*Research Article*

# Penalized Maximum Likelihood Method to a Class of Skewness Data Analysis

## Xuedong Chen,[1] Qianying Zeng,[2] and Qiankun Song[2]

[1] *School of Science, Huzhou Teachers College, Huzhou 313000, China*
[2] *School of Management, Chongqing Jiaotong University, Chongqing 400074, China*

Correspondence should be addressed to Qiankun Song; qiankunsong@163.com

An extension of some standard likelihood and variable selection criteria based on procedures of linear regression models under the skew-normal distribution or the skew-$t$ distribution is developed. This novel class of models provides a useful generalization of symmetrical linear regression models, since the random term distributions cover both symmetric as well as asymmetric and heavy-tailed distributions. A generalized expectation-maximization algorithm is developed for computing the $\ell_1$ penalized estimator. Efficacy of the proposed methodology and algorithm is demonstrated by simulated data.

## 1. Introduction

In recent years, nonnormal distributions have received substantial interest in the statistics literature. The growing need for more flexible tools to analyze data that exhibit nonnormal features, including asymmetry, multimodality, and heavy tails, has led to intense development in nonnormal model-based methods. Many such distributions have been successfully applied to numerous datasets from a wide range of fields, including medical sciences, bioinformatics, environmetrics, engineering, economics, and financial sciences. Some recent applications of skew-normal and skew-$t$ models include [1–5].

The rich literature and active discussion of skew distributions as well as regression analysis of skew distributions were initiated by the pioneering work of [6], in which the univariate skew distribution was introduced. Following its generalization to the multivariate skew-normal distribution in [7], the number of contributions has grown rapidly. The concept of introducing additional parameters to regulate skewness in a distribution was subsequently extended to other parametric families. For a comprehensive survey of skew distributions, see, for example, the articles [4, 8–10].

Besides the skew-normal distribution, which plays a central role in these developments, the skew-$t$ distribution has also received much attention. Being a natural extension of the $t$ distribution, the skew-$t$ distribution retains reasonable tractability and is more robust against outliers than the skew-normal distribution.

On the other hand, the Lasso is a popular method for regression analysis that uses an $\ell_1$ penalty to achieve a spares solution. This idea has been broadly applied, for example, to generalized linear models by [11] and Cox's proportional hazard models for survival data by [12]. Inspired by an important work of [13], we will introduce a new approach to regression analysis based on a class skew distribution by taking advantage of sparse penalization methods. We propose to fit an $\ell_1$ penalized linear model with an $\ell_1$ penalty imposed on the location parameters of the skew distribution. Due to the sparse shrinkage property of the $\ell_1$ penalty, some components of regression coefficients are estimated by exact zero when the tuning parameter is properly chosen.

The rest of the paper is organized as follows. In Section 2, we introduce the skew-normal and skew-$t$ distribution and present some of their properties. Section 3 outlines the EM-type algorithm and a penalty estimate of the proposed model of the skew distribution. In Section 4, a method of choice of tuning parameter is developed. The methodology proposed is illustrated in Section 5 by analyzing a simulated study.

## 2. The Skew-Normal Distribution and the Skew-$t$ Distribution

*2.1. The Skew-Normal Distribution.* As developed by [6], a random variable $Y$ follows a univariate skew-normal distribution with location parameter $\mu$, scale parameter $\omega$, and shape parameter $\alpha \in \mathbb{R}$, denoted by $Y \sim \mathrm{SN}(\mu, \omega, \alpha)$, if it has the density

$$f_{\mathrm{SN}}\left(y \mid \mu, \omega, \alpha\right) = \frac{2}{\omega}\phi\left(\frac{y-\mu}{\omega}\right)\Phi\left(\alpha\frac{y-\mu}{\omega}\right), \quad (1)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density function and cumulative function, respectively. Note that if $\alpha = 0$, the density of $Y$ reduces to the $N(\mu, \omega)$ density.

A stochastic representation of the $\mathrm{SN}(\mu, \omega, \alpha)$ distribution can be given by

$$Y = \mu + \omega\delta z + \omega\sqrt{1-\delta^2}\varepsilon, \quad (2)$$

where $z \sim \mathrm{TN}_{[0,\infty)}(0,1)$ and $\mathrm{TN}_{[a,b]}(\mu, \sigma^2)$ denotes the univariate truncated normal distribution with mean $\mu$, variance $\sigma^2$ in interval $[a,b]$. $\varepsilon \sim N(0,1)$, independent of $z$, and $\delta = \alpha/\sqrt{1+\alpha^2}$; hence $E(Y) = \mu + \omega\delta\sqrt{2/\pi}$.

