (BIC)

(Schwarz, 1978) to choose the tuning parameter in this article. The formula of BIC is

$$\text{BIC} = -2\,\ell_n\,(\boldsymbol{\beta}, \varphi) + d_\lambda \frac{\log\,(n)}{n},$$

where d_λ is the number of nonzero regression coefficients. We also apply BIC to select the tuning parameter λ for NORM and LAD.

For comparison, we consider the following four different symmetric error distributions: (I) $N(0, 2)$, (II) t_4, (III) Laplace(0,1), and (IV) $\sqrt{40}\{\text{Beta}(2, 2) - 0.5\}$. Figure 1 shows the estimated error distributions along with the probability histogram of one simulated set of errors for normal, t, Laplace, and Beta distributions. In this figure, the black solid line represents the true error pdf. The black dotted line shows the estimated error density function based on NORM, red dashed for SLC, and blue dot-dashed for LAD. Estimated distributions from both NORM and SLC are close to the true density when the error is normal. SLC and LAD are close to the true density for the case of Laplace distribution. For the Beta distribution which has bounded support, SLC provides a better distribution estimator than those of the other methods. It seems that the proposed method provides quite reasonable estimated error distributions for all four cases.

For simulation, we generate samples from the model $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \epsilon_i$ for $i = 1, \ldots, n$ where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^\top$. We generate covariates $\boldsymbol{x}_i$ from a multivariate normal distribution $N(\mathbf{0}, \Omega)$ where $(j, k)^{th}$ element of Ω is $\Omega_{jk} = 0.5^{|j-k|}$ for $1 \le j, k \le p$. We repeat the simulation 200 times. We calculate the initial parameter for $\boldsymbol{\beta}$ from the linear regression with the smoothed version of estimated log-concave density as described in Section 3. For each error distribution, Table 1 shows the median of model errors over 200 simulated datasets with $n = 100$, 250, and 500. The model error is defined by $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top E(\mathbf{X}^\top \mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ (Tibshirani, 1996). The median absolute deviation (MAD) (Fan and Li, 2001) of the model errors over 200 simulated datasets is also summarized in Table 1. The average number of 0 coefficients is also reported restricted only to the true zero coefficients. We omit the average number of zero coefficients erroneously set to 0 as there was none. The tuning parameter λ is selected for each simulated sample.

From Table 1, with the increase of the sample size, the median and MAD of the model error decrease in all four settings. It is natural to expect that NORM is the best for Case (I) and LAD is the best for Case (III). For Case (I), SLC has a comparable performance with NORM when the sample size is large enough. For Case (III), when the sample size is large, SLC has a smaller model error than LAD. It is because estimated φ is a piecewise function that SLC could estimate Laplace well for a large sample size. Even with t_4 (Case (II)), SLC provides the best performance in the view of the median of the model error even though the t-distribution does not belong to the family of log-concave distributions. NORM generally works well with finding exact zero coefficients. SLC takes second place in variable selection performance except for Case (III) with $n = 250$ and 500 with a slight difference. SLC works well in Case (II) and (IV) relative to other cases.

4.2. Real data

In this subsection, we apply the proposed method to two real datasets: Ozone data and Plasma Beta-Carotene Level data. For these datasets, we consider the linear regression model and apply NORM, SLC, LAD, and OLS to analyze the datasets. To estimate the standard errors of estimated regression coefficients from NORM, SLC, and LAD, we use the bootstrap method based on 1,000 bootstrap samples.

Example 1. The Ozone data, originally used in (Breiman and Friedman, 1985), consists of the daily maximum one-hour-average ozone reading and eight meteorological variables in Los Angeles

Table 1: Result with ME × 100 for n = 100, 250, and 500

n	Error	Method	Number of true zeros	Model error	
				Median	MAD
100	(I)	NORM	3.915	11.635	4.876
		SLC	3.755	13.305	5.945
		LAD	2.980	16.435	6.749
	(II)	NORM	4.105	12.511	6.019
		SLC	3.855	10.370	4.521
		LAD	3.065	11.205	4.694
	(III)	NORM	4.030	10.056	4.599
		SLC	3.930	6.239	3.521
		LAD	3.200	6.959	3.545
	(IV)	NORM	4.025	9.898	4.489
		SLC	3.615	10.412	5.278
		LAD	2.935	21.914	10.578
250	(I)	NORM	4.115	4.741	2.347
		SLC	4.025	5.063	2.099
		LAD	3.745	6.105	2.927
	(II)	NORM	4.270	4.664	2.468
		SLC	4.125	3.597	1.622
		LAD	3.985	3.655	1.756
	(III)	NORM	4.045	4.749	2.068
		SLC	4.000	2.860	1.316
		LAD	4.025	2.762	1.514
	(IV)	NORM	4.215	4.637	1.980
		SLC	3.955	3.363	1.701
		LAD	3.685	8.453	3.846
500	(I)	NORM	4.250	2.209	1.097
		SLC	4.145	2.468	1.272
		LAD	3.940	3.322	1.664
	(II)	NORM	4.190	2.543	1.196
		SLC	4.180	1.916	0.974
		LAD	4.070	1.993	0.913
	(III)	NORM	4.260	2.514	1.178
		SLC	4.140	1.245	0.609
		LAD	4.150	1.265	0.645
	(IV)	NORM	4.150	2.602	1.135
		SLC	3.855	1.727	0.859
		LAD	3.755	4.458	2.084

