RMI
Advancing Risk Management
for Singapore and Beyond
National University of Singapore

Penalized High Dimensional Empirical Likelihood

Cheng Yong Tang and Chenlei Leng

RMI Working Paper No. 10/06
Submitted: June 21, 2010

Abstract:

We propose the penalized empirical likelihood (PEL) method for parameter estimation and variable selection for problems with diverging numbers of parameters. Our results are demonstrated for estimating the mean vector in multivariate analysis and regression coefficients in linear models. By using an appropriate penalty function, we show that PEL has the oracle property. That is, with probability tending to one, PEL identifies the true model and estimates the nonzero coefficients as efficiently as if the sparsity of the true model were known in advance. The advantage of PEL as a nonparametric likelihood approach is illustrated in testing hypothesis and constructing confidence sets. Numerical simulations confirm our theoretical findings.

Keywords: Confidence set; Empirical likelihood; High dimensional data analysis; Smoothly clipped absolute deviation; Variable selection

Cheng Yong Tang & Chenlei Leng
Department of Statistics and Applied Probability
Risk Management Institute
National University of Singapore
6 Science Drive 2
Singapore 117546
Phone: 65-65168949
Email: statc@nus.edu.sg stalc@nus.edu.sg

Penalized High Dimensional Empirical Likelihood [*]

By Cheng Yong Tang and Chenlei Leng

Department of Statistics and Applied Probability
National University of Singapore

June 21, 2010

Abstract

We propose the penalized empirical likelihood (PEL) method for parameter estimation and variable selection for problems with diverging numbers of parameters. Our results are demonstrated for estimating the mean vector in multivariate analysis and regression coefficients in linear models. By using an appropriate penalty function, we show that PEL has the oracle property. That is, with probability tending to one, PEL identifies the true model and estimates the nonzero coefficients as efficiently as if the sparsity of the true model were known in advance. The advantage of PEL as a nonparametric likelihood approach is illustrated in testing hypothesis and constructing confidence sets. Numerical simulations confirm our theoretical findings.

Some key words: Confidence set; Empirical likelihood; High dimensional data analysis; Smoothly clipped absolute deviation; Variable selection.

1 Introduction

The empirical likelihood (EL) method introduced by Owen (1988) is an influential nonparametric statistical instrument and has successful implementations in various areas (Owen, 2001). Through optimizing data driven nonparametric likelihood functions, the EL method is free from stringent parametric distributional assumptions. The resulting advantages include, among others, objective determination of the shape of the confidence region (Owen, 1988), seamless incorporation of auxiliary or prior information (Qin and Lawless, 1994) and Bartlett correctability (DiCiccio et al., 1991, Chen and Cui, 2006).

Recently, high dimensional data analysis has become an area of active research. When the data dimensionality diverges, variable selection through regularization is proven ef-

[*] We thank the Editor and two referees for their insightful comments and constructive suggestions. Research supports from National University of Singapore research grants are gratefully acknowledged.

fective. As argued in Hastie, Tibshirani and Friedman (2009) and Fan and Lv (2008), the penalized likelihood method can properly adjust bias-variance trade-off so that performance improvement can be achieved; see also series of works (Tibshirani, 1996; Fan and Li, 2001; Zou, 2006; Wang, Li and Tsai, 2007; Zhang and Lu, 2007; Candes and Tao, 2007; Fan and Lv, 2008; Lv and Fan, 2009) for penalized likelihood approaches and discussions.

The changing landscape of dimensionality from fixed (low) to growing (high) brings new interests and challenges to the EL method (Chen and Van Keilogom, 2009). Hjort, McKeague and Van Keilogom (2009) and Chen, Peng and Qin (2009) showed that the EL method continues to work when data dimensionality is growing. Nonetheless, the study of regularized EL method remains less studied, especially for high dimensional data analysis (McKeague, 2009). An exploration to this context could further broaden and extend the applications of EL.

Our study is motivated by the interest in exploring growing dimensional data analysis based on the EL method. We propose a unified framework for variable selection, parameter estimation and its inference via the penalized empirical likelihood (PEL) with the SCAD penalty (Fan and Li, 2001). Our main finding is that the PEL method has the oracle property and is advantageous in producing sparse models without specifying a parametric likelihood. We show that the profiled PEL ratio is asymptotically χ^2 distributed so that it facilitates testing of hypothesis and constructing range respecting confidence regions. Though (penalized) full parametric likelihood ratio test formulated in Fan and Peng (2004) shares this property, specifying a high dimensional distribution can be very challenging. In this respect, the PEL method provides a robust alternative to the parametric likelihood ratio approach in high dimensional statistical inference. Complementary to the penalized likelihood approach, the PEL method has the merits in both efficiency and adaptivity stemmed from a nonparametric likelihood approach (Owen, 2001).

The rest of this paper is structured as follows. Sections 2 and 3 discuss the PEL method for mean vectors and linear models respectively. Section 4 outlines computational aspect of the PEL method. Section 5 demonstrates the empirical performance of PEL by simulations. Some discussions are given in Section 6. All technical proofs are deferred to the Appendix.

2 Penalized Empirical Likelihood for Mean

We first study the PEL estimates of a mean vector μ from a collection of independent multivariate random vectors $\{X_i\}_{i=1}^n$, where $X_i \in \mathcal{R}^p$ with p diverging as $n \to \infty$. The diverging rate of p is discussed later. Throughout this paper we assume that $E(X_i) = \mu_0$ and $var(X_i) = \Sigma = (\sigma_{jk})_{j,k=1}^p$. The EL for μ is given by

$$L(\mu) = \sup\left\{\prod_{i=1}^n w_i : w_i \geq 0, \sum_{i=1}^n w_i = 1 \text{ and } \sum_{i=1}^n w_i X_i = \mu\right\}. \tag{2.1}$$

It is seen that $L(\mu)$ is maximized when $w_i = n^{-1}$. Thus the maximum empirical likelihood estimate of μ is identical to the sample mean $\bar{X}$.

Two elements are important in studying the properties of the high dimensional EL method. One is the rate of divergence of p as $n \to \infty$. The other is the probability distribution of X_i. In high dimensional data setup, the properties of EL are studied under various scenarios of the two elements. Define the log EL ratio as $\ell(\mu) = -\log\{L(\mu)\} - n\log(n)$. By assuming boundedness of X_i and $p^3/n \to \infty$, Hjort et al. (2009) showed that $(2p)^{-1/2}\{2\ell(\mu_0) - p\} \xrightarrow{d} N(0,1)$. If the boundedness of X_i is relaxed to moment conditions that both $E(||p^{-1/2}X_i||^q)$ and $p^{-1}\sum_{j=1}^{p} E|X_i^{(j)} - \mu_j|^q$ are bounded for some $q \geq 4$, then $p^{3+6/(q-2)}/n \to 0$ is needed to ensure this result (Hjort et al. 2009). Under milder restrictions on the diverging rate of p, Chen et al. (2009) established the same result. They utilized the following model permitting more detailed analysis:

$$\begin{aligned}
&X_i = \Gamma Z_i + \mu \text{ where } \Gamma \text{ is a } p \times m \text{ matrix, } m \geq p \text{ and } \Gamma\Gamma^T = \Sigma,\ Z_i \in \mathcal{R}^m \text{ satisfying} \\
&E(Z_i) = 0, var(Z_i) = I_m, E(Z_{il})^{4k} = m_{4k} \in (0, \infty),\ \text{for } l_1 \neq l_2 \neq \cdots \neq l_q \\
&E(Z_{il_1}^{\alpha_1} Z_{il_2}^{\alpha_2} \ldots Z_{il_q}^{\alpha_q}) = E(Z_{il_1}^{\alpha_1})E(Z_{il_2}^{\alpha_2}) \ldots E(Z_{il_q}^{\alpha_q}) \text{ and } \textstyle\sum_{l=1}^{q} \alpha_l \leq 4k. \qquad (2.2)
\end{aligned}$$

Here k is some positive integer and I_m is the m-dimensional identity matrix. Chen et al. (2009) argued that $p = o(n^{1/2})$ is the best rate under which $(2p)^{-1/2}\{2\ell(\mu_0) - p\} \xrightarrow{d} N(0,1)$.

We define the PEL estimator $\hat{\mu}$ to be the maximizer of

$$\log\{L(\mu)\} - n\sum_{i=1}^{p} p_\tau(|\mu_i|), \qquad (2.3)$$

where $p_\tau(t)$ is a penalty function with tuning parameter τ. See Fan and Li (2001), Zou (2006), Fan and Lv (2008) and Lv and Fan (2009) for examples of this function. In this paper, we use the smoothly clipped absolute deviation (SCAD) penalty whose first derivative satisfies

$$p'_\tau(t) = \tau\left\{I(t \leq \tau) + \frac{(a\tau - t)_+}{(a-1)\tau} I(t > \tau)\right\} \qquad (2.4)$$

for some $a > 2$ (Fan and Li, 2001). Following the convention in Fan and Li (2001), we set $a = 3.7$ in our work. The penalty (2.4) is a spline function on an interval near zero and constant outside, so that it can shrink small value of an estimate to zero while having no impact on a large one. As illustrated in Fan and Li (2001), this penalty function satisfies three requirements for variable selection, namely, asymptotic unbiasedness, sparsity and continuity of the estimated parameters. The penalized parametric likelihood using the SCAD penalty has the oracle property (Fan and Li, 2001; Fan and Peng, 2004)in identifying correctly the model structure and estimating the nonzero coefficients with optimal efficiency.

An application of the Lagrange multiplier method on (2.1) leads to

$$w_i = \frac{1}{n}\frac{1}{1 + \lambda_\mu^T(X_i - \mu)} \text{ where } \lambda_\mu \text{ solves } n^{-1}\sum_{i=1}^{n} \frac{X_i - \mu}{1 + \lambda_\mu^T(X_i - \mu)} = 0. \qquad (2.5)$$

By substituting (2.5) in $L(\mu)$, we see that maximizing (2.3) is equivalent to minimizing

$$\ell_p(\mu) = \sum_{i=1}^{n} \log\{1 + \lambda_\mu^T (X_i - \mu)\} + n \sum_{i=1}^{p} p_\tau(|\mu_i|). \tag{2.6}$$

For (2.1) and (2.3) to have solutions, μ needs to be in the convex hull formed by the data $\{X_i\}_{i=1}^n$. Therefore, the PEL estimator $\hat{\mu}$ essentially respects the range of the data.