There is an alternative definition as discussed by [2, 9], which can be viewed as a new parameterization of above parameters; that is, $\psi = \omega\delta$, $\sigma^2 = \omega^2(1-\delta^2)$; furthermore, suppose $u \sim N(0,1)$ and $\eta \sim N(0, \sigma^2)$, independently; then (2) can be represented as

$$Y = \mu + \psi\left|u\right| + \eta, \quad (3)$$

which defines a univariate skew-normal distribution $\mathrm{SN}(\mu, \omega, \alpha)$, where $\omega^2 = \psi^2 + \sigma^2$ and $\alpha = \psi/\sigma$, so (3) can be viewed as a reparametrization of (2). So it follows that the skew-normal distribution admits a convenient hierarchical characterization given by

(i) $u \sim N(0,1)$;

(ii) $Y \mid u \sim N(\mu + \psi|u|, \sigma^2)$.

*2.2. The Skew-$t$ Distribution.* To achieve a higher degree of excess kurtosis and accommodate to data of a heavy tail, skew-$t$ distributions have been introduced by [14]. A univariate random variable $Y$ follows the skew-$t$ distribution, $Y \sim \mathrm{ST}(\mu, \omega, \alpha, \nu)$; the density function reads

$$f_{\mathrm{ST}}\left(y \mid \mu, \omega, \alpha, \nu\right) = \frac{2}{\omega}t_{\nu}\left(d_y\right)T_{\nu+1}\left(\alpha d_y\sqrt{\frac{\nu+1}{\nu+d_y^2}}\right), \quad (4)$$

where $d_y = (y-\mu)/\omega$ and $t_{\nu}$ and $T_{\nu}$ denote, respectively, the density function and distribution function of a standard Student's $t$ distribution with $\nu$ degrees of freedom, and parameters $\mu$, $\omega$, and $\alpha$ are the same as above. Clearly, if $Y \sim \mathrm{ST}(\mu, \omega, \alpha, \nu)$, it has the following stochastic representation:

$$Y = \mu + \omega\frac{X}{\sqrt{W}}, \quad (5)$$

where $X \sim \mathrm{SN}(0,1,\alpha)$ and $W \sim \Gamma(\nu/2, \nu/2)$, independently and $\Gamma(a,b)$ denotes the Gamma distribution with parameters $a$ and $b$.

Similarly, the skew-$t$ distribution has the convolution-type representation as

$$Y = \mu + \psi\left|u\right| + \eta, \quad (6)$$

where $u \sim t_1(0,1,\nu)$, $\eta \sim t_1(0, \sigma^2, \nu)$, independently, and $\psi$ and $\sigma^2$ are the same as above. Meanwhile, the skew-$t$ distribution also has the hierarchical representation as

(i) $w \sim \Gamma(\nu/2, \nu/2)$;

(ii) $u \mid w \sim N(0, 1/w)$;

(iii) $Y \mid u, w \sim N(\mu + \psi|u|, \sigma^2/w)$.

## 3. Regression Model of Skew Distribution and EM Algorithm

In this section, we consider the linear regression model where the observation follows skew-normal distributions or skew-$t$ distributions. In general, a linear regression model is defined as

$$y_i = x_i\beta + \omega\delta\sqrt{\frac{2}{\pi}} + \varepsilon_i, \quad i = 1, \ldots, n, \quad (7)$$

or

$$y_i = x_i\beta + \omega\delta\sqrt{\frac{2}{\pi}}\frac{\Gamma\left(\left(\nu-1\right)/2\right)}{\Gamma\left(\nu/2\right)} + \varepsilon_i, \quad i = 1, \ldots, n, \quad (8)$$

where $y_i$ are responses, parameter vector $\beta = (\beta_1, \ldots, \beta_p)^T$, $x_i$ is a vector of explanatory variable values, and the random error $\varepsilon_i$ follows $\mathrm{SN}(0, \omega, \alpha)$ or $\mathrm{ST}(0, \omega, \alpha, \nu)$, respectively, which corresponds to the regression model with an error distribution of which the mean is zero.

