Pak. J. Statist.
2015 Vol. 31(5), 447-460

VARIABLE SELECTION IN JOINT MEANS AND VARIANCE MODELS OF THE PARETO DISTRIBUTION

Ying Dong[1,2], Lixin Song[1], Muhammad Amin[1,3] and **Xinyong Shi[4]**
[1] School of Mathematical Sciences, Dalian University of Technology, Dalian, 116023, P. R. China. Email: lxsong@dlut.edu.cn
[2] Faculty of Science, Dalian Nationalities University, Dalian, 116600, P. R. China. Email: yingd 66@163.com
[3] Nuclear Institute for Food and Agriculture (NIFA), 446, Peshawar, Pakistan. Email: aminkanju@gmail.com
[4] Troops 68048, The Chinese People's Liberation Army, Baoji, 721013, P. R. China. Email: sxy74@163.com

ABSTRACT

Pareto Distribution is a powerful law of probability distribution that accords with city populations, actuarial, geophysical, scientific, and others. Variable selection is vital in the modeling of statistics and also for Pareto distribution. It is a general understanding that most of the existing variable selection methods are confined to mean explanation variables only. In this sense, various joint mean and variance models are also investigated to redefine Pareto Distribution Model. A unified penalized likelihood method which can simultaneously select significant variables in the mean and variance models is proposed. The consistency and the oracle property of the regularized estimators is also established with the help of apt mode of selecting tuning parameters. Finite sample performance of the proposed variable selection procedure is assessed using different simulation studies.

KEYWORDS

Bayesian information criterion (BIC); Joint mean and variance models; Oracle property; Variable selection

1. INTRODUCTION

The Pareto Distribution, named after the Italian economist Vilfredo Pareto, was proposed first as a model for the distribution of city populations within a given area. One of its modern uses is to utilize as model for the distribution of incomes. Lomax (1954) employed it in the analysis of business failure data while Balkema and De Haan (1974) presented that it arises as a limit distribution of lingering lifetime at great age. Bryson (1974) endorsed its use as a heavy tailed alternative to the exponential. On the other, it is also a powerful law of the probability distribution that coincides with actuarial, geophysical, scientific, and many other types of observable singularities.

Variable selection has the vital significance in statistical modeling. Usually, investigators introduce a large number of predictors in order to reduce possible model biases, but in many cases, the number of important covariates is relatively small, so it is

reasonable to assume a sparse model. Therefore, it is a need of variable selection to identify most important variables that provides more interpretable models with better prediction power. Most existing variable selection procedures are only limited to select the mean explanation variables. Nevertheless, modeling the variance will be of direct interest in its own right to identify the source of variability in the observations in many situations such as industrial quality improvement experiments and econometric sector. Thus, it is as important as that of the mean. In present era, a colossal consideration is made to joint mean and variance model. In this view, Park (1966) proposed to a log linear model for the variance parameter and practiced a two stage process to estimate the Gaussian model. Harvey (1976) deliberated Maximum Likelihood (ML) estimation of the location and scale effects and the succeeding likelihood ratio test under the general conditions. Aitkin (1987) projected the ML estimation for a joint mean and variance models and applied it to the commonly cited Minitab tree data. Outliers are common to be in observable, so their accommodation is of interest rather than deletion. Taylor and Verbyla (2004) proposed joint modeling of location and scale parameters of the t-distribution. Generally, distributions from the family of generalized linear models are considered by Lee and Nelder (1998), Smyth and Verbyla (1999) and by Wang and Zhang (2009) as well. All these concluded to estimate the mean and dispersion parameters of the distribution under the double generalized linear models. Wu and Li (2012) conferred the variable selection for joint mean and dispersion models of the inverse Gaussian distribution. Wu et al. (2012) delivered into the model of Box-Cox transformation about the joint mean and variance and do the same on the skew-normal distribution in the year 2013 as well.

Wu et al. (2012) inspired us to formulate the selection of vital explanatory variables which is the backbone of joint mean and variance models of the Pareto distribution. The proposed model is equipped with this requirement. The consistency and the oracle property of the regularized estimators with the help of apt mode of selecting tuning parameters is established. Finite sample performance of the proposed variable selection procedure is assessed through simulation studies. Some of these developments are very close to the research work by Wu (2014).

The contents of this article are organized as follows. First, the joint mean and variance models of the Pareto Distribution are proposed in Section 2, and then discuss the variable selection method for these models via the penalized likelihood function. Furthermore, some statistical properties of our variable selection procedure are presented. The iterative algorithm to compute the penalized maximum likelihood estimators under the proposed models is presented in Section 3. The simulation studies are discussed in Section 4 to illustrate the proposed methodologies.

2. VARIABLE SELECTION IN JOINT MEAN AND VARIANCE MODELS OF THE PARETO DISTRIBUTION VIA PENALIZED MAXIMUM LIKELIHOOD

2.1 Joint Mean and Variance Models of the Pareto Distribution

Consider the following joint mean and variance models of the Pareto distribution:

$$\begin{cases} y_i \sim Pareto(a_i, d_i), \\ \mu_i = E(y_i) = e^{x_i^T \beta} = \dfrac{\alpha_i d_i}{\alpha_i - 1}, \\ \sigma_i^2 = Var(y_i) = e^{z_i^T} \gamma = \dfrac{\alpha_i d_i^2}{(\alpha_i - 1)^2 (\alpha_i - 2)}, \\ i = 1, 2, \cdots, n. \end{cases} \tag{2.1}$$