Let $\mathcal{A} = \{j : \mu_{0j} \neq 0\}$ be the set of none-zero components of the true mean vector μ_0 and denote the cardinality of $\mathcal{A}$ as $|\mathcal{A}| = d$. Here we allow d to grow at the same rate of p as $n \to \infty$ without imposing any specific restriction. Without loss of generality, denote $\mu = (\mu_1^T, \mu_2^T)^T$ where $\mu_1 \in \mathcal{R}^d$ and $\mu_2 \in \mathcal{R}^{p-d}$ correspond to the zero and nonzero components respectively, i.e. $\mu_0 = (\mu_{10}^T, 0)^T$. Correspondingly, we decompose the variance-covariance matrix of X as $\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. The following regularity conditions are assumed.

A.1 The observations $\{X_i\}_{i=1}^n$ are iid following model (2.2) for some $k \geq 3$.

A.2 The eigenvalues of Σ satisfy that, $C_1 \leq \gamma_1(\Sigma) \leq \gamma_2(\Sigma) \leq \cdots \leq \gamma_p(\Sigma) \leq C_2$ for some $C_2 > C_1 > 0$.

A.3 As $n \to \infty$, $p \to \infty$, $p^2/n^{1-1/(4k)} \to 0$, $p^{1-2\delta}/n^{1/2-2\delta} \to 0$ for the δ specified in Lemma 1 in the Appendix, and $d \leq p$.

Model (2.2) in A.1 is used to characterize the tail probability behavior of X_i with correlated components (Bai and Saranadasa, 1996; Chen et al., 2009), which is crucial in high dimensional statistical analysis. The model in (2.2) includes the elliptical contoured distribution and the Gaussian family as special cases, where the latter is often assumed for high dimensional data, for example in Meinshausen and Bühlmann (2006) and Fan and Lv (2008). Alternative conditions to (2.2) such as in Hjort et al. (2009) can be used at the expense of a slower growing p. Both p and d are allowed to diverge as long as A.3 is satisfied.

For the penalty function $p_\tau(\cdot)$, we make the following assumptions.

B.1 As $n \to \infty$, the tuning parameter τ satisfies $\tau \to 0$ and $\tau(n/p)^{1/2-\delta} \to \infty$ for the δ specified in condition A.3. The nonzero components satisfy $\min_{j \in \mathcal{A}} |\mu_{0j}|/\tau \to \infty$.

B.2 $\max_{j \in \mathcal{A}} p'_\tau(|\mu_{0j}|) = o\{(\sqrt{np})^{-1}\}$, $\max_{j \in \mathcal{A}} p''_\tau(|\mu_{0j}|) = o(p^{-1/2+\delta} n^{-\delta})$.

Condition B.1 states that the weakest signal should dominate the penalty parameter τ. This assumption is routinely made to ensure the recovery of signals. Condition B.2 is used in controlling the impact of the penalty on the nonzero component. For the SCAD penalty (2.4), B.2 is satisfied because $\max_{j \in \mathcal{A}} p'(|\mu_j|) = 0$ for n large enough given B.1; see also Fan and Li (2001).

We show the property of PEL estimator $\hat{\mu} = (\hat{\mu}_1^T, \hat{\mu}_2^T)^T$ in the following theorem.

Theorem 1. *Under regularity conditions A.1, A.2, A.3, B.1 and B.2, as $n \to \infty$,*

1. *(Selection consistency) with probability tending to 1, $\hat{\mu}_2 = 0$;*
2. *(Asymptotic efficiency)*

$$\sqrt{n} W_n I_{\mathcal{A}}^{-1/2}(\hat{\mu}_1 - \mu_{10}) \xrightarrow{d} N(0, G)$$

where $W_n \in \mathcal{R}^{q\times d}$ such that $W_n W_n^T \to G$ for $G \in \mathcal{R}^{q\times q}$ with fixed q and $I_{\mathcal{A}} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

In part 2 of Theorem 1, W_n represents the projection of a diverging dimensional vector μ_1 to any fixed dimensionality q so that the limiting distribution can be characterized by a multivariate normal. Remarks of Theorem 1 are made in the following.

1. Theorem 1 states the oracle property of PEL in the sense of Fan and Peng (2004). Namely, PEL is consistent in model selection, and is as efficient as the EL estimates assuming the true sparse model ($\mu_2 = 0$) were known. In addition, we observe that the PEL estimator of the nonzero component $\hat{\mu}_1$ is more efficient than the corresponding subvector $\bar{X}_1$ in $\bar{X}$, whose asymptotic distribution is characterized by $\sqrt{n} W_n \Sigma_{11}^{-1/2}(\bar{X}_1 - \mu_{10}) \xrightarrow{d} N(0, G)$. Because the correlations are implicitly incorporated in its formulation, PEL can improve the efficiency in estimating μ_1 by accounting for $\mu_2 = 0$.

2. Although the PEL method imposes no distributional assumption, it can be as efficient as a parametric likelihood based approach. For example when $X_i \sim N(\mu, \Sigma)$, the oracle efficiency of the parametric likelihood approach knowing the true zero components agrees with the result in Theorem 1 Part 2.

Next we show that the PEL method provides a unified framework for the testing hypothesis and constructing confidence sets. In practice, confidence sets for a finite dimensional subset of the nonzero parameters is of great interest. To this end, we consider the following hypothesis testing

$$H_0 : L_n\mu_{10} = 0 \text{ vs } H_1 : L_n\mu_{10} \neq 0,$$

where $L_n \in \mathcal{R}^{q\times d}$ such that $L_n L_n^T = I_q$ for a fixed q and I_q is the q dimensional identity matrix. This type of hypotheses cover testing for both individual and multiple components, and can be extended to linear functions of μ. A similar type of hypothesis testing is considered in Fan and Peng (2004) under a parametric likelihood framework. Given the PEL formulation, a nonparametric profiled likelihood ratio statistic is constructed as

$$\tilde{\ell}(L_n) = -2\left\{\ell_p(\hat{\mu}) - \min_{\mu, L_n\mu_1 = 0} \ell_p(\mu)\right\}. \qquad (2.7)$$

We show the asymptotic property of the PEL ratio in the following theorem.

Theorem 2. *Under the null hypothesis and the conditions in Theorem 1, as* $n \to \infty$, $\tilde{\ell}(L_n) \xrightarrow{d} \chi^2_q$.

Therefore, a $(1-\alpha)$-level confidence set for $L_n\mu_1$ can be constructed as

$$V_\alpha = \left[v : -2 \left\{ \ell_p(\hat{\mu}) - \min_{\mu, L_n\mu_1 = v} \ell_p(\mu) \right\} \le \chi^2_{q,1-\alpha} \right] \tag{2.8}$$

where $\chi^2_{q,1-\alpha}$ is the $1-\alpha$ level quantile of χ^2_q distribution. As a direct result of Theorem 2, we have that $P(L_n\mu_{10} \in V_\alpha) \to 1-\alpha$ as $n \to \infty$.

Theorem 2 shows that the well known Wilks's phenomenon for the parametric likelihood and EL (Owen, 2001) continues to apply to PEL. Results parallel to Theorem 2 can also be established for the parametric likelihood ratio test under the usual parametric likelihood framework (Fan and Peng, 2004). Nevertheless, the PEL method is free of any distributional assumption and does not require estimating the variance covariance matrix of the parameters.

3 Penalized Empirical Likelihood for Linear Models

EL for linear models was first considered in Owen (1991). Chen (1994) studied the EL confidence region for regression coefficients in linear models. Existing works of EL method mainly focus on fixed dimensional regressions; see Owen (2001) and Chen and Van Keilegom (2009) for comprehensive overviews. We consider the linear model of the following form

$$Y_i = X_i^T\beta + \epsilon_i, \ \ i = 1, \ldots, n, \tag{3.1}$$

were $X_i \in \mathcal{R}^p$, and $\beta \in \mathcal{R}^p$. We assume that $\{X_i\}_{i=1}$ is a sequence of independent and identically distributed random vectors following some distribution, and without loss of generality $E(X_i) = 0$. The random vector can be centered to have zero mean if otherwise. We also assume error ϵ_i to be independent and identically distributed with mean zero and finite variance σ^2. Thus $E(Y_i|X_i) = X_i^T\beta_0$ is the conditional mean function and $var(Y_i|X_i) = \sigma^2$.

Following Owen (1991), the EL for linear models can be constructed based on the normal equations. Let $U_i(\beta) = X_i(Y_i - X_i^T\beta)$, we define the EL of β by

$$L(\beta) = \sup \left\{ \prod_{i=1}^n w_i : w_i \ge 0, \sum_{i=1}^n w_i = 1 \text{ and } \sum_{i=1}^n w_i U_i(\beta) = 0 \right\}. \tag{3.2}$$

By the Lagrange multiplier method,

$$w_i = \frac{1}{n}\frac{1}{1+\lambda_\beta^T U_i(\beta)} \text{ where } \lambda_\beta \text{ solves } n^{-1}\sum_{i=1}^n \frac{U_i(\beta)}{1+\lambda_\beta^T U_i(\beta)} = 0.$$

The PEL estimator $\hat{\beta}$ is then defined as the minimizer of

$$\ell_p(\beta) = \sum_{i=1}^{n} \log\{1 + \lambda_\beta^T U_i(\beta)\} + n \sum_{i=1}^{p} p_\tau(|\beta_i|). \tag{3.3}$$

Let $\mathcal{B} = \{j : \beta_{0j} \neq 0\}$ be the set of none-zero components of the true mean vector β_0 and the cardinality $|\mathcal{B}| = d$ where d satisfies Condition A.3 and is allowed to grow as $n \to \infty$. Following the conventions in the mean case, we write $\beta = (\beta_1^T, \beta_2^T)^T$ where $\beta_1 \in \mathcal{R}^d$ and $\beta_2 \in \mathcal{R}^{p-d}$ and hence the true parameter is $\beta_0 = (\beta_{10}^T, 0)^T$.

We study the PEL for linear models with random design and we assume that the random regressors X_i follows the multivariate model in (2.2) and the same regularity conditions as in the mean case. Condition A.1 on the design variable X_i is the counterpart of the assumptions made on the design matrix, for instance in Zou and Zhang (2009) and Lv and Fan (2009). Since we allow a general correlation matrix of X_i, Condition A.1 is not restrictive compared to those commonly assumed on the design matrix. In addition, we assume for the random errors:

A.4 $\{\epsilon_i\}_{i=1}^n$ is iid and $E(\epsilon_i^{4k}) < \infty$ for the same k in A.1.