It follows that the log-likelihood function for original parameter $\theta = (\beta, \omega, \alpha)$ or $\theta = (\beta, \omega, \alpha, \nu)$, given the observed sample $y = (y_1, \ldots, y_n)^t$ from SN or ST, respectively, is given by

$$\mathrm{LL}\left(\theta\right) = n\log\left(2\right) - \frac{n}{2}\log\left(2\pi\right) - \frac{n}{2}\log\left(\omega^2\right) + \sum_{i=1}^{n}\log\left(K_i\right), \quad (9)$$

where $K_i$ is the function of $\phi(\cdot)\Phi(\cdot)$ or $t(\cdot)T(\cdot)$ in (1) or (4), according to the SN or ST distribution, respectively. Note that the Newton-Raphson method can be directly applied to get the ML estimate of the above log-likelihood, but in order to obtain a more efficient approach of estimation and variable selection of skew models, we discuss a more elaborate technique to find the ML estimate based on an EM algorithm.

Firstly, for skew-normal distributions, let $u = (u_1, \ldots, u_n)^t$; then under the hierarchical representation of Section 2.1, it follows that the complete log-likelihood

function of modified parameter $\varphi = (\beta, \sigma, \psi)$ associated with $y_c = (y, u)^t$ is

$$\ell_c(\varphi) = \ell(\varphi \mid y_c)$$
$$= C - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(y_i - x_i\beta - \psi u_i)^2 - \sigma^2 u_i^2\right], \tag{10}$$

where $C$ is a constant that is independent of $\varphi$. Then consider the penalized log-likelihood defined by

$$\text{PLL}(\varphi) = \ell_c(\varphi) - \frac{1}{2}\sum_{j=1}^{p}P_\lambda\left(|\beta_j|\right)$$
$$= C - \frac{n}{2}\log(\sigma^2)$$
$$- \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(y_i - x_i\beta - \psi u_i)^2 - \sigma^2 u_i^2\right] \tag{11}$$
$$- \frac{1}{2}\sum_{j=1}^{p}P_\lambda\left(|\beta_j|\right),$$

where $P_\lambda(\cdot)$ is a nonnegative penalty function and we use $P_\lambda(|\beta_j|) = \lambda|\beta_j|$, which is the Lasso penalty, where $\lambda$ is the tuning parameter not included in $\varphi$.

The Lasso estimator is then defined as $\text{argmaxPLL}(\varphi)$; that is,

$$\widehat{\varphi} = \text{argmin}\left\{\frac{n}{2}\log(\sigma^2)\right.$$
$$- \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(y_i - x_i\beta - \psi u_i)^2 - \sigma^2 u_i^2\right] \tag{12}$$
$$\left. + \frac{\lambda}{2}\sum_{j=1}^{p}|\beta_j|\right\}.$$

Clearly, $u = (u_1, \ldots, u_n)$ can be seen as "missing data"; in the following, we will derive a generalized expectation-maximization algorithm for computing the $\ell_1$ penalized estimator. It turns out that the penalized estimator can be computed by iterative Lasso-penalized least square.

Set

$$Q\left(\varphi \mid \widehat{\varphi}^{(k)}\right) = E\left[\ell_c(\varphi) \mid y, \widehat{\varphi}^{(k)}\right] = E_{u|y}\left[\ell(\varphi \mid y_c) \mid y, \widehat{\varphi}^{(k)}\right] \tag{13}$$

as the $Q$-function over $\varphi$ at iteration $k$ step and where $E_{u|y}$ denotes the conditional expectation of $u$ given $y$.

At the $E$-step, by using the properties of conditional expectation of the truncated normal distribution, we obtain

$$\widehat{e}_{1,i}^{(k)} = E\left[u_i \mid y_i, \widehat{\varphi}^{(k)}\right] = \widehat{\mu}_{1,i}^{(k)} + \widehat{\sigma}_1^{(k)}\frac{\phi\left(\widehat{\mu}_{1,i}^{(k)}/\widehat{\sigma}_1^{(k)}\right)}{\Phi\left(\widehat{\mu}_{1,i}^{(k)}/\widehat{\sigma}_1^{(k)}\right)},$$
$$\widehat{e}_{2,i}^{(k)} = E\left[u_i^2 \mid y_i, \widehat{\varphi}^{(k)}\right] \tag{14}$$
$$= \left(\widehat{\mu}_{1,i}^{(k)}\right)^2 + \left(\widehat{\sigma}_1^{(k)}\right)^2 + \widehat{\mu}_{1,i}^{(k)}\widehat{\sigma}_1^{(k)}\frac{\phi\left(\widehat{\mu}_{1,i}^{(k)}/\widehat{\sigma}_1^{(k)}\right)}{\Phi\left(\widehat{\mu}_{1,i}^{(k)}/\widehat{\sigma}_1^{(k)}\right)},$$

