Hacettepe Journal of Mathematics and Statistics
Volume 41 (2) (2012), 307 – 320

VARIABLE SELECTION FOR JOINT MEAN AND DISPERSION MODELS OF THE LOGNORMAL DISTRIBUTION

Liu-Cang Wu*†‡, Zhong-Zhan Zhang* and Deng-Ke Xu*

Received 24 : 11 : 2010 : Accepted 25 : 09 : 2011

Abstract

The lognormal distribution is widely used in applications. Variable selection is an important issue in all regression analysis and in this paper we investigate simultaneous variable selection in the joint mean and dispersion models of the lognormal distribution. We propose a unified penalized likelihood method which can simultaneously select significant variables in the mean and dispersion models. Furthermore, the proposed variable selection method can simultaneously perform parameter estimation and variable selection in the mean and dispersion models. With appropriate selection of the tuning parameters, we establish the consistency and the oracle property of the regularized estimators. Some simulation studies and a real example are used to illustrate the proposed method.

Keywords: Joint mean and dispersion models of the lognormal distribution, LASSO, Penalized maximum likelihood, SCAD, Variable selection.

2000 AMS Classification: 62 F 12, 62 H 12.

*College of Applied Sciences, Beijing University of Technology, Beijing 100124, People's Republic of China. E-mail: (L.-C Wu) wuliucang@163.com (Z.-Z. Zhang) zzhang@bjut.edu.cn (D.-K. Xu) xudengke1983@emails.bjut.edu.cn.

†Faculty of Science, Kunming University of Science and Technology, Kunming 650093, People's Republic of China.

‡Corresponding Author.

1. Introduction

The lognormal distribution is widely used in geology, hydrology, biology, hylology, industry, economics and so on, occurring in the literature, e.g., see Shimizu *et.al.* [15], Crow and Shimizu [5], Limpert, Stahel and Abbt [12] and references therein, where the basic process under consideration leads to a skewed distribution.

It is well known that efficient estimation of mean parameters in regression depends on a correct modeling of the dispersion. The loss of efficiency may be substantial if the working dispersion model deviates far from the underlying true dispersion model. Modeling of the dispersion is also necessary to obtain correct standard errors and confidence intervals, as well as for many other applications such as prediction, estimation of detection limits or immunoassay (Carroll[3]; Carroll and Rupert[4]). In many studies, modeling the dispersion will be of direct interest in its own right, to identify the sources of variability in the observations [17].

Joint mean and dispersion models have been received a lot of attention in recent years. For example, for joint mean and dispersion models of the normal distribution, Park [14] proposed a log linear model for the variance parameter and described the Gaussian model using a two stage process to estimate the parameters. Harvey [8] discussed maximum likelihood (ML) estimation of the mean and variance effects and the subsequent likelihood ratio test under general conditions. Aitkin [1] provided ML estimation for a joint mean and variance model and applied it to the commonly cited Minitab tree data. Verbyla [20] estimated the parameters using restricted maximum likelihood (REML) and provided leverage and influence diagnostics for ML and REML. It is common for the observables to contain outliers. If the outliers are considered to be genuine then for their accommodation, rather than deletion, Taylor and Verbyla [18] proposed joint modeling of location and scale parameters of the t distribution. More general distributions from the family of generalized linear models are considered by Smyth [16], Nelder and Lee [13], Lee and Nelder[9], Smyth and Verbyla [17] and Wang and Zhang [21]. In these papers the mean and dispersion parameters of the distribution are estimated using double generalized linear models.

It is common for the observables to have a skewed distribution. If present, the normality assumption of the error distribution for the model is questionable and estimates of the parameters may be misleading. In this paper, the problem of interest is joint modeling of mean and dispersion models of the lognormal distribution. We consider the following joint mean and dispersion models of the lognormal distribution:

$$\begin{cases} y_i \sim \mathrm{LN}(\mu_i, \sigma_i^2) \\ \log(\eta_i) = x_i^T \beta \\ \log(\phi_i) = z_i^T \gamma \\ i = 1, 2, \ldots, n, \end{cases} \tag{1.1}$$

where $\eta_i = \mathrm{E}(y_i) = e^{\mu_i + \frac{1}{2}\sigma_i^2}$, $\mathrm{Var}(y_i) = (e^{\sigma_i^2} - 1)e^{2\mu_i + \sigma_i^2} = \phi_i \eta_i^2$, $\phi_i = e^{\sigma_i^2} - 1$, $y = (y_1, \ldots, y_n)^T$ is a vector of n independent responses, n is the sample size. $x_i = (x_{i1}, \ldots, x_{ip})^T$ and $z_i = (z_{i1}, \ldots, z_{iq})^T$ are observed covariates corresponding to y_i, $\beta = (\beta_1, \ldots, \beta_p)^T$ is a $p \times 1$ vector of unknown parameters in the mean model, and $\gamma = (\gamma_1, \ldots, \gamma_q)^T$ is a $q \times 1$ vector of unknown parameters in the dispersion model. z_i may contain some or all of the variables in x_i and other variables not included in x_i, that is, the mean model and the dispersion model may incorporate different covariates, or some of the same covariates, and may depend on common covariates in different ways. Denote by $x = (x_1, \ldots, x_n)^T$ and $z = (z_1, \ldots, z_n)^T$ the covariate matrices.

We aim to remove the unnecessary explanatory variables from model (1.1).

Variable selection is an important issue in all regression analysis. To the best of our knowledge, most existing variable selection procedures are limited to selecting only the mean explanation variables, e.g., Fan and Lv [7], Li, Peng and Zhu [10] and references therein. However, little work has been done on selecting the dispersion explanation variables. Wang and Zhang [21] proposed criterion EAIC for selection of only the mean explanation variable based on extended quasi-likelihood, which is used in joint generalized linear models with structured dispersions.

