Statistics and Its Interface Volume 10 (2017) 217–227

# Variable selection in joint location, scale and skewness models with a skew-t-normal distribution

Liucang Wu[*,†], Guo-Liang Tian, Yan-Qing Zhang, and Ting Ma

Although there are many papers on variable selection methods in the modeling of the mean and/or variance parameters, little work has been done on how to select significant explanatory variables in the modeling of the skewness parameter. In this article, we propose a unified penalized likelihood method to simultaneously select significant variables and estimate unknown parameters in a joint location, scale and skewness model with a *skew-t-normal* (StN) distribution when outliers and asymmetrical outcomes are present. With an appropriate selection of the tuning parameters, we establish the consistency and the oracle property of the regularized estimators. Simulation studies are conducted to assess the finite sample performance of the proposed variable selection procedure. A real example is used to illustrate the proposed method.

AMS 2000 subject classifications: Primary 62F12; Secondary 62H12.

Keywords and phrases: Heteroscedastic regression models, Joint location, scale and skewness models, Penalized maximum likelihood estimator, Skew-t-normal distribution, Variable selection.

## 1. INTRODUCTION

Statistical distributions are basic tools for describing some random phenomena and are also the foundation of statistical inferences. In recent years, the quantity of data collected and requiring statistical analysis have been increasing rapidly, allowing the fitting of more complex and potentially more realistic models. Because skew distributions compared with symmetrical distributions can release timely and accurate information, the application of skew distributions is very extensive in the fields of finance, economics, biomedicine and so on. For this reason, skew distributions have received considerable attention in recent years. For example, Gómez et al. (2007) introduced a so-called *skew-t-normal* (StN) distribution and claimed that it is a robust alternative to the skew-normal (SN) distribution (Azzalini, 2005) in modeling heavy tailed data with outliers and strong degrees of asymmetry. They also showed that the StN distribution has a wider range of skewness than the SN distribution. Cabral et al. (2008) suggested a Bayesian approach to modeling mixtures of StN distributions by the Markov chain Monte Carlo algorithm. Lin et al. (2009) investigated statistical diagnostics for nonlinear regression models by replacing the normal error distribution with an StN error distribution. Ho et al. (2011) provided a maximum likelihood inference for mixtures of StN distributions through practical EM-type algorithms.

On the other hand, similar to the modeling of the mean and variance parameters, the modeling of the skewness parameter itself may be of statistical interest. To achieve the goal of effectively controlling the skewness, it may be helpful to understand what predictors/factors affect the skewness. Thus, the modeling of the skewness parameter is of same importance as that of the mean and variance parameters. This motivates us to develop a joint location, scale and skewness model with an StN distribution.

Although there are many papers on variable selection methods in the modeling of the mean and/or variance parameters, these papers cannot directly be used to select important variables in joint mean, variance and skewness models. For example, Fan and Lv (2010), Hu and Lian (2013) and references therein only provided methods for variable selection in the modeling of the mean of the responses. In recent years, variable selection in the modeling of the variance has gained popularity. Based on the adjusted profile likelihood, Zhang and Wang (2011) proposed a new criterion, called as PICa, to simultaneously select explanatory variables in the modeling of mean and variance parameters for heteroscedastic linear regression models. Wu, Zhang and Xu (2012a) proposed methods to simultaneously select significant variables in joint mean and variance models, providing a useful extension of the classical normal regression models. Wu, Zhang and Xu (2012b) proposed a hybrid strategy, in which variable selection is employed to reduce the dimension of the explanatory variables in joint mean and variance models, and Box–Cox transformation is made to remedy the response. Wu and Li (2012) considered variable selection for joint mean and dispersion models with the inverse Gaussian distribution. Wu, Zhang and Xu (2013) investi-

[*] Corresponding author.

[†] The first author's research is fully supported by the National Natural Science Foundation of China (No. 11261025, 11126309).

gated the simultaneous variable selection in joint location and scale models with an SN distribution.

However, little work has been done on how to select significant explanatory variables in the modeling of the skewness parameter. In practice, it is also important to determine what variables have significant impact on the skewness. This motivates us to develop a unified penalized likelihood method to simultaneously select significant variables and estimate unknown parameters in the joint location, scale and skewness models with an StN distribution when outliers and asymmetrical outcomes are present.

The rest of this paper is organized as follows. In Section 2, we first propose a variable selection method in the joint location, scale and skewness models with an StN distribution. Then, penalized maximum likelihood estimators are derived. Finally, we present some theoretical properties on the proposed variable selection procedure, including the consistency and the oracle property of the regularized estimators. In Section 3, based on the local quadratic approximations, we provide an iterative algorithm for finding the penalized maximum likelihood estimators. The choice of the tuning parameters is also presented. In Section 4, some simulation studies are performed and a real data set on the *body mass index* (BMI) is analyzed to demonstrate the proposed methods. Some concluding remarks are given in Section 5. Some technical proofs are put in the two appendices.

# 2. VARIABLE SELECTION IN JOINT LOCATION, SCALE AND SKEWNESS MODELS

## 2.1 The joint model with skew-t-normal distribution

A random variable $Y$ is said to have an StN distribution with location $\mu$, scale $\sigma^2$, skewness $\lambda$ and the degrees of freedom $\nu$, denoted by $Y \sim \mathrm{StN}(\mu, \sigma^2, \lambda, \nu)$, if its density function is (Gómez et al., 2007)

$$f(y) = \frac{2}{\sigma} t_\nu\left(\frac{y-\mu}{\sigma}\right) \Phi\left(\lambda \frac{y-\mu}{\sigma}\right), \qquad (1)$$

where

$$t_\nu(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

is the density function of the $t$ distribution with $\nu$ degrees of freedom and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. In Appendix A, we will show that the density function $f(y)$ is unimodal. The expectation and variance of the random variable $Y$ are

$$E(Y) = \mu + \sqrt{\frac{2}{\pi}}\sigma\lambda\left(\frac{\nu}{2}\right)^{\frac{1}{2}} \frac{\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})} E_V(V+\lambda^2)^{-\frac{1}{2}}$$

and

$$\mathrm{Var}(Y) = \sigma^2 \left\{ \frac{\nu}{\nu-2} - \frac{\lambda^2\nu}{\pi}\left[\frac{\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})}\right]^2 [E_V(V+\lambda^2)^{-\frac{1}{2}}]^2 \right\},$$

respectively, where $V \sim \mathrm{Gamma}((\nu-1)/2, \nu/2)$. If $\lambda = 0$, then $\mathrm{StN}(\mu, \sigma^2, \lambda, \nu)$ becomes $t_\nu(\mu, \sigma^2)$. That is, if the skewness $\lambda = 0$, then the density of $\mathrm{StN}(\mu, \sigma^2, \lambda, \nu)$ is symmetric. If $\nu \to \infty$, then $\mathrm{StN}(\mu, \sigma^2, \lambda, \nu)$ approaches to the skew-normal $\mathrm{SN}(\mu, \sigma^2, \lambda)$ considered by Azzalini (1985). If $\lambda = 0$ and $\nu \to \infty$, then $\mathrm{StN}(\mu, \sigma^2, \lambda, \nu)$ reduces to $N(\mu, \sigma^2)$. In this paper, we employ the parameter $\nu$, degrees of freedom in (1), to measure the level of robustness that adapts to the noise, contamination, and outliers in both theoretical and practical aspects. Theoretically, the $\nu$ acts as a robustness parameter to tune the heaviness of the tails and to down-weight the effect of the outliers on the estimation of parameters. In practice, like Lucas (1997) and Lange et al. (1989), we also suggest taking $\nu = 3, 4$ or 5.

We consider the following joint location, scale and skewness model with the StN distribution:

$$\begin{cases} Y_i & \overset{\text{ind}}{\sim} & \mathrm{StN}(\mu_i, \sigma_i^2, \lambda_i, \nu), \quad i = 1, \ldots, n, \\ \mu_i & = & x_i^\top \beta, \\ \log(\sigma_i^2) & = & z_i^\top \gamma, \\ \lambda_i & = & w_i^\top \alpha, \end{cases} \qquad (2)$$

where $\{Y_i\}_{i=1}^n$ are independent responses, $\nu\,(> 2)$ is a known degree of freedom [the reason for considering a known $\nu$ is similar to those as shown in Lucas (1997) and Lange et al. (1989)], $x_i = (x_{i1}, \ldots, x_{ip})^\top$, $z_i = (z_{i1}, \ldots, z_{iq})^\top$ and $w_i = (w_{i1}, \ldots, w_{ir})^\top$ are three covariate vectors for subject $i$, $\beta = (\beta_1, \ldots, \beta_p)^\top$ is a $p \times 1$ vector of regression coefficients in the location model, $\gamma = (\gamma_1, \ldots, \gamma_q)^\top$ is a $q \times 1$ vector of unknown parameters in the scale model and $\alpha = (\alpha_1, \ldots, \alpha_r)^\top$ is an $r \times 1$ vector of unknown parameters in the skewness model. The covariate vectors $x_i$, $z_i$ and $w_i$ are not necessarily identical. The objective is to simultaneously select significant variables and estimate parameters in the model (2).