for 330 days in 1976. This data has been used in various studies. The goal is to model the relationship between the daily maximum one-hour-average ozone reading (ozone) and the other eight variables: temperature (temp), inverse base height (invHt), pressure gradient (press), visibility (vis), 500 millibar pressure height (milPress), humidity (hum), inversion base temperature (invTemp), and wind speed (wind). This data is publicly available in R package `cosso` (Zhang and Lin, 2013). We plot a histogram of ozone in Figure 2 which indicates that the distribution of ozone is seriously skewed. We use the log-transformed ozone to release its skewness.

Table 2 compares the estimated results of the regression coefficients from NORM, SLC, and LAD besides the OLS method. The standard errors are given in their corresponding parentheses. All four methods provide quite similar results based on the selected variables and signs of the significant variables. SLC provides the most sparse model as they select temp, invHT, vis, and hum as the significant variables in the model. NORM and LAD select five variables: temp, invHT, press, vis, and hum. Figure 3 shows the correlation between the standardized covariates. MillPress and invTemp

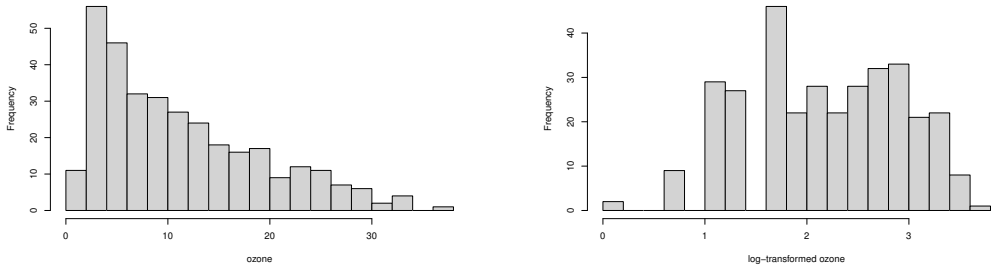

Figure 2: *The histograms of ozone and log-transformed ozone.*

Table 2: Estimated parameters and standard errors from the Ozone Data

Variable	Method			
	NORM	SLC	LAD	OLS
temp	0.551 (0.068)	0.534 (0.067)	0.556 (0.087)	0.574 (0.089)
invHt	−0.268 (0.047)	−0.232 (0.051)	−0.240 (0.061)	−0.285 (0.063)
press	0.044 (0.041)	0 (-)	0.023 (0.054)	0.052 (0.050)
vis	−0.056 (0.035)	−0.040 (0.031)	−0.028 (0.050)	−0.072 (0.037)
milPress	0 (-)	0 (-)	0 (-)	−0.024 (0.070)
hum	0.110 (0.047)	0.126 (0.054)	0.218 (0.051)	0.103 (0.046)
invTemp	0 (-)	0 (-)	0 (-)	0.002 (0.129)
wind	0 (-)	0 (-)	0 (-)	0.017 (0.034)

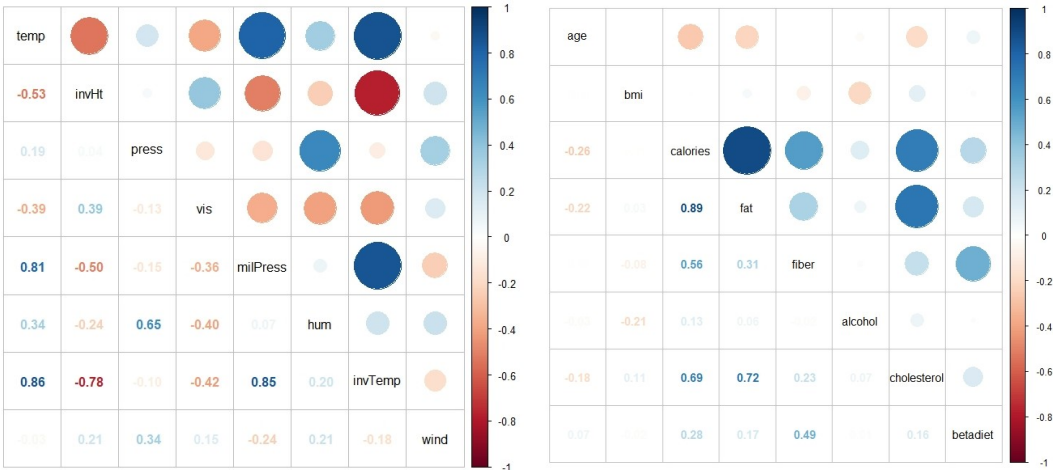

Figure 3: *The correlation plot between standardized covariates for Ozone and Plasma Beta-Carotene Level data.*

