Journal of Statistical Planning and Inference 186 (2017) 42–57

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Penalized empirical likelihood for semiparametric models with a diverging number of parameters

Jianglin Fang [a,b], Wanrong Liu [b,*], Xuewen Lu [c]

[a] College of Science, Hunan Institute of Engineering, Xiangtan, Hunan, 411104, China

[b] College of Mathematics and Computer Science, Hunan Normal University, Changsha, 410081, China

[c] Department of Mathematics and Statistics, University of Calgary, Calgary, Alberta, Canada T2N 1N4

ARTICLE INFO

Article history:
Received 16 January 2016
Received in revised form 22 October 2016
Accepted 15 February 2017
Available online 24 February 2017

Keywords:
Empirical likelihood
Penalized empirical likelihood
High-dimensional data
Semiparametric models
Estimating equation

ABSTRACT

We apply empirical likelihood (EL) for high-dimensional semiparametric models and propose penalized empirical likelihood (PEL) method for parameter estimation and variable selection. It is shown that the estimator based on EL has the asymptotic consistent property, and that the limit distribution of the EL ratio statistic for the parameters θ is asymptotic normal distribution. Furthermore, in a high-dimensional setting, we prove that PEL in semiparametric models has the oracle property, that is, with probability tending to 1, the estimator based on PEL for the nonzero coefficients is efficient. Moreover, the PEL ratio statistic for the parameters θ is a χ^2_q distribution under the true null hypothesis. The performance of the proposed method is illustrated via a real data application and numerical simulations.

1. Introduction

The empirical likelihood (EL) method, as a powerful nonparametric method of statistical inference, has sampling properties similar to those of the bootstrap and has many advantages compared with the usual parametric procedures. For example, the confidence region constructed by the EL method can be calibrated by using a nonparametric version of Wilks' Theorem involving a chi-squared limiting distribution, and the shape and orientation of the confidence region are determined by the data. Therefore, since the method of EL was proposed by Owen (1988), this has been extended to complex inference problems in various areas. Owen (1990) applied it to construct confidence regions for multivariate mean; Owen (1991) extended the EL method to linear regression problem; Diciccio et al. (1991) studied the Bartlett adjustment for EL; Qin and Lewless (1994) derived EL inference procedures for general estimating equations, Xue and Zhu (2007a,b) developed this method for varying coefficient models and semiparametric regression analysis with longitudinal data; etc.

High-dimensional data, whose dimension p tends to infinity as the sample size $n \to \infty$, becomes more and more popular in many areas, such as financial and statistical applications, hyperspectral imagery, internet portals, high-throughput genomic data analysis and other areas of computational biology; see, e.g., Bai and Aranadasa (1996), Ledoit and Wolf (2002) and Hjort et al. (2009). The method of EL has been applied to some high-dimensional problems and its asymptotic behavior under the setting where n and p both tend to infinity has also been carefully studied. Hjort et al. (2009) derived the limit distribution of the EL ratio statistic based on p-dimensional estimating equations when $p \to \infty$ with n at the rate $p = o\left(n^{1/3}\right)$; Chen et al. (2009) improved upon the rate restriction in Hjort et al. (2009) and established a nondegenerate limit distribution of the EL ratio statistic, allowing $p = o\left(n^{1/2}\right)$ under suitable regularity conditions; Tang and Leng (2010)

* Corresponding author.
E-mail addresses: fangjlin@foxmail.com (J. Fang), liuwr08@aliyun.com (W. Liu).

http://dx.doi.org/10.1016/j.jspi.2017.02.006
0378-3758/

evaluated a penalized EL when the underlying parameter is sparse in the context of the mean parameters and linear models; Lahiri and Mukhopadhyay (2012) proposed a modified EL method by adding a penalty term to the original EL criterion for estimating the high-dimensional mean parameters with $p > n$.