where $\widehat{\mu}_{1,i}^{(k)} = \widehat{\psi}^{(k)}(y_i - x_i\widehat{\beta}^{(k)})/[(\widehat{\psi}^{(k)})^2 + (\widehat{\sigma}^{(k)})^2]$ and $\widehat{\sigma}_1^{(k)} = \widehat{\sigma}^{(k)}/\sqrt{(\widehat{\psi}^{(k)})^2 + (\widehat{\sigma}^{(k)})^2}$.

Then, combining the $Q$-function and the Lasso penalty, we set

$$\widetilde{Q}\left(\varphi \mid \widehat{\varphi}^{(k)}\right) = Q\left(\varphi \mid \widehat{\varphi}^{(k)}\right) - \frac{\lambda}{2}\sum_{j=1}^{p}|\beta_j|, \tag{15}$$

as the modified $Q$-function.

At the $M$-step update $\varphi$ by maximizing $\widetilde{Q}(\varphi \mid \widehat{\varphi}^{(k)})$ over $\varphi$ in a sequence of conditional maximization steps. Firstly, we get the following expression:

$$\widehat{\beta}^{(k+1)} = \text{argmin}\left\{\frac{1}{2(\widehat{\sigma}^{(k)})^2}\sum_{i=1}^{n}\left[y_i - x_i\beta - \widehat{\psi}^{(k)}\widehat{e}_{1,i}^{(k)}\right]\right.$$
$$\left. + \frac{\lambda}{2}\sum_{j=1}^{p}|\beta_j|\right\}; \tag{16}$$

then, we can obtain the closed expression of other parameters,

$$\widehat{\psi}^{(k+1)} = \frac{\sum_{i=1}^{n}\left(y_i - x_i\widehat{\beta}^{(k+1)}\right)\widehat{e}_{1,i}^{(k)}}{\sum_{i=1}^{n}\widehat{e}_{2,i}^{(k)}},$$
$$\widehat{\sigma^2}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\left[\left(y_i - x_i\widehat{\beta}^{(k+1)}\right)^2 + \left(\widehat{\psi}^{(k+1)}\right)^2\widehat{e}_{2,i}^{(k)}\right. \tag{17}$$
$$\left. - 2\left(y_i - x_i\widehat{\beta}^{(k+1)}\right)\widehat{\psi}^{(k+1)}\widehat{e}_{2,i}^{(k)}\right].$$

So, the above procedures can be summarized as follows.

Step 1: set initial values for $\beta$, $\psi$, $\sigma$, and tuning parameter $\lambda$.

Step 2: compute the value of $\widehat{e}_{1,i}$ and $\widehat{e}_{2,i}$ at current iteration step as the "working observed value."

Step 3: solve the penalized least squares problem about parameter $\beta$ and obtain its new Lasso estimator.

Step 4: compute the new estimator of $\psi$ and $\sigma$.

Step 5: repeat steps 2–4 till convergence.

Secondly, for skew-$t$ distributions, let $Y = (y_1, \ldots, y_n)$, $u = (u_1, \ldots, u_n)$, $w = (w_1, \ldots, w_n)$; then, based on the hierarchical representation of Section 2.2, we can get

the complete log-likelihood function associated with $y_c = (y, u, w)$ as

$$\ell_{ST_c}(\varphi) = \ell_{ST}(\varphi \mid y_c)$$

$$= C - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}w_i(y_i - x_i\beta - \psi u_i)^2$$

$$- \frac{1}{2}\sum_{i=1}^{n}w_i u_i^2 + \frac{n\nu}{2}\log\left(\frac{\nu}{2}\right) - n\log\Gamma\left(\frac{\nu}{2}\right) \tag{18}$$

$$+ \frac{\nu}{2}\sum_{i=1}^{n}\log(w_i) - \frac{\nu}{2}\sum_{i=1}^{n}w_i.$$

Similarly, setting

$$\text{PLL}_{ST}(\varphi) = \ell_{ST_c}(\varphi) - \frac{1}{2}\sum_{j=1}^{p}P_\lambda\left(|\beta_j|\right), \tag{19}$$