have serious correlations with other variables, and it seems that they are removed after the variable selection as other variables could explain these two variables.

Example 2. The Plasma Beta-Carotene Level dataset collected by a cross-sectional study (Nierenberg *et al.*, 1989) is available on `http://lib.stat.cmu.edu/datasets/Plasma_Retinol`. This dataset consists of 315 samples and in this study, we only use 273 female patients. Our interest is in the relationship between the plasma beta-carotene level (betaplasma) and the following 8 covariates: age,

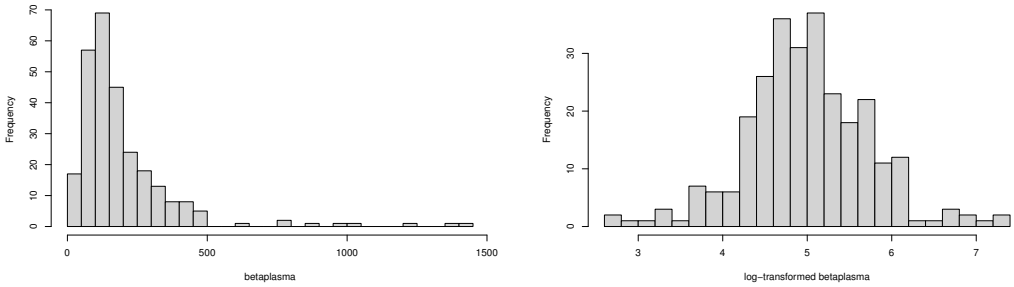

Figure 4: *The histograms of betaplasma and log-transformed betaplasma.*

Table 3: Estimated parameters and standard errors from the Plasma Beta-Carotene Level data

Variable	Method			
	NORM	SLC	LAD	OLS
age	0.114 (0.056)	0.168 (0.055)	0.161 (0.060)	0.135 (0.059)
bmi	−0.236 (0.057)	−0.265 (0.053)	−0.245 (0.072)	−0.265 (0.057)
calories	0 (-)	−0.106 (0.078)	−0.060 (0.069)	−0.085 (0.174)
fat	−0.069 (0.052)	0 (-)	0 (-)	−0.049 (0.148)
fiber	0.176 (0.074)	0.203 (0.067)	0.139 (0.071)	0.246 (0.086)
alcohol	0 (-)	0 (-)	0 (-)	0.030 (0.059)
cholesterol	0 (-)	0 (-)	0 (-)	−0.013 (0.082)
betadiet	0.100 (0.062)	0.177 (0.065)	0.248 (0.075)	0.128 (0.063)

quetelet index (bmi), number of calories (calories), grams of fat (fat), grams of fiber (fiber), number of alcoholic drinks per week (alcohol), cholesterol and dietary beta-carotene (betadiet). Figure 4 displays a histogram of betaplasma whose distribution is quite skewed with the unusual points. To release its skewness, we use the log-transformation on betaplasma. Before taking the log scale, we remove the 217^{th} observation with the zero value which would lead to negative-infimum in log-transformation.

Signs of the significant variables are all matched among the comparison models, but selected variables and corresponding coefficients are slightly different. Table 3 reveals that SLC and LAD discard fat, alcohol, and cholesterol but NORM discards calories instead of fat in the model. Figure 3 shows that there is a serious correlation between calories, fat, and cholesterol. After variable selection, variables among these three have been discarded.

5. Conclusion

Although assuming normality in the mean linear regression is common and mostly used, the estimation may not be reliable when the true error distribution has a heavy tail or unique shape form. To relax such misspecification problem, in this article, we studied the estimation of the regression coefficients and error distributions using log-concave densities with L_1 penalty on the regression coefficients. We used the CNM algorithm to nonparametrically estimate the log-concave error distribution, and the estimation of regression coefficients was derived by turning the maximization problem into the LP problem with constraints. To obtain a suitable initial value, we also developed a method using a smoothed log-concave estimator to stably estimate regression parameters from a smooth likelihood. Although this method is helpful in greatly reducing the computing time and stably searching the global maximizer of the likelihood, there is no guarantee that the estimator is the global maximizer. Further

study is required to improve the computational accuracy and efficiency.