In this paper we aim to develop an efficient unified penalized likelihood method to select important explanatory variables that make a significant contribution to the model (1.1). We propose a unified penalized likelihood method which can simultaneously select significant variables in the mean and dispersion model. Furthermore, the proposed variable selection method can simultaneously perform parameter estimation and variable selection in the mean and dispersion model. With appropriate selection of the tuning parameters, we establish the consistency and the oracle property of the regularized estimators. Some simulations are carried out to assess the finite sample performance of the proposed method. A real example is used to illustrate the proposed method.

The rest of this paper is organized as follows. In Section 2, we first propose a variable selection method for model (1.1) via a penalized likelihood function. Then, we present some theoretical properties of this procedure, including the consistency and the oracle property of the regularized estimators. The standard error formula of the parameter estimators and the choice of the tuning parameters are provided. In Section 3, based on local quadratic approximations, we propose an iterative algorithm for finding the penalized maximum likelihood estimators. In Section 4, some simulations and a real example are used to illustrate the proposed method. Some concluding remarks are given in Section 5. The technical proofs of all asymptotic results are provided in the Appendix.

2. Variable selection via penalized maximum likelihood

2.1. Penalized maximum likelihood. Many traditional variable selection criteria can be considered as a penalized likelihood which balances modeling biases and estimation variances (Fan and Li, [6]). Suppose that we have a random sample $(y_i, x_i, z_i), i = 1, 2, \ldots, n$, from model (1.1). Let $\ell(\beta, \gamma)$ denote the log-likelihood function. Then, similar to Fan and Li [6], we define the penalized likelihood function

$$\mathcal{L}(\beta, \gamma) = \ell(\beta, \gamma) - n\sum_{j=1}^{p} p_{\lambda_{1j}}(|\beta_j|) - n\sum_{k=1}^{q} p_{\lambda_{2k}}(|\gamma_k|), \tag{2.1}$$

where $p_{\lambda_{1j}}(\cdot)$ and $p_{\lambda_{2k}}(\cdot)$ are pre-specified general penalty functions with regularization parameters λ_{1j} and λ_{2k} which can be chosen by a data-driven criterion such as cross-validation (CV) or generalized cross-validation (GCV, Fan and Li[6]; Tibshirani[19]), respectively. In this paper, Section 4.1, we consider three penalty functions: SCAD (Fan and Li[6]), LASSO (Tibshirani[19]) and Hard (Antoniadis[2]). In Section 2.4, we use BIC to choose the tuning parameters. Note that the penalty functions and regularization parameters are not necessarily the same for all j, k. For example, we wish to keep some important variables in the final model and therefore do not want to penalize their coefficients.

Let $\theta = (\theta_1, \ldots, \theta_s)^T = (\beta_1, \ldots, \beta_p; \gamma_1, \ldots, \gamma_q)^T$ with $s = p + q$. We use the following penalized likelihood function

$$\mathcal{L}(\theta) = \ell(\theta) - n\sum_{j=1}^{p} p_{\lambda_{1j}}(|\beta_j|) - n\sum_{k=1}^{q} p_{\lambda_{2k}}(|\gamma_k|), \tag{2.2}$$

where $\ell(\theta) = -\sum_{i=1}^{n} \ln y_i - \frac{1}{2}\sum_{i=1}^{n} \ln \ln(e^{z_i^T\gamma}+1) - \frac{1}{2}\sum_{i=1}^{n} \frac{(\ln y_i - x_i^T\beta + \frac{1}{2}\ln(e^{z_i^T\gamma}+1))^2}{\ln(e^{z_i^T\gamma}+1)}$.

The penalized maximum likelihood estimator of θ, denoted by $\hat{\theta}$, maximizes the function $\mathcal{L}(\theta)$ in (2.2). With appropriate penalty functions, maximizing $\mathcal{L}(\theta)$ 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 $\mathcal{L}(\theta)$ we achieve the goal of selecting important variables and obtaining the parameter estimators, simultaneously. In Section 3, we provide the technical details and an algorithm to calculate the penalized maximum likelihood estimator $\hat{\theta}$.

2.2. Asymptotic properties. In this subsection, we consider the consistency and asymptotic normality of the resulting penalized likelihood estimator. We first introduce some notation. Let θ_0 denote the true value of θ. Furthermore, let $\theta_0 = (\theta_{01}, \ldots, \theta_{0s})^T = (\theta_0^{(1)T}, \theta_0^{(2)T})^T$. Without loss of generality, it is assumed that $\theta_0^{(1)}$ consists of all nonzero components of θ_0 and that $\theta_0^{(2)} = 0$. In addition, we suppose the tuning parameters have been rearranged with respect to the elements of θ_0. Let s_1 be the dimension of $\theta_0^{(1)}$. Let

$$a_n = \max_{1\le j\le s}\{p'_{\lambda_n}(|\theta_{0j}|) : \theta_{0j} \neq 0\},$$
$$b_n = \max_{1\le j\le s}\{|p''_{\lambda_n}(|\theta_{0j}|)| : \theta_{0j} \neq 0\},$$

where we write λ as λ_n to emphasize that λ_n depends the sample size n.

2.1. Theorem. *Assume $a_n = O_p(n^{-\frac{1}{2}}), b_n \to 0$ and $\lambda_n \to 0$ as $n \to \infty$. λ_n is equal to either λ_{1n} or λ_{2n}, depending on whether θ_{0j} is a component of β_0, γ_0, $(1 \le j \le s)$. Under the conditions (C1)-(C2) in the Appendix, there exists a local maximizer $\hat{\theta}_n$ of the penalized likelihood function $\mathcal{L}(\theta)$ in (2.2) such that $\|\hat{\theta} - \theta_0\| = O_p(n^{-1/2})$ with probability tending to 1.*

The proof is given in the Appendix. We now consider the asymptotic normality of $\hat{\theta}_n$. Let

$$A_n = \mathrm{diag}(p''_{\lambda_n}(|\theta_{01}^{(1)}|), \ldots, p''_{\lambda_n}(|\theta_{0s_1}^{(1)}|)),$$
$$c_n = (p'_{\lambda_n}(|\theta_{01}^{(1)}|)\mathrm{sgn}(\theta_{01}^{(1)}), \ldots, p'_{\lambda_n}(|\theta_{0s_1}^{(1)}|)\mathrm{sgn}(\theta_{s_1}^{(1)}))^T,$$

where λ_n has the same definition as that in Theorem 2.1, and $\theta_{0j}^{(1)}$ is the jth component of $\theta_0^{(1)}$ $(1 \le j \le s_1)$. We denote the Fisher information matrix of θ by $\mathcal{I}_n(\theta)$.