where $f(y_i | \alpha_i, d_i) = \dfrac{\alpha_i d_i^{\alpha_i}}{y_i^{\alpha_i + 1}}$, $\alpha_i > 0$, $0 < d_i \le y_i < +\infty$. $y = (y_1, \cdots, y_n)^T$ is a vector of n independent responses, and n represents the sample size. $X = (x_1, x_2, \ldots, x_n)^T$ and $Z = (z_1, z_2, \ldots, z_n)^T$ are covariates, where $x_i = (x_{i1}, \ldots, x_{ip})^T$, and $z_i = (z_{i1}, \ldots, z_{iq})^T$. The z_i may contain some or all of the variables in x_i and other variables which are not included in x_i. $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is a $p \times 1$ vector of unknown parameters, $\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_p)^T$ is $q \times 1$ vector of unknown parameters. In this paper, new procedure is proposed to remove the unnecessary explanatory variables from joint mean and variance models of the Pareto distribution.

2.2 Penalized Maximum Likelihood

Many traditional variable selection methods can be considered as a penalized likelihood to balance the model biases and estimation variances (Fan and Li, 2001). Suppose that we have a random sample $(y_i, x_i, z_i), (i = 1, 2, \cdots n)$ from the joint mean and variance models of the Pareto distribution. Let $L(\beta, \gamma)$ denote the log-likelihood function. Then we have

$$\begin{aligned} L(\beta, \gamma) &= \sum_{i=1}^{n} \{\ln \alpha_i + \alpha_i \ln d_i - (\alpha_i + 1) \ln y_i\} \\ &= \sum_{i=1}^{n} \ln \left(\frac{\sqrt{\mu_i^2 + \sigma_i^2}}{\sigma_i} + 1 \right) + \sum_{i=1}^{n} \left(\frac{\sqrt{\mu_i^2 + \sigma_i^2}}{\sigma_i} + 1 \right) \ln \left(\frac{\mu_i^2 + \sigma_i^2 - \sigma_i \sqrt{\mu_i^2 + \sigma_i^2}}{\mu_i} \right) \\ &\quad - \sum_{i=1}^{n} \left(\frac{\sqrt{\mu_i^2 + \sigma_i^2}}{\sigma_i} + 2 \right) \ln y_i \\ &= \sum_{i=1}^{n} \ln \left(\sqrt{e^{2 x_i^T \beta} + e^{z_i^T \gamma}} + e^{\frac{z_i^T \gamma}{2}} \right) - \sum_{i=1}^{n} \frac{z_i^T \gamma}{2} \end{aligned}$$

$$+\sum_{i=1}^{n}\left(\frac{\sqrt{e^{2x_i^T\beta}+e^{z_i^T\gamma}}}{e^{\frac{z_i^T\gamma}{2}}}+1\right)\left[\ln\left(e^{2x_i^T\beta}+e^{z_i^T\gamma}-e^{\frac{z_i^T\gamma}{2}}\sqrt{e^{2x_i^T\beta}+e^{z_i^T\gamma}}\right)-x_i^T\beta\right]$$

$$-\sum_{i=1}^{n}\left(\frac{\sqrt{e^{2x_i^T\beta}+e^{z_i^T\gamma}}}{e^{\frac{z_i^T\gamma}{2}}}+2\right)\ln y_i$$

$$=\sum_{i=1}^{n}\ln\left(k_i+e^{\frac{z_i^T\gamma}{2}}\right)-\sum_{i=1}^{n}\frac{z_i^T\gamma}{2}+\sum_{i=1}^{n}\left(\frac{k_i}{e^{\frac{z_i^T\gamma}{2}}}+1\right)\left[\ln\left(k_i^2-e^{\frac{z_i^T\gamma}{2}}k_i\right)-x_i^T\beta\right]$$

$$-\sum_{i=1}^{n}\left(\frac{k_i}{e^{\frac{z_i^T\gamma}{2}}}+2\right)\ln y_i,$$

where $k_i=\sqrt{e^{2x_i^T\beta}+e^{z_i^T\gamma}}$.

Similar to Fan and Li (2001), we define the penalized likelihood function as follows

$$Q(\beta,\gamma)=L(\beta,\gamma)-n\sum_{j=1}^{p}P_{\lambda_{1j}}\left(|\beta_j|\right)-n\sum_{k=1}^{q}P_{\lambda_{2k}}\left(|\gamma_k|\right), \tag{2.2}$$

where $P_\lambda(\cdot)$ is a pre-specified penalty function (such as LASSO and SCAD) with a regularization parameter λ, which can be chosen by a data-driven criterion such as cross-validation (CV), generalized cross-validation (GCV, Fan and Li, 2001; Tibshirani, 1996) and Bayes information criterion (BIC). In this paper, we consider three penalty functions: least absolute shrinkage and selection operator (LASSO), smoothly clipped absolute deviation (SCAD), and CP (Wang et al., 2010; Amin et al., 2015; Dong et al., 2014, here we used the combination of SCAD with Ridge). We used BIC to choose the tuning parameters in this paper.