In PEL for regression models, the μ_{0j} in Condition B.1 is replaced by β_{0j}. In particular,

B.1' As $n \to \infty$, the tuning parameter τ satisfies $\tau \to 0$ and $\tau(n/p)^{1/2-\delta} \to \infty$ for the δ specified in condition A.3. The nonzero components satisfy $\min_{j \in \mathcal{A}} |\beta_{0j}|/\tau \to \infty$.

The property of the PEL estimator $\hat{\beta}$ is given in the following theorem.

Theorem 3. *Under the regularity conditions A.1-A.4, B.1' and B.2, as $n \to \infty$,*

1. *(Selection consistency) with probability tending to 1, $\hat{\beta}_2 = 0$;*
2. *(Asymptotic efficiency)*

$$\sqrt{n} W_n I_{\mathcal{B}}^{-1/2}(\hat{\beta}_1 - \beta_{10}) \xrightarrow{d} N(0, G).$$

where $W_n \in \mathcal{R}^{q \times d}$ such that $W_n W_n^T \to G$ for $G \in \mathcal{R}^{q \times q}$ matrix with fixed q and $I_{\mathcal{B}} = \sigma^2 \Sigma_{11}^{-1}$.

Theorem 3 states the oracle property of PEL in linear regression. To see it, consider the EL to estimate β subject to the constraint $\beta_2 = 0$. The same analysis as that in proving Theorem 3 shows that the maximizer $\tilde{\beta}$ of (3.2) subject to $\beta_2 = 0$ follows the same asymptotic distribution given in Part 2 of the theorem.

Similar to the mean case, we consider the following test for the nonzero regression coefficient β_1:

$$H_0 : L_n \beta_{10} = 0 \text{ vs } H_1 : L_n \beta_{10} \neq 0,$$

where L_n is a $q \times d$ matrix such that $L_n L_n^T = I_q$ for a fixed and finite q. Then the PEL ratio test statistic is formulated in the same as that in the mean case:

$$\tilde{\ell}(L_n) = -2\left\{\ell_p(\hat{\beta}) - \min_{\beta : L_n\beta_1 = 0} \ell_p(\beta)\right\}. \tag{3.4}$$

We summarize the property of the test statistic in the following theorem.

Theorem 4. *Under the null hypothesis and assumptions in Theorem 3, as $n \to \infty$, $\tilde{\ell}(L_n) \xrightarrow{d} \chi_q^2$.*

Thus, a $(1-\alpha)$-level confidence set for $L_n\beta$ is constructed as

$$V_\alpha = \left[v : -2\left\{\ell_p(\hat{\beta}) - \min_{\beta, L_n\beta_1 = v} \ell_p(\beta)\right\} \le \chi^2_{q,1-\alpha}\right] \tag{3.5}$$

where $\chi^2_{q,1-\alpha}$ is the $1-\alpha$ level quantile of χ_q^2 distribution. Again, we have that $P(L_n\beta_{10} \in V_\alpha) \to 1-\alpha$ as $n \to \infty$.

For linear regression models, Theorem 4 again gives a convenient approach for hypothesis testing and constructing data oriented confidence sets free of any shape constraint. In high dimensional data analysis, such feature is desirable. This together with the oracle property of the PEL method by Theorem 3 demonstrate that in high dimensional data analysis, the PEL method is indeed appealing due to its robustness and efficiency.

4 Tuning and Computation

To choose the penalty parameter τ, we use the following BIC motivated by Wang, Li and Leng (2009)

$$BIC(\tau) = 2\ell_p(\theta_\tau) + C_n \cdot \log(n) \cdot df_\tau \tag{4.6}$$

where θ_τ is the PEL estimate of θ with tuning parameter τ; df_τ is the number of nonzero coefficient in θ_τ; C_n is a scaling factor diverging to infinity at a slow rate (Wang, et al., 2009) for $p \to \infty$. When p is fixed, we can simply take $C_n = 1$, otherwise, $C_n = \max\{\log\log p, 1\}$ seems to be a good choice. A rigorous proof of the consistency of this BIC for PEL is much more challenging than that for the linear regression considered in Wang, et al. (2009), and merits further investigation for future work.

Due to the non-quadratic nature of the objective function in PEL, an iterative algorithm must be used to solve the minimization. In this paper, we propose to use iterative nonlinear optimization algorithm together with the local quadratic approximation (LQA) studied in Fan and Li (2001) to minimize PEL ratio defined by (2.6). Specifically, we approximate $p_\tau(|\theta_j|)$ by $p_\tau(|\theta_j^{(k)}|) + \frac{1}{2}\{p'_\tau(|\theta_j^{(k)}|)/|\theta_j^{(k)}|\}\{\theta_j^2 - (\theta_j^{(k)})^2\}$, where $\theta_j^{(k)}$ is the kth step estimate of θ_j. We then make use of the algorithm discussed in Owen (2001, Chapter 12) to obtain the minimum through nonlinear optimization. The procedure is repeated until convergence.

During the iteration, we follow the strategy in Fan and Li (2001) to set $\hat{\theta}_j^{(k)}$ as zero whenever the jth component of the kth step minimizer $\hat{\theta}^{(k)}$ of (2.6) is very close to zero. The data analysis and the simulation in the next section suggest that this algorithm converges quickly, usually in a few iterations given a good initial value. We have also examined the local linear approximation in Zou and Li (2008) and the results obtained were similar to those obtained by LQA presented in Section 5.

5 Simulation

We present extensive simulation studies to illustrate the usefulness of PEL. First, we want to assess the performance of PEL in terms of variable selection and the accuracy of resulting estimates. Second, we evaluate the Wilks' phenomenon of PEL as in Theorem 2 and Theorem 4. The tuning parameter τ is taken from a fine grid and the one which minimizes $BIC(\tau)$ is used for estimation and constructing confidence regions.

5.1 PEL for Mean

We generate iid $Z_i \in \mathcal{R}^p$ whose components follow χ_1^2 distribution independently and take $X_i = \mu_0 + \Sigma^{1/2}(Z_i - 1)$, where the true parameter was set as $\mu_0 = (1, 0.6, 0.3, 0, \dots, 0)^T$ in all simulation settings. We considered $\Sigma = (\sigma_{ij})$ in settings with $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.3$ or 0.7 respectively for $i \neq j$. By increasing the correlation coefficient, we planned to examine the performance of the PEL estimator under different degrees of dependency. Note that in this case the property of X_i satisfies the factor model (2.2).

We conducted simulations for different dimensionality p and sample size n, and the simulation was repeated 1000 times for each designed case. For each replication, after generating $\{X_i\}_{i=1}^n$, $\hat{\mu}$ was obtained by minimizing PEL given by (2.6). The initial value of the minimization was naturally chosen as $\bar{X}$. The optimization was conducted through an iterative nonlinear optimization procedure and LQA described in Section 4. For each given μ in a neighborhood of $\bar{X}$, λ_μ was obtained by solving (2.1). Then (2.6) was minimized via nonlinear minimization procedure. During the iterative minimization, we followed the convention in Fan and Li (2001) to set a component of $\hat{\mu}$ to be zero if it is less than some threshold level very close to zero. In our simulation study, the threshold level was 0.001.

For comparison, we also computed the sample average $\bar{X}$, the soft-thresholded estimator $\hat{\mu}_{ST,j} = \text{sign}(\bar{X}_j)[|\bar{X}_j| - \lambda_1]_+$ where $[s]_+ = s$ for $s > 0$ and 0 otherwise; the hard-thresholded estimator $\hat{\mu}_{HT,j} = \bar{X}_j I(|\bar{X}_j| > \lambda_2)$; and finally a quadratic loss based estimator

$$\hat{\mu}_{QL} = \text{argmin}_\mu \ (\bar{X} - \mu)^T S_n^{-1} (\bar{X} - \mu) + \lambda_3 \sum_{j=1}^{p} |\mu_j|$$

where $S_n = n^{-1} \sum (X_i - \bar{X})(X_i - \bar{X})^T$. Here $\lambda_k, \ k = 1, 2, 3$ are the corresponding tuning parameters and are chosen by 5 fold cross validation to minimize the squared predictive error for the mean vector. The quadratic loss estimator was motivated by the need to incorporate the correlations among the components of X. We compared these estimators in terms of the root mean square error (RMSE) for the nonzero components. For model selection, we recorded the average numbers of zero coefficients which were correctly (True) and incorrectly (False) identified.

The results were summarized in Table 1. We see that PEL had small average false estimated zeros and the average numbers of zero components are close to $p-3$. This demonstrates the good performance of the proposed PEL approach in variable selection. The results also illustrate that when the correlation between components of X were not zero, the PEL estimator consistently had smaller RMSEs than other approaches, especially for highly correlated data. This agrees with the results in Theorem 1 that by incorporating the correlations among components, the PEL estimates can achieve improved efficiency. The PEL estimator was also seen to outperform other the estimators also in variable selection, usually by a large margin especially for highly correlated data.

The performance of the PEL confidence region was also evaluated. Setting $L_n = (1, 0, \ldots, 0)$ in (2.8) leads to a confidence set for μ_1, the first component of μ, at $1-\alpha$ level. For the nominal size $\alpha = 0.05$, we report the empirical frequencies of $\mu_1 \notin V_\alpha$ for a sequence of μ_1 values in Table 2. We note that at the true value of $\mu_1 = 1$, the frequency in rejecting the null hypothesis was close to the nominal level $\alpha = 0.05$. This is the case for all settings of the simulation and indicates that the proposed test maintained the size and confirmed the result in Theorem 2. When the discrepancy between μ_1 and the true value was larger, the rejection frequency increased. In particular, when the difference between μ_1 and the true value was 0.2, the rejection rate was close to one for $p = 40, n = 500$. This shows that the proposed PEL-based test had a good power for testing the null hypothesis. In addition, we observe that the power of the test also increased when the correlation became stronger.