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

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

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 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 $\hat{\theta}_n = ((\hat{\theta}_n^{(1)})^T, (\hat{\theta}_n^{(2)})^T)^T$ in Theorem 2.1 must satisfy

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

(ii) **(Asymptotic normality)**

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

where "$\xrightarrow{\mathcal{L}}$" stands for convergence in distribution, and $\hat{\theta}_n = ((\hat{\theta}_n^{(1)})^T, (\hat{\theta}_n^{(2)})^T)^T$ is the penalized maximum likelihood estimator of θ. To emphasize its dependence on the sample size n, we also denote it by $\hat{\theta}_n$. $\bar{\mathcal{I}}_n^{(1)}$ is the $(s_1 \times s_1)$ submatrix of

$\bar{\mathcal{I}}_n$ corresponding to nonzero components of $\theta_0^{(1)}$ and I_{s_1} is the $(s_1 \times s_1)$ identity matrix.

2.3. Standard error formula of $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}^{(\mathbf{1})}$. As a consequence of Theorem 2.2, the asymptotic covariance matrix of $\hat{\theta}_n^{(1)}$ is

$$\mathrm{Cov}(\hat{\theta}_n^{(1)}) = \frac{1}{n}(\bar{\mathcal{I}}_n^{(1)} + A_n)^{-1}\bar{\mathcal{I}}_n^{(1)}(\bar{\mathcal{I}}_n^{(1)} + A_n)^{-1}. \tag{2.3}$$

So the asymptotic standard error for $\hat{\theta}_n^{(1)}$ is straightforward. However, $\bar{\mathcal{I}}_n^{(1)}$ and A_n are evaluated at the true value $\theta_0^{(1)}$, which is unknown. A natural choice is to evaluate $\bar{\mathcal{I}}_n^{(1)}$ and A_n at the estimator value $\hat{\theta}_n^{(1)}$ so that the estimator of the asymptotic covariance matrix of $\hat{\theta}_n^{(1)}$ is obtained by (2.3).

Corresponding to the partition of θ_0, we assume $\theta = ((\theta^{(1)})^T, (\theta^{(2)})^T)^T$. Set $\ell'(\theta_0^{(1)}) = [\frac{\partial \ell(\theta)}{\partial \theta^{(1)}}]_{\theta=\theta_0}$ and $\ell''(\theta_0^{(1)}) = [\frac{\partial^2 \ell(\theta)}{\partial \theta^{(1)} \partial \theta^{(1)T}}]_{\theta=\theta_0}$, where $\theta_0 = ((\theta^{(1)})^T, 0^T)^T$. Also, let

$$\Sigma_{\lambda_n}(\theta_0^{(1)}) = \mathrm{diag}\left\{ \frac{p'_{\lambda_{n1}}(|\theta_{01}^{(1)}|)}{|\theta_{01}^{(1)}|}, \ldots, \frac{p'_{\lambda_{ns_1}}(|\theta_{0s_1}^{(1)}|)}{|\theta_{0s_1}^{(1)}|} \right\}.$$

By using the observed information matrix to approximate the Fisher information matrix, the covariance matrix of $\hat{\theta}_n^{(1)}$ can be estimated by

$$\hat{\mathrm{Cov}}(\hat{\theta}_n^{(1)}) = \{\ell''(\hat{\theta}_n^{(1)}) - n\Sigma_{\lambda_n}(\hat{\theta}_n^{(1)})\}^{-1}\hat{\mathrm{Cov}}\{\ell'(\hat{\theta}_n^{(1)})\}\{\ell''(\hat{\theta}_n^{(1)}) - n\Sigma_{\lambda_n}(\hat{\theta}_n^{(1)})\}^{-1},$$

where $\hat{\mathrm{Cov}}\{\ell'(\hat{\theta}_n^{(1)})\}$ is the covariance of $\ell'(\theta^{(1)})$ evaluated at $\hat{\theta}_n^{(1)}$.

2.4. Selection of the tuning parameters. Many selection criteria, such as cross validation (CV), generalized cross validation (GCV), AIC and BIC selection can be used for tuning parameters. Wang *et al.* [22] suggested using the BIC for the SCAD estimator in linear models and partially linear models, and proved its model selection consistency property, i.e., the optimal parameter chosen by BIC can identify the true model with probability tending to one. We will also use the BIC to select the optimal λ:

$$BIC(\lambda) = -\frac{2}{n}\ell(\hat{\theta}) + df_\lambda \times \frac{\log(n)}{n},$$

where $0 \le df_\lambda \le s$ is simply the number of nonzero coefficients of $\hat{\theta}$. Except for a constant,

$$\begin{aligned}\ell(\hat{\theta}) &= \ell(\hat{\beta}, \hat{\gamma}) \\ &= -\sum_{i=1}^n \ln y_i - \frac{1}{2}\sum_{i=1}^n \ln\ln(e^{z_i^T\hat{\gamma}} + 1) - \frac{1}{2}\sum_{i=1}^n \frac{(\ln y_i - x_i^T\hat{\beta} + \frac{1}{2}\ln(e^{z_i^T\hat{\gamma}} + 1))^2}{\ln(e^{z_i^T\hat{\gamma}} + 1)},\end{aligned}$$

where $\hat{\beta}$ and $\hat{\gamma}$ are the penalized maximum likelihood estimators.