## 2.2 Penalized maximum likelihood estimators

Let $Y_{\text{obs}} = \{(y_i, x_i, z_i, w_i)\colon i = 1, \ldots, n\}$ denote the observed data in the model (2). For convenience, let $\theta = (\theta_1, \ldots, \theta_s)^\top = (\beta^\top, \gamma^\top, \alpha^\top)^\top$, where $s = p + q + r$. Then, the observed data log-likelihood function of $\theta$ is given by

$$\ell(\theta) \propto -\frac{1}{2}\sum_{i=1}^n z_i^\top\gamma - \frac{\nu+1}{2}\sum_{i=1}^n \log\left\{\nu + \frac{(y_i - x_i^\top\beta)^2}{\exp(z_i^\top\gamma)}\right\} + \sum_{i=1}^n \log\Phi(k_i),$$

where

$$k_i = \frac{w_i^\top \alpha (y_i - x_i^\top \beta)}{\exp(z_i^\top \gamma/2)}.$$

Similar to Wu, Zhang and Xu (2013), we define the penalized log-likelihood function as

$$\begin{aligned}\mathcal{L}(\theta) &= \ell(\theta) - n\sum_{j=1}^{p} p_{\tau_{1j}}(|\beta_j|) - n\sum_{k=1}^{q} p_{\tau_{2k}}(|\gamma_k|) \\ &\quad - n\sum_{m=1}^{r} p_{\tau_{3m}}(|\alpha_m|) \\ &= \ell(\theta) - n\sum_{j=1}^{s} p_{\tau_j}(|\theta_j|), \qquad (3)\end{aligned}$$

where $p_{\tau_j}(\cdot)$ is a pre-specified penalty function with a regularization or tuning parameter $\tau_j$, which can be determined by one of the data-driven criteria such as cross-validation (CV), generalized cross-validation (GCV, see, e.g., Fan and Li, 2001; Tibshirani, 1996), and Bayesian information criterion (BIC, see, e.g., Wang et al, 2007). In Section 3.2, we use BIC to choose these tuning parameters. Note that the penalty functions and regularization parameters are not necessarily identical for all $j$. For example, if we wish to keep some important variables in the final model, we would not penalize their coefficients.

The penalized maximum likelihood estimators $\hat{\theta}_n$ of $\theta$ can be obtained by maximizing $\mathcal{L}(\theta)$ specified by (3). With appropriate penalty functions, maximizing $\mathcal{L}(\theta)$ with respect to $\theta$ leads to certain parameter estimators vanishing from the initial model so that the corresponding explanatory variables are automatically removed. Hence, by maximizing $\mathcal{L}(\theta)$ we can achieve the goal of selecting important variables and obtaining the parameter estimators, simultaneously. In Section 3.1, some technical details and an algorithm are provided to calculate the penalized maximum likelihood estimators $\hat{\theta}_n$.

## 2.3 Asymptotic properties

To study the consistency and asymptotic normality of the resulting penalized likelihood estimators, we let $\theta_0$ denote the true value of $\theta$. Furthermore, let $\theta_0 = (\theta_{01}, \ldots, \theta_{0s})^\top = (\theta_0^{(1)\top}, \theta_0^{(2)\top})^\top$. Without loss of generality, suppose that all nonzero components of $\theta_0$ are included in $\theta_0^{(1)}$ and $\theta_0^{(2)} = 0$. In addition, we assume that the tuning parameters are rearranged corresponding to the elements of $\theta_0$. Let $s_1$ be the dimension of $\theta_0^{(1)}$,

$$a_n = \max_{1\leqslant j\leqslant s}\{p'_{\tau_n}(|\theta_{0j}|)\colon \theta_{0j} \neq 0\}$$

and

$$b_n = \max_{1\leqslant j\leqslant s}\{|p''_{\tau_n}(|\theta_{0j}|)|\colon \theta_{0j} \neq 0\},$$

where $\tau$ is denoted by $\tau_n$ for emphasizing its dependency on the sample size $n$.

To prove the theorems below (see Appendix B), we require the following regularity conditions:

(C1) The three covariate vectors $x_i$, $z_i$ and $w_i$ are fixed for $i = 1, \ldots, n$;

(C2) The parameter space is compact and the true value $\theta_0$ is located in the interior of the parameter space;

(C3) All $x_i$, $z_i$ and $w_i$ are bounded, i.e., all elements in them are bounded by a single finite real number;

(C4) $y_i \overset{\text{ind}}{\sim} \mathrm{StN}(\mu_i, \sigma_i^2, \lambda_i, \nu)$, where $\mu_i = x_i^\top\beta_0$, $\log\sigma_i^2 = z_i^\top\gamma_0$ and $\lambda_i = w_i^\top\alpha_0$, $i = 1, \ldots, n$.

**Theorem 1** (Consistency). *Assume that* $a_n = O_p(n^{-\frac{1}{2}}), b_n \to 0$ *and* $\tau_n \to 0$ *as* $n \to \infty$. *Let* $\tau_n$ *be* $\tau_{1n}$, $\tau_{2n}$ *or* $\tau_{3n}$, *depending on whether* $\theta_{0j}$ *is a component of* $\beta_0$, $\gamma_0$ *or* $\alpha_0\,(1 \leqslant j \leqslant s)$. *Under Conditions* (C1)–(C4), *with probability tending to 1 there exists a local maximizer* $\hat{\theta}_n$ *of the penalized log-likelihood function* $\mathcal{L}(\theta)$ *in* (3) *such that* $\|\hat{\theta}_n - \theta_0\| = O_p(n^{-1/2})$.

Next we consider the asymptotic normality of $\hat{\theta}_n$. Let

$$\begin{aligned} A_n &= \mathrm{diag}\Big(p''_{\tau_n}(|\theta_{01}^{(1)}|), \ldots, p''_{\tau_n}(|\theta_{0s_1}^{(1)}|)\Big), \\ c_n &= \Big(p'_{\tau_n}(|\theta_{01}^{(1)}|)\mathrm{sgn}(\theta_{01}^{(1)}), \ldots, p'_{\tau_n}(|\theta_{0s_1}^{(1)}|)\mathrm{sgn}(\theta_{0s_1}^{(1)})\Big)^\top, \end{aligned}$$

where $\tau_n$ is identical to that in Theorem 1 and $\theta_{0j}^{(1)}$ is the $j$-th component of $\theta_0^{(1)}$ $(1 \leqslant j \leqslant s_1)$. Denote the Fisher information matrix of $\theta$ by $\mathcal{I}_n(\theta)$.

**Theorem 2** (Oracle property). *Assume that the penalty function* $p_{\tau_n}(t)$ *satisfies*

$$\liminf_{n\to\infty}\liminf_{t\to 0^+}\frac{p'_{\tau_n}(t)}{\tau_n} > 0$$

*and* $\bar{\mathcal{I}}_n = \mathcal{I}_n(\theta_0)/n$ *converges to a finite and positive definite matrix* $\mathcal{I}(\theta_0)$ *as* $n \to \infty$. *Under the conditions specified in Theorem 1, if* $\tau_n \to 0$ *and* $\sqrt{n}\tau_n \to \infty$ *as* $n \to \infty$, *then the* $\sqrt{n}$*-consistent estimator* $\hat{\theta}_n = ((\hat{\theta}_n^{(1)})^\top, (\hat{\theta}_n^{(2)})^\top)^\top$ *in Theorem 1 must satisfy*

(*i*) (Sparsity) $\hat{\theta}_n^{(2)} = 0$;

(*ii*) (Asymptotic normality)

$$\sqrt{n}(\bar{\mathcal{I}}_n^{(1)})^{-1/2}(\bar{\mathcal{I}}_n^{(1)} + A_n)\{(\hat{\theta}_n^{(1)} - \theta_0^{(1)}) + (\bar{\mathcal{I}}_n^{(1)} + A_n)^{-1}c_n\} \xrightarrow{\mathrm{L}} N_{s_1}(0, I_{s_1}),$$

*where "*$\xrightarrow{\mathrm{L}}$*" means convergence in distribution,* $\bar{\mathcal{I}}_n^{(1)}$ *is the* $s_1 \times s_1$ *sub-matrix of* $\bar{\mathcal{I}}_n$ *corresponding to* $\theta_0^{(1)}$ *and* $I_{s_1}$ *is the* $s_1 \times s_1$ *identity matrix.*

# 3. COMPUTATION

## 3.1 Algorithm

First, we note that the first two derivatives of the log-likelihood function $\ell(\theta) = \ell(\beta, \gamma, \alpha)$ are continuous. For a given point $\theta^{(t)}$, $\ell(\theta)$ can be approximated by

$$\begin{aligned}\ell(\theta) &\approx \ell(\theta^{(t)}) + \left[\frac{\partial \ell(\theta^{(t)})}{\partial \theta}\right]^\top (\theta - \theta^{(t)}) \\ &\quad + \frac{1}{2}(\theta - \theta^{(t)})^\top \left[\frac{\partial^2 \ell(\theta^{(t)})}{\partial \theta \partial \theta^\top}\right] (\theta - \theta^{(t)}).\end{aligned}$$

Also, given a scalar $\phi_0$ we can approximate the penalty function $p_\tau(\phi)$ by a quadratic function (Fan and Li, 2001)

$$p_\tau(|\phi|) \approx p_\tau(|\phi_0|) + \frac{1}{2}\frac{p'_\tau(|\phi_0|)}{|\phi_0|}(\phi^2 - \phi_0^2), \qquad \text{for} \quad \phi \approx \phi_0.$$