In statistical analysis, there are many situations in which researchers do not have enough knowledge to construct a parametric likelihood function, but misspecification of the distributional or functional form may give us biased estimates for the parametric model. On the one hand, to avoid the misspecification bias for the distribution form, it is common to use estimating equations, which allow us to estimate unknown parameters without using parametric likelihood functions; on the other hand, to avoid the misspecification bias for the functional form, it is common to use semiparametric models or nonparametric models. Semiparametric models are a useful compromise between parametric and nonparametric models to mitigate the curse of dimensionality but still allow reasonable flexibility to specify functional form; see, e.g., Engle et al. (1986), Ichimura (1993), Härdle et al. (2000) and Du et al. (2010). Since EL was proposed, this has been extended to statistic inference for semiparametric models, for example, Shi and Lau (2000) studied partial linear models by using the method of EL; Xue and Zhu (2006) showed properties of EL for single-index models; Bertail (2006) gave the properties of EL for some semiparametric models; etc. In high-dimensional setting, Li et al. (2012) considered EL for a varying coefficient partially linear model with diverging number of parameters; Wang et al. (2013) studied EL inference for semiparametric estimating equations; Tang et al. (2013) investigated EL for partially linear proportional hazards models with growing dimensions.

An interesting problem is variable selection in statistical analysis. Traditional procedures such as Akaike information criterion (AIC) and Bayesian information criterion (BIC) for variable selection were studied by Breiman (1996). To overcome the inefficiency of traditional variable selection procedures, Fan and Li (2001) proposed approach by nonconcave penalized likelihood to select variables and estimate the coefficients of variables. Variable selection procedures in high-dimensional setting, where the dimension p of the observations increases with the sample size n, are given by Fan and Peng (2004), Tang and Leng (2010) and Leng and Tang (2012), etc. Fan and Peng (2004) extended nonconcave penalized likelihood to diverging number of parameters setting; Tang and Leng (2010), developed PEL for high-dimensional linear models, Leng and Tang (2012) proposed PEL for growing dimensional general estimating equations.

Motivated by the EL method for high-dimensional data in Hjort et al. (2009), we aim to extend EL method for semiparametric models that enables us to deal with the case of the data dimension p depending on the sample size n and growing to infinity as $n \to \infty$. The main contributions of this work are:

1. As for high-dimensional semiparametric models, we show that empirical likelihood gives consistent estimators for parameter θ. Furthermore, we prove that the limit distribution of the EL ratio statistic is asymptotic normal distribution, which extends the results in Leng and Tang (2012) to semiparametric models, and also extends the results in Wang et al. (2013) to high-dimensional setting, i.e., as $n \to \infty$, $k \to \infty$ and $p \to \infty$, where k and p are the dimension of nonparametric component and parametric component of semiparametric models, respectively.
2. We present variable selection, parameter estimation and inference for semiparametric models by PEL, as $n \to \infty, k \to \infty$ and $p \to \infty$. For high-dimensional sparse semiparametric models, this method can select the none-zero coefficients with probability converging to 1. Furthermore, PEL ratio statistic shows the well-known Wilks' phenomenon, facilitating hypothesis testing and constructing confidence regions.

The remainder of this paper is organized as follows. In Section 2, we prove the asymptotic properties of EL estimators for high-dimensional semiparametric models under some regular conditions, and establish the limit distribution of the EL ratio statistic. In Section 3, we show variable selection, parameter estimation and hypothesis testing for high-dimensional semiparametric models by PEL. Computational algorithm and simulation results are reported in Section 4, and one real data example is presented in Section 5. Finally, the technical proofs of main results are stated in the Appendix.