It is expected that the choice of λ_{1j} and λ_{2k} should be such that the tuning parameter for a zero coefficient is larger than that for a nonzero coefficient. Thus we can simultaneously unbiasedly estimate the larger coefficient, and shrink the small coefficient towards zero. Hence, in practice, we suggest

(i) $\lambda_{1j} = \frac{\lambda}{|\hat{\beta}_j^0|}, j = 1, \ldots, p,$

(ii) $\lambda_{2k} = \frac{\lambda}{|\hat{\gamma}_k^0|}, k = 1, \ldots, q,$

where $\hat{\beta}_j^0$ and $\hat{\gamma}_k^0$ are initial estimators of β_j and γ_k respectively obtained by using unpenalized maximum likelihood estimators of β and γ. $0 \le df_\lambda \le s$ is simply the number of nonzero coefficients of $\hat{\theta}$.

The tuning parameter can be obtained as

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

From our simulation study, we found that this method works well.

3. Algorithm

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

$$\ell(\theta) \approx \ell(\theta_0) + \left[\frac{\partial \ell(\theta_0)}{\partial \theta}\right]^T (\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^T \left[\frac{\partial^2 \ell(\theta_0)}{\partial \theta \partial \theta^T}\right] (\theta - \theta_0).$$

Also, given an initial value θ_0 we can approximate the penalty function $p_\lambda(\theta)$ by a quadratic function (Fan and Li, [6])

$$p_\lambda(|\theta|) \approx p_\lambda(|\theta_0|) + \frac{1}{2}\frac{p'_\lambda(|\theta_0|)}{|\theta_0|}(\theta^2 - \theta_0^2), \text{ for } \theta \approx \theta_0.$$

Therefore, the penalized likelihood function (2.2) can be local approximated by

$$\mathcal{L}(\theta) \approx \ell(\theta_0) + \left[\frac{\partial \ell(\theta_0)}{\partial \theta}\right]^T (\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^T \left[\frac{\partial^2 \ell(\theta_0)}{\partial \theta \partial \theta^T}\right] (\theta - \theta_0) - \frac{n}{2}\theta^T \Sigma_\lambda(\theta_0)\theta,$$

where

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

where

$$\theta = (\theta_1, \ldots, \theta_s)^T = (\beta_1, \ldots, \beta_p; \gamma_1, \ldots, \gamma_q)^T \text{ and}$$
$$\theta_0 = (\theta_{01}, \ldots, \theta_{0s})^T = (\beta_{01}, \ldots, \beta_{0p}; \gamma_{01}, \ldots, \gamma_{0q})^T.$$

So the quadratic maximization problem for $\mathcal{L}(\theta)$ leads to a solution iterated by

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

Secondly, when the data satisfy the lognormal distribution, the log-likelihood function $\ell(\theta)$ can be written as

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

Therefore, the resulting functions are

$$U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = (U_1^T(\beta), U_2^T(\gamma))^T,$$

where

$$U_1(\beta) = \frac{\partial \ell}{\partial \beta} = \sum_{i=1}^n \frac{(\ln y_i - x_i^T\beta + \frac{1}{2}\ln(e^{z_i^T\gamma} + 1))x_i}{\ln(e^{z_i^T\gamma} + 1)},$$

$$U_2(\gamma) = \frac{\partial \ell}{\partial \gamma} = -\frac{1}{2}\sum_{i=1}^{n} \frac{e^{z_i^T\gamma} z_i}{(e^{z_i^T\gamma}+1)\ln(e^{z_i^T\gamma}+1)} - \frac{1}{2}\sum_{i=1}^{n} \frac{(\ln y_i - x_i^T\beta + \frac{1}{2}\ln(e^{z_i^T\gamma}+1))e^{z_i^T\gamma} z_i}{(e^{z_i^T\gamma}+1)\ln(e^{z_i^T\gamma}+1)} + \frac{1}{2}\sum_{i=1}^{n} \frac{(\ln y_i - x_i^T\beta + \frac{1}{2}\ln(e^{z_i^T\gamma}+1))^2 e^{z_i^T\gamma} z_i}{(e^{z_i^T\gamma}+1)\ln^2(e^{z_i^T\gamma}+1)},$$

and write

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

where

$$\frac{\partial^2 \ell}{\partial\beta\partial\beta^T} = -\sum_{i=1}^{n} \frac{x_i x_i^T}{\ln(e^{z_i^T\gamma}+1)},$$

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

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

$$\begin{aligned}\frac{\partial^2 \ell}{\partial\gamma\partial\gamma^T} = &-\frac{1}{2}\sum_{i=1}^{n} \frac{e^{z_i^T\gamma} z_i z_i^T}{(e^{z_i^T\gamma}+1)\ln(e^{z_i^T\gamma}+1)} + \frac{1}{2}\sum_{i=1}^{n} \left(\frac{e^{z_i^T\gamma}}{(e^{z_i^T\gamma}+1)\ln(e^{z_i^T\gamma}+1)}\right)^2 z_i z_i^T \\ &+ \frac{1}{2}\sum_{i=1}^{n} \frac{[(\ln y_i - x_i^T\beta + \frac{1}{2}\ln(e^{z_i^T\gamma}+1)) + \frac{1}{2}](e^{z_i^T\gamma})^2 z_i z_i^T}{(e^{z_i^T\gamma}+1)^2\ln(e^{z_i^T\gamma}+1)} \\ &- \frac{1}{2}\sum_{i=1}^{n} \frac{(\ln y_i - x_i^T\beta + \frac{1}{2}\ln(e^{z_i^T\gamma}+1))e^{z_i^T\gamma} z_i z_i^T}{(e^{z_i^T\gamma}+1)\ln(e^{z_i^T\gamma}+1)} \\ &+ \frac{1}{2}\sum_{i=1}^{n} \frac{(\ln y_i - x_i^T\beta + \frac{1}{2}\ln(e^{z_i^T\gamma}+1))^2 e^{z_i^T\gamma} z_i z_i^T}{(e^{z_i^T\gamma}+1)\ln^2(e^{z_i^T\gamma}+1)} \\ &+ \sum_{i=1}^{n} \frac{(\ln y_i - x_i^T\beta + \frac{1}{2}\ln(e^{z_i^T\gamma}+1))(e^{z_i^T\gamma})^2 z_i z_i^T}{(e^{z_i^T\gamma}+1)^2\ln^2(e^{z_i^T\gamma}+1)} \\ &- \frac{1}{2}\sum_{i=1}^{n} \frac{(\ln y_i - x_i^T\beta + \frac{1}{2}\ln(e^{z_i^T\gamma}+1))^2(e^{z_i^T\gamma})^2 z_i z_i^T}{(e^{z_i^T\gamma}+1)^2\ln^2(e^{z_i^T\gamma}+1)} \\ &- \sum_{i=1}^{n} \frac{(\ln y_i - x_i^T\beta + \frac{1}{2}\ln(e^{z_i^T\gamma}+1))^2(e^{z_i^T\gamma})^2 z_i z_i^T}{(e^{z_i^T\gamma}+1)^2\ln^3(e^{z_i^T\gamma}+1)}.\end{aligned}$$