Therefore, the penalized log-likelihood function (3) can be locally approximated by

$$\begin{aligned}\mathcal{L}(\theta) &\approx \ell(\theta^{(t)}) + \left[\frac{\partial \ell(\theta^{(t)})}{\partial \theta}\right]^\top (\theta - \theta^{(t)}) \\ &+ \frac{1}{2}(\theta - \theta^{(t)})^\top \left[\frac{\partial^2 \ell(\theta^{(t)})}{\partial \theta \partial \theta^\top}\right] (\theta - \theta^{(t)}) - \frac{n}{2}\theta^\top \Sigma_\tau(\theta^{(t)})\theta,\end{aligned}$$

where

$$\begin{aligned}\Sigma_\tau(\theta) = \operatorname{diag}\Bigg(&\frac{p'_{\tau_{11}}(|\beta_1|)}{|\beta_1|}, \ldots, \frac{p'_{\tau_{1p}}(|\beta_p|)}{|\beta_p|}, \frac{p'_{\tau_{21}}(|\gamma_1|)}{|\gamma_1|}, \ldots, \\ &\frac{p'_{\tau_{2q}}(|\gamma_q|)}{|\gamma_q|}, \frac{p'_{\tau_{31}}(|\alpha_1|)}{|\alpha_1|}, \ldots, \frac{p'_{\tau_{3r}}(|\alpha_r|)}{|\alpha_r|}\Bigg).\end{aligned}$$

Thus, the Newton–Raphson algorithm can be used to iteratively calculate

$$\begin{aligned}\theta^{(t+1)} = \theta^{(t)} + &\left\{\frac{\partial^2 \ell(\theta^{(t)})}{\partial \theta \partial \theta^\top} - n\Sigma_\tau(\theta^{(t)})\right\}^{-1} \\ &\times \left\{n\Sigma_\tau(\theta^{(t)})\theta^{(t)} - \frac{\partial \ell(\theta^{(t)})}{\partial \theta}\right\}.\end{aligned}$$

In the follows, we calculate the score vector and the Hessian matrix. The score vector is given by

$$U(\theta) \hat{=} \frac{\partial \ell(\theta)}{\partial \theta} = \left(U_1^\top(\beta), U_2^\top(\gamma), U_3^\top(\alpha)\right)^\top,$$

where

$$\begin{aligned}U_1(\beta) &= \frac{\partial \ell(\theta)}{\partial \beta} \\ &= (\nu+1)\sum_{i=1}^n \left\{\nu + \frac{(y_i - x_i^\top\beta)^2}{e^{z_i^\top\gamma}}\right\}^{-1} \frac{(y_i - x_i^\top\beta)x_i}{e^{z_i^\top\gamma}} \\ &\quad -\sum_{i=1}^n \frac{\varphi(k_i)}{\Phi(k_i)} \frac{w_i^\top\alpha}{e^{\frac{1}{2}z_i^\top\gamma}} x_i,\end{aligned}$$

$$\begin{aligned}U_2(\gamma) &= \frac{\partial \ell(\theta)}{\partial \gamma} \\ &= -\frac{1}{2}\sum_{i=1}^n z_i + \frac{\nu+1}{2}\sum_{i=1}^n \left\{\nu + \frac{(y_i - x_i^\top\beta)^2}{e^{z_i^\top\gamma}}\right\}^{-1} \\ &\quad \times \frac{(y_i - x_i^\top\beta)^2 z_i}{e^{z_i^\top\gamma}} - \frac{1}{2}\sum_{i=1}^n \frac{\varphi(k_i)k_i}{\Phi(k_i)} z_i,\end{aligned}$$

$$U_3(\alpha) = \frac{\partial \ell(\theta)}{\partial \alpha} = \sum_{i=1}^n \frac{\varphi(k_i)}{\Phi(k_i)} \frac{y_i - x_i^\top\beta}{e^{z_i^\top\gamma}} w_i,$$

and $\varphi(\cdot)$ is the density function of $N(0,1)$. The Hessian matrix is

$$H(\theta) \hat{=} \frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta^\top} = \begin{pmatrix} \frac{\partial^2 \ell(\theta)}{\partial \beta \partial \beta^\top} & \frac{\partial^2 \ell(\theta)}{\partial \beta \partial \gamma^\top} & \frac{\partial^2 \ell(\theta)}{\partial \beta \partial \alpha^\top} \\ \frac{\partial^2 \ell(\theta)}{\partial \gamma \partial \beta^\top} & \frac{\partial^2 \ell(\theta)}{\partial \gamma \partial \gamma^\top} & \frac{\partial^2 \ell(\theta)}{\partial \gamma \partial \alpha^\top} \\ \frac{\partial^2 \ell(\theta)}{\partial \alpha \partial \beta^\top} & \frac{\partial^2 \ell(\theta)}{\partial \alpha \partial \gamma^\top} & \frac{\partial^2 \ell(\theta)}{\partial \alpha \partial \alpha^\top} \end{pmatrix},$$

where

$$\begin{aligned}&\frac{\partial^2 \ell(\theta)}{\partial \beta \partial \beta^\top} \\ &= 2(\nu+1)\sum_{i=1}^n \left\{\nu + \frac{(y_i - x_i^\top\beta)^2}{e^{z_i^\top\gamma}}\right\}^{-2} \frac{(y_i - x_i^\top\beta)^2}{e^{2z_i^\top\gamma}} x_i x_i^\top \\ &\quad - (\nu+1)\sum_{i=1}^n \left\{\nu + \frac{(y_i - x_i^\top\beta)^2}{e^{z_i^\top\gamma}}\right\}^{-1} \frac{x_i x_i^\top}{e^{z_i^\top\gamma}} \\ &\quad - \sum_{i=1}^n \frac{\varphi^2(k_i)}{\Phi^2(k_i)} \frac{(w_i^\top\alpha)^2 x_i x_i^\top}{e^{z_i^\top\gamma}} - \sum_{i=1}^n \frac{\varphi(k_i)k_i}{\Phi(k_i)} \frac{(w_i^\top\alpha)^2 x_i x_i^\top}{e^{z_i^\top\gamma}},\end{aligned}$$

$$\begin{aligned}&\frac{\partial^2 \ell(\theta)}{\partial \beta \partial \gamma^\top} \\ &= (\nu+1)\sum_{i=1}^n \left\{\nu + \frac{(y_i - x_i^\top\beta)^2}{e^{z_i^\top\gamma}}\right\}^{-2} \frac{(y_i - x_i^\top\beta)^3 x_i z_i^\top}{e^{2z_i^\top\gamma}} \\ &\quad - (\nu+1)\sum_{i=1}^n \left\{\nu + \frac{(y_i - x_i^\top\beta)^2}{e^{z_i^\top\gamma}}\right\}^{-1} \frac{(y_i - x_i^\top\beta)x_i z_i^\top}{e^{z_i^\top\gamma}} \\ &\quad - \frac{1}{2}\sum_{i=1}^n \left[\frac{\varphi^2(k_i)k_i}{\Phi^2(k_i)} + \frac{\varphi(k_i)(k_i - 1)}{\Phi(k_i)}\right] \frac{w_i^\top\alpha}{e^{\frac{1}{2}z_i^\top\gamma}} x_i z_i^\top\end{aligned}$$

$$\begin{aligned}&\frac{\partial^2 \ell(\theta)}{\partial \beta \partial \alpha^\top} \\ &= -\sum_{i=1}^n \left[\frac{\varphi(k_i)(1 - k_i^2)}{\Phi(k_i)} - \frac{\varphi^2(k_i)k_i}{\Phi^2(k_i)}\right] \frac{x_i w_i^\top}{e^{\frac{1}{2}z_i^\top\gamma}},\end{aligned}$$

$$
\begin{aligned}
&\frac{\partial^2 \ell(\theta)}{\partial\gamma\partial\gamma^\top}\\
= \;& \frac{\nu+1}{2}\sum_{i=1}^{n}\left\{\nu+\frac{(y_i-x_i^\top\beta)^2}{\mathrm{e}^{z_i^\top\gamma}}\right\}^{-2}\frac{(y_i-x_i^\top\beta)^4 z_i z_i^\top}{\mathrm{e}^{2z_i^\top\gamma}}\\
&-\frac{\nu+1}{2}\sum_{i=1}^{n}\left\{\nu+\frac{(y_i-x_i^\top\beta)^2}{\mathrm{e}^{z_i^\top\gamma}}\right\}^{-1}\frac{(y_i-x_i^\top\beta)^2 z_i z_i^\top}{\mathrm{e}^{z_i^\top\gamma}}\\
&-\frac{1}{4}\sum_{i=1}^{n}\frac{\varphi^2(k_i)k_i^2}{\Phi^2(k_i)}z_i z_i^\top-\frac{1}{4}\sum_{i=1}^{n}\frac{\varphi(k_i)k_i(k_i-1)}{\Phi(k_i)}z_i z_i^\top,\\
&\frac{\partial^2 \ell(\theta)}{\partial\gamma\partial\alpha^\top}\\
= \;& -\frac{1}{2}\sum_{i=1}^{n}\left[\frac{\varphi(k_i)(1-k_i^2)}{\Phi(k_i)}-\frac{\varphi^2(k_i)k_i}{\Phi^2(k_i)}\right]\frac{y_i-x_i^\top\beta}{\mathrm{e}^{\frac{1}{2}z_i^\top\gamma}}z_i w_i^\top,\\
&\frac{\partial^2 \ell(\theta)}{\partial\alpha\partial\alpha^\top}\\
= \;& -\sum_{i=1}^{n}\left[\frac{\varphi(k_i)k_i}{\Phi(k_i)}+\frac{\varphi^2(k_i)}{\Phi^2(k_i)}\right]\left(\frac{y_i-x_i^\top\beta}{\mathrm{e}^{\frac{1}{2}z_i^\top\gamma}}\right)^2 w_i w_i^\top.
\end{aligned}
$$

We summarize the calculation of the penalized maximum likelihood estimators of the parameters in the following algorithm.