2. Empirical likelihood for high-dimensional semiparametric models

We consider semiparametric models by using the estimating equations which contain unknown functions framework as the following:

$$E\{g(\mathbf{X}, H(T), \theta)\} = 0, \tag{1}$$

where $\mathbf{X} = (X^T, Y, Z^T)^T$ is a random vector, T is an associated variable with a bounded support, $H(T) \in \mathbb{R}^k$ is an unknown smooth function with

$$H(t) = (H_1(t), H_2(t), \ldots, H_k(t))^T = E(\varphi(\cdot) \mid T = t),$$

$\varphi(\cdot) = (\varphi_1(\cdot), \varphi_2(\cdot), \ldots, \varphi_k(\cdot))^T$ are known measurable functions, g is a r dimensional vector of known functions, $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^T \in \Theta_\theta$ is a vector of unknown parameters and $\Theta_\theta \in \mathbb{R}^p$. The model (1) includes various existing semiparametric specifications, such as the partially linear model: $g(\mathbf{X}, H(T), \theta) = Y - X^T\theta - H(T)$; the varying-coefficient partially linear model: $g(\mathbf{X}, H(T), \theta) = Y - X^T\theta - H(T)Z$; the single index regression model: $g(\mathbf{X}, H(T), \theta) = Y - H(X^T\theta)$. In addition, the linear errors-in-variable model, the partially linear model with the covariables missing at random, the partially linear errors-in-variables model and the semiparametric varying-coefficient linear model all have the corresponding semiparametric estimating equations as defined in (1).

Assume that $\{(\mathbf{X}_i^T, T_i^T)^T\}_{i=1}^n$ are independent and identically distributed random samples from the unknown distribution of $F \in \mathcal{F}$, where $\mathcal{F} = \mathcal{F}_x \times \mathcal{F}_t \subset \mathbb{R}^{p+1+k_1} \times [a, b]$, where k_1 denotes a positive integer, and a and b denote generic constants. We have the semiparametric estimating equation framework as follows:

$$E\{g(\mathbf{X}, H(T), \theta)\} = 0 (r \geq p),$$

where

$$g(\mathbf{X}, H(T), \theta) = (g_1(\mathbf{X}, H(T), \theta), g_2(\mathbf{X}, H(T), \theta), \ldots, g_r(\mathbf{X}, H(T), \theta))^T.$$

When k and p are fixed, Wang et al. (2013) showed that EL optimally combines information, and the limit distribution of the EL ratio statistic is the χ_p^2 distribution under null hypothesis.

In this paper, we first extend the fixed-dimensional results in Wang et al. (2013) to cases with diverging dimensionality, i.e., $k, p, r \to \infty$ as $n \to \infty$. The traditional EL function for θ is defined as follows:

$$L(\theta) = \sup \left\{ \prod_{i=1}^n (nq_i) : \sum_{i=1}^n q_i = 1, q_i \geq 0, \sum_{i=1}^n q_i g(\mathbf{X}_i, H(T_i), \theta) = 0 \right\}. \tag{2}$$

Because (2) contains unknown functions $H(t)$, it cannot be used directly to make inference on θ. To solve this problem, a natural method is to replace $H(t)$ by its estimator $\hat{H}(t)$. We take its kernel estimator $\hat{H}(t) = \sum_{j=1}^n \mathcal{W}_{nj}(t)\varphi(X_j)$ with $\mathcal{W}_{nj}(t) = K((t - T_j)/h) / \sum_{j=1}^n K((t - T_j)/h)$, where $K(\cdot)$ is a kernel function with bandwidth $h \to 0$. Define an estimated EL function for θ as

$$\tilde{L}(\theta) = \sup \left\{ \prod_{i=1}^n (nq_i) : \sum_{i=1}^n q_i = 1, q_i \geq 0, \sum_{i=1}^n q_i g(\mathbf{X}_i, \hat{H}(T_i), \theta) = 0 \right\}. \tag{3}$$

The estimated EL ratio is

$$\tilde{l}(\theta) = -2[\log\{\tilde{L}(\theta)\} - n \log(n)]. \tag{4}$$

By using the Lagrange multiplier method, $\{q_i\}_{i=1}^n$ in (3) are

$$q_i = \frac{1}{n} \frac{1}{1 + \lambda^T g(\mathbf{X}_i, \hat{H}(T_i), \theta)},$$

with the restriction of λ that

$$\frac{1}{n} \sum_{i=1}^n \frac{g(\mathbf{X}_i, \hat{H}(T_i), \theta)}{1 + \lambda^T g(\mathbf{X}_i, \hat{H}(T_i), \theta)} = 0. \tag{5}$$