$$Q\left(\varphi \mid \widehat{\varphi}^{(k)}\right) = E\left[\ell_{ST_c}(\varphi) \mid y, \widehat{\varphi}^{(k)}\right],$$

then we can combine the $E$-step, $M$-step, and the Lasso estimator together as

$$\widehat{\beta}^{(k+1)} = \operatorname*{argmin}\left\{\frac{1}{2(\widehat{\sigma}^{(k)})^2}\sum_{i=1}^{n}\widehat{s}_{1,i}^{(k)}\left[y_i - x_i\beta - \widehat{\psi}^{(k)}\widehat{se}_1^{(k)}\right]\right.$$

$$\left. + \frac{\lambda}{2}\sum_{j=1}^{p}\left|\beta_j\right|\right\},$$

$$\widehat{\psi}^{(k+1)} = \frac{\sum_{i=1}^{n}\left(y_i - x_i\widehat{\beta}^{(k+1)}\right)\widehat{se}_{1,i}^{(k)}}{\sum_{i=1}^{n}\widehat{se}_{2,i}^{(k)}},$$

$$\widehat{\sigma^2}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}\left[\widehat{s}_{1,i}^{(k)}\left(y_i - x_i\widehat{\beta}^{(k+1)}\right)^2 + \left(\widehat{\psi}^{(k+1)}\right)^2\widehat{se}_{2,i}^{(k)}\right.$$

$$\left. - 2\left(y_i - x_i\widehat{\beta}^{(k+1)}\right)\widehat{\psi}^{(k+1)}\widehat{e}_{2,i}^{(k)}\right], \tag{20}$$

and parameter $\nu^{(k+1)}$ can be obtained via estimating equation as follows:

$$n\left[\log\left(\frac{\nu}{2}\right) - \text{digamma}\left(\frac{\nu}{2}\right) + 1\right] + \sum_{i=1}^{n}\left(\widehat{s}_{2,i}^{(k)} - \widehat{s}_{1,i}^{(k)}\right) = 0, \tag{21}$$

where

$$\widehat{s}_{1,i}^{(k)} = E_\varphi\left[w_i \mid y_i, \widehat{\varphi}^{(k)}\right] = \left(\frac{\nu+1}{\nu+dy_i}\right)\frac{T_{\nu+3}\left(y_{2,i}/\alpha\right)}{T_{\nu+1}\left(y_{2,i}/\alpha\right)},$$

$$\widehat{s}_{2,i}^{(k)} = E_\varphi\left[\log(w_i) \mid y_i, \widehat{\varphi}^{(k)}\right]$$

$$= \widehat{s}_{1,i}^{(k)} - \log\left(\frac{\nu+dy_i}{2}\right) - \text{digamma}\left(\frac{\nu+1}{2}\right) - 1,$$

$$\widehat{se}_{1,i}^{(k)} = E_\varphi\left[w_i u_i \mid y_i, \widehat{\varphi}^{(k)}\right] = \alpha\sqrt{\frac{\nu+1}{\nu+dy_i}}\frac{t_{\nu+1}\left(y_{2,i}/\alpha\right)}{T_{\nu+1}\left(y_{2,i}/\alpha\right)} + \widehat{s}_{2,i}q_i,$$

$$\widehat{se}_{2,i}^{(k)} = E_\varphi\left[w_i u_i^2 \mid y_i, \widehat{\varphi}^{(k)}\right]$$

$$= (\alpha)^2 + \alpha q_i\frac{t_{\nu+1}\left(y_{2,i}/\alpha\right)}{T_{\nu+1}\left(y_{2,i}/\alpha\right)} + \widehat{s}_{2,i}(q_i)^2, \tag{22}$$

and $dy_i = (y_i - x_i\beta)^2/(\sigma^2 + \psi^2)$, $q_i = \psi(y_i - x_i\beta)/(\sigma^2 + \psi^2)$, $\alpha = \psi/\sqrt{\sigma^2 + \psi^2}$, $y_{1,i} = q_i\sqrt{(\nu+1)/(\nu+dy_i)}$, and $y_{2,i} = q_i\sqrt{(\nu+3)/(\nu+dy_i)}$.

Similarly, we can also summarize the above procedures as follows.

Step 1: set initial values for $\beta$, $\psi$, $\sigma$, $\nu$, and tuning parameter $\lambda$.

Step 2: compute the value of $\widehat{s}_{1,i}$, $\widehat{s}_{2,i}$ and $\widehat{se}_{1,i}$, $\widehat{se}_{2,i}$ at current iteration step as the "working observed value."