In this study, let $\theta=(\theta_1,\theta_2,\cdots,\theta_s)^T=(\beta_1,\beta_2,\cdots,\beta_p,\gamma_1,\gamma_2,\cdots,\gamma_q)^T$ with $s=p+q$; we often use the following penalized likelihood function:

$$Q(\theta)=L(\theta)-n\sum_{j=1}^{s}P_{\lambda_{1j}}\left(|\theta_j|\right), \tag{2.3}$$

where

$$L(\theta)=\sum_{i=1}^{n}\ln\left(k_i+e^{\frac{z_i^T\gamma}{2}}\right)-\sum_{i=1}^{n}\frac{z_i^T\gamma}{2}+\sum_{i=1}^{n}\left(\frac{k_i}{e^{\frac{z_i^T\gamma}{2}}}+1\right)\left[\ln\left(k_i^2-e^{\frac{z_i^T\gamma}{2}}k_i\right)-x_i^T\beta\right]$$

$$-\sum_{i=1}^{n}\left(\frac{k_i}{e^{\frac{z_i^T\gamma}{2}}}+2\right)\ln y_i,$$

and $k_i = \sqrt{e^{2x_i^T\beta} + e^{z_i^T\gamma}}$. $P_{\lambda_{1j}}\left(|\theta_j|\right)$ is a pre-specified penalty function (such as LASSO and SCAD).

The penalized maximum likelihood estimator of θ, denoted by $\hat{\theta}$ maximizes the function $Q(\theta)$ in (2.3) except for a constant term. With appropriate penalty functions, maximizing $Q(\theta)$ with respect to θ leads to certain parameter estimators vanishing from the initial models so that the corresponding explanatory variables are automatically removed. Hence, through maximizing $Q(\theta)$, we achieve the goal of selecting important variables and obtaining the parameter estimates, simultaneously.

2.3 Theoretical Properties

We consider the consistency and asymptotic normality of the penalized likelihood estimator in this subsection. Firstly, we introduce some notations. Let θ_0 denote the true value of θ. Furthermore, let $\theta_0 = (\theta_{01}, \theta_{02}, \ldots, \theta_{0s})^T = \left(\left(\theta_0^{(1)}\right)^T, \left(\theta_0^{(2)}\right)^T\right)^T$, without loss of generality, it is assumed that $\theta_0^{(1)}$ consists of all non-zero components of θ_0 and that $\theta_0^{(2)} = 0$. In addition, we suppose that the tuning parameters have been rearranged with respect to the elements of θ_0. Let s_1 be the dimension of $\theta_0^{(1)}$,

$$a_n = \max_{1\le j\le s}\left\{P_\lambda'\left(\left|\theta_{0j}\right|\right), \theta_{0j} \neq 0\right\}, b_n = \max_{1\le j\le s}\left\{P_\lambda''\left(\left|\theta_{0j}\right|\right), \theta_{0j} \neq 0\right\}$$

To obtain the property of consistency and asymptotic normality, we require the following regularity conditions on our model.

(A): The covariate vectors, $x_i = \left(x_{i1}, \cdots, x_{ip}\right)^T$, and $z_i = \left(z_{i1}, \cdots, z_{iq}\right)^T$, $\left(i = 1, 2, \cdots, n\right)$ are fixed and bounded.,

(B): The true value θ_0 is in the interior of the parameter space Θ.

(C): The y_i, $\left(i = 1, 2, \cdots, n\right)$ are independent in our model.

Theorem 2.1 (Consistency).

Assume $a_n = O_p(n^{-\frac{1}{2}})$, $n \to \infty$, $b_n \to 0$ and $\lambda_n \to 0$. λ_n is equal to either λ_{1n} or λ_{2n}, depending on whether θ_{0j} is a component of β_0 or γ_0, $\left(i = 1, 2, \cdots, n\right)$. Under conditions (A)-(C), with probability tending to 1, there exists a local maximizer $\hat{\theta}_n$ of the penalized likelihood function $Q(\theta)$ in equation (2.3), such that

$$\left\|\hat{\theta}_n - \theta_0\right\| = O_p(n^{-\frac{1}{2}}).$$

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

$$A_n = diag\left\{P''_{\lambda_n}\theta_{01}^{(1)},\cdots,P''_{\lambda_n}\theta_{s1}^{(1)}\right\}$$

$$d_n = diag\left\{P'_{\lambda_n}\theta_{01}^{(1)}\operatorname{sgn}\left(\theta_{01}^{(1)}\right),\cdots,P'_{\lambda_n}\theta_{s_1}^{(1)}\operatorname{sgn}\left(\theta_{s_1}^{(1)}\right)\right\}^T$$

where $\theta_{0j}^{(1)}$ is the j-th component of $\theta_0^{(1)}$, $\left(j=1,2,\cdots,s_1\right)$, denote the Fisher information matrix of θ by $I_n(\theta)$.

Theorem 2.2 (Oracle property).

Assume that the penalty function $P'_{\lambda_n}(\theta)$ satisfies

$$\liminf_{n\to\infty}\liminf_{\theta\to 0^+} P'_{\lambda_n}(\theta)/\lambda_n > 0,$$

and when $n\to\infty$, $\tilde{I}_n = I_n\left(\theta_0\right)/n$ converges to a finite and positive definite matrix $I_n\left(\theta_0\right)$. Meanwhile, under the conditions of Theorem 2.1, if $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$ as $n\to\infty$, then the $\sqrt{n}$-consistent estimator $\hat{\theta}_n = \left(\left(\hat{\theta}_n^{(1)}\right)^T, \left(\hat{\theta}_n^{(2)}\right)^T\right)^T$ in Theorem 2.1 must satisfy

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

(ii) (Asymptotic normality)

$$\sqrt{n}\left(\tilde{I}_n^{(1)}\right)^{-\frac{1}{2}}\left(\tilde{I}_n + A_n\right)\times\left\{\left(\hat{\theta}_n^{(1)} - \theta_0^{(1)}\right) + \left(\tilde{I}_n^{(1)} + A_n\right)^{-1} d_n\right\} \xrightarrow{L} N_{s_1}\left(0, I_{s_1}\right)$$

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

Remark:

The Theorem 2.2 stands for the Oracle property of the estimator under the model of (2.1). Proofs of Theorems 2.1 and Theorem 2.2 are essentially the same as Fan and Li (2001). To save space, the proofs are omitted.