The Fisher information matrix is

$$\mathcal{I}_n(\theta) = E\left(-\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^T}\right) = \begin{pmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{pmatrix},$$

where

$$\mathcal{I}_{11} = \sum_{i=1}^{n} \frac{x_i x_i^T}{\ln(e^{z_i^T\gamma}+1)},$$

$$\mathcal{I}_{12} = -\frac{1}{2}\sum_{i=1}^{n} \frac{e^{z_i^T\gamma} x_i z_i^T}{(e^{z_i^T\gamma}+1)\ln(e^{z_i^T\gamma}+1)},$$

$$\mathcal{I}_{21} = -\frac{1}{2}\sum_{i=1}^{n} \frac{e^{z_i^T\gamma} z_i x_i^T}{(e^{z_i^T\gamma}+1)\ln(e^{z_i^T\gamma}+1)},$$

$$\mathcal{I}_{22} = \frac{1}{2}\sum_{i=1}^{n} \Big(\frac{e^{z_i^T\gamma}}{(e^{z_i^T\gamma}+1)\ln(e^{z_i^T\gamma}+1)}\Big)^2 z_i z_i^T + \frac{1}{4}\sum_{i=1}^{n} \frac{(e^{z_i^T\gamma})^2 z_i z_i^T}{(e^{z_i^T\gamma}+1)^2\ln(e^{z_i^T\gamma}+1)}.$$

By using the Fisher information matrix to approximate the observed information matrix, we obtain the following iteration solution:

$$\begin{aligned}\theta_1 &\approx \theta_0 + \left\{\frac{\partial^2\ell(\theta_0)}{\partial\theta\partial\theta^T} - n\Sigma_\lambda(\theta_0)\right\}^{-1}\left\{n\Sigma_\lambda(\theta_0)\theta_0 - \frac{\partial\ell(\theta_0)}{\partial\theta}\right\} \\ &\approx \theta_0 + \{\mathcal{I}_n(\theta_0) + n\Sigma_\lambda(\theta_0)\}^{-1}\{U(\theta_0) - n\Sigma_\lambda(\theta_0)\theta_0\} \\ &= \{\mathcal{I}_n(\theta_0) + n\Sigma_\lambda(\theta_0)\}^{-1}\{U(\theta_0) + \mathcal{I}_n(\theta_0)\theta_0\}.\end{aligned}$$

Finally, the following algorithm summarizes the computation of the penalized maximum likelihood estimators of the parameters in model (1.1).

3.1. Algorithm. ***Step 1****. Take the ordinary maximum likelihood estimators (without penalty)* $\beta^{(0)}, \gamma^{(0)}$ *of* β, γ *as their initial values, that is,* $\theta^{(0)} = ((\beta^{(0)})^T, (\gamma^{(0)})^T)^T$.

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

$$\theta^{(m+1)} = \{\mathcal{I}_n(\theta^{(m)}) + n\Sigma_\lambda(\theta^{(m)})\}^{-1}\{U(\theta^{(m)}) + \mathcal{I}_n(\theta^{(m)})\theta^{(m)}\}.$$

Step 3*. Repeat Step 2 until a certain convergence criteria is satisfied.*

4. Monte Carlo simulations and a real example

In this section, some simulation studies and a real example from Land Rent data are used to illustrate the proposed method.

4.1. Simulation study. In this subsection, we conduct some Monte Carlo simulations to evaluate the finite sample performance of the proposed method. As in Li, Peng and Zhu [10], Li and Liang [11] and Zhao and Xue [24] the performance of the estimators $\hat\beta$ and $\hat\gamma$ will be assessed using the generalized mean square error (GMSE), defined as

$$\mathrm{GMSE}(\hat\beta) = (\hat\beta - \beta_0)^T \mathrm{E}(XX^T)(\hat\beta - \beta_0),$$

$$\mathrm{GMSE}(\hat\gamma) = (\hat\gamma - \gamma_0)^T \mathrm{E}(ZZ^T)(\hat\gamma - \gamma_0).$$

We simulate data from model (1.1), that is,

$$\begin{cases} y_i \sim LN(\mu_i, \sigma_i^2) \\ \log(\eta_i) = x_i^T\beta \\ \log(\phi_i) = z_i^T\gamma \\ i = 1, 2, \ldots, n, \end{cases}$$

where $\beta_0 = (1, 1, 0, 0, 1, 0, 0, 0)^T$ and $\gamma_0 = (0.8, 0.8, 0, 0, 0.8, 0, 0, 0)^T$. To perform this simulation, we take the covariates $x_i \sim U(-1, 1), z_i \sim U(-1, 1), i = 1, \ldots, 8$, y_i is generated according to model (1.1). In addition, we make 1000 simulation runs in each simulation. The average number of estimated zero coefficients for parameter of the mean model and dispersion model, with 1000 simulation runs, is reported in Table 1. In Table 1, the column labeled "C" gives the average number of coefficients, of the true zero, correctly set to zero, and the column label "IC" gives the average number of the true nonzeros incorrectly set to zero. Furthermore, the column labeled GMSE gives the generalized mean square errors of the estimator $\hat{\beta}$ and $\hat{\gamma}$.