Step 3: solve the penalized least squares problem about parameter $\beta$ and obtain its new Lasso estimator.

Step 4: compute the new estimates of $\psi$ and $\sigma$ directive and then get the new estimates of $\nu$.

Step 5: get the tuning parameters $\lambda$ by GCV.

Step 6: repeat steps 2–5 till convergence.

## 4. Choice of Tuning Parameters

We now consider the choice of tuning parameters. In application, the cross-validation (CV), or generalized cross-validation (GCV), is often used for choosing tuning parameters. Following the examples of [11, 15], we develop a componentwise deviance-based GCV criterion for the proposed models.

Let $\widetilde{\varphi}$ be the EM estimates of parameters $\varphi$ without penalty functions. For a given value of tuning parameter of $\lambda$, let $\widehat{\beta}$ be the maximum penalized likelihood estimates of the parameters of the model by fixing the rest of components of $\varphi$ at $\widetilde{\varphi}$. Denote the deviance function as

$$D\left(\widehat{\beta}\right) = E\left[\ell_c(\varphi) \mid \widehat{\beta}\right] - E\left[\ell_c(\varphi) \mid \widetilde{\varphi}\right]. \tag{23}$$

Further, let $\ell''(\widehat{\beta})$ be the second derivative of the log-likelihood function with respect to $\beta$ evaluated at $\widehat{\beta}$. We define a GCV criterion of the model as

$$\text{GCV}(\lambda) = \frac{D\left(\widehat{\beta}\right)}{n(1 - e(\lambda)/n)^2}, \tag{24}$$
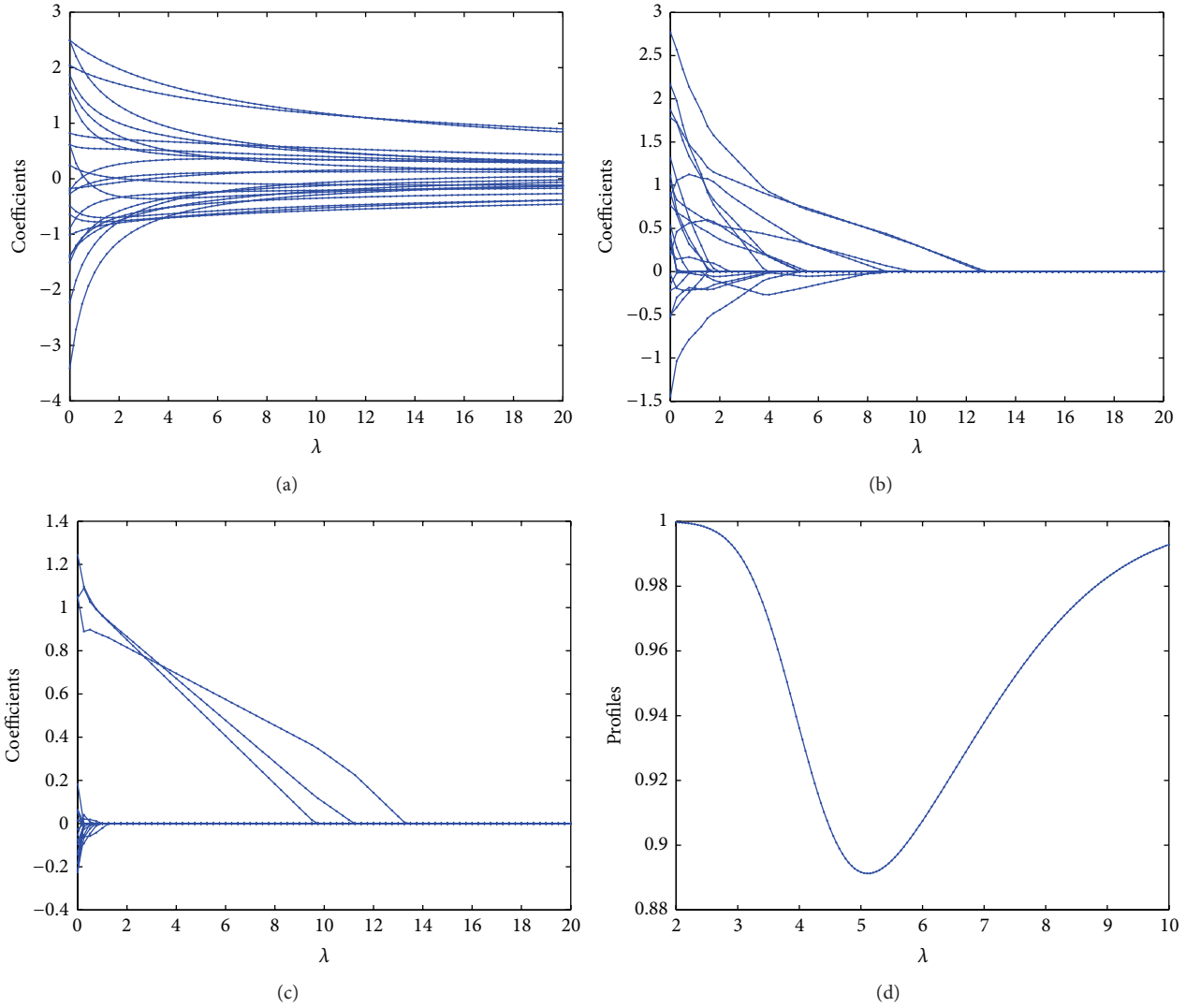
FIGURE 1: (a)–(c) Trace plot of regression coefficients and profile of tuning parameter.