3. COMPUTATION

We employ an algorithm to obtain the likelihood estimation in joint mean and variance models of the Pareto distribution in this subsection. We also give the method of how to choose the tuning parameters.

3.1 Computation of the Likelihood Estimation in Joint Mean and Variance Models of the Pareto Distribution

Firstly, we find that the first two derivatives of the log-likelihood function $L(\theta)$ are continuous. For a given point θ_0, the log-likelihood function can be approximated by

$$L(\theta) \approx L(\theta_0) + \left(\frac{\partial L(\theta_0)}{\partial \theta}\right)^T (\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^T \left(\frac{\partial^2 L(\theta_0)}{\partial \theta \partial \theta^T}\right)(\theta - \theta_0),$$

For the given θ_0, $P_\lambda(|\theta|)$ can also be locally approximated by a quadratic function as $P_\lambda(|\theta|) \approx P_\lambda(|\theta_0|) + \frac{1}{2}\{P_\lambda'(|\theta_0|)/\theta_0\}(\theta^2 - \theta_0^2)$, for $\theta \approx \theta_0$.

Therefore, the penalized likelihood function (2.3) can be locally approximated by

$$Q(\theta) \approx L(\theta_0) + \left(\frac{\partial L(\theta_0)}{\partial \theta}\right)^T (\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^T \left(\frac{\partial^2 L(\theta_0)}{\partial \theta \partial \theta^T}\right)(\theta - \theta_0) - \frac{n}{2}\theta^T \sum_\lambda(\theta_0)\theta,$$

where

$$\sum_\lambda(\theta_0) = diag\left\{\frac{P'_{\lambda_{11}}(|\beta_{01}|)}{\beta_{01}}, \cdots, \frac{P'_{\lambda_{1p}}(|\beta_{0p}|)}{\beta_{0p}}, \frac{P'_{\lambda_{21}}(|\gamma_{01}|)}{\gamma_{01}}, \cdots, \frac{P'_{\lambda_{2q}}(|\gamma_{0q}|)}{\gamma_{0q}}\right\},$$

$$\theta = (\theta_1, \theta_2, \ldots, \theta_s)^T = (\beta_1, \beta_2, \ldots, \beta_p, \gamma_1, \gamma_2, \ldots, \gamma_q)^T$$

and $\theta_0 = (\theta_{01}, \theta_{02}, \ldots, \theta_{0s})^T = (\beta_{01}, \beta_{02}, \ldots, \beta_{0p}, \gamma_{01}, \gamma_{02}, \ldots, \gamma_{0q})^T$.

Accordingly, the quadratic maximization problem for $Q(\theta)$ leads to a solution interacted by

$$\theta_1 \approx \theta_0 + \left\{\frac{\partial^2 L(\theta_0)}{\partial \theta \partial \theta^T} - n\sum_\lambda(\theta_0)\right\}^{-1} \left\{n\sum_\lambda(\theta_0)\theta_0 - \frac{\partial L(\theta_0)}{\partial \theta}\right\}$$

Secondly, under the model of Pareto distribution, the log-likelihood function $L(\theta)$ can be written as

$$L(\theta) = L(\beta, \gamma) = \sum_{i=1}^n \ln\left(k_i + e^{\frac{z_i^T\gamma}{2}}\right) - \sum_{i=1}^n \frac{z_i^T\gamma}{2}$$
$$+ \sum_{i=1}^n \left(\frac{k_i}{e^{\frac{z_i^T\gamma}{2}}} + 1\right)\left[\ln\left(k_i^2 - e^{\frac{z_i^T\gamma}{2}}k_i\right) - x_i^T\beta\right] - \sum_{i=1}^n \left(\frac{k_i}{e^{\frac{z_i^T\gamma}{2}}} + 2\right)\ln y_i,$$

where $k_i = \sqrt{e^{2x_i^T\beta} + e^{z_i^T\gamma}}$. Therefore, the resulting functions are

$$U(\theta)=\frac{\partial L(\theta)}{\partial \theta}=\left(U_1^T(\beta),U_2^T(\gamma)\right)^T,\ \text{where}$$

$$U_1(\beta)=\frac{\partial L}{\partial \beta}=\sum_{i=1}^{n}\frac{e^{2x_i^T\beta}}{k_i\left(k_i+e^{\frac{1}{2}z_i^T\gamma}\right)}x_i+\sum_{i=1}^{n}\left(\frac{e^{2x_i^T\beta-\frac{1}{2}z_i^T\gamma}}{k_i}\right)\left[\ln\left(\frac{k_i^2-k_ie^{\frac{1}{2}z_i^T\gamma}}{y_i}\right)-x_i^T\beta\right]x_i$$

$$+\sum_{i=1}^{n}\left(\frac{k_i}{e^{\frac{1}{2}z_i^T\gamma}}+1\right)\left[\frac{e^{2x_i^T\beta}\left(2k_i-e^{\frac{1}{2}z_i^T\gamma}\right)}{k_i^2\left(k_i-e^{\frac{1}{2}z_i^T\gamma}\right)}-1\right]x_i,$$