5.2 PEL for Linear Models

We consider the following linear model (Tibshirani, 1996)

$$Y = X^T\beta + \sigma\epsilon,$$

where $X \in \mathcal{R}^p$ follows a multivariate normal distribution with unit variance and the correlation between the ith and jth components satisfies $\rho_{ij} = 0.5^{|i-j|}$. The true $\beta \in \mathcal{R}^p$ is given by $\beta_0 = (3, 1.5, 0, 0, 2, 0, \ldots, 0)^T$. We set $\sigma = 1$ and let ϵ follow standard normal distribution. We studied a few cases with varying p and n.

After generating $\{X_i, Y_i\}_{i=1}^n$ independently from the model, the PEL estimator was obtained by minimizing (3.3) where solving EL in (3.2) is implemented by an approach

						Avg. No. of 0 Coefficients	
				RMSE		True	False
	ρ		μ_1	μ_2	μ_3		
$p = 10$, $n = 50$		$\bar{X}$	0.197	0.201	0.197	-	-
	0.3	$\hat{\mu}$	0.190	0.218	0.243	6.038	0.426
		$\hat{\mu}_{ST}$	0.281	0.279	0.225	4.345	0.335
		$\hat{\mu}_{HT}$	0.205	0.324	0.284	6.056	0.882
		$\hat{\mu}_{QL}$	0.255	0.261	0.225	3.269	0.299
	0.7	$\hat{\mu}$	0.137	0.149	0.175	5.858	0.194
		$\hat{\mu}_{ST}$	0.296	0.287	0.238	4.466	0.328
		$\hat{\mu}_{HT}$	0.208	0.308	0.284	5.730	0.817
		$\hat{\mu}_{QL}$	0.180	0.181	0.175	3.019	0.100
$p = 20$, $n = 100$		$\bar{X}$	0.143	0.141	0.142	-	-
	0.3	$\hat{\mu}$	0.133	0.147	0.185	15.623	0.194
		$\hat{\mu}_{ST}$	0.222	0.225	0.202	12.424	0.224
		$\hat{\mu}_{HT}$	0.142	0.195	0.257	15.594	0.676
		$\hat{\mu}_{QL}$	0.187	0.189	0.183	10.984	0.116
	0.7	$\hat{\mu}$	0.089	0.095	0.117	15.670	0.050
		$\hat{\mu}_{ST}$	0.227	0.227	0.206	11.790	0.189
		$\hat{\mu}_{HT}$	0.143	0.182	0.245	14.060	0.564
		$\hat{\mu}_{QL}$	0.129	0.131	0.128	10.231	0.009
$p = 30$, $n = 200$		$\bar{X}$	0.100	0.101	0.097	-	-
	0.3	$\hat{\mu}$	0.095	0.092	0.134	26.229	0.056
		$\hat{\mu}_{ST}$	0.165	0.161	0.160	20.654	0.044
		$\hat{\mu}_{HT}$	0.099	0.097	0.208	25.225	0.408
		$\hat{\mu}_{QL}$	0.141	0.139	0.141	20.030	0.015
	0.7	$\hat{\mu}$	0.062	0.061	0.078	26.311	0.011
		$\hat{\mu}_{ST}$	0.166	0.162	0.159	19.521	0.038
		$\hat{\mu}_{HT}$	0.101	0.098	0.183	22.998	0.288
		$\hat{\mu}_{QL}$	0.095	0.093	0.096	18.595	0.001
$p = 40$, $n = 500$		$\bar{X}$	0.061	0.064	0.062	-	-
	0.3	$\hat{\mu}$	0.057	0.056	0.065	36.809	0.000
		$\hat{\mu}_{ST}$	0.108	0.111	0.109	29.743	0.000
		$\hat{\mu}_{HT}$	0.063	0.063	0.080	35.099	0.031
		$\hat{\mu}_{QL}$	0.093	0.096	0.093	29.788	0.000
	0.7	$\hat{\mu}$	0.037	0.037	0.039	36.809	0.000
		$\hat{\mu}_{ST}$	0.108	0.109	0.109	27.947	0.003
		$\hat{\mu}_{HT}$	0.063	0.063	0.076	31.531	0.022
		$\hat{\mu}_{QL}$	0.062	0.064	0.062	28.486	0.000

Table 1: Simulation Results of PEL method in estimating the mean vector. $\bar{X}$: the sample mean; $\hat{\mu}$: our approach; $\hat{\mu}_{ST}$: soft-threshold estimator; $\hat{\mu}_{HT}$: hard-threshold estimator;$\hat{\mu}_{QL}$: the estimator using the quadratic loss.

μ_1	ρ	0.8	0.9	1.0	1.1	1.2
$p = 10$, $n = 50$	0.3	0.170	0.068	0.062	0.113	0.211
	0.7	0.302	0.107	0.059	0.148	0.311
$p = 20$, $n = 100$	0.3	0.364	0.120	0.056	0.135	0.316
	0.7	0.672	0.204	0.055	0.181	0.493
$p = 30$, $n = 200$	0.3	0.668	0.206	0.069	0.229	0.517
	0.7	0.936	0.386	0.071	0.319	0.758
$p = 40$, $n = 500$	0.3	0.978	0.508	0.064	0.400	0.878
	0.7	0.983	0.793	0.065	0.636	0.962

Table 2: The empirical frequency that a given value of μ_1 does not fall in the 95% confidence set constructed by (2.8). The truth is $\mu_1 = 1$.

					Avg. No. of 0 Coefficients	
			RMSE		True	False
		β_1	β_2	β_5		
$p = 10$, $n = 50$	$\hat{\beta}$	0.179	0.190	0.180	6.108	0
	$\hat{\beta}_{or}$	0.168	0.169	0.147	-	-
$p = 20$, $n = 100$	$\hat{\beta}$	0.132	0.144	0.122	15.624	0
	$\hat{\beta}_{or}$	0.114	0.120	0.100	-	-
$p = 30$, $n = 200$	$\hat{\beta}$	0.088	0.092	0.081	26.082	0
	$\hat{\beta}_{or}$	0.080	0.082	0.070	-	-
$p = 40$, $n = 500$	$\hat{\beta}$	0.054	0.055	0.049	36.685	0
	$\hat{\beta}_{or}$	0.051	0.052	0.046	-	-

Table 3: Simulation Results of PEL method in estimating the regression coefficient in linear models. $\hat{\beta}$: our method; $\hat{\beta}_{or}$: the oracle estimator assuming the true sparsity is known.

similar to that in the mean case. The results were summarized in Tables 3 and 4, in a similar manner to Tables 1 and 2 for the mean vector case.

From Table 3, PEL did not incorrectly set nonzero coefficients as zero, showing its attractiveness in minimizing the false negative rate. The average numbers of the estimated zero coefficients were also close to $p-3$, implying that the selected model was very close to the true model in terms of nonzero coefficients. The PEL estimator $\hat{\beta}$ was compared with the oracle estimator $\hat{\beta}_{or}$ given by the least square estimates knowing the true sparsity of the model beforehand. We see that the RMSEs of PEL estimates were slightly larger but quite close to those of the Oracle ones. For large sample size ($n = 500, p = 40$), the RMSEs of $\hat{\beta}$ and $\hat{\beta}_{or}$ were almost identical. This confirmed the theoretical results in Theorem 3 that the PEL estimates achieve the efficiency of the Oracle estimator asymptotically.

For $L_n = (1, 0, \ldots, 0)^T$ in (3.4), we also examined the empirical rejection frequency for $\beta_1 \notin V_\alpha$ constructed by (3.5), where β_1 is the first component of β. From Table 4, it is

β_1	2.8	2.9	3.0	3.1	3.2
$p = 10$, $n = 50$	0.272	0.139	0.085	0.115	0.254
$p = 20$, $n = 100$	0.393	0.173	0.086	0.172	0.401
$p = 30$, $n = 200$	0.602	0.226	0.082	0.238	0.641
$p = 40$, $n = 500$	0.937	0.496	0.073	0.476	0.948

Table 4: The empirical frequency that a given value of β_1 does not fall in the 95% confidence constructed by (3.5). The truth is $\beta_1 = 3$.

seen that the empirical rejection rates in the simulation were close to the nominal level 0.05 when β_1 were set as the true value in the simulation. This confirmed our result in Theorem 4. The power increased when the sample sizes became larger or β_1 took values departing further away from the true value in the data model. This demonstrates the promising performance of PEL in hypothesis testing and confidence set construction.

6 Discussion

For high dimensional data analysis, we have proposed the penalized empirical likelihood (PEL) method for variable selection and coefficient estimation, and illustrated its usefulness in estimating a mean vector and regression coefficients in linear models. We have shown that the PEL ratio statistic is valid for hypothesis tests and constructing confidence intervals. The use of a penalty function produces sparse models, which help construct the confidence region in at least two aspects. First, the complexity of profiling is reduced due to sparsity. Second, sparse models is useful for constructing tighter confidence regions. This is analogous to the smaller standard errors if the penalized likelihood is used instead of the usual likelihood (Fan and Li, 2001).

The PEL method is computation intensive. In return, PEL shares merits stemmed from a nonparametric likelihood based approach, for example in terms of robustness to distributional assumptions, automatic determination of the shape of confidence regions; automatic incorporation of information and many others. Our results show that some of the desirable properties of empirical likelihood continue to apply in regularized high dimensional data analysis, and that PEL is an effective yet robust approach for variable selection. We have confirmed these results by numerical simulation. In this paper, we have limited our discussion to two illustrative scenarios. It is of great interest to explore the usefulness of PEL in a wider context.

Appendix

We first prove the results in the mean case. For (2.5), Theorem 1 of Chen et al. (2009) shows that $||\lambda_{\mu_0}|| = O_p\{(p/n)^{1/2}\}$ assuring the asymptotic expansion of $\ell(\mu_0)$. In the following Lemma, we extend the result for λ_μ where μ is in a large enough neighborhood of μ_0. For notational purposes, we define $I_p = (H_1^T, H_2^T)$ where $H_1 \in \mathcal{R}^{d\times p}$ and $H_2 \in \mathcal{R}^{(p-d)\times p}$. We present the following lemmas before proving Theorem 1.