Step 1. Calculate the maximum likelihood estimators of $\theta$ without penalty, denoted by $\hat\theta_{\mathrm{MLE}} = (\hat\beta_{\mathrm{MLE}}^\top, \hat\gamma_{\mathrm{MLE}}^\top, \hat\alpha_{\mathrm{MLE}}^\top)^\top$. Set the initial values $\theta^{(0)} = \hat\theta_{\mathrm{MLE}}$ and $t = 0$.

Step 2. Given the $t$-th approximation $\theta^{(t)} = (\beta^{(t)\top}, \gamma^{(t)\top}, \alpha^{(t)\top})^\top$, update $\theta^{(t+1)} = \theta^{(t)} + \{H(\theta^{(t)}) - n\Sigma_\tau(\theta^{(t)})\}^{-1}\{n\Sigma_\tau(\theta^{(t)})\theta^{(t)} - U(\theta^{(t)})\}$.

**Remark 1:** Note that the skew $t$-normal density given by (1) is unimodal. In addition, the parameter space $\Theta$ is compact and the true value $\theta_0$ of the parameter vector $\theta$ is located in the interior of the parameter space $\Theta$, which guarantees the convergence of the above algorithm.

**Remark 2:** In the derivation of the proposed algorithm, we employed the local quadratic function to approximate the penalty function $p_\tau(\phi)$. A better approximation could be achieved by using the local linear function as in Zou and Li (2008). However, different penalties will result in different weighting schemes in the local linear approximation, e.g., the LASSO gives a constant weighting scheme. For simplicity, in this paper, we choose to use the local quadratic approximation.

## 3.2 Choosing the tuning parameters

There are several criteria, e.g., CV, GCV, AIC and BIC, which can be used to determine optimal tuning parameters. Wang et al. (2007) suggested using BIC to choose the optimal tuning parameter in linear models and partially linear models with smoothly clipped absolute deviation (SCAD) penalty, and proved its model selection consistency property; that is, the optimal parameter chosen based on BIC can identify the true model with probability tending to one. We also adopt BIC to select the optimal $\tau$'s.

It is expected that the choice of $\tau_{1j}$, $\tau_{2k}$ and $\tau_{3m}$ should satisfy that the tuning parameter for a zero coefficient is larger than those for nonzero coefficients. Thus we can unbiasedly estimate larger coefficients, and shrink the small coefficients towards zero simultaneously. Hence, similar to Wu, Zhang and Xu (2013), we suggest

(i) $\tau_{1j} = \tau/|\hat\beta_j^0|$, $j = 1, \ldots, p$,
(ii) $\tau_{2k} = \tau/|\hat\gamma_k^0|$, $k = 1, \ldots, q$,
(iii) $\tau_{3m} = \tau/|\hat\alpha_m^0|$, $m = 1, \ldots, r$,

where $\hat\beta_j^0$, $\hat\gamma_k^0$ and $\hat\alpha_m^0$ are initial estimators of $\beta_j$, $\gamma_k$ and $\alpha_m$ respectively by maximizing the un-penalized likelihood. Define

$$\mathrm{BIC}(\tau) = -\frac{2}{n}\ell(\hat\theta_n) + df_\tau \times \frac{\log(n)}{n},$$

where $\hat\theta_n$ is the maximum penalized likelihood estimator of $\theta$, $df_\tau$ $(0 \leqslant df_\tau \leqslant s)$ denotes the number of nonzero components of $\hat\theta_n$, and $\ell(\theta)$ is defined in (3). The optimal tuning parameter $\hat\tau$ can be determined by minimizing $\mathrm{BIC}(\tau)$ over $\tau$.

# 4. SIMULATION STUDIES AND A REAL EXAMPLE

In this section, simulation studies are performed and a real data set on the body mass index (BMI) is analyzed to demonstrate the proposed methods.

## 4.1 Simulation studies

To evaluate the finite sample performance of the proposed unified penalized likelihood method, we conduct some Monte Carlo simulations. For the sake of comparison, we consider three different penalties; that is, SCAD (Fan and Li, 2001), least absolute shrinkage and selection operator (LASSO, Tibshirani, 1996) and the hard thresholding penalty (abbreviated as "HARD" in the follows, see Antoniadis, 1997). The performance of the penalized MLEs $\hat\beta_n$, $\hat\gamma_n$ and $\hat\alpha_n$ are assessed by using the mean square error (MSE):

$$
\begin{aligned}
\mathrm{MSE}(\hat\beta_n) &= E(\hat\beta_n - \beta)^\top(\hat\beta_n - \beta),\\
\mathrm{MSE}(\hat\gamma_n) &= E(\hat\gamma_n - \gamma)^\top(\hat\gamma_n - \gamma),\\
\mathrm{MSE}(\hat\alpha_n) &= E(\hat\alpha_n - \alpha)^\top(\hat\alpha_n - \alpha).
\end{aligned}
$$

In the simulations, let $\beta = (1, 1, 0, 0, 1, 0, 0, 0)^\top$, $\gamma = (0.7, 0.7, 0, 0, 0.7, 0, 0, 0)^\top$ and $\alpha = (0.5, 0.5, 0, 0, 0.5, 0, 0, 0)^\top$. All components in covariate vectors $x_i$, $z_i$ and $w_i$ are independently generated from the uniform distribution $U(-1, 1)$. We independently generate $y_i$ from $\mathrm{StN}(x_i^\top\beta, \exp(z_i^\top\gamma), w_i^\top\alpha, \nu)$ for $i = 1, \ldots, n$. For a given degree of freedom $\nu$ $(\nu = 3, 5)$, a given sample size $n$

*Table 1. Comparisons with $\nu = 3$ and different combinations of sample size and penalty*

| Model | $n$ | SCAD | | | LASSO | | | HARD | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | C | IC | MSE | C | IC | MSE | C | IC | MSE |
| Location Model | 20 | 4.4720 | 0.6640 | 1.2968 | 3.9680 | 0.6440 | 1.2542 | 3.7100 | 0.3940 | 1.6165 |
| | 50 | 4.7040 | 0.2260 | 0.5235 | 4.0660 | 0.1540 | 0.6123 | 4.3920 | 0.1680 | 0.6150 |
| | 100 | 4.8700 | 0.0120 | 0.1703 | 4.2540 | 0.0100 | 0.2811 | 4.8340 | 0.0120 | 0.1837 |
| | 200 | 4.9250 | 0 | 0.1030 | 4.6000 | 0 | 0.1758 | 4.9300 | 0 | 0.0998 |
| | 250 | 4.9450 | 0 | 0.0748 | 4.7450 | 0 | 0.1202 | 4.9600 | 0 | 0.0749 |
| | 300 | 4.9650 | 0 | 0.0630 | 4.8450 | 0 | 0.1099 | 4.9620 | 0 | 0.0604 |
| Scale Model | 20 | 4.2980 | 1.8540 | 1.8959 | 4.2420 | 1.9100 | 1.3432 | 3.2820 | 1.2600 | 2.3897 |
| | 50 | 4.4780 | 1.4460 | 1.2575 | 4.3380 | 1.3780 | 1.0151 | 4.0300 | 1.1200 | 1.2881 |
| | 100 | 4.6660 | 0.9600 | 0.7654 | 4.4400 | 0.7480 | 0.7251 | 4.5600 | 0.7920 | 0.7212 |
| | 200 | 4.7050 | 0.5600 | 0.4742 | 4.6150 | 0.4450 | 0.6076 | 4.7200 | 0.5350 | 0.4634 |
| | 250 | 4.7400 | 0.4300 | 0.3771 | 4.6550 | 0.4200 | 0.6071 | 4.7600 | 0.4350 | 0.3782 |
| | 300 | 4.7950 | 0.2600 | 0.2561 | 4.8350 | 0.2300 | 0.5412 | 4.8650 | 0.2700 | 0.2479 |
| Skewness Model | 20 | 3.5080 | 0.9580 | 2.3753 | 2.7580 | 0.5640 | 2.7502 | 2.3880 | 0.7260 | 3.0622 |
| | 50 | 4.3260 | 0.8540 | 0.9844 | 2.9560 | 0.2620 | 0.5540 | 3.9020 | 0.6420 | 1.3357 |
| | 100 | 4.6720 | 0.6380 | 0.3803 | 3.3080 | 0.1040 | 0.2196 | 4.7300 | 0.6360 | 0.3805 |
| | 200 | 4.7900 | 0.3600 | 0.2264 | 3.5950 | 0.0200 | 0.1351 | 4.8600 | 0.3750 | 0.2102 |
| | 250 | 4.8100 | 0.1350 | 0.1271 | 3.8400 | 0.0100 | 0.0955 | 4.9100 | 0.1650 | 0.1151 |
| | 300 | 4.8400 | 0.0900 | 0.0829 | 4.0500 | 0.0050 | 0.0798 | 4.9300 | 0.0750 | 0.0772 |

Note: "C" denotes the average number of zero regression coefficients that are correctly estimated as zero and "IC" denotes the average number of non-zero regression coefficients that are erroneously set to zero.