$$U_2(\gamma)=\frac{\partial L}{\partial \gamma}=\frac{1}{2}\sum_{i=1}^{n}\frac{e^{\frac{1}{2}z_i^T\gamma}}{k_i}z_i-\frac{1}{2}\sum_{i=1}^{n}z_i-\sum_{i=1}^{n}\frac{e^{2x_i^T\beta}}{k_i}\left[\frac{\ln\left(k_i^2-k_ie^{\frac{1}{2}z_i^T\gamma}\right)-\ln y_i-x_i^T\beta}{e^{\frac{1}{2}z_i^T\gamma}}+\frac{1}{k_i}\right]z_i,$$

and we denote

$$H(\theta)=\frac{\partial^2 L(\theta)}{\partial\theta\partial\theta^T}=\begin{pmatrix}\frac{\partial^2 L}{\partial\beta\partial\beta^T} & \frac{\partial^2 L}{\partial\beta\partial\gamma^T}\\ \frac{\partial^2 L}{\partial\gamma\partial\beta^T} & \frac{\partial^2 L}{\partial\gamma\partial\gamma^T}\end{pmatrix},\ \text{where}$$

$$\frac{\partial^2 L}{\partial\beta\partial\beta^T}=\sum_{i=1}^{n}\frac{2k_i^2\left(k_i+e^{\frac{1}{2}z_i^T\gamma}\right)-e^{2x_i^T\beta}\left(2k_i+e^{\frac{1}{2}z_i^T\gamma}\right)}{k_i^3\left(k_i+e^{\frac{1}{2}z_i^T\gamma}\right)^2}e^{2x_i^T\beta}x_ix_i^T$$

$$+\sum_{i=1}^{n}\frac{2k_i^2-e^{2x_i^T\beta}}{k_i^3e^{\frac{1}{2}z_i^T\gamma}}\left[\ln\left(\frac{k_i^2-k_ie^{\frac{1}{2}z_i^T\gamma}}{y_i}\right)-x_i^T\beta\right]e^{2x_i^T\beta}x_ix_i^T$$

$$+2\sum_{i=1}^{n}\frac{1}{k_i\,e^{\frac{1}{2}z_i^T\gamma}}\left[\frac{e^{2x_i^T\beta}\left(2k_i-k_ie^{\frac{1}{2}z_i^T\gamma}\right)}{k_i^2\left(k_i-e^{\frac{1}{2}z_i^T\gamma}\right)}-1\right]e^{2x_i^T\beta}x_ix_i^T$$

$$+\sum_{i=1}^{n}\left[\frac{4k_i^4-6k_i^3e^{\frac{1}{2}z_i^T\gamma}-4k_i^2e^{2x_i^T\beta}+2k_i^2e^{z_i^T\gamma}+5k_ie^{2x_i^T\beta+\frac{1}{2}z_i^T\gamma}-2e^{2x_i^T\beta+\frac{1}{2}z_i^T\gamma}}{k_i^4\left(k_i-e^{\frac{1}{2}z_i^T\gamma}\right)^2}\right]$$

$$\cdot\left(\frac{k_i}{e^{\frac{1}{2}z_i^T\gamma}}+1\right)e^{2x_i^T\beta}x_ix_i^T,$$

$$\frac{\partial^2 L}{\partial\beta\partial\gamma^T} = -\sum_{i=1}^{n}\frac{\left(e^{2x_i^T\beta+\frac{1}{2}z_i^T\gamma}\right)}{2k_i^3}x_i z_i^T - \sum_{i=1}^{n}\frac{\left(e^{2x_i^T\beta-\frac{1}{2}z_i^T\gamma}\right)\left(k_i - e^{\frac{1}{2}z_i^T\gamma}\right)}{2k_i^3}x_i z_i^T e^{\frac{1}{2}z_i^T\gamma}$$
$$-\sum_{i=1}^{n}\frac{\left(k_i^2 + e^{z_i^T\gamma}\right)}{2k_i^3}\left[\ln\left(\frac{k_i^2 - k_i e^{\frac{1}{2}z_i^T\gamma}}{y_i}\right) - x_i^T\beta\right]x_i z_i^T e^{2x_i^T\beta-\frac{1}{2}z_i^T\gamma}$$
$$+\sum_{i=1}^{n}\left(\frac{k_i}{e^{\frac{1}{2}z_i^T\gamma}}+1\right)\left(\frac{k_i - 2e^{\frac{1}{2}z_i^T\gamma}}{2k_i^4}\right)x_i z_i^T e^{2x_i^T\beta+\frac{1}{2}z_i^T\gamma}$$
$$+\sum_{i=1}^{n}\frac{e^{\frac{1}{2}z_i^T\gamma} - k_i^2 e^{-\frac{1}{2}z_i^T\gamma}}{2k_i}\left[\frac{e^{x_i^T\beta}\left(2k_i - e^{\frac{1}{2}z_i^T\gamma}\right)}{k_i^2\left(k_i - e^{\frac{1}{2}z_i^T\gamma}\right)} - 1\right]x_i z_i^T,$$

$$\frac{\partial^2 L}{\partial\gamma\,\partial\beta^T} = -\frac{1}{2}\sum_{i=1}^{n}\frac{e^{2x_i^T\beta+\frac{1}{2}z_i^T\gamma}}{k_i^3}z_i x_i^T - \sum_{i=1}^{n}\frac{e^{4x_i^T\beta}}{k_i}\left[\frac{\frac{2k_i - e^{\frac{1}{2}z_i^T\gamma}}{k_i^2\left(k_i - e^{\frac{1}{2}z_i^T\gamma}\right)} - 1}{e^{\frac{1}{2}z_i^T\gamma}} - \frac{1}{k_i^3}\right]z_i x_i^T$$
$$-\sum_{i=1}^{n}\frac{\left(-2k_i + e^{2x_i^T\beta}\right)}{k_i^3}\left[\frac{\ln\left(k_i^2 - k_i e^{\frac{1}{2}z_i^T\gamma}\right) - \ln y_i - x_i^T\beta}{e^{\frac{1}{2}z_i^T\gamma}} + \frac{1}{k_i}\right]z_i x_i^T e^{2x_i^T\beta},$$