Lemma 1. *Let $a_n = (p/n)^{1/2-\delta}$, $D_n = \{\mu : ||\mu - \mu_0|| \le ca_n\}$ where $\delta, c > 0$ are constants being strictly positive and δ satisfies $p^{1-\delta}/n^{1/2-\delta} \to 0$, then $||\lambda_\mu|| = O_p(a_n)$ for $\mu \in D_n$.*

Proof of Lemma 1

For $\mu \in D_n$, let $S_\mu = n^{-1}\sum_{i=1}^n (X_i - \mu)(X_i - \mu)^T$ and $\lambda_\mu = \rho\theta$ where $||\theta|| = 1$ is a unit vector. Following the argument in Owen (2001, P220), we have

$$\rho\{\theta^T S_\mu \theta - \max_{i=1,\ldots,n} ||X_i - \mu|| n^{-1} |\sum_{i=1}^n \theta^T (X_i - \mu)|\} \le n^{-1} |\sum_{i=1}^n \theta^T (X_i - \mu)|.$$

As $\mu \in D_n$, it is seen $n^{-1}|\sum_{i=1}^n \theta^T(X_i - \mu)| = O_p(a_n)$. We may decompose S_μ as

$$S_\mu = S_{\mu_0} + n^{-1}\sum_{i=1}^n (X_i - \mu_0)(\mu_0 - \mu)^T + n^{-1}\sum_{i=1}^n (\mu_0 - \mu)(X_i - \mu_0)^T + (\mu_0 - \mu)(\mu_0 - \mu)^T.$$

Using the Frobenius norm of a matrix A defined as $||A|| = \sqrt{\mathrm{tr}(A^T A)}$, we note that S_{μ_0} converges to Σ and $||S_n - \Sigma|| = O_p(\sqrt{p/n})$ by results in Chen et al. (2009) under model (2.2). In addition, by the fact that $||n^{-1}\sum_{i=1}^n (X_i - \mu_0)|| = O_p(\sqrt{p/n})$, we have $||S_\mu - S_{\mu_0}|| = o_p(\sqrt{p/n})$. Hence for $\mu \in D_n$, S_μ also converges to Σ under the Frobenius norm. Therefore, $\theta^T S_\mu \theta = O_p(1)$. By Lemma 2 of Chen et al. (2009), $\max_i ||X_i - \mu_0|| = O_p\{p^{1/2}\} + o_p\{p^{-(2k-1)/(4k)} n^{1/(4k)}\}$. Thus

$$\max_{i=1,\ldots,n} ||X_i - \mu|| \le ca_n + \max_{i=1,\ldots,n} ||X_i - \mu_0|| = O(a_n) + O_p\{p^{1/2}\} + o_p\{p^{-(2k-1)/(4k)} n^{1/(4k)}\}.$$

By condition A.3 and the specification of δ, Lemma 1 is established by noting $\max_{i=1,\ldots,n} ||X_i - \mu|| n^{-1} |\sum_{i=1}^n \theta^T (X_i - \mu)| = o_p(1)$. As a consequence, Lemma 1 implies that $\max_{i=1\ldots,n} \lambda_\mu^T (X_i - \mu) = o_p(1)$. In the rest part of the proofs, we use Frobenius norm for matrices.

Lemma 2. *As $n \to \infty$, with probability tending to 1, $\ell_p(\mu)$ given by (2.6) has a minimum in D_n.*

Proof: For $\mu \in D_n$, by definition of $L(\mu)$,

$$Q_{1n}(\mu, \lambda) = n^{-1}\sum_{i=1}^n \frac{X_i - \mu}{1 + \lambda^T (X_i - \mu)} = 0.$$

By Lemma 1, we have that $\lambda_\mu^T(X_i-\mu)$ is stochastically small uniformly for $\mu\in D_n$. Applying Taylor's expansion on $Q_{1n}(\mu,\lambda_\mu)$, we have $0=\bar{X}-\mu-S_\mu\lambda_\mu+r_n$ where $r_n=n^{-1}\sum_{i=1}^n\left[(X_i-\mu)\{\lambda_\mu^T(X_i-\mu)\}^2(1+\xi_i)^{-3}\right]$ is the remainder term and $|\xi_i|\le|\lambda_\mu^T(X_i-\mu)|$. Inverting the expansion, we have $\lambda_\mu=S_\mu^{-1}(\bar{X}-\mu)+S_\mu^{-1}r_n$. Substituting λ_μ into $\ell(\mu)$,we establish that

$$2\ell(\mu)=n(\bar{X}-\mu)^TS_\mu^{-1}(\bar{X}-\mu)-nr_n^TS_\mu^{-1}r_n+2/3\sum_{i=1}^n\{\lambda_\mu^T(X_i-\mu)\}^3(1+\xi_i)^{-4}. \quad \text{(A.1)}$$

For $\mu\in\partial D_n$ where ∂D_n denotes the boundary of D_n, we write $\mu=\mu_0+ca_n\theta_\mu$ where θ_μ is a unit vector, we have a decomposition as $2\ell(\mu)=T_0+T_1+T_2$ where $T_0=n(\bar{X}-\mu_0)^TS_{\mu_0}^{-1}(\bar{X}-\mu_0)$, $T_1=n(\mu-\mu_0)^TS_\mu^{-1}(\mu-\mu_0)$ and

$$\begin{aligned}T_2=&\,n\left\{(\bar{X}-\mu_0)^T(S_\mu^{-1}-S_{\mu_0}^{-1})(\bar{X}-\mu_0)+2(\bar{X}-\mu_0)^TS_\mu^{-1}(\mu-\mu_0)\right\}\\&-nr_n^TS_\mu^{-1}r_n+2/3\sum_{i=1}^n\{\lambda_\mu^T(X_i-\mu)\}^3(1+\xi_i)^{-4}.\end{aligned}$$

As $n\to\infty$, we see that $T_1=c^2na_n^2\theta_\mu S_\mu^{-1}\theta_\mu\ge c^2na_n^2\gamma_p^{-1}(S_\mu)=O_p(na_n^2)$, $na_n^2\to\infty$, $T_2/T_1\xrightarrow{p}0$ and $2\ell(\mu_0)-T_0=o_p(1)$. This implies that for any C given, as $n\to\infty$ $P[\{2\ell(\mu)-2\ell(\mu_0)\}>C]\to1$. Further, note that for n large,

$$\begin{aligned}\ell_p(\mu)-\ell_p(\mu_0)&=\ell(\mu)-\ell(\mu_0)+n\sum_j\{p_\tau(|\mu_j|)-p_\tau(|\mu_{0j}|)\}\\&\ge\ell(\mu)-\ell(\mu_0)+n\sum_{j\in\mathcal{A}}\{p_\tau(|\mu_j|)-p_\tau(|\mu_{0j}|)\}\ge\ell(\mu)-\ell(\mu_0)\end{aligned}$$

where the last inequality holds due to Condition B.1 and the unbiased property of the SCAD penalty so that for $j\in\mathcal{A}$, $p_\tau(|\mu_{0j}|)=p_\tau(|\mu_j|)$ when n is large. Hence, with probability tending to 1, $\ell_p(\mu)>\ell_p(\mu_0)$ for $\mu\in\partial D_n$, which establishes Lemma 2.

Proof of Theorem 1

By Lemma 2, we note that the minimizer of $\ell_p(\mu)$ is in D_n. Considering $\mu\in D_n$, we have that for each of its component

$$\frac{1}{n}\frac{\partial\ell_p(\mu)}{\partial\mu_j}=n^{-1}\sum_{i=1}^n\frac{-\lambda_{\mu j}}{1+\lambda_\mu^T(x_i-\mu)}+p'_\tau(|\mu_j|)\mathrm{sign}(\mu_j)=I_j+II_j.$$

First, we show that

$$\max_{j\notin\mathcal{A}}|I_j| \quad\le\quad \max_{j\notin\mathcal{A}}|\lambda_{\mu j}|\left|n^{-1}\sum_{i=1}^n\frac{1}{1+\lambda_\mu^T(x_i-\mu)}\right|=O_p(a_n)O_p(1)=o_p(\tau),$$

because $\|\lambda_\mu\|=O_p(a_n)$ by Lemma 1 and $\tau(n/p)^{1/2-\delta}\to\infty$ by condition B.1 on $p_\tau(\cdot)$. This implies that $P(\max_{j\notin\mathcal{A}}|I_j|>\tau/2)\to0$. In addition, we note that $|\mu_j|\le ca_n$ for $j\notin\mathcal{A}$ and $\tau/a_n\to\infty$. Thus for n large enough and $j\notin\mathcal{A}$, $p'(|\mu_j|)=\tau$ and hence

$$II_j=p'(|\mu_j|)\mathrm{sign}(\mu_j)=\tau\mathrm{sign}(\mu_j).$$

Therefore, the sign of μ_j dominates $\partial\ell_p/\partial\mu_j$ asymptotically for $j \notin \mathcal{A}$. Namely, as $n \to \infty$, for any $j \notin \mathcal{A}$, with probability tending to one,

$$\frac{1}{n}\frac{\partial\ell_p(\mu)}{\partial\mu_j} < 0 \text{ for } \mu_j \in (0, ca_n) \text{ and } \frac{1}{n}\frac{\partial\ell_p(\mu)}{\partial\mu_j} > 0 \text{ for } \mu_j \in (-ca_n, 0).$$

Therefore $\hat\mu_2 = 0$ with probability tending to one and we conclude part 1 of Theorem 1.