($n = 20, 50, 100, 200, 250, 300$), and a given penalty (SCAD, LASSO, HARD), we calculate the penalized MLEs $\hat{\beta}_n$, $\hat{\gamma}_n$ and $\hat{\alpha}_n$ and repeat the experiment 1,000 times. The simulation results are reported in Tables 1 and 2, where "C" denotes average number of zero regression coefficients that are correctly estimated as zero and "IC" denotes the average number of non-zero regression coefficients that are erroneously set to zero.

From Tables 1 to 2, we can clearly see the following conclusions:

(a) For a given degree of freedom $\nu$ and a given penalty, as expected, the variable selection performs better as the sample size $n$ increases. The MSEs of estimators $\hat{\beta}_n$, $\hat{\gamma}_n$ and $\hat{\alpha}_n$ also become smaller as the sample size $n$ increases.

(b) For a given sample size $n$ and a given $\nu$, the performances of both SCAD and HARD procedures are similar in terms of model error and model complexity. Furthermore, the performances of both SCAD and HARD are significantly better than that of LASSO.

(c) For a given penalty and a given sample size $n$, especially for the scale model, the variable selection performs better as the degree of freedom $\nu$ increases. The MSEs of estimators $\hat{\gamma}_n$ also become smaller as the degree of freedom $\nu$ increases.

(d) For a given combination of $\nu$, sample size $n$ and penalty, the performance of variable selection in the location model is significantly better than that in the scale model and in the skewness model, the skewness model is significantly better than that in the scale model in the sense of model error and model complexity.

## 4.2 Application to the body mass index data

Now, we illustrate the proposed variable selection procedure by using the body mass index data for 102 male and 100 female athletes collected at Australian Institute of Sport (Cook and Weisberg, 1994). The BMI data set consists of the response variable $Y$—BMI in weight/(height)$^2$ and eight predictors: $X_1$—the red cell count; $X_2$—the white cell count; $X_3$—the Hematocrit; $X_4$— the Hemoglobin; $X_5$—the plasma ferritin concentration; $X_6$—the sum of skin folds; $X_7$—the body fat percentage and $X_8$—the lean body mass. We are interested in establishing the relationship between the body mass index $Y$ and the important predictors.

In practice, we may treat the degrees of freedom $\nu$ as an additional unknown parameter and its MLE can be obtained from the following profile log-likelihood

$$\begin{aligned}
\ell_p(\nu) &= \ell(\tilde{\beta}(\nu), \tilde{\gamma}(\nu), \tilde{\alpha}(\nu), \nu) \\
&= \frac{n\nu}{2}\log(\nu) + n\log\Gamma\left(\frac{\nu+1}{2}\right) - n\log\Gamma\left(\frac{\nu}{2}\right) \\
&\quad -\frac{\nu+1}{2}\sum_{i=1}^{n}\log\left\{\nu + \frac{[y_i - x_i^\top\tilde{\beta}(\nu)]^2}{\exp[z_i^\top\tilde{\gamma}(\nu)]}\right\} \\
&\quad -\frac{1}{2}\sum_{i=1}^{n} z_i^\top\tilde{\gamma}(\nu) + \sum_{i=1}^{n}\log\Phi(k_i(\nu)),
\end{aligned}$$

where $\tilde{\beta}(\nu)$, $\tilde{\gamma}(\nu)$ and $\tilde{\alpha}(\nu)$ denote the restricted MlEs of $\beta$,

*Table 2. Comparisons with $\nu = 5$ and different combinations of sample size and penalty*

| Model | $n$ | SCAD | | | LASSO | | | HARD | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | C | IC | MSE | C | IC | MSE | C | IC | MSE |
| Location Model | 20 | 4.4760 | 0.7100 | 1.3923 | 3.8740 | 0.7640 | 1.3830 | 3.4300 | 0.3920 | 1.9189 |
| | 50 | 4.6960 | 0.2780 | 0.6169 | 3.8840 | 0.2640 | 0.7687 | 4.1560 | 0.1680 | 0.8013 |
| | 100 | 4.8780 | 0.0260 | 0.2033 | 4.1520 | 0.0400 | 0.3571 | 4.8100 | 0.0220 | 0.2343 |
| | 200 | 4.8950 | 0 | 0.1272 | 4.5050 | 0.0050 | 0.2461 | 4.9100 | 0 | 0.1295 |
| | 250 | 4.9250 | 0 | 0.0921 | 4.6200 | 0.0050 | 0.1646 | 4.9450 | 0 | 0.0901 |
| | 300 | 4.9650 | 0 | 0.0688 | 4.8050 | 0 | 0.1517 | 4.9650 | 0 | 0.0662 |
| Scale Model | 20 | 4.4880 | 1.9080 | 1.6243 | 4.1560 | 1.8060 | 1.2478 | 3.5360 | 1.2600 | 1.9244 |
| | 50 | 4.6100 | 1.4960 | 1.1184 | 4.2840 | 1.1820 | 0.9296 | 4.2280 | 1.1140 | 1.1058 |
| | 100 | 4.7300 | 0.9480 | 0.6928 | 4.4280 | 0.6380 | 0.6527 | 4.6640 | 0.7880 | 0.6363 |
| | 200 | 4.8200 | 0.5400 | 0.4200 | 4.6250 | 0.3450 | 0.5394 | 4.8150 | 0.5000 | 0.4016 |
| | 250 | 4.8300 | 0.3400 | 0.2854 | 4.7250 | 0.2250 | 0.4614 | 4.8550 | 0.3200 | 0.2738 |
| | 300 | 4.8550 | 0.1914 | 0.2561 | 4.7950 | 0.1050 | 0.4082 | 4.8800 | 0.1500 | 0.1780 |
| Skewness Model | 20 | 3.4500 | 1.1120 | 2.8505 | 2.7040 | 0.7280 | 2.9804 | 2.2620 | 1.0180 | 2.9804 |
| | 50 | 4.2080 | 1.0400 | 1.0795 | 2.9020 | 0.3500 | 0.6703 | 3.7080 | 0.9200 | 1.5409 |
| | 100 | 4.5880 | 0.8460 | 0.5020 | 3.2000 | 0.1660 | 0.2934 | 4.5780 | 0.7560 | 0.5149 |
| | 200 | 4.7650 | 0.4750 | 0.2759 | 3.5300 | 0.0550 | 0.1539 | 4.8000 | 0.5150 | 0.2791 |
| | 250 | 4.7900 | 0.3300 | 0.1998 | 3.7800 | 0.0250 | 0.1268 | 4.8850 | 0.3800 | 0.1894 |
| | 300 | 4.8250 | 0.1450 | 0.1146 | 4.0300 | 0.0200 | 0.0990 | 4.8900 | 0.1750 | 0.1158 |

$\gamma$ and $\alpha$ when $\nu$ is fixed,

$$k_i(\nu) = \frac{w_i^\top \tilde{\alpha}(\nu)[y_i - x_i^\top \tilde{\beta}(\nu)]}{\exp(z_i^\top \tilde{\gamma}(\nu)/2)}.$$

In the current real application, we obtain $\hat{\nu} = 3.86$.

Figure 1 plots the histogram and the probability density function curve of $Y$, indicating that the BMI approximately follows an StN distribution. The purpose of our modeling may be helpful to understand what predictors/factors affecting the location, scale and skewness of $Y$.

Therefore, we could model the BMI data by the following joint location, scale and skewness model with an StN distribution:

$$\begin{cases} y_i \overset{\text{ind}}{\sim} \text{StN}(\mu_i, \sigma_i^2, \lambda_i, \nu), \quad \hat{\nu} = 3.86, \quad i = 1, \ldots, 202, \\ \mu_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_8 X_{i8}, \\ \log \sigma_i^2 = \gamma_0 + \gamma_1 X_{i1} + \gamma_2 X_{i2} + \cdots + \gamma_8 X_{i8}, \\ \lambda_i = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \cdots + \alpha_8 X_{i8}. \end{cases}$$

We apply the proposed variable selection procedure based on the SCAD, LASSO and HARD penalties in Section 2 to the above model. The results are displayed in Table 3.

From Table 3, we notice that in this data example, the SCAD and HARD based methods perform very similarly in terms of the selected variables. We can see that our procedure identified seven nonzero regression coefficients $\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$, $\beta_5$, $\beta_7$ and $\beta_8$ in the location model, four nonzero regression coefficients $\gamma_1$, $\gamma_2$, $\gamma_3$ and $\gamma_7$ in the scale model and seven nonzero regression coefficients $\alpha_1$, $\alpha_2$, $\alpha_3$, $\alpha_4$, $\alpha_6$, $\alpha_7$ and $\alpha_8$ in the skewness model.