We compare the performance of variable selections in model (1.1) with different sample sizes and penalties.

Table 1. Comparisons with different penalties and sample sizes

Model	n	SCAD			LASSO			Hard		
		C	IC	GMSE	C	IC	GMSE	C	IC	GMSE
Mean	100	4.9910	0	0.0121	4.9420	0	0.0246	4.9990	0	0.0103
Model	200	5.0000	0	0.0052	4.9790	0	0.0116	5.0000	0	0.0045
	300	5.0000	0	0.0029	4.9890	0	0.0078	5.0000	0	0.0027
Dispersion	100	4.8660	0.3850	0.3459	4.6300	0.2340	0.1753	4.8300	0.1010	0.3876
Model	200	4.9730	0.0470	0.1826	4.7700	0.0090	0.0696	4.8600	0.0040	0.1893
	300	4.9920	0.0020	0.1569	4.8790	0	0.0390	4.9640	0	0.1572

From Table 1, we can make the following observations:

(1) We can see that the variable selection method based on SCAD, LASSO and Hard become better in terms of model error and model complexity as the sample size n increases.
(2) For the given penalty function, the performance of variable selection become better and better as the sample size n increases. The GMSE of estimators $\hat{\beta}$ and $\hat{\gamma}$ also become smaller and smaller as the sample size n increases.
(3) For a given sample size n, the performances of both the SCAD and Hard procedures are similar. Furthermore, the performances of both SCAD and Hard are significantly better than that of LASSO.
(4) For a given penalty function and sample size n, the performance of variable selection in the mean model are significantly better than that of the dispersion model in the sense of model error and model complexity.

4.2. A real example. In this subsection, Land Rent data [23] are used to illustrate the proposed method in Section 2. Land rent data reported a dataset about Y, the average rent per acre planted to alfalfa, four predictors: X_1, the average rent paid for all tillable land; X_2, the density of dairy cows (number per square mile); X_3 the proportion of farmland used pasture, and $X_4 = 1$ if liming is required to grow alfalfa; 0, otherwise. Alfalfa is a high protein crop that is suitable feed for dairy cows. It is thought that rent for land planted to alfalfa relative to rent for other agriculture purposes would be higher in areas with a high density of dairy cows and rents would be lower in counties where liming is required, since that would mean additional expense. The unit of analysis was a county in Minnesota, the 67 counties with appreciable rented farmland were included. The data were collected to study the variation in rent paid for agricultural land planted to alfalfa.

For our purposes, we study the following joint mean and dispersion models of the lognormal distribution.

$$\begin{cases} y_i \sim LN(\mu_i, \sigma_i^2) \\ \log(\mu_i) = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 + X_{i4}\beta_4 \\ \log(\phi_i) = \gamma_0 + X_{i1}\gamma_1 + X_{i2}\gamma_2 + X_{i3}\gamma_3 + X_{i4}\gamma_4 \\ i = 1, 2, \ldots, 67. \end{cases}$$

We apply the variable selection procedure based on the SCAD and Hard proposed in section 2 to the above model.

Table 2. Variable selection for Land rent data

Model	Method	Const.	X_1	X_2	X_3	X_4
Mean Model	SCAD	2.3409	0.0232	0.0057	0	0
	Hard	2.3949	0.0239	0.0062	0	0
Dispersion Model	SCAD	0	0	0	-2.5963	-0.5257
	Hard	0	0	0	-2.2261	-2.0054

From Table 2, 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 two nonzero regression coefficients β_1 and β_2 in the mean model, two nonzero regression coefficients γ_3 and γ_4 in the dispersion model. This indicates that the X_3 (the proportion of farmland used pasture) and X_4 ($X_4 = 1$, if liming is required to grow alfalfa; 0, otherwise) have no significant impact on the mean of Y (the average rent per acre planted to alfalfa), X_1 (the average rent paid for all tillable land) and X_2 (the density of dairy cows, number per square mile) have no significant impact on the variance of Y (the average rent per acre planted to alfalfa).

5. Conclusion and discussion

We have proposed a unified penalized likelihood method which can simultaneously select significant variables and estimate regression coefficients in the mean model and dispersion model. Furthermore, with appropriate selection of the tuning parameters, we establish the consistency and the oracle property of the regularized estimators. Simulation studies and a real data example clearly show that the proposed method can simultaneously select significant variables and estimate parameters in joint mean and dispersion models of the lognormal distribution.

Similar ideas can be further extended to other exponential family models within the join mean and dispersion framework. Due to the fact that exponential family distribution contains the lognormal distribution as a special case, we are currently studying variable selection in joint semiparametric modeling of mean and dispersion of the exponential family distribution.

To conclude this article, we would like to discuss some interesting topics for future study. The proposed method is valid for a fixed number of parameters. It would be interesting to consider the case when the number of parameters goes to infinity.

Acknowledgments

This research is supported by the National Natural Science Foundation of China (10971007, 11126309), the Funding Project of Science and Technology Research Plan of Beijing Education Committee(JC006790201001), the Natural Science Foundation of Yunnan Province (2009ZC039M) and the Doctoral Foundation of Kunming University of

Science and Technology (2009-024). We would like to thank the editor and referees for their valuable comments which have greatly improved the manuscript.

Appendix. Proofs of the theorems

For convenience and simplicity, let C denote a positive constant that may be different at each appearance throughout this paper. Before we prove our main theorems, we list some regularity conditions that are used in this paper.