Next we show part 2 of Theorem 1. By result from part 1 and the definition of PEL, the estimator $\hat\mu$ based on PEL is the constrained minimizer of (2.6) subject to $H_2\mu = 0$. According to Qin and Lawless (1995), by the Lagrange multiplier method, it is equivalent to minimizing a new objective function

$$\tilde\ell(\mu,\lambda,\nu) = n^{-1}\sum_{i=1}^{n}\log\{1+\lambda^T(X_i-\mu)\} + \sum_{j=1}^{p} p_\tau(|\mu_j|) + \nu^T H_2\mu, \tag{A.2}$$

where $\nu \in \mathcal{R}^{p-d}$ is the vector of extra Lagrange multiplier. Define

$$\tilde Q_{1n}(\mu,\lambda,\nu) = n^{-1}\sum_{i=1}^{n}\frac{X_i-\mu}{1+\lambda^T(X_i-\mu)},$$
$$\tilde Q_{2n}(\mu,\lambda,\nu) = -n^{-1}\sum_{i=1}^{n}\frac{\lambda}{1+\lambda^T(X_i-\mu)} + b(\mu) + H_2^T\nu \text{ and } \tilde Q_{3n}(\mu,\lambda,\nu) = H_2\mu$$

where $b(\mu) = \{p'_\tau(|\mu_1|)\text{sign}(\mu_1), p'_\tau(|\mu_2|)\text{sign}(\mu_2), \ldots, p'_\tau(|\mu_d|)\text{sign}(\mu_d), 0, \ldots, 0\}^T$. Note that the minimizer $(\hat\mu, \hat\lambda, \hat\nu)$ satisfies that $0 = \tilde Q_{jn}(\hat\mu, \hat\lambda, \hat\nu)$ for $j = 1, 2, 3$. We have from Lemma 1 that $\hat\lambda = O_p(a_n)$ is stochastically small. And from Lemma 2, $||\hat\mu - \mu_0|| = O_p(a_n)$. Therefore, similar to the argument in Qin and Lawless (1995), from $0 = \tilde Q_{2n}(\hat\mu, \hat\lambda, \hat\nu)$ we conclude that $||\hat\nu|| = O_p(a_n)$. Hence, we can use stochastic expansions of $\tilde Q_{jn}$ around value $(\mu_0, 0, 0)$ for $j = 1, 2, 3$. This leads to

$$\begin{pmatrix} -\tilde Q_{1n}(\mu_0,0,0) \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -\Sigma & -I & 0 \\ -I & 0 & H_2^T \\ 0 & H_2 & 0 \end{pmatrix}\begin{pmatrix} \hat\lambda - 0 \\ \hat\mu - \mu_0 \\ \hat\nu - 0 \end{pmatrix} + R_n^{(1)} + R_n^{(2)} + R_n^{(3)} + R_n^{(4)} \tag{A.3}$$

where $R_n^{(1)} = (R_{1n}^{T(1)}, R_{2n}^{T(1)}, 0)^T$, $R_{1n}^{(1)} = n^{-1}\sum_{i=1}^n\{\hat\lambda^T(X_i-\hat\mu)\}^2(X_i-\hat\mu)/\{1+\hat\lambda(X_i-\hat\mu)\} - (S_{\hat\mu} - S_{\mu_0})\hat\lambda$, $R_{2n}^{(1)} = -n^{-1}\sum_{i=1}^n\hat\lambda\hat\lambda^T(X_i-\hat\mu)/\{1+\hat\lambda(X_i-\hat\mu)\}$. Following Theorem 3 in Chen et al. (2009), $||R_{1n}^{(1)}|| = o_p(1/\sqrt{n})$ given Condition A.3. It is clear that $||R_{2n}^{(1)}|| = O_p(a_n^2) = o_p(1/\sqrt{n})$. The other three terms are given by $R_n^{(2)} = \{0, b^T(\mu_{0\mathcal{A}}), 0\}^T$, $R_n^{(3)} = \{0, \{b'(\mu^*)(\hat\mu - \mu_0)\}^T, 0\}^T$ and $R_n^{(4)} = -\{\{(S_{\mu_0} - \Sigma)\hat\lambda\}^T, 0, 0\}^T$. By the implication of SCAD penalty function in B.2, we see that $||R_n^{(2)}|| = o_p(1/\sqrt{n})$ and $||R_n^{(3)}|| = o_p(1/\sqrt{n})$. Further, since $b(\cdot)$ and $b'(\cdot)$ only incur at those nonzero components, $R_n^{(2)} = R_n^{(3)} = 0$ for n large enough. As $||S_{\mu_0} - \Sigma|| = O_p(\sqrt{p/n})$, $||R_n^{(4)}|| = O_p(\sqrt{p/n})O_p(a_n) = o_p(1/\sqrt{n})$.

Define $K_{11} = -\Sigma$, $K_{12} = (-I, 0)$, $K_{21} = K_{12}^T$, $K_{22} = \begin{pmatrix} 0 & H_2^T \\ H_2 & 0 \end{pmatrix}$ and $K = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}$

and let $\vartheta = (\mu^T, \nu^T)^T$. By inverting (A.3), we have

$$\begin{pmatrix} \hat{\lambda} - 0 \\ \hat{\vartheta} - \vartheta_0 \end{pmatrix} = K^{-1} \left\{ \begin{pmatrix} -Q_{1n}(\mu_0, 0, 0) \\ 0 \\ 0 \end{pmatrix} + R_n \right\} \tag{A.4}$$

where $R_n = \sum_{k=1}^4 R_n^{(k)}$ and $||R_n|| \leq \sum_{k=1}^4 ||R_n^{(k)}|| = o_p(1/\sqrt{n})$ by above arguments. Applying matrix inverse by blocks, we have that

$$K^{-1} = \begin{pmatrix} K_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} -K_{11}^{-1} K_{12} \\ I \end{pmatrix} A^{-1} \begin{pmatrix} -K_{21} K_{11}^{-1} & I \end{pmatrix}$$

where $A = K_{22} - K_{21} K_{11}^{-1} K_{12} = \begin{pmatrix} \Sigma^{-1} & H_2^T \\ H_2 & 0 \end{pmatrix}$. Then $\hat{\vartheta} - \vartheta_0 = A^{-1} K_{21} K_{11}^{-1} \tilde{Q}_{1n}(\mu_0, 0, 0) + o_p(1/\sqrt{n})$. Another matrix inverse by blocks on A gives

$$A^{-1} = \begin{pmatrix} \Sigma - \Sigma H_2^T (H_2 \Sigma H_2^T)^{-1} H_2 \Sigma & -\Sigma H_2^T (H_2 \Sigma H_2^T)^{-1} \\ -(H_2 \Sigma H_2^T)^{-1} H_2 \Sigma & (H_2 \Sigma H_2^T)^{-1} \end{pmatrix}.$$

This implies

$$\hat{\mu} - \mu_0 = \{I - \Sigma H_2^T (H_2 \Sigma H_2^T)^{-1} H_2\}\{(\bar{X} - \mu_0) + R_{1n}\} \tag{A.5}$$

where R_{1n} is the corresponding component in vector $K^{-1} R_n$ and $||R_{1n}|| = o_p(1/\sqrt{n})$. It is clear that the expansion of the nonzero component μ_1 is given by

$$\hat{\mu}_1 - \mu_{10} = \{H_1 - H_1 \Sigma H_2^T (H_2 \Sigma H_2^T)^{-1} H_2\}\{(\bar{X} - \mu_0) + R_{1n}\}. \tag{A.6}$$

Let $I_{\mathcal{A}} = H_1 \Sigma H_1^T - H_1 \Sigma H_2^T (H_2 \Sigma H_2^T)^{-1} H_2 \Sigma H_1^T = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$,

$$Y_{ni} = \frac{1}{\sqrt{n}} Z_{ni} \text{ where } Z_{ni} = W_n I_{\mathcal{A}}^{-1/2} \{H_1 - H_1 \Sigma H_2^T (H_2 \Sigma H_2^T)^{-1} H_2\}(X_i - \mu_0).$$

It is straightforward to verify that

$$P(||Y_{ni}|| > \epsilon) \leq n^{-1} \epsilon^{-2} E||Z_{ni}||^2 = O(1/n) \text{ and } E||Y_{ni}||^4 = n^{-2} E(Z_{ni}^T Z_{ni})^2 = O(p^2/n^2).$$

Hence, $\sum_{i=1}^n E||Y_{ni}||^2 I(||Y_{ni}|| > \epsilon) \leq n\sqrt{E||Y_{n1}||^4}\sqrt{P(||Y_{n1}|| > \epsilon)} \to 0$. As $W_n W_n^T \to G$, by Lindeberg-Feller's CLT (Van Der Vaart, 1998), we have

$$\sqrt{n} W_n I_{\mathcal{A}}^{-1/2} \{H_1 - H_1 \Sigma H_2^T (H_2 \Sigma H_2^T)^{-1} H_2\}(\bar{X} - \mu_0) \xrightarrow{d} N(0, G).$$

Finally, by noting that $||\sqrt{n} W_n I_{\mathcal{A}}^{-1/2} H_1 R_{1n}||^2 \leq n \gamma_d(W_n W_n^T) \gamma_1^{-1}(I_{\mathcal{A}}) \gamma_1^{-2}(K) ||R_n||^2 = o_p(1)$, the proof of Theorem 1 is completed.

Proof of Theorem 2

First, we present the asymptotic expansion of $\ell(\hat{\mu})$ where $\hat{\mu}$ is the minimizer of (2.6), i.e. the EL ratio evaluated at the PEL estimates. Let $z_i = \hat{\lambda}^T(X_i - \hat{\mu})$. Since $\max_i |\hat{\lambda}^T(X_i - \hat{\mu})| = o_p(1)$ as implied by Lemma 1, by Taylor's expansion and results in proof of Theorem 1, we have

$$\ell(\hat{\mu}) = \sum_{i=1}^{n} z_i - \sum_{i=1}^{n} z_i^2/2 + \sum_{i=1}^{n} z_i^3/\{3(1+\xi_i)^4\} + o_p(1), \tag{A.7}$$

where $|\xi_i| < |\hat{\lambda}^T(X_i - \hat{\mu})|$ and the $o_p(1)$ term is due to the penalty function. Recall in Lemma 2, we have shown the expansion for $\mu \in D_n$ as

$$\lambda_\mu = S_\mu^{-1}(\bar{X} - \mu) + S_\mu^{-1} r_n,$$

where $r_n = n^{-1}\sum_{i=1}^n \left[(X_i - \mu)\{\lambda_\mu^T(X_i - \mu)\}^2(1+\xi_i)^{-3}\right]$ and $|\xi_i| \le |\lambda_\mu^T(X_i - \mu)|$. Substituting the expansion of $\hat{\lambda}$ and $\hat{\mu}$ given by (A.5) into z_i we show that

$$2\ell(\hat{\mu}) = n(\bar{X} - \mu_0)^T H_2^T (H_2 \Sigma H_2^T)^{-1} H_2 (\bar{X} - \mu_0) + o_p(1). \tag{A.8}$$

Next we develop the expansion of $\ell_p(\hat{\mu})$ under the null hypothesis $H_0 : L_n\mu_1 = 0$. Under the null hypothesis, since $L_n L_n^T = I_q$, there exists $\tilde{H}_2$ such that $\tilde{H}_2\mu = 0$ and $\tilde{H}_2\tilde{H}_2^T = I_{p-d+q}$. Now by repeating the proof of Theorem 1, we establish that under the null hypothesis, the estimation of μ can be obtained by minimizing

$$\tilde{\ell}_p(\mu, \lambda, \nu) = \sum_{i=1}^{n} \log\{1 + \lambda^T(X_i - \mu)\} + n\sum_{i=1}^{p} p_\tau(|\mu_i|) + \nu^T \tilde{H}_2 \mu, \tag{A.9}$$