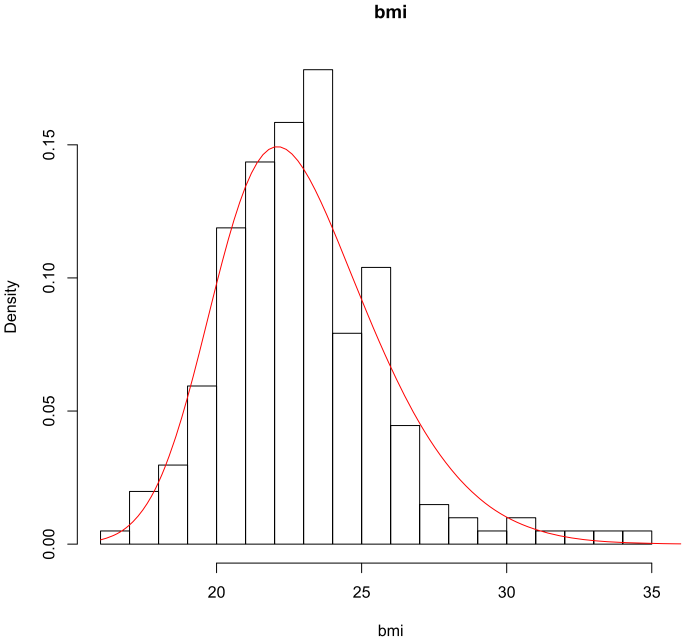

*Figure 1. Histogram of the body mass index and fitted StN density function.*

This indicates that the $X_6$ (the sum of skin folds) has no significant impact on the location of $Y$ (the body mass index, weight/(height)$^2$ ); $X_4$( the Hemoglobin), $X_5$(the plasma ferritin concentration), $X_6$(the sum of skin folds), $X_8$(the lean body mass) have also no significant impact on the scale of $Y$ (the body mass index). Furthermore, $X_5$(the plasma ferritin concentration) has also no significant impact on the skewness of $Y$ (the body mass index).

*Table 3. Penalized maximum likelihood estimators of parameters for the BMI data*

| Model | Method | Const | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ |
|---|---|---|---|---|---|---|---|---|---|---|
| Location Model | SCAD | −31.4833 | 1.8674 | −5.8179 | 0.3654 | 0.9318 | −0.0110 | 0 | 0.5360 | 0.2492 |
| | LASSO | −23.0420 | 1.5784 | −1.9311 | 0.1173 | 0.5428 | −0.0072 | 0 | 0.2581 | 0.1206 |
| | HARD | −31.4916 | 1.7673 | −5.7813 | 0.3722 | 0.9381 | −0.0109 | 0 | 0.5371 | 0.2471 |
| Scale Model | SCAD | 7.0534 | −0.1333 | 0.3244 | −0.0284 | 0 | 0 | 0 | −0.0158 | 0 |
| | LASSO | 6.5806 | −0.1492 | 0.1300 | −0.0191 | 0.0289 | 0 | 0 | 0 | 0 |
| | HARD | 7.0544 | −0.1271 | 0.3224 | −0.0289 | 0 | 0 | 0 | −0.0159 | 0 |
| Skewness Model | SCAD | 13.9388 | 21.8682 | −0.7605 | 1.6106 | −12.0149 | 0 | −0.0622 | 0.0624 | 0.4234 |
| | LASSO | 10.6491 | −0.3135 | 0.0893 | 0.2663 | −0.2804 | 0 | 0.0335 | −0.1760 | 0 |
| | HARD | 13.7397 | 19.7302 | −0.6301 | 1.4854 | −10.9999 | 0 | −0.0556 | 0.0479 | 0.3857 |

## 5. CONCLUSION

We proposed an efficient and unified penalized likelihood procedure which can simultaneously select significant variables and estimate unknown regression coefficients in the joint location, scale and skewness models with the StN distribution when the data under consideration involve heavy tail and asymmetric outcomes. Furthermore, with proper choice of tuning parameters, we showed that this variable selection procedure is consistent, and the estimators of regression coefficients have oracle property. Simulation studies are performed and a real data set on the BMI data is analyzed to illustrate the proposed methods.

The proposed method is valid for the fixed number of parameters. It would be interesting to consider the case when the number of parameters goes to infinity. In some applications, it is necessary to develop some new theories and methods to obtain the variable selection in semiparametric joint location, scale and skewness models with the StN distribution. The proposed method for variable selection in joint location, scale and skewness models with a skew-t-normal distribution could be formulated under the generalized additive model for location, scale and shape (GAMLSS) framework (Rigby and Stasinopoulos, 2005).

The MATLAB codes of this paper are available on request.

## ACKNOWLEDGEMENTS

The authors would like to thank the Associate Editor and one referee for their valuable comments which have greatly improved the paper. The first author's research is fully supported by the National Natural Science Foundation of China (No. 11261025, 11126309).

## APPENDIX A: PROOF OF THE DENSITY FUNCTION (1) BEING UNIMODAL

The density function of $Y \sim \mathrm{StN}(\mu, \sigma^2, \nu, \lambda)$ is given by

$$\begin{aligned} f(y) &= \frac{2}{\sigma} t_\nu\left(\frac{y-\mu}{\sigma}\right) \Phi\left(\lambda \frac{y-\mu}{\sigma}\right) \\ &= \frac{2\Gamma((\nu+1)/2)}{\sigma\sqrt{\pi\nu}\Gamma(\nu/2)} \left[1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right]^{-(\nu+1)/2} \\ &\quad \times \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\lambda\frac{y-\mu}{\sigma}} \exp(-x^2/2)\,\mathrm{d}x, \quad y \in \mathbb{R}. \end{aligned}$$

When $\lambda = 0$, $f(y)$ reduces to the density of the univariate $t$-distribution, which is symmetrical and unimodal. We first consider the case of $\lambda > 0$.

When $\lambda > 0$, $f(y)$ is unimodal if and only if the root of the equation $f'(y) = 0$ exists and is unique, where

$$\begin{aligned} & f'(y) \\ &= \frac{2\Gamma((\nu+1)/2)}{\sigma\sqrt{\pi\nu}\Gamma(\nu/2)} \left[1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right]^{-(\nu+1)/2} \frac{1}{\sigma\sqrt{2\pi}} \\ &\quad \times \left\{ -\frac{\nu+1}{2}\left[1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right]^{-1} \frac{2(y-\mu)}{\nu\sigma} \right. \\ &\quad \left. \times \int_{-\infty}^{\lambda\frac{y-\mu}{\sigma}} \mathrm{e}^{-x^2/2}\,\mathrm{d}x + \lambda \exp\left[-\frac{\lambda^2}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right] \right\}. \end{aligned}$$

Let $z = (y-\mu)/\sigma \in \mathbb{R}$. That there exists a $y_0 \in \mathbb{R}$ such that $f'(y_0) = 0$ is equivalent to that there is a $z_0 = (y_0 - \mu)/\sigma \in \mathbb{R}$ such that $g(z) = 0$, where

$$\begin{aligned} g(z) \hat{=} & -\frac{\nu+1}{2}\left(1 + \frac{z^2}{\nu}\right)^{-1} \frac{2z}{\nu} \int_{-\infty}^{\lambda z} \mathrm{e}^{-x^2/2}\,\mathrm{d}x \\ & + \lambda \exp\left(-\frac{\lambda^2 z^2}{2}\right). \end{aligned} \tag{A.1}$$

It is easy to see that $g(0) = \lambda > 0$ and $g(z) > 0$ when $z < 0$. When $z > 0$, we have $\lambda z > 0$. Since

$$\begin{aligned} & \int_{-\infty}^{\lambda z} \mathrm{e}^{-x^2/2}\,\mathrm{d}x \\ &= \int_{-\infty}^{0} \mathrm{e}^{-x^2/2}\,\mathrm{d}x + \int_{0}^{\lambda z} \mathrm{e}^{-x^2/2}\,\mathrm{d}x \\ &= \sqrt{\frac{\pi}{2}} + \int_{0}^{\lambda z} \mathrm{e}^{-x^2/2}\,\mathrm{d}x \end{aligned}$$

$$
\begin{aligned}
&> \int_0^{\lambda z} \mathrm{e}^{-x^2/2}\,\mathrm{d}x > \int_0^{\lambda z} \min_{x\in[0,\lambda z]} \mathrm{e}^{-x^2/2}\,\mathrm{d}x \\
&= \int_0^{\lambda z} \mathrm{d}x \cdot \left( \min_{x\in[0,\lambda z]} \mathrm{e}^{-x^2/2} \right) = \lambda z \cdot \mathrm{e}^{-\lambda^2 z^2/2},
\end{aligned}
$$

from (A.1), we have

$$
\begin{aligned}
g(z) \;&<\; -\frac{\nu+1}{2}\left(1+\frac{z^2}{\nu}\right)^{-1}\frac{2z}{\nu}\cdot\lambda z\cdot\exp\left(-\frac{\lambda^2 z^2}{2}\right) \\
&\qquad +\lambda\exp\left(-\frac{\lambda^2 z^2}{2}\right) \\
&= \lambda\left[-(\nu+1)\left(1+\frac{z^2}{v}\right)^{-1}\frac{z^2}{\nu}+1\right]\exp\left(-\frac{\lambda^2 z^2}{2}\right) \\
&= \frac{\lambda\nu(1-z^2)}{\nu+z^2}\exp\left(-\frac{\lambda^2 z^2}{2}\right) \\
&\leq 0, \quad \text{if } \; z \geq 1.
\end{aligned}
$$

In particular, $g(1) < 0$. Since $g(z)$ is a continuous function, based on the zero point theorem in mathematical analysis, there exists a $z_0 \in (0,1)$ such that $g(z_0) = 0$. In other words, there is a $y_0 \in (\mu, \mu+\sigma)$ satisfying $f'(y_0) = 0$. Therefore, the existence is verified.