To prove the theorems in the paper, we require the following regularity conditions:

(C1) The parameter space is compact and the true value θ_0 is in the interior of the parameter space.

(C2) The design matrices x_i and z_i in model (1.1) are all bounded, meaning that all the elements of the matrices are bounded by a single finite real number.

Proof of Theorem 2.1. For any given $\varepsilon > 0$, we will prove there exists a sufficiently large constant C such that

$$P\Big\{ \sup_{||v||=C} \mathcal{L}(\theta_0 + n^{-\frac{1}{2}}v) < \mathcal{L}(\theta_0)\Big\} \geq 1 - \varepsilon.$$

Note that $p_{\lambda_n}(0) = 0$ and $p_{\lambda_n}(\cdot) > 0$. Obviously, we have

$$\begin{aligned}
&\mathcal{L}\left(\theta_0 + n^{-\frac{1}{2}}v\right) - \mathcal{L}(\theta_0) \\
&\quad = [\ell(\theta_0 + n^{-\frac{1}{2}}v) - n\sum_{j=1}^{s} p_{\lambda_n}(|\theta_{0j} + n^{-\frac{1}{2}}v_j|)] - [\ell(\theta_0) - n\sum_{j=1}^{s} p_{\lambda_n}(|\theta_{0j}|)] \\
&\quad \leq [\ell(\theta_0 + n^{-\frac{1}{2}}v) - \ell(\theta_0)] - n\sum_{j=1}^{s_1}[p_{\lambda_n}(|\theta_{0j} + n^{-\frac{1}{2}}v_j|) - p_{\lambda_n}(|\theta_{0j}|)] \\
&\quad = k_1 + k_2,
\end{aligned}$$

where

$$\begin{aligned}
k_1 &= \ell(\theta_0 + n^{-\frac{1}{2}}v) - \ell(\theta_0); \\
k_2 &= -n\sum_{j=1}^{s_1}[p_{\lambda_n}(|\theta_{0j} + n^{-\frac{1}{2}}v_j|) - p_{\lambda_n}(|\theta_{0j}|)].
\end{aligned}$$

We first consider k_1. Using Taylor's expansion, we know

$$\begin{aligned}
k_1 &= [\ell(\theta_0 + n^{-\frac{1}{2}}v) - \ell(\theta_0)] \\
&= n^{-\frac{1}{2}}v^T\ell'(\theta_0) + \frac{1}{2}n^{-1}v^T\ell''(\theta^*)v \\
&= k_{11} + k_{12},
\end{aligned}$$

where θ^* lies between θ_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

$$k_{11} = n^{-\frac{1}{2}}v^T\ell'(\theta_0) \leq n^{-\frac{1}{2}}||\ell'(\theta_0)||||v|| = O_p(1).$$

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

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

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

$$\begin{aligned} k_{12} &= \frac{1}{2}n^{-1}v^T\ell''(\theta^*)v = \frac{1}{2}v^T[n^{-1}\ell''(\theta_0)]v[1+o_p(1)] \\ &= \frac{1}{2}v^T\{n^{-1}[\ell''(\theta_0) - E\ell''(\theta_0) - \mathcal{I}(\theta_0)]\}v[1+o_p(1)] \\ &= -\frac{1}{2}v^T\mathcal{I}(\theta_0)v[1+o_p(1)]. \end{aligned}$$

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

Next we study the term k_2. It follows from Taylor's expansion and the Cauchy-Schwartz inequality that

$$\begin{aligned} k_2 &= -n\sum_{j=1}^{s_1}[p_{\lambda_n}(|\theta_{0j} + n^{-\frac{1}{2}}v_j|) - p_{\lambda_n}(|\theta_{0j}|)] \\ &= -n\sum_{j=1}^{s_1}\{n^{\frac{1}{2}}p'_{\lambda_n}(|\theta_{0j}|)\mathrm{sgn}(\theta_{0j})v_j + \frac{1}{2}p''_{\lambda_n}(|\theta_{0j}|)v_j^2[1+O_p(1)]\} \\ &\le \sqrt{s_1}n^{\frac{1}{2}}||v|| \max_{1\le j\le s}\{p'_{\lambda_n}(|\theta_{j0}|), \theta_{j0}\neq 0\} + \frac{1}{2}||v||^2 \max_{1\le j\le s}\{|p''_{\lambda_n}(|\theta_{j0}|)| : \theta_{j0}\neq 0\} \\ &= \sqrt{s_1}n^{\frac{1}{2}}||v||a_n + \frac{1}{2}||v||^2 b_n. \end{aligned}$$

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 sufficiently large constant C such that

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

implying that there exists a local maximizer $\hat{\theta}_n$ such that $\hat{\theta}_n$ is a $\sqrt{n}$-consistent estimator of θ_0. The proof of Theorem 2.1 is completed. □

Proof of Theorem 2.2. We first prove part (i). From $\lambda_{max} \to 0$, it is easy to show that $a_n = 0$ for large n. Secondly, we prove that any given $\theta^{(1)}$ satisfying $\theta^{(1)} - \theta_0^{(1)} = O_p(n^{-1/2})$ and any constant $C > 0$, we have

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

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

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

where θ^* is between θ and θ_0. By a standard argument, we have

$$\frac{1}{n}\frac{\partial\ell(\theta_0)}{\partial\theta_j} = O_p\big(n^{-1/2}\big) \text{ and } \frac{1}{n}\bigg\{\frac{\partial^2\ell(\theta_0)}{\partial\theta_j\partial\theta_l} - \mathrm{E}\Big(\frac{\partial^2\ell(\theta_0)}{\partial\theta_j\partial\theta_l}\Big)\bigg\} = O_p(1).$$

Note that $\|\hat{\theta} - \theta_0\| = O_p(n^{-1/2})$. We then have

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

According to the assumption in Theorem 2.2, we obtain

$$\liminf_{n\to\infty}\liminf_{\theta\to o^+}\lambda_n^{-1}p'_{\lambda_n}(\theta)>0 \text{ and } \lambda_n^{-1}n^{-1/2}\to 0,$$

so that

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

and

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

Therefore, $\mathcal{L}(\theta)$ achieve its maximum at $\theta=((\theta^{(1)})^T,0^T)^T$ and the first part of Theorem 2.2 has been proved.