Therefore, the estimated EL ratio function for θ defined in (4) is given by

$$\tilde{l}(\theta) = -2[\log\{\tilde{L}(\theta)\} - n \log(n)] = 2 \sum_{i=1}^n \log\{1 + \lambda^T g(\mathbf{X}_i, \hat{H}(T_i), \theta)\}. \tag{6}$$

Because maximizing (3) is equivalent to minimizing (6), we can minimize $\tilde{l}(\theta)$ in (6) to obtain an estimator $\hat{\theta}$ of the parameter θ, which implies that $\hat{\theta}$ can be obtained as

$$\hat{\theta} = \arg \min_{\theta \in \Theta_\theta} \max_{\lambda \in \hat{\Lambda}_n(\theta)} 2 \sum_{i=1}^n \log\{1 + \lambda^T g(\mathbf{X}_i, \hat{H}(T_i), \theta)\},$$

where $\hat{\Lambda}_n(\theta) = \{\lambda \in \mathbb{R}^r | \lambda^T g(\mathbf{X}_i \hat{H}(T_i), \theta) > -1, i = 1, 2, \ldots, n\}$ for any $\theta \in \Theta_\theta$. In method of EL for high-dimensional semiparametric models, the magnitude of $\|\lambda\|$ is no longer $O_p(n^{-1/2})$, as in the fixed dimensional case (Wang et al., 2013). To develop the asymptotic distribution of $\tilde{l}(\theta)$ in (6), we need make the following assumptions.

Assumption 1. $\{H_j(t)\}_{j=1}^k$ satisfy Lipschitz condition of order 1;

Assumption 2. $K(t)$ is a symmetric density function defined in a bounded and compact support with the bandwidth $h = O_p(n^{-1/3})$;

Assumption 3. The density of T, say $r(t)$, exists and satisfies

$$0 < \inf_{a \leq t \leq b} r(t) \leq \sup_{a \leq t \leq b} r(t) < \infty;$$

Assumption 4. $\sup_t E\{\|\varphi_l(X)\|^2|T=t\} < \infty$, for $l = 1, 2, \ldots, k$;

Assumption 5. There exist $u(\mathbf{X}, T)$ and C_1 such that

$$\frac{\partial^2 g(\mathbf{X}, H(T), \theta)}{\partial H_j \partial H_l} \le u(X, T), E\{u^2(\mathbf{X}, T)\} \le C_1 < \infty \quad (j, l = 1, 2, \ldots, k),$$

and

$$\sup_t E\left\{\left\|\frac{\partial g(\mathbf{X}, H(T), \theta)}{\partial H_l}\right\|^2 \middle| T = t\right\} \le C_1 < \infty \quad \text{for } l = 1, 2, \ldots, k;$$

Assumption 6. There exist $v_1(\mathbf{X}, T)$ and C_2 such that

$$\frac{\partial g(\mathbf{X}, H(T), \theta)}{\partial \theta_l} < v_1(X, T), E\{v_1^2(\mathbf{X}, T)\} \le C_2 < \infty \quad (l = 1, 2, \ldots, p),$$

and $v_2(\mathbf{X}, T)$ exists such that

$$\frac{\partial^2 g(\mathbf{X}, H(T), \theta)}{\partial \theta_i \partial \theta_j} \le v_2(X, T), E\{v_2^2(\mathbf{X}, T)\} \le C_2 < \infty \quad (i, j = 1, 2, \ldots, p);$$

Assumption 7. The support Θ_θ of θ is a compact set in $\mathbb{R}^p$, and $\theta_0 \in \Theta_\theta$ is the unique solution to $E\{g(\mathbf{X}, H(T), \theta)\} = 0$;

Assumption 8. Assume $E\{\sup_{\theta\in\Theta_\theta} |g(\mathbf{X}, H(T), \theta)_t|\} = o_p(n^{1/\alpha})$ for some $\alpha \ge 12$ with a large n, where $g(\mathbf{X}, H(T), \theta)_t$ is the tth component of $g(\mathbf{X}, H(T), \theta)$, and $t = 1, 2, \ldots, r$;