Denote the minimizer of (A.9) by $(\breve{\mu}, \breve{\lambda}, \breve{\nu})$. By the proof of part 1 in Theorem 1, $\breve{\mu}_2 = 0$ with probability tending to 1. Thus $n\{\sum_{i=1}^p p_\tau(|\hat{\mu}_j|) - \sum_{i=1}^p p_\tau(|\breve{\mu}_j|)\} = 0$ with probability tending to 1. By replacing H_2 in (A.5) by $\tilde{H}_2$ and by expansion (A.7), we establish that

$$2\ell(\hat{\mu})_{L_n\mu_1=0} = 2\ell(\breve{\mu}) = n(\bar{X} - \mu_0)^T \tilde{H}_2^T (\tilde{H}_2 \Sigma \tilde{H}_2^T)^{-1} \tilde{H}_2 (\bar{X} - \mu_0) + o_p(1). \tag{A.10}$$

Combining (A.8) and (A.10), we have

$$\tilde{\ell}(L_n) = n(\bar{X} - \mu_0)\Sigma^{-1/2}\{P_1 - P_2\}\Sigma^{-1/2}(\bar{X} - \mu_0) + o_p(1)$$

where $P_1 = \Sigma^{1/2}\tilde{H}_2^T(\tilde{H}_2\Sigma\tilde{H}_2^T)^{-1}\tilde{H}_2\Sigma^{1/2}$ and $P_2 = \Sigma^{1/2}H_2^T(H_2\Sigma H_2^T)^{-1}H_2\Sigma^{1/2}$. As $P_1 - P_2$ is an idempotent matrix of rank q, $P_1 - P_2$ can be written as $\Xi_n^T\Xi_n$ where Ξ_n is a $q \times p$ matrix such that $\Xi_n\Xi_n^T = I_q$ (Fan and Peng, 2004). Further we see that $\sqrt{n}\Xi_n\Sigma^{-1/2}(\bar{X} - \mu_0) \xrightarrow{d} N(0, I_q)$ can be straightforwardly established. Then $n(\bar{X} - \mu_0)^T\Sigma^{-1/2}(P_1 - P_2)\Sigma^{-1/2}(\bar{X} - \mu_0) \xrightarrow{d} \chi_q^2$ and Theorem 2 follows.

To establish Theorem 3 for linear models, we need the following Lemmas. We use the same set of notations as those in the mean case in the following proofs.

Lemma 3. *Let $a_n = (p/n)^{1/2-\delta}$, $D_n = \{\beta : ||\beta - \beta_0|| \le ca_n\}$ where $\delta, c > 0$ are constants and δ satisfies $p^{1-\delta}/n^{1/2-\delta} \to 0$, then $||\lambda_\beta|| = O_p(a_n)$ for $\beta \in D_n$.*

Proof of Lemma 3

For simplicity, we assume $E(X_i) = 0$ in the following proof. It can be extended to case $E(X_i) = \mu$ by centering the random vector. We start from the following inequality similar to that in Lemma 1. Let $\lambda_\beta = \rho\theta$, where $\rho = ||\lambda_\beta||$ and θ is a unit vector

$$\rho\{\theta^T T_n(\beta)\theta - \max_{1\leq i\leq n} ||U_i(\beta)||n^{-1}|\sum_{i=1}^n \theta^T U_i(\beta)|\} \leq n^{-1}|\sum_{i=1}^n \theta^T U_i(\beta)|,$$

where $T_n(\beta) = n^{-1}\sum_{i=1}^n U_i(\beta)U_i^T(\beta)$. Note that $U_i(\beta_0) = X_i\epsilon_i$ and X_i satisfies the factor model (2.2) and ϵ_i is independent of X_i satisfying condition A.4. Therefore it is straightforward to show that $||n^{-1}\sum_{i=1}^n U_i(\beta_0)|| = O_p(\sqrt{p/n})$. And hence $n^{-1}|\sum_{i=1}^n \theta^T U_i(\beta_0)| = O_p(\sqrt{p/n})$. As $T_n(\beta_0) = n^{-1}\sum_{i=1}^n X_iX_i^T\epsilon_i^2$, it is seen that that $\theta^T T_n(\beta_0)\theta = O_p(1)$. Further, by noting condition A.4 that $m^{4k} = E(\epsilon_i^{4k}) < \infty$, we show that

$$\begin{aligned}\max_i ||U_i(\beta_0)|| &= \left\{\max_i (X_i^T X_i\epsilon_i^2)^k - m^{2k}E(X_i^T X_i)^k + m^{2k}E(X_i^T X_i)^k\right\}^{1/(2k)}\\ &= \left[var\{(X_i^T X_i\epsilon_i^2)^k\}\max_i\left\{\frac{(X_i^T X_i\epsilon_i^2)^k - m^{2k}E(X_i^T X_i)^k}{var\{(X_i^T X_i\epsilon_i^2)^k}\right\} + m^{2k}E(X_i^T X_i)^k\right]^{1/(2k)}\\ &= O(p^{1/2}) + o_p\{p^{-(2k-1)/(4k)}n^{1/(4k)}\}.\end{aligned}$$

Therefore $\max_{1\leq i\leq n}||U_i(\beta_0)||n^{-1}|\sum_{i=1}^n \theta^T U_i(\beta_0)| = o_p(1)$. And we conclude that $||\lambda_{\beta_0}|| = O_p(\sqrt{p/n})$. For other $\beta \in D_n$, we note that $n^{-1}|\sum_{i=1}^n \theta^T U_i(\beta)| = n^{-1}|\sum_{i=1}^n \theta^T\{U_i(\beta_0) + X_iX_i^T(\beta - \beta_0)\}|$. Since

$$T_n(\beta) - T_n(\beta_0) = -2n^{-1}\sum_{i=1}^n X_iX_i^TX_i^T(\beta-\beta_0)\epsilon_i + n^{-1}\sum_{i=1}^n X_iX_i^T\{X_i^T(\beta-\beta_0)\}^2,$$

we have $\theta^T T_n(\beta)\theta = \theta^T T_n(\beta_0)\theta + o_p(1) = O_p(1)$. By results in the proof of Lemma 1, since $E(X_i) = 0$, we have $\max_i ||X_i|| = O_p(p^{1/2}) + o_p\{p^{-(2k-1)/(4k)}n^{1/(4k)}\}$. Further we note that for $\beta \in D_n$, $\max_i ||X_iX_i^T(\beta - \beta_0)|| \leq ca_n \max_i ||X_i||^2 = o_p(1)$. Then $\max_i ||U_i(\beta)||n^{-1}|\sum_{i=1}^n \theta^T U_i(\beta)| = o_p(1)$ by Condition A.3 and thus Lemma 3 follows. As consequences of Lemmas 1 and 3, $\max_i |\lambda_\beta^T X_i| = o_p(1)$ and $\max_i |\lambda_\beta^T U_i(\beta)| = o_p(1)$.

Before proving Theorem 3, we present the following lemma.

Lemma 4. *As $n \to \infty$, with probability tending to 1, $\ell_p(\beta)$ given by (3.3) has a minimum in D_n.*

Proof of Lemma 4

We sketch the proof of Lemma 4 as follows. By Lemma 3 and the definition of EL, for $\beta \in D_n$,

$$0 = Q_{1n}(\beta, \lambda_\beta) = n^{-1}\sum_{i=1}^n \frac{U_i(\beta)}{1 + \lambda_\beta^T U_i(\beta)}$$

can be expanded as $0 = n^{-1}\sum_{i=1}^n U_i(\beta) - \lambda_\beta^T T_n(\beta) + r_n$, where the remainder term $r_n = n^{-1}\sum_{i=1}^n U_i(\beta)\{\lambda_\beta^T U_i(\beta)\}^2(1+\xi_i)^{-3}$ and $|\xi_i| \le |\lambda_\beta^T U_i(\beta)|$. This implies $\lambda_\beta = T_n^{-1}(\beta)\bar{U}(\beta) + T_n^{-1}(\beta)r_n$ where $\bar{U}(\beta) = n^{-1}\sum_{i=1}^n U_i(\beta)$. Substituting this into $\ell(\beta)$, we show, similar to that in proof of Lemma 2, that

$$2\ell(\beta) = n\bar{U}(\beta)^T T_n^{-1}(\beta)U(\beta) - nr_n^T T_n^{-1}(\beta)r_n + 2/3\sum_{i=1}^n\{\lambda_\beta^T U_i(\beta)\}^3(1+\xi_i)^{-4}.$$

Note that $U_i(\beta_0) = X_i\epsilon_i$ and X_i follows model (2.2). Also, for β in in boundary set ∂D_n, $2\ell(\beta) - 2\ell(\beta_0)$ is dominated by a positive term similar to that in proving Lemma 2

$$n\{n^{-1}\sum_{i=1}^n(X_iX_i^T)(\beta-\beta_0)\}^T T_n^{-1}(\beta)\{n^{-1}\sum_{i=1}^n(X_iX_i^T)(\beta-\beta_0)\} = O_p(na_n^2).$$

Therefore, as $n \to \infty$, $P[2\{\ell(\beta) - \ell(\beta_0)\} > C] \to 1$ for any $C > 0$. , Hence, we conclude the existence of a minimum in D_n similar to that in proving Lemma 2.

Proof of Theorem 3

By Lemma 4, we denote the solution to $\ell_p(\beta)$ as $\hat{\beta}$ satisfying $||\hat{\beta} - \beta_0|| \le ca_n$. We shall again show that for any $j \notin \mathcal{B}$, with probability tending to one,

$$\frac{1}{n}\frac{\partial\ell_p(\beta)}{\partial\beta_j} < 0 \text{ for } \beta_j \in (0, \epsilon_n) \text{ and } \frac{1}{n}\frac{\partial\ell_p(\beta)}{n\partial\beta_j} > 0 \text{ for } \beta_j \in (-\epsilon_n, 0) \text{ and } \epsilon_n = ca_n.$$

Firstly, we have

$$\frac{1}{n}\frac{\partial\ell_p(\beta)}{\partial\beta_j} \quad = \quad n^{-1}\sum_{i=1}^n\frac{-\lambda_\beta^T X_iX_{ij}}{1+\lambda_\beta^T U_i(\beta)} + p'_\tau(|\beta_j|)\text{sign}(\beta_j) =: I_j + II_j,$$

Note that $\max_i|\lambda_\beta^T X_i| = o_p(1)$ from Lemma 3, for n large enough we have

$$\max_{j\notin\mathcal{B}}|I_j| \le \max_i|\lambda_\beta^T X_i|\max_{j\notin\mathcal{B}}\left|n^{-1}\sum_{i=1}^n X_{ij}\right| \le \max_i|\lambda_\beta^T X_i|\left|\left|n^{-1}\sum_{i=1}^n X_i\right|\right| = o_p(\sqrt{p/n}).$$

This implies that $P(\max_{j\notin A}|I_j| > \tau/2) \to 0$. Then the conclusion of part 1 follows.