Secondly, we discuss the asymptotic normality of $\hat{\theta}_n^{(1)}$. From Theorem 2.1 and the first part of Theorem 2.2, there exists a penalized maximum likelihood estimator $\hat{\theta}_n^{(1)}$ that is the $\sqrt{n}$-consistent local maximizer of the function $\mathcal{L}\big\{((\theta^{(1)})^T,0^T)^T\big\}$. The estimator $\hat{\theta}_n^{(1)}$ must satisfy

$$\begin{aligned}0&=\frac{\partial\mathcal{L}(\theta)}{\partial\theta_j}|_{\theta=(\hat{\theta}^{(1)})^T)^T,0^T)^T}-np'_{\lambda_n}(|\hat{\theta}_{nj}^{(1)})|)\mathrm{sgn}(\hat{\theta}_{nj}^{(1)})\\&=\frac{\partial\ell(\theta_0)}{\partial\theta_j}+\sum_{l=1}^{s_1}\Big\{\frac{\partial^2\ell(\theta_0)}{\partial\theta_j\partial\theta_l}+O_p(1)\Big\}\big(\hat{\theta}_{nl}^{(1)}-\theta_{0l}^{(1)}\big)\\&\qquad-np'_{\lambda_n}\big(|\theta_{0j}^{(1)}|\big)\mathrm{sgn}\big(\hat{\theta}_{0j}^{(1)}\big)-n\big\{p''_{\lambda_n}(|\theta_{0j}^{(1)}|)+O_p(1)\big\}\big(\hat{\theta}_{nj}^{(1)}-\theta_{0j}^{(1)}\big).\end{aligned}$$

In other words, we have

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

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

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

Note that

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

so it follows immediately by using Slustsky's theorem that

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

The second part of Theorem 2.2 has been proved. □

References

[1] Aitkin , M. *Modelling variance heterogeneity in normal regression using GLIM*, Applied Statistics **36**, 332–339, 1987.

[2] Antoniadis, A. *Wavelets in statistics: a review (with discussion)*, Journal of the Italian Statistical Association **6**, 97–144, 1997.

[3] Carroll, R. J. *The effect of variance function estimating on prediction and calibration: an example*, In Statistical Decision Theory and Related Topics IV (eds J. O. Berger and S. S.Gupta) vol.II. (Springer, Heidelberg, 1987).

[4] Carroll, R. J. and Rupert, D. *Transforming and weighting in regression* (Chapman and Hall, London, 1988).

[5] Crow, E. and Shimizu, K. *Lognormal distributions: theory and practice* (Marcel Decker, New York, 1988).

[6] Fan, J. Q. and Li, R. *Variable selection via nonconcave penalized likelihood and its oracle properties*, Journal of American Statistical Association **96**, 1348–1360, 2001.
[7] Fan, J. Q. and Lv, J. C. *A selective overview of variable selection in high dimensional feature space*, Statistica Sinica **20**, 101–148, 2010.
[8] Harvey, A. C. *Estimating regression models with multiplicative heteroscedasticity*, Econometrica **44**, 460–465, 1976.
[9] Lee, Y. and Nelder, J. A. *Generalized linear models for the analysis of quality improvement experiments*, The Canadian Journal of Statistics **26** (1), 95–105, 1998.
[10] Li, G. R., Peng, H. and Zhu, L. X. *Nonconcave penalized M-estimation with a diverging number of parameters*, Statistica Sinica **21**, 391–419, 2011.
[11] Li, R. and Liang, H. *Variable selection in semiparametric regression modeling*, The Annals of Statistics **36**, 261–286, 2008.
[12] Limpert, E., Stahel, W. A. and Abbt, M. *Lognormal distributions across the sciences: Keys and clues*, BioScience **51**, 341–352, 2001.
[13] Nelder, J. A. and Lee, Y. *Generalized linear models for the analysis of Taguchi-type experiments*, Applied Stochastic Models and Data Analysis **7**, 107–120, 1991.
[14] Park, R. E. *Estimation with heteroscedastic error terms*, Econometrica **34**, 888, 1966.
[15] Shimizu, K. *et.al. Lognormal distribution and its applications* (John Wiley and Sons, New York, 1988).
[16] Smyth, G. K. *Generalized linear models with varying dispersion*, Journal of the Royal Statistical Society, Series B **51**, 47–60, 1989.
[17] Smyth, G. K. and Verbyla, A. P. *Adjusted likelihood methods for modelling dispersion in generalized linear models*, Environmetrics **10**, 696–709, 1999.
[18] Taylor, J. T. and Verbyla, A. P. *Joint modelling of location and scale parameters of the t distribution*, Statistical Modelling **4**, 91–112, 2004.
[19] Tibshirani, R. *Regression shrinkage and selection via the LASSO*, Journal of the Royal Statistical Society, Series B **58**, 267–288, 1996.
[20] Verbyla, A. P. *Variance heterogeneity: residual maximum likelihood and diagnostics*, Journal of the Royal Statistical Society, Series B **52**, 493–508, 1993.
[21] Wang, D. R. and Zhang, Z. Z. *Variable selection in joint generalized linear models*, Chinese Journal of Applied Probability and Statistics **25**, 245–256, 2009.
[22] Wang, H., Li, R. and Tsai, C. *Tuning parameter selectors for the smoothly clipped absolute deviation method*, Biometrika **94**, 553–568, 2007.
[23] Weisberg, S. *Applied Linear Regression* (Wiley, New York, 1985).
[24] Zhao, P. X. and Xue, L. G. *Variable selection for semiparametric varying coefficient partially linear errors-in-variables models*, Journal of Multivariate Analysis **101**, 1872–1883, 2010.