Assumption 9. Let $\Sigma(\theta) = E\{g(\mathbf{X}, H(T), \theta)g(\mathbf{X}, H(T), \theta)^T\}$. The eigenvalues of $\Sigma(\theta)$ are bounded away from zero and infinity;

Assumption 10. $k \to \infty, p \to \infty, pn^{-(1/6)} \to 0, kn^{-(1/6)} \to 0$ and $p/r \to c_0$ $(0 < c_0 < 1)$ as $n \to \infty$.

Assumptions 1–5 ensure that the function $H(t)$ is estimated with retain precision and the nonparametric estimation does not affect the asymptotic result of the estimated EL ratio, i.e., the estimated EL ratio $\tilde{l}(\theta)$ has the same asymptotic distribution as the ordinary EL ratio. Assumptions 7–8 ensure the existence and consistency of the minimizer of (6) and control the tail probability behavior of the estimating equation, and Assumption 9 ensures that there exists an asymptotic variance for the EL estimator of the high-dimensional parameters θ. Assumptions 6 and 10, similar to those contained (Leng and Tang, 2012), guarantee the asymptotic normal distribution property of the EL ratio $\tilde{l}(\theta)$. Because no particular structural information is available on $g(\mathbf{X}, H(T), \theta)$, establishing the asymptotic theoretical results for EL approach is very challenging. Therefore, so strong Assumptions 6–10 are needed and the bounds in the stochastic analysis are conservative. This is also the case in Leng and Tang (2012), which studied the EL method for growing dimensional general estimating equations. When specific model structure is available, the restriction on the sample dimension p can be relaxed.

We now present asymptotic normality of the high-dimensional EL estimator for θ and its rate of convergence.

Theorem 1. *Under Assumptions 1–10, as $n \to \infty$, the minimizer $\hat{\theta}$ (EL estimator) of (6) satisfies the following results:*

(1) *(Asymptotic consistency) with probability tending to 1, $\|\hat{\theta} - \theta_0\| = O_p\{(\frac{p}{n})^{\frac{1}{2}}\}$;*

(2) *(Asymptotic normality) as $n \to \infty$, $\sqrt{n}B_n V^{-1/2}(\hat{\theta} - \theta_0) \xrightarrow{L} N(0, G)$, where $\xrightarrow{L}$ represents the convergence in distribution, $B_n \in \mathbb{R}^{q\times p}$ such that $B_n B_n^T \to G$ and G is a $q \times q$ matrix with fixed q, and V is given by*

$$V = \left[E\left\{\frac{\partial g(\mathbf{X}, H(T), \theta)}{\partial \theta}\right\}^T E\{g(\mathbf{X}, H(T), \theta)g(\mathbf{X}, H(T), \theta)^T\}^{-1} E\left\{\frac{\partial g(\mathbf{X}, H(T), \theta)}{\partial \theta}\right\}\right]^{-1}.$$

In Theorem 1, B_n represents a projection of the diverging dimensional vector to a fixed dimension q, and the limiting distribution of the projected vector of $(\hat{\theta} - \theta_0)$ can be described by a multivariate normal distribution. This theorem not only provides the consistency and normality of projected vector of the EL estimator $\hat{\theta}$ for semiparametric models, but also states the effective property of EL in semiparametric models.

Next, we show the asymptotic properties of EL ratio for high-dimensional semiparametric models.