To verify that $z_0 \in (0,1)$ is the unique solution to the equation $g(z) = 0$, we only need to show that $g(z)$ is monotone in the unit interval $(0,1)$. In fact, we have

$$
\begin{aligned}
g'(z) \;&=\; \frac{(\nu+1)(z^2-\nu)}{(z^2+\nu)^2}\int_{-\infty}^{\lambda z}\mathrm{e}^{-x^2/2}\,\mathrm{d}x \\
&\qquad -\lambda z\cdot\mathrm{e}^{-\lambda^2 z^2/2}\left(\frac{\nu+1}{\nu+z^2}+\lambda^2\right) \\
&\hat{=}\; I_1 + I_2.
\end{aligned}
$$

If $0 < z < 1$, we have $z^2 - \nu \leq 0$ and $\lambda z > 0$, implying that $I_1 \leq 0$ and $I_2 < 0$. Hence, $g'(z) < 0$ for any $z \in (0,1)$. That is, $g(z)$ is monotone decreasing in the unit interval $(0,1)$. The uniqueness is verified.

When $\lambda < 0$, the conclusion can be proved similarly.

## APPENDIX B: PROOFS OF THE THEOREMS

**Proof of Theorem 1.** For any given $\varepsilon > 0$, we first prove that there exists a large constant $C$ such that

$$
\Pr\left\{\sup_{\|v\|=C}\mathcal{L}(\theta_0+n^{-\frac{1}{2}}v) < \mathcal{L}(\theta_0)\right\} \geqslant 1-\varepsilon.
$$

Note that $p_{\tau_{jn}}(0) = 0$ and $p_{\tau_{jn}}(\cdot) > 0$, we obtain

$$
\begin{aligned}
&\mathcal{L}\left(\theta_0+n^{-\frac{1}{2}}v\right)-\mathcal{L}(\theta_0) \\
&= \left[\ell(\theta_0+n^{-\frac{1}{2}}v) - n\sum_{j=1}^{s} p_{\tau_{jn}}(|\theta_{0j}+n^{-\frac{1}{2}}v_j|)\right] \\
&\quad - \left[\ell(\theta_0) - n\sum_{j=1}^{s} p_{\tau_{jn}}(|\theta_{0j}|)\right] \\
&\leqslant \left[\ell(\theta_0+n^{-\frac{1}{2}}v)-\ell(\theta_0)\right] \\
&\quad - n\sum_{j=1}^{s_1}\left[p_{\tau_{jn}}(|\theta_{0j}+n^{-\frac{1}{2}}v_j|)-p_{\tau_{jn}}(|\theta_{0j}|)\right] \\
&\hat{=}\; K_1 + K_2.
\end{aligned}
$$

We first consider $K_1$. Using the Taylor expansion, we have

$$
\begin{aligned}
K_1 \;&=\; \ell(\theta_0+n^{-\frac{1}{2}}v)-\ell(\theta_0) \\
&=\; n^{-\frac{1}{2}}v^\top\ell'(\theta_0)+\tfrac{1}{2}n^{-1}v^\top\ell''(\theta^*)v \\
&\hat{=}\; K_{11}+K_{12},
\end{aligned}
$$

where $\theta^*$ lies between $\theta_0$ and $\theta_0 + n^{-\frac{1}{2}}v$. Note that $n^{-\frac{1}{2}}\|\ell'(\theta_0)\| = O_p(1)$. By applying the Cauchy–Schwartz inequality, we obtain

$$
K_{11} = n^{-\frac{1}{2}}v^\top\ell'(\theta_0) \leqslant n^{-\frac{1}{2}}\|\ell'(\theta_0)\|\cdot\|v\| = O_p(1).
$$

According to Chebyshev's inequality, we know that for any $\varepsilon > 0$,

$$
\begin{aligned}
&\Pr\left\{\frac{1}{n}\|\ell''(\theta_0)-E\ell''(\theta_0)\|\geqslant\varepsilon\right\} \\
&\leqslant\; \frac{1}{n^2\varepsilon^2}E\left\{\sum_{j=1}^{s}\sum_{l=1}^{s}\left(\frac{\partial^2\ell(\theta_0)}{\partial\theta_j\partial\theta_l}-E\frac{\partial^2\ell(\theta_0)}{\partial\theta_j\partial\theta_l}\right)^2\right\} \\
&\leqslant\; \frac{Cs^2}{n\varepsilon^2} = o(1),
\end{aligned}
$$

which implies $\frac{1}{n}\|\ell''(\theta_0)-E\ell''(\theta_0)\| = o_p(1)$ and

$$
\begin{aligned}
&K_{12} \\
&=\; \frac{1}{2}n^{-1}v^\top\ell''(\theta^*)v = \frac{1}{2}v^\top\Big[n^{-1}\ell''(\theta_0)\Big]v\Big[1+o_p(1)\Big] \\
&=\; \frac{1}{2}v^\top\Big\{n^{-1}\Big[\ell''(\theta_0)-E\ell''(\theta_0)-\mathcal{I}(\theta_0)\Big]\Big\}v\Big[1+o_p(1)\Big] \\
&=\; -\frac{1}{2}v^\top\mathcal{I}(\theta_0)v\Big[1+o_p(1)\Big].
\end{aligned}
$$

Therefore, we conclude that $K_{12}$ uniformly dominates $K_{11}$ in $\|v\| = C$ if the constant $C$ is sufficiently large.

Next, we study the term $K_2$. It follows from the Taylor expansion and the Cauchy–Schwartz inequality that

$$
\begin{aligned}
K_2 \;&=\; -n\sum_{j=1}^{s_1}\Big[p_{\tau_{jn}}(|\theta_{0j}+n^{-\frac{1}{2}}v_j|)-p_{\tau_{jn}}(|\theta_{0j}|)\Big] \\
&=\; -n\sum_{j=1}^{s_1}\Big\{n^{\frac{1}{2}}p'_{\tau_{jn}}(|\theta_{0j}|)\mathrm{sgn}(\theta_{0j})v_j
\end{aligned}
$$

$$
\begin{aligned}
& +\frac{1}{2}p''_{\tau_{jn}}(|\theta_{0j}|)v_j^2\Big[1+O_p(1)\Big]\Big\} \\
\leqslant\ & \sqrt{s_1}n^{\frac{1}{2}}\|v\|\max_{1\leqslant j\leqslant s}\Big\{p'_{\tau_{jn}}(|\theta_{j0}|),\theta_{j0}\neq 0\Big\} \\
& +\frac{1}{2}\|v\|^2\max_{1\leqslant j\leqslant s}\Big\{|p''_{\tau_{jn}}(|\theta_{j0}|)|\colon\ \theta_{j0}\neq 0\Big\} \\
=\ & \sqrt{s_1}n^{\frac{1}{2}}\|v\|a_n+\frac{1}{2}\|v\|^2 b_n.
\end{aligned}
$$

Since it is assumed that $a_n = O_p(n^{-\frac{1}{2}})$ and $b_n \to 0$, we conclude that $K_{12}$ dominates $K_2$ if we choose a sufficiently large $C$. Therefore, for any given $\varepsilon > 0$, there exists a large constant $C$ such that

$$
\Pr\left\{\sup_{\|v\|=C}\mathcal{L}(\theta_0+n^{-\frac{1}{2}}v)<\mathcal{L}(\theta_0)\right\}\geqslant 1-\varepsilon,
$$

implying that there exists a local maximizer $\hat{\theta}_n$ such that $\hat{\theta}_n$ is a $\sqrt{n}$-consistent estimator of $\theta_0$. This completed the proof of Theorem 1.