From our simulation studies, the proposed method showed comparable performance with existing methods that used correctly specified error distributions. When the true error distribution is different from the specified error distribution, the proposed method showed much better performance than existing methods. Even when the true distribution was not log-concave, the proposed method also showed better performance than existing methods. This is because many distributions which are not log-concave can be closely approximated by the log-concave distributions.

Although we only considered the LASSO penalty to the likelihood, the proposed method can be used with other penalties such as SCAD and MCP. In this case, however, a suitable estimating algorithm should be developed. A further extension to make our model more robust can also be studied. For example, Hu *et al.* (2017) proposed to use a least trimmed squares method to make their model robust to outliers and our proposed method can also be extended in a similar direction. We leave these extensions to future research.

References

Bagnoli M and Bergstrom T (2005). Log-concave probability and its applications, *Economic Theory*, **26**, 445–469.

Balabdaoui F, Rufibach K, and Wellner JA (2009). Limit distribution theory for maximum likelihood estimation of a log-concave density, *The Annals of Statistics*, **37**, 1299.

Breiman L and Friedman JH (1985). Estimating optimal transformations for multiple regression and correlation, *Journal of the American Statistical Association*, **80**, 580–598.

Chen Y and Samworth RJ (2013). Smoothed log-concave maximum likelihood estimation with applications, *Statistica Sinica*, **23**, 1373–1398.

Dharmadhikari S and Joag-Dev K (1988). *Unimodality, Convexity, and Applications*, Academic Press Boston, Boston.

Dümbgen L, Hüsler A, and Rufibach K (2007). Active set and EM algorithms for log-concave densities based on complete and censored data. Available from: arXiv preprint arXiv:0707.4643

Dümbgen L and Rufibach K (2009). Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency, *Bernoulli*, **15**, 40–68.

Dümbgen L and Rufibach K (2011). Logcondens: Computations related to univariate log-concave density estimation, *Journal of Statistical Software*, **39**, 1–28.

Dümbgen L, Samworth R, and Schuhmacher D (2011). Approximation by log-concave distributions, with applications to regression, *The Annals of Statistics*, **39**, 702–730.

Fan J and Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96**, 1348–1360.

Friedman J, Hastie T, and Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software*, **33**, 1.

Hu H, Yao W, and Wu Y (2017). The robust EM-type algorithms for log-concave mixtures of regression models, *Computational Statistics & Data Analysis*, **111**, 14–26.

Kim S and Seo B (2018). Linear regression under log-concave and Gaussian scale mixture errors: comparative study, *Communications for Statistical Applications and Methods*, **25**, 633–645.

Kim S and Seo B (2021). Modal linear regression using log-concave distributions, *Journal of the Korean Statistical Society*, **50**, 479–494.

Konis K and Konis MK (2020). lpSolveAPI: R Interface to 'lp_solve' Version 5.5.2.0, R package version 5.5.2.0-17.7.

Lange KL, Little RJ, and Taylor JM (1989). Robust statistical modeling using the t-distribution, *Journal of the American Statistical Association*, **84**, 881–896.

Liu Y and Wang Y (2018). A fast algorithm for univariate log-concave density estimation, *Australian & New Zealand Journal of Statistics*, **60**, 258–275.

Nierenberg DW, Stukel TA, Baron JA, Dain BJ, Greenberg ER, and Skin Cancer Prevention Study Group (1989). Determinants of plasma levels of beta-carotene and retinol, *American Journal of Epidemiology*, **130**, 511–521.

Prékopa and András (1971). Logarithmic concave measures with application to stochastic programming, *Acta Scientiarum Mathematicarum*, **32**, 301–316.

Rousseeuw PJ and Leroy AM (1987). *Robust Regression and Outlier Detection*, Wiley, New York.

Schwarz G (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**, 461–464.

Sherwood B and Maidman A (2020). rqPen: Penalized Quantile Regression, R package version 2.2.2.

Silverman BW (1982). On the estimation of a probability density function by the maximum penalized likelihood method, *The Annals of Statistics*, **10**, 795–810.

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

Wang H, Li G, and Jiang G (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso, *Journal of Business & Economic Statistics*, **25**, 347–355.

Wang Y, Lawson CL, and Hanson RJ (2020). lsei: Solving Least Squares or Quadratic Programming Problems under Equality/Inequality Constraints, R package version 1.3-0.

Zhang CH (2010). Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics*, **38**, 894–942.

Zhang HH and Lin CY (2013). Cosso: Fit Regularized Nonparametric Regression Models Using COSSO Penalty, R package version 2.1-1.

Zou H (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, **101**, 1418–1429.

Received March 22, 2022; Revised May 04, 2022; Accepted May 16, 2022