Theorem 2. *Let θ_0 be the true value of parameter vector θ, under assumptions in Theorem 1 and $\alpha \geq 36$ in Assumption 8, $\tilde{l}(\theta_0)$ has an asymptotic standard normal distribution, i.e.,*

$$(2p)^{-\frac{1}{2}}(\tilde{l}(\theta_0) - p) \xrightarrow{L} N(0, 1). \tag{7}$$

On the one hand, Theorem 2 can be used to test the hypothesis

$$H_0 : \theta = \theta_0 \quad \text{Vs} \quad H_1 : \theta \neq \theta_0.$$

On the other hand, it can also be used to construct confidence regions for θ. Let

$$I_\alpha(\theta) = \{\theta : \tilde{l}(\theta) \leq p + z_{\alpha/2}\sqrt{2p}\},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$-quantile of the standard normal distribution. Therefore, by Theorem 2, $I_\alpha(\theta)$ gives an approximate confidence region for θ with asymptotically correct coverage probability $1 - \alpha$, i.e.,

$$P(\theta \in I_\alpha(\theta)) = 1 - \alpha + o_p(1).$$

3. Penalized empirical likelihood for high-dimensional semiparametric models

In high dimensional data analysis, when the dimension of parameters is large, i.e., $p \to \infty$, a reasonable assumption is that only a subset of the parameters is nonzero. In order to select the subset of influential covariates and to carry out the statistical inference on θ under the sparsity assumption, we add a penalty function term in (3) and the penalized EL ratio is defined as

$$\breve{l}(\theta) = \sum_{i=1}^{n} \log\{1 + \lambda^T g(\mathbf{X}_i, \hat{H}(T_i), \theta)\} + n\sum_{j=1}^{p} p_\nu(|\theta_j|), \tag{8}$$

where $p_\nu(|\theta_j|)$ is some penalty function with a tuning parameter ν.

There are many commonly used penalties in the literature. For example, L_2 penalty in Hoerl and Kennard (1970); L_1 penalty in Donoho and Johnstone (1994a); LASSO penalty in Tibshirani (1997); Smoothly Clipped Absolute Deviation Penalty (SCAD) in Fan and Li (2001); adaptive LASSO in Zhang and Lu (2007). In this section, we study PEL for semiparametric models by using SCAD penalty in Fan and Li (2001), whose first derivative satisfies

$$p'_\nu(t) = \nu\{I(t \leq \nu) + \frac{(a\nu - t)_+}{(a-1)\nu} I(t > \nu)\},$$

where $I(\cdot)$ is the indicator function and $a > 2$. Similar to Fan and Li (2001), we set $a = 3.7$ in this paper.

Write $\theta_0 = (\theta_{01}, \theta_{02}, \ldots, \theta_{0p})^T \in \mathbb{R}^p$ and define $\mathcal{A} = \{j : \theta_{0j} \neq 0\}$ with its cardinality $s = |\mathcal{A}|$ which is unknown. Without loss of generality, we let $\theta = (\theta^{(1)^T}, \theta^{(2)^T})^T$, where $\theta^{(1)\in\mathbb{R}^s}$ and $\theta^{(2)\in\mathbb{R}^{p-s}}$ correspond to the nonzero and zero components respectively such as $\theta_0 = (\theta_0^{(1)^T}, 0^T)^T$. For the penalty function $p_\nu(\cdot)$, we make the following assumptions.

Assumption 11. As $n \to \infty$, $\nu(p/n)^{\frac{1}{2}} \to \infty$, and $\min_{j\in\mathcal{A}} \theta_{0j}/\nu \to \infty$;

Assumption 12. Assume $\max_{j\in\mathcal{A}} P'_\nu(|\theta_{0j}|) = o\{(np)^{-1/2}\}$ and $\max_{j\in\mathcal{A}} P''_\nu(|\theta_{0j}|) = o\{(p)^{-1/2}\}$.