$$\frac{\partial^2 L}{\partial\gamma\,\partial\gamma^T} = \frac{1}{4}\sum_{i=1}^{n}\frac{e^{2x_i^T\beta+\frac{1}{2}z_i^T\gamma}}{k_i^3}z_i z_i^T$$
$$+\sum_{i=1}^{n}\frac{1}{2k_i^3}\left[\frac{\ln\left(k_i^2 - k_i e^{\frac{1}{2}z_i^T\gamma}\right) - \ln y_i - x_i^T\beta}{e^{\frac{1}{2}z_i^T\gamma}} + \frac{1}{k_i}\right]z_i z_i^T e^{2x_i^T\beta+\frac{1}{2}z_i^T\gamma}$$
$$-\frac{1}{2}\sum_{i=1}^{n}\frac{e^{2x_i^T\beta}}{k_i}\left\{\frac{\ln\left(k_i^2 - k_i e^{\frac{1}{2}z_i^T\gamma}\right) - \ln y_i - x_i^T\beta}{e^{\frac{1}{2}z_i^T\gamma}} + \frac{k_i - e^{\frac{1}{2}z_i^T\gamma}}{k_i^2} - \frac{e^{z_i^T\gamma}}{k_i^3}\right\}z_i z_i^T,$$

where $k_i = \sqrt{e^{2x_i^T\beta+z_i^T\gamma}}$.

Finally, the following algorithm summarizes the computation of penalized maximum likelihood estimators of the parameters in model (2.1).

Algorithm:

Step 1. Take the ordinary maximum likelihood estimators (without penalty) $\hat{\beta}_{MLE}$ and $\hat{\gamma}_{MLE}$. $\beta^{(0)}=\hat{\beta}_{MLE}$, $\gamma^{(0)}=\hat{\gamma}_{MLE}$ of β,γ as their initial values, that is, $\theta^{(0)}=\left(\left(\beta^{(0)}\right)^T,\left(\gamma^{(0)}\right)^T\right)^T$.

Step 2. Given the current values $\beta^{(l)}$, $\gamma^{(l)}$, $\theta^{(l)}=\left(\left(\beta^{(l)}\right)^T,\left(\gamma^{(l)}\right)^T\right)^T$, update

$$\theta^{(l+1)}\approx\theta^{(l)}+\left\{H\left(\theta^{(l)}\right)-n\sum\nolimits_{\lambda}\left(\theta^{(l)}\right)\right\}^{-1}\left\{n\sum\nolimits_{\lambda}\left(\theta^{(l)}\right)\theta^{(l)}-U\left(\theta^{(l)}\right)\right\}$$

Step 3. Repeat step 2 until certain convergence criteria are satisfied.

3.2 Selection of the Tuning Parameters λ_n

Implementing the methods described above, we need to estimate the threshold parameters λ_n. Wang et al. (2007) found that the BIC-type criterion is consistent in model selection, and verified that the penalized estimator with the tuning parameter selected, can identify the true model consistently. Following this idea, the BIC is used to choose the tuning parameters. The formula of BIC is $BIC=-2L\left(\hat{\theta}_n\right)+d\ln(n)$, where d is the number of nonzero coefficients of $\hat{\theta}_n$, and

$$L\left(\hat{\theta}_n\right)=L\left(\hat{\beta}_n,\hat{\gamma}_n\right)=\sum_{i=1}^{n}\ln\left(k_i+e^{\frac{z_i^T\hat{\gamma}_n}{2}}\right)-\sum_{i=1}^{n}\frac{z_i^T\hat{\gamma}_n}{2}$$
$$+\sum_{i=1}^{n}\left(\frac{k_i}{e^{\frac{1}{2}z_i^T\hat{\gamma}_n}}+1\right)\left[\ln\left(k_i^2-e^{\frac{z_i^T\hat{\gamma}_n}{2}}k_i\right)-x_i^T\hat{\beta}_n\right]-\sum_{i=1}^{n}\left(\frac{k_i}{e^{\frac{1}{2}z_i^T\hat{\gamma}_n}}+1\right)\ln y_i,$$

where $k_i=\sqrt{e^{2x_i^T\hat{\beta}_n}+e^{z_i^T\hat{\gamma}_n}}$, $\hat{\beta}_n$ and $\hat{\gamma}_n$ are the penalized maximum likelihood estimators. Fan and Li (2001) numerically showed that $a=3.7$ minimizes the Bayesian risk and recommended its use in practice. Thus, we set $a=3.7$. It is expected that the choice of λ_{1j} and λ_{2k} should satisfy the tuning parameter for zero coefficient is larger than for non-zero coefficient. Thus, we can simultaneously unbiasedly estimate a larger than that for non-zero coefficient and shrink the smaller coefficient towards zero. Hence, in practice, we suggest taking $\lambda_{1j}=\lambda\Big/\left|\hat{\beta}_j^0\right|$, $\lambda_{2k}=\lambda\Big/\left|\hat{\gamma}_k^0\right|$, where $\hat{\beta}_j^0$ and $\hat{\gamma}_k^0$ are the initial estimators of β_j and γ_k, $\left(j=1,2,\cdots,p;k=1,2,\cdots,q\right)$, respectively, by using un-penalized maximum likelihood estimators of β and γ. The tuning parameter can be obtained as