where $e(\lambda)$ is the effective number of regression coefficients. It is given by

$$e(\lambda) = \text{trace}\left\{\left[\ell''\left(\widehat{\beta}\right) - \Sigma\left(\widehat{\beta}\right)\right]^{-1}\ell''\left(\widehat{\beta}\right)\right\}, \qquad (25)$$

where $\Sigma(\widehat{\beta}) = \text{diag}\{1/|\widehat{\beta}_1|, \ldots, 1/|\widehat{\beta}_p|\}$. The tuning parameter $\lambda$ is chosen by minimizing GCV($\lambda$).

## 5. Simulation Study

In order to investigate the experimental performance of our methodology, we undertake a simulation study to investigate the effect of estimation and variable selection. The dataset is generated as follows: let $x_i = (x_{1i}, \ldots, x_{pi})^T$ be covariate, where $p = 30$ and $x_{1i,\ldots,p1}$ are independent and distributed as uniform distribution $U(-1, 1)$. Moreover, assume response $y_i$ is formulated as

$$y_i = x_i^T \beta + \varepsilon_i, \qquad (26)$$

where $\varepsilon_i$ is distributed as the following five cases.

Case (1): $\varepsilon_i \sim N(0, \sigma^2)$; Case (2): $\varepsilon_i \sim SN(0, \sigma_1^2, \alpha_1)$; Case (3): $\varepsilon_i \sim SN(0, \sigma_1^2, \alpha_2)$; Case (4): $\varepsilon_i \sim ST(0, \sigma_2^2, \alpha_1, \nu)$; and Case (5): $\varepsilon_i \sim ST(0, \sigma_2^2, \alpha_2, \nu)$.

We take the sample size $n$ to be 100. The true values for $\beta$ are $\beta_1 = \beta_2 = \beta_3 = 1$ and the other components of $\beta$ are equal to zero. Further, let $\sigma_1 = 2.5$, $\sigma_2 = 1.0$, $\alpha_1 = 1.5$, $\alpha_2 = -0.8$, and $\nu = 3$.

Under above settings, we use three linear regression models based on normal, skew-normal, and skew-$t$ distribution to fit each dataset, respectively.

Table 1 shows the results of the simulation study, where No. denotes the number of nonzero components of $\beta$, which can determine the effect of variable selection, and where AE denotes the average of absolute error; that is, $(1/p)\sum_{j=1}^{p}|\beta_j^0 - \widehat{\beta}_j|$, which can show the accuracy of the estimates, and the $\lambda$ is the tuning parameter.