**Proof of Theorem 2.** We first prove Part (i). From $\tau_{\max} \to 0$, it is easy to show that $a_n = 0$ for large $n$. Second, we prove that for any given $\theta^{(1)}$ satisfying $\theta^{(1)} - \theta_0^{(1)} = O_p(n^{-1/2})$ and any constant $C > 0$, we have

$$
\mathcal{L}(\theta^{(1)},0)=\max_{\|\theta^{(1)}\|\leqslant Cn^{-1/2}}\mathcal{L}(\theta^{(1)},\theta^{(2)}).
$$

In fact, for any $\theta_j (j = s_1 + 1, \ldots, s)$, using the Taylor expansion, we obtain

$$
\begin{aligned}
& \frac{\partial\mathcal{L}(\theta)}{\partial\theta_j} \\
=\ & \frac{\partial\ell(\theta)}{\partial\theta_j}-np'_{\tau_{jn}}(|\theta_j|)\mathrm{sgn}(\theta_j) \\
=\ & \frac{\partial\ell(\theta_0)}{\partial\theta_j}+\sum_{l=1}^{s}\frac{\partial^2\ell(\theta^*)}{\partial\theta_j\partial\theta_l}(\theta_l-\theta_{0l})-np'_{\tau_{jn}}(|\theta_j|)\mathrm{sgn}(\theta_j),
\end{aligned}
$$

where $\theta^*$ is a point between $\theta$ and $\theta_0$. By the standard argument, we have

$$
\frac{1}{n}\frac{\partial\ell(\theta_0)}{\partial\theta_j}=O_p(n^{-1/2})
$$

and

$$
\frac{1}{n}\left\{\frac{\partial^2\ell(\theta_0)}{\partial\theta_j\partial\theta_l}-E(\frac{\partial^2\ell(\theta_0)}{\partial\theta_j\partial\theta_l})\right\}=O_p(1).
$$

Note that $\|\hat{\theta}_n - \theta_0\| = O_p(n^{-1/2})$, we have

$$
\frac{\partial\mathcal{L}(\theta)}{\partial\theta_j}=-n\tau_{jn}\left\{\tau_{jn}^{-1}p'_{\tau_{jn}}(|\theta_j|)\mathrm{sgn}(\theta_j)+O_p(\tau_{jn}^{-1}n^{-1/2})\right\}.
$$

According to the assumption in Theorem 2, we obtain

$$
\liminf_{n\to\infty}\liminf_{\theta\to 0^+}\tau_{jn}^{-1}p'_{\tau_{jn}}(\theta)>0\quad\text{and}\quad\tau_{jn}^{-1}n^{-1/2}\to 0.
$$

So that

$$
\frac{\partial\mathcal{L}(\theta)}{\partial\theta_j}<0,\quad\text{for}\quad 0<\theta_j<Cn^{-1/2}
$$

and

$$
\frac{\partial\mathcal{L}(\theta)}{\partial\theta_j}>0,\quad\text{for}\quad -Cn^{-1/2}<\theta_j<0.
$$

Therefore, $\mathcal{L}(\theta)$ achieve its maximum at $\theta = ((\theta^{(1)})^\top, 0^\top)^\top$ and this completes the proof of the first part of Theorem 2.

Second, we study the asymptotic normality of $\hat{\theta}_n^{(1)}$. From Theorem 1 and the first part of Theorem 2, there exists a penalized maximum likelihood estimator $\hat{\theta}_n^{(1)}$ that is the $\sqrt{n}$-consistent local maximizer of the function $\mathcal{L}(\theta^{(1)}, 0)$. The estimator $\hat{\theta}_n^{(1)}$ must satisfy

$$
\begin{aligned}
0\ =\ & \left.\frac{\partial\mathcal{L}(\theta)}{\partial\theta_j}\right|_{\theta=(\hat{\theta}^{(1)\top},\,0^\top)^\top}-np'_{\tau_{jn}}(|\hat{\theta}_{nj}^{(1)})|)\mathrm{sgn}(\hat{\theta}_{nj}^{(1)}) \\
=\ & \frac{\partial\ell(\theta_0)}{\partial\theta_j}+\sum_{l=1}^{s_1}\left\{\frac{\partial^2\ell(\theta_0)}{\partial\theta_j\partial\theta_l}+O_p(1)\right\}(\hat{\theta}_{nl}^{(1)}-\theta_{0l}^{(1)}) \\
& -np'_{\tau_{jn}}(|\theta_{0j}^{(1)}|)\mathrm{sgn}(\hat{\theta}_{0j}^{(1)}) \\
& -n\left\{p''_{\tau_{jn}}(|\theta_{0j}^{(1)}|)+O_p(1)\right\}\times(\hat{\theta}_{nj}^{(1)}-\theta_{0j}^{(1)}).
\end{aligned}
$$

In other words, we have

$$
\left\{\frac{\partial^2\ell(\theta_0)}{\partial\theta^{(1)}\partial(\theta^{(1)})^\top}+nA_n+O_p(1)\right\}\left(\hat{\theta}_n^{(1)}-\theta_0^{(1)}\right)+c_n=\frac{\partial\ell(\theta_0)}{\partial\theta^{(1)}}.
$$

Using the Liapounov form of the multivariate central limit theorem, we obtain

$$
\frac{1}{\sqrt{n}}\frac{\partial\ell(\theta_0)}{\partial\theta^{(1)}}\xrightarrow{\mathrm{L}}N(0,\mathcal{I}^{(1)}).
$$

Note that

$$
\frac{1}{n}\left\{\frac{\partial^2\ell(\theta_0)}{\partial\theta^{(1)}\partial(\theta^{(1)})^\top}-E(\frac{\partial^2\ell(\theta_0)}{\partial\theta^{(1)}\partial(\theta^{(1)})^\top}\right\}=O_p(1),
$$

it follows immediately by using Slustsky's Theorem that

$$
\begin{aligned}
\sqrt{n}(\bar{\mathcal{I}}_n^{(1)})^{-1/2}(\bar{\mathcal{I}}_n^{(1)}+A_n)\{(\hat{\theta}_n^{(1)}-\theta_0^{(1)})+(\bar{\mathcal{I}}_n^{(1)}+A_n)^{-1}c_n\} \\
\xrightarrow{\mathrm{L}}N_{s_1}(0,I_{s_1}).
\end{aligned}
$$

The second part of Theorem 2 is proved.

*Received 2 March 2015*

## REFERENCES

[1] Antoniadis, A. (1997). Wavelets in statistics: A review (with discussions). *Journal of the Italian Statistical Society* **6**, 97–144.

[2] Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* **12**, 171–178. MR0808153

[3] Azzalini, A. (2005). The skew-normal distribution and related multivariate families (with discussions). *Scandinavian Journal of Statistics* **32**, 159–188. MR2188669

[4] Cabral, C. R. B., Bolfarine, H. and Pereira, J. R. G. (2008). Bayesian density estimation using skew Student-t-normal mixtures. *Computational Statistics & Data Analysis* **52**, 5075–5090. MR2526576

[5] Cook, R. D. and Weisberg, S. (1994). Bayesian density estimation using skew Student-t-normal mixtures. *An Introduction to Regression Graphics*. John Wiley and Sons, New York. MR1285353

[6] Fan, J. Q. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of American Statistical Association* **96**, 1348–1360. MR1946581

[7] Fan, J. Q. and Lv, J. C. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20**, 101–148. MR2640659

[8] Gómez, H. W., Venegas, O. and Bolfarine, H. (2007). Skew-symmetric distributions generated by the distribution function of the normal distribution. *Environmetrics* **18**, 395–407. MR2370731

[9] Ho, H. J., Pyne, S. and Lin, T. I. (2012). Maximum likelihood inference for mixtures of skew Student-t-normal distributions through practical EM-type algorithms. *Statistics and Computing* **22**, 287–299. MR2865071

[10] Hu, Y. and Lian, H. (2013). Variable selection in a partially linear proportional hazards model with a diverging dimensionality. *Statistics and Probability Letters* **83**, 61–69. MR2998724

[11] Lange, K. L, Little, R. J. A. and Taylor, J. M. G. (1989). Robust statistical modelling using the $t$ distribution. *Journal of American Statistical Association* **84**, 881–896. MR1134486

[12] Lin, J. G., Xie, F. C. and Wei, B. C. (2009). Statistical diagnostics for skew-t-normal nonlinear models. *Communications in Statistics-Simulation and Computation* **38**, 2096–2110. MR2751190

[13] Lucas, A. (1997). Robustness of the Student $t$ based M-estimator. *Communications in Statistics-Theory and Methods* **26**, 1165–1182. MR1450228

[14] Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**, 507–554. MR2137253

[15] Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of Royal Statistical Society, Series B* **58**, 267–288. MR1379242

[16] Wang, H., Li, R. and Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–568. MR2410008

[17] Wu, L. C. and Li, H. Q. (2012). Variable selection for joint mean and dispersion models of the inverse Gaussian distribution. *Metrika* **75**, 795–808. MR2956276

[18] Wu, L. C., Zhang, Z. Z. and Xu, D. K. (2012a). Variable selection in joint mean and variance models. *System Engineering—Theory and Practice* **32**, 1754–1760.

[19] Wu, L. C., Zhang, Z. Z. and Xu, D. K. (2012b). Variable selection in joint mean and variance models of Box–Cox transformation. *Journal of Applied Statistics* **39**, 2543–2555. MR2993302

[20] Wu, L. C., Zhang, Z. Z. and Xu, D. K. (2013). Variable selection in joint location and scale models of the skew-normal distribution. *Journal of Statistical Computation and Simulation* **83**, 1266–1278. MR3169234

[21] Zhang, Z. Z. and Wang, D. R. (2011). Simultaneous variable selection for heteroscedastic regression models. *Science China Mathematics* **54**, 515–530. MR2775427

[22] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussions). *The Annals of Statistics* **36**, 1509–1566. MR2435443

Liucang Wu
Faculty of Science
Kunming University of Science and Technology
Kunming
People's Republic of China
E-mail address: wuliucang@163.com

Guo-Liang Tian
Department of Statistics and Actuarial Science
The University of Hong Kong
Hong Kong
People's Republic of China
E-mail address: gltian@hku.hk

Yan-Qing Zhang
Department of Statistics
Yunnan University
Kunming
People's Republic of China
E-mail address: zyqznl2010@126.com

Ting Ma
Faculty of Science
Kunming University of Science and Technology
Kunming
People's Republic of China
E-mail address: mt.silence.0208@163.com