$$\hat{\lambda}=\arg\min_{\lambda}BIC(\lambda).$$

4. SIMULATION STUDY

In this section, we conduct some Monte Carlo simulations with joint mean and variance models of the Pareto distribution to evaluate the finite sample performance of the proposed methodologies. We simulate data from model (2.1)

$$\begin{cases} y_i \sim Pareto(\alpha_i, d_i), \\ \mu_i = E(y_i) = e^{x_i^T\beta} = \dfrac{\alpha_i d_i}{\alpha_i - 1}, \\ \delta_i^2 = Var(y_i) = e^{z_i^T\gamma} = \dfrac{\alpha_i d_i^2}{(\alpha_i - 1)^2(\alpha_i - 2)}, \\ i = 1, 2, \cdots, n. \end{cases}$$

The coefficients α_i and k_i can be computed in the formula of $\beta_0 = (1.5, 0, 1, 0, 2, 0, 0, 0, 0, 0)^T$.

To perform this simulation, we take $\gamma_0 = (0, 1, 1, 0, 0, 1.5, 0, 0, 0, 0, 0, 0, 0, 0, 0)^T$ and $\beta_0 = (1.5, 0, 1, 0, 2, 0, 0, 0, 0, 0)^T$. All of the simulation results are based on 1000 independent repetitions. The average number of the estimated zero coefficients for parameters in model (2.1), with 1000 simulation runs, is reported in Table 1. In Table 1, the column labeled "C$\left(\hat{\beta}_n\right)$ and $C\left(\hat{\gamma}_n\right)$" gives the average number of zero coefficients correctly set to zero, and the column "I$\left(\hat{\beta}_n\right)$ *and* $I\left(\hat{\gamma}_n\right)$" gives the average number of nonzero coefficients incorrectly set to zero. Furthermore, the column labeled "GMSE" gives the generalized mean square error of $\hat{\beta}_n$ and $\hat{\gamma}_n$. Similar to what Li and Liang (2008), Zhao and Xue (2010) had done, the performance of estimators $\hat{\beta}_n$ and $\hat{\gamma}_n$ will be assessed by using the generalized mean square error (GMSE), defined as

$$GMSE\left(\hat{\beta}_n\right) = \left(\hat{\beta}_n - \hat{\beta}_0\right)^{\mathrm{T}} \mathrm{E}\left(XX^{\mathrm{T}}\right)\left(\hat{\beta}_n - \hat{\beta}_0\right),$$

$$GMSE\left(\hat{\gamma}_n\right) = \left(\hat{\gamma}_n - \hat{\gamma}_0\right)^{\mathrm{T}} \mathrm{E}\left(ZZ^{\mathrm{T}}\right)\left(\hat{\gamma}_n - \hat{\gamma}_0\right).$$

The sample size in the simulations is $n = 200$. As same as Huang, Ma and Zhang (2008), we consider two cases, which are exhibited in the following two examples.

Example 4.1 *(General)*

The covariates x_i are the multivariate normal distributions with mean 0 and covariance between the i-th and j-th elements being $r^{|i-j|}$ with $r = 0.1$, and $r = 0.9$. And the covariates z_i have the same distribution with x_i. y_i is generated according to model (2.1). The results of Mean and Variance Model are shown in Table 1 and Table 2, respectively.

Table 1
The Simulation Result for Mean Model in Example 4.1

n	Method	$r=0.1$			$r=0.9$		
		$GMSE\left(\hat{\beta}_n\right)$	$C\left(\hat{\beta}_n\right)$	$I\left(\hat{\beta}_n\right)$	$GMSE\left(\hat{\beta}_n\right)$	$C\left(\hat{\beta}_n\right)$	$I\left(\hat{\beta}_n\right)$
200	Lasso	0.654	4.870	0.000	0.488	4.630	0.000
	SCAD	0.377	5.918	0.000	0.394	6.620	0.000
	CP	0.405	5.535	0.000	0.335	6.931	0.000

Table 2
The Simulation Result for Variance Model in Example 4.1

n	Method	$r=0.1$			$r=0.9$		
		$GMSE\left(\hat{\gamma}_n\right)$	$C\left(\hat{\gamma}_n\right)$	$I\left(\hat{\gamma}_n\right)$	$GMSE\left(\hat{\gamma}_n\right)$	$C\left(\hat{\gamma}_n\right)$	$I\left(\hat{\gamma}_n\right)$
200	Lasso	0.549	8.870	0.000	0.744	8.125	0.002
	SCAD	0.398	9.905	0.000	0.398	9.913	0.010
	CP	0.402	9.422	0.000	0.375	10.001	0.017

From the Table 1 and Table 2, it can be shown that the performance of SCAD is much better than the Lasso and CP in the two models, when $r=0.1$. But, when the correlation is get higher, the performance of the CP is a litter better than the Lasso and SCAD when $r=0.9$. So, we can find that the CP and SCAD are always better than the Lasso in the two models, because the Lasso does not have the oracle property.