Examination of Table 1 reveals that (i) when the true distribution is normal, the difference of the three models is very small and the coefficient $\beta$ can be found correctly; (ii)

TABLE 1: Results of the simulation study.

| Case | N | | | SN | | | ST | | |
|---|---|---|---|---|---|---|---|---|---|
| | No. | AE | $\lambda$ | No. | AE | $\lambda$ | No. | AE | $\lambda$ |
| 1 | 3 | 0.6238 | 9.45 | 3 | 0.7323 | 7.46 | 3 | 0.7431 | 6.91 |
| 2 | 7 | 2.3536 | 12.20 | 3 | 1.0913 | 6.38 | 3 | 1.0396 | 6.22 |
| 3 | 11 | 3.3893 | 14.80 | 3 | 1.0488 | 6.37 | 3 | 1.1158 | 6.12 |
| 4 | 8 | 2.8939 | 18.92 | 3 | 1.1261 | 6.12 | 3 | 1.0135 | 6.25 |
| 5 | 10 | 3.8273 | 16.82 | 3 | 1.1542 | 5.12 | 3 | 1.0363 | 5.05 |

when the true distribution is nonnormal, the models based on normal fit the data badly; not only does the proportion of nonzero increase but also the average error becomes significant; (iii) for the cases of the true distribution being nonnormal, the models based on the skew distribution can improve the fitting effect better than the ones of the normal distribution.

Figures 1(a)–1(c) present the trace plots of coefficients versus tuning parameters: (a) Case (5) fitted by normal distribution (N) model; (b) Case (5) fitted by skew-normal distributions (SN) model; (c) Case (5) fitted by skew-$t$ distributions (ST) model. Moreover, (d) shows profiles of the tuning parameter of (c).

From Figure 1(c), it is clear that the ST model is most appropriate for fitting this dataset.

## 6. Conclusion

In this paper, we have proposed a class of linear regression models based on the asymmetric distribution. An EM-type algorithm and the Lasso penalty estimates are developed by exploring the statistical properties of the model that can be implemented efficiently in types of software such as SAS, R, and Matlab. Meanwhile, we have shown that the $\ell_1$ penalized maximum likelihood estimation can be done via a generalized EM algorithm that is equivalent to iterative $\ell_1$ penalized least square for some parameters. We have observed that if the data are generated from a true model of the skew distribution, the $\ell_1$ penalized models based on EM algorithm not only identify zero coefficient but also are significantly more accurate than the general MLE models.

On the other hand, due to recent advances in computational technology, it is worthwhile to carry out Bayesian treatments via Markov chain Monte Carlo (MCMC) sampling methods in the context of this paper.

## Conflict of Interests

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

## Acknowledgments

## References

[1] T. I. Lin, J. C. Lee, and W. J. Hsieh, "Robust mixture modeling using the skew $t$ distribution," *Statistics and Computing*, vol. 17, no. 2, pp. 81–92, 2007.

[2] S. K. Sahu, D. K. Dey, and M. D. Branco, "A new class of multivariate skew distributions with applications to Bayesian regression models," *The Canadian Journal of Statistics*, vol. 31, no. 2, pp. 129–150, 2003.

[3] V. H. Lachos, P. Ghosh, and R. B. Arellano-Valle, "Likelihood based inference for skew-normal independent linear mixed models," *Statistica Sinica*, vol. 20, no. 1, pp. 303–322, 2010.

[4] R. B. Arellano-Valle and A. Azzalini, "On the unification of families of skew-normal distributions," *Scandinavian Journal of Statistics: Theory and Applications*, vol. 33, no. 3, pp. 561–574, 2006.

[5] I. Vrbik and P. D. McNicholas, "Analytic calculations for the EM algorithm for multivariate skew-$t$ mixture models," *Statistics & Probability Letters*, vol. 82, no. 6, pp. 1169–1174, 2012.

[6] A. Azzalini, "A class of distributions which includes the normal ones," *Scandinavian Journal of Statistics: Theory and Applications*, vol. 12, no. 2, pp. 171–178, 1985.

[7] A. Azzalini and A. Dalla-Valle, "The multivariate skew-normal distribution," *Biometrika*, vol. 83, no. 4, pp. 715–726, 1996.

[8] A. Azzalini, "The skew-normal distribution and related multivariate families," *Scandinavian Journal of Statistics: Theory and Applications*, vol. 32, no. 2, pp. 159–200, 2005.

[9] A. Gupta, "Multivariate skew $t$-distribution," *Statistics*, vol. 37, no. 4, pp. 359–363, 2003.

[10] M. Genton, *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, Chapman & Hall, Boca Raton, Fla, USA, 2004.

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

[12] R. Tibshirani, "The lasso method for variable selection in the Cox's model," *Statistics in Medicine*, vol. 16, no. 2, pp. 385–395, 1997.

[13] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[14] A. Azzalini and A. Capitanio, "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew $t$-distribution," *Journal of the Royal Statistical Society: Series B Methodological*, vol. 65, no. 2, pp. 367–389, 2003.

[15] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.