The rest part of the proof is similar to that in proving Theorem 1, sketched as follows. For simplicity in presentation, we assume $\sigma^2 = 1$. Since $\hat{\beta}_2 = 0$ with probability tending to one, we consider the constrained minimization of $\ell(\beta)$ subject to $H_2\beta = 0$. By the Lagrange multiplier method, this leads to minimizing the objective function

$$\tilde{\ell}(\beta,\lambda,\nu) = n^{-1}\sum_{i=1}^n\log\{1+\lambda^T U_i(\beta)\} + \sum_{j=1}^d p_\tau(|\beta_j|) + \nu^T H_2\beta.$$

For $\beta \in D_n$, $||\lambda||$ and $||\nu||$ are $o_p(1)$ by Lemma 3 and $0 = \tilde{Q}_{1n}(\beta,\lambda,\nu) := n^{-1}\sum_{i=1}^n U_i(\beta)/\{1+\lambda^T U_i(\beta)\}$. Define

$$\tilde{Q}_{2n}(\beta,\lambda,\nu) = -n^{-1}\sum_{i=1}^n\frac{X_iX_i^T\lambda}{1+\lambda^T U_i(\beta)} + b(\beta) + H_2^T\nu \text{ and } \tilde{Q}_{3n}(\beta,\lambda,\nu) = H_2\beta$$

where $b(\beta) = \{p'_\tau(|\beta_1|)\text{sign}(\beta_1), p'_\tau(|\beta_2|)\text{sign}(\beta_2), \ldots, p'_\tau(|\beta_d|)\text{sign}(\beta_d), 0, \ldots, 0\}^T$. We expand $\tilde{Q}_{jn}(\beta, \lambda, \nu)$ at $(\beta_0, 0, 0)^T$ as follows:

$$\begin{pmatrix} \tilde{Q}_{1n}(\beta_0, 0, 0) \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -\Sigma & -\Sigma & 0 \\ -\Sigma & 0 & H_2^T \\ 0 & H_2 & 0 \end{pmatrix} \begin{pmatrix} \hat{\lambda} - 0 \\ \hat{\beta} - \beta_0 \\ \hat{\nu} - 0 \end{pmatrix} + R_n \tag{A.11}$$

where $R_n = \sum_{k=1}^{5} R_n^{(k)}$, $R_n^{(1)} = (R_{1n}^{T(1)}, R_{2n}^{T(1)}, 0)^T$, $R_{1n}^{(1)} \in \mathcal{R}^p$ and $R_{1n}^{(2)} \in \mathcal{R}^p$ and the kth component of $R_{jn}^{(1)}$, $j = 1, 2$, is given by

$$R_{jn,k}^{(1)} = \frac{1}{2}(\hat{\eta} - \eta_0)^T \frac{\partial^2 Q_{jn,k}(\eta^*)}{\partial \eta \partial \eta^T}(\hat{\eta} - \eta_0),$$

$\eta = (\beta, \lambda)^T$, $\eta^* = (\beta^*, \lambda^*)$ satisfying $||\beta^* - \beta_0|| \leq ||\hat{\beta} - \beta_0||$ and $||\lambda^*|| \leq ||\hat{\lambda}||$, $R_n^{(2)} = \{0, b^T(\beta_0), 0\}^T$, $R_n^{(3)} = \{0, \{b'(\mu_0)(\hat{\beta} - \beta_0)\}^T, 0\}^T$, $R_n^{(4)} = \{\{(T_n(\beta_0) - \Sigma)\hat{\lambda}\}^T, 0, 0\}^T$ and $R_n^{(5)} = \{\{(T_n(\beta_0) - \Sigma)\hat{\lambda}\}^T, \{(T_n(\beta_0) - \Sigma)(\hat{\beta} - \beta_0)\}^T, 0)\}^T$. For $k = 1, \ldots, 4$, $R_n^{(k)} = o_p(1/\sqrt{n})$ following the arguments in proving Theorem 1. Since $||T_n(\beta_0) - \Sigma|| = O_p(\sqrt{p/n})$, $||\hat{\beta} - \beta_0|| = O_p(a_n)$ and $||\hat{\lambda}_\beta|| = O_p(a_n)$, $||R_n^{(5)}|| = o_p(1/\sqrt{n})$ as well. By (A.11) and matrix inverting by blocks as those in proving Theorem 1, we have

$$\hat{\beta} - \beta_0 = \{\Sigma^{-1} - \Sigma^{-1} H_2^T (H_2 \Sigma^{-1} H_2^T)^{-1} H_2 \Sigma^{-1}\} \left(n^{-1} \sum_{i=1}^{n} X_i \epsilon_i + R_{1n} \right).$$

where R_{1n} is the corresponding component in R_n. The asymptotic covariance of $\sqrt{n}\hat{\beta}_1 = \sqrt{n} H_1 \hat{\beta}$ is given by $I_{\mathcal{B}} = H_1 \Sigma^{-1} H_1^T - H_1 \Sigma^{-1} H_2^T (H_2 \Sigma^{-1} H_2^T)^{-1} H_2 \Sigma^{-1} H_1^T$. By decomposition $\Sigma^{-1} = \begin{pmatrix} \dot{\Sigma}_{11} & \dot{\Sigma}_{12} \\ \dot{\Sigma}_{21} & \dot{\Sigma}_{22} \end{pmatrix}$ where $\dot{\Sigma}_{ij} = H_i \Sigma^{-1} H_j^T$, we have $I_{\mathcal{B}} = \dot{\Sigma}_{11} - \dot{\Sigma}_{12} \dot{\Sigma}_{22}^{-1} \dot{\Sigma}_{21}$. Applying matrix inverse by blocks on Σ, we note that

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_{11}^{-1} + \Sigma_{11}^{-1} \Sigma_{12} F^{-1} \Sigma_{21} \Sigma_{11}^{-1} & -\Sigma_{11}^{-1} \Sigma_{12} F^{-1} \\ -F^{-1} \Sigma_{21} \Sigma_{11}^{-1} & F^{-1} \end{pmatrix}$$

where $F = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$. This implies $I_{\mathcal{B}} = \Sigma_{11}^{-1}$. Since $||\sqrt{n} W_n I_{\mathcal{B}}^{-1/2} R_{1n}|| = o_p(1)$, applying Lindeberg-Feller's CLT, part 2 of Theorem 3 follows.

Proof of Theorem 4

The proof of Theorem 4 follows that of Theorem 2. Under the null hypothesis, there exist $\tilde{H}_2$ such that $\tilde{H}_2 \beta_0 = 0$ and $\tilde{H}_2 \tilde{H}_2^T = I_{p-d+q}$. Define $\bar{U} = n^{-1} \sum_{i=1}^{n} X_i \epsilon_i$. Essentially, we have shown in the proof of Theorem 2 that

$$\tilde{\ell}(L_n) = n \bar{U} \Sigma^{-1/2} \{P_1 - P_2\} \Sigma^{-1/2} \bar{U} + o_p(1)$$

where $P_1 = \Sigma^{-1/2} \tilde{H}_2^T (\tilde{H}_2 \Sigma^{-1} \tilde{H}_2^T)^{-1} \tilde{H}_2 \Sigma^{-1/2}$ and $P_2 = \Sigma^{-1/2} H_2^T (H_2 \Sigma^{-1} H_2^T)^{-1} H_2 \Sigma^{-1/2}$. Then Theorem 4 can be verified by noting that $P_1 - P_2$ is an idempotent matrix with rank q. The rest of the proof follows that of Theorem 2 straightforwardly.

References

Bai, Z. and Saranadasa, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica* 6 311 - 329.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *The Annals of Statistics* 35 2313-2351.

Chen, S. X. (1994). Empirical likelihood confidence intervals for linear regression coefficients. *Journal of Multivariate Analysis* 49 24-40.

Chen, S. X. and Cui, H-J. (2006). On Bartlett correction of empirical likelihood in the presence of nuisance parameters. *Biometrika* 16 1101-1115.

Chen, S. X. and Van Keilegom, I. (2009). A review on empirical likelihood methods for regression (with discussion). *Test* 18 415-447.

Chen, S. X., Peng, L. and Qin, Y-L. (2009). Effects of Data Dimension on Empirical Likelihood. *Biometrika* 96 711-722.

DiCiccio, T. J., Hall, P. and Romano, J.P. (1991). Empirical likelihood is Bartlett-correctable. *The Annals of Statistics* 19 1053-1061.

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

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society Series B* 70 849-911.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* 32 928-961.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning Data Mining, Inference, and Prediction. 2nd Edition.* Springer-New York.

Hjort, N. L., McKeague, I., and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *The Annals of Statistics* 37 1079-1111.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics* 37 3498-3528.

McKeague, I. W. (2009). Comments on: A review on empirical likelihood methods for regression. *Test* 18 461-462.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* 34 1436-1462.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single function. *Biometrika* 75 237-249.

Owen, A. (1991). Empirical likelihood for linear models. *The Annals of Statistics* 19 1725-1747.

Owen, A. (2001). *Empirical Likelihood*. Chapman and Hall-CRC.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* 22 300-325.

Qin, J. and Lawless, J. (1995). Estimating equations, empirical likelihood and constraints on parameter. *The Canadian Journal of Statistics* 23 145-159.

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

Van Der Varrt, A. W. (1998). *Asymptotic Statistics*, Cambridge Univ. Press.

Wang, H., Li, R., and Tsai, C. L (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 94 553-568.

Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of Royal Statistical Society, Series B* 71 671-683.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for Cox's proportional hazard model. *Biometrika* 94 691-703.

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

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *The Annals of Statistics* 36 1509-1533.