Example 4.2 *(Group Structure)*

The covariates $x_i(i=1,\ldots,n)$ are generated as follows: $x_{ik}\sim N(0,1)$, $(k=1,\ldots,6)$, $x_{ik}=x_{ik-4}+\eta_k$ when $k=7,\ldots,10$, where η_k are i.i.d. $N(0,0.01)$. The covariates $z_j(j=1,\ldots,n)$ are generated as follows: $z_{jt}\sim N(0,1)(t=1,\ldots,9)$, $z_{jt}=z_{jt-6}+\eta_t$ when $t=10,\ldots,15$, where η_k are i.i.d. $N(0,0.01)$. y_i is generated according to model (2.1). The results of Mean and Variance Model are shown in Table 3 and Table 4, respectively.

Table 3
The Simulation Result for Mean Model in Example 4.2

n	Method	$GMSE\left(\hat{\beta}_n\right)$	$C\left(\hat{\beta}_n\right)$	$I\left(\hat{\beta}_n\right)$
200	Lasso	1.134	5.270	0.000
	SCAD	1.006	6.111	0.000
	CP	0.988	6.535	0.000

Table 4
The Simulation Result for Variance Model in Example 4.2

n	**Method**	$GMSE(\hat{\gamma}_n)$	$C(\hat{\gamma}_n)$	$I(\hat{\gamma}_n)$
200	Lasso	1.454	8.870	0.002
	SCAD	1.377	10.005	0.000
	CP	1.371	10.435	0.001

It is obvious that the correlation is very high in this example. So, from the Table 3 and Table 4, the performance of the CP is a little better than the Lasso and SCAD, while the SCAD is better than Lasso in the two models. The reason is that the Lasso does not have the oracle property.

5. CONCLUSION AND DISCUSSION

In this paper, a new procedure is proposed which can select and estimate the significant variables simultaneously in joint mean and variance models of the Pareto distribution. Meanwhile, the consistency and the oracle property of the regularized estimators is also established with the appropriate method of selecting the tuning parameters. The finite sample performance of the proposed model through simulation studies is assessed. The conclusion evolves the futuristic views of the proposed model in research fields. It will be more beneficial to utilize it for infinite number of parameters as this model is valid only for fixed number of parameters. Now, as the knowledge is broadening day by day, so it is essential to link new doors of theories to old ones. Hence, it is indeed a need to develop some new methods to obtain the variable selection in the joint mean and variance models with different distribution.

REFERENCES

1. Aitkin, M. (1987). Modelling variance heterogeneity in normal regression using GLIM. *Appl. Stat.*, 36, 332-339.
2. Amin, M., Song, L., Thorlie, M.A. and Wang, X. (2015). Combined penalized quantile regression in high dimensional models. *Pak. J. Statist.*, 31, 49-70.
3. Balkema, A.A. and De Haan, L. (1974). Residual life time at great age. *Ann. Probab*., 2, 792-804.
4. Bryson, M.C. (1974). Heavy tailed distributions: Properties and tests. *Technometrics*, 16, 61-68.
5. Dong, Y., Song L., Wang, M. and Xu, Y. (2014). Combined-penalized likelihood estimations with a diverging number of parameters. *Journal of Applied Statistics*, 41, 1274-1285.
6. Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96, 1348-1360.
7. Harvey, A.C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica*, 44, 460-465.
8. Huang, J., Ma, S.C. and Zhang, H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica*, 18, 1603-1618.

9. Lee, Y. and Nelder J.A. (1998). Generalized linear models for the analysis of quality improvement experiments. *Can. J. Stat.*, 26, 95-105.
10. Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *Ann. Stat*. 36, 261-286.
11. Lomax, K.S. (1954). Business failures. Another example of the analysis of failure data. *J. Amer. Statist. Assoc.*, 49, 847-852.
12. Park, R.E. (1966). Estimation with heteroscedastic error terms. *Econometrica*, 34, 888.
13. Smyth G.K. and Verbyla, A.P. (1999). Adjusted likelihood methods for modelling dispersion in generalized linear models. *Environmetrics*, 10, 696-709.
14. Taylor, J.T. and Verbyla, A.P. (2004). Joint modelling of location and scale parameters of the t distribution. *Stat. Model*, 4, 91-112.
15. Tibshirani, R.J. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. B*., 58, 267-288.
16. Wang, D.R. and Zhang, Z.Z. (2009) Variable selection in joint generalized linear models. *Chin. J. Appl. Probab. Stat*., 25, 245-256.
17. Wang, H., Li, R. and Tsai, C.L. (2007). On the consistency of SCAD tuning parameter selector. *Biometrika*, 94, 553-568.
18. Wang, X.M., Park, T. and Carriere K.C. (2010). Variable selection via combined penalization for high-dimensional data analysis. *Comput. Stat. Data Anal*., 54, 2230-2243.
19. 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.
20. Wu, L.C. and Zhang, Z.Z and Xu, D.K. (2012). Variable selection in joint mean and variance models of Box-Cox transformation. *Journal of Applied Statistics*, 39, 2543-2555.
21. Wu, L.C. and 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.
22. Wu, L.C. (2014). Variable selection in joint location and scale models of the skew-t-normal distribution. *Communications in Statistics - Simulation and Computation*. 43(3), 615-630.
23. Wu, L.C., Zhang, Z.Z., Tian, G.L. and Xu, D.K. (2014). A robust variable selection to t-type joint generalized linear models via penalized t-type pseudo-likelihood, *Commun. in Statist.–Simul. and Compu.*, DOI: 10.1080/03610918.2014.901358.
24. Zhao, P.X. and Xue, L.G. (2010). Variable selection for semiparametric varying coefficient partially linear errors-in-variables models. *J. Multivariate Anal*., 101, 1872-1883.