Assumption 11 states that the weakest signal should dominate the penalty parameter, and Assumption 12 can be used to control the impact of the penalty on the nonzero component. They hold for many penalty functions such as that in Fan and Li (2001). For notational purposes, we define $\mathcal{I}_p = (D_1^T, D_2^T)$, where $\mathcal{I}_p$ is the p-dimensional identity matrix, $D_1 \in \mathbb{R}^{s\times p}$ and $D_2 \in \mathbb{R}^{(p-s)\times p}$, and use the Frobenius norm of a matrix D, defined as $\|D\| = \{tr(D^T D)\}^{\frac{1}{2}}$. We can minimize $\breve{l}(\theta)$ in (8) to obtain a PEL estimator $\breve{\theta}$ of the parameter θ, and the PEL is defined as

$$\breve{\theta} = \arg\min_{\theta\in\Theta_\theta} \max_{\lambda\in\hat{\Lambda}_n(\theta)} \sum_{i=1}^{n} \log\{1 + \lambda^T g(\mathbf{X}_i, \hat{H}(T_i), \theta)\} + n\sum_{j=1}^{p} p_\nu(|\theta_j|).$$

Write the PEL estimator $\breve{\theta} = (\breve{\theta}^{(1)^T}, \breve{\theta}^{(2)^T})^T$, and the following theorem describes the basic properties of the PEL estimator.

Theorem 3. *Under assumptions in Theorem 1 and Assumptions 11–12, as $n \to \infty$, the PEL estimator $\breve{\theta}$ has the following results:*

(1) *(Selection consistency)* $\lim_{n\to\infty} p(\breve{\theta}^{(2)} = 0) = 1$;

(2) *(Asymptotic normality)* $\sqrt{n}B_n V_p^{-1/2}(\breve{\theta}^{(1)} - \theta_0^{(1)}) \xrightarrow{L} N(0, G)$, *where* $V_p = D_1 V - D_1 V D_2^T (D_2 V D_2^T)^{-1} D_2 V$ *and V is given by Theorem 1, and $B_n \in \mathbb{R}^{q\times p}$ such that $B_n B_n^T \to G$ for $G \in \mathbb{R}^{q\times q}$ with fixed q.*

Theorem 3 not only says that the PEL estimator $\breve{\theta}$ is consistent and converging to the true parameter θ_0, but also states the oracle property of PEL in semiparametric models. Furthermore, Theorem 3(2) shows that the PEL estimator $\breve{\theta}^{(1)}$ of nonzero parametric components $\theta^{(1)}$ is $\sqrt{n}$-consistent and asymptotically normal.

Next, we consider the following hypothesis testing:

$$H_0 : L_n\theta_0 = 0 \quad \text{vs} \quad H_1\, L_n\theta_0 \neq 0,$$

where $L_n \in \mathbb{R}^{q\times p}$ such that $L_nL_n^T = I_q$ for a fixed q, and I_q is the q-dimensional identity matrix. Based on (8), a PEL ratio test statistic is constructed as

$$\breve{l}(L_n) = -2\{\breve{l}(\breve{\theta}) - \min_{\theta, L_n\theta_0=0} \breve{l}(\theta)\}. \tag{9}$$

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

Theorem 4. *Under the null hypothesis and assumptions in Theorem 3, As* $n \to \infty$, $\breve{l}(L_n) \xrightarrow{L} \chi_q^2$.

Theorem 4 extends the results in Leng and Tang (2012) to high-dimensional semiparametric models. By using Theorem 4, we can construct asymptotically confidence region for $L_n\theta_0$, that is,

$$\mathcal{I}_\alpha(L_n\theta) = \{\theta : -2\{\breve{l}(\breve{\theta}) - \min_{\theta, L_n\theta=0} \breve{l}(\theta)\} \le \chi^2_{q,(1-\alpha)}\}, \tag{10}$$

where $\chi^2_{q,(1-\alpha)}$ is the $(1-\alpha)$ quantile of χ_q^2 distribution. $\mathcal{I}_\alpha(L_n\theta)$ gives an approximate confidence region for $L_n\theta$ with asymptotically correct coverage probability $(1-\alpha)$, i.e., as $n \to \infty$, $P(L_n\theta \in \mathcal{I}_\alpha(L_n\theta)) \to 1-\alpha$.

Remark 1. In this paper, we investigate EL for high-dimensional semiparametric models and propose PEL method for parameter estimation and variable selection, and the framework presented here is applicable only where the sample size is larger than the dimension of the parameter. When that is violated, preliminary methods such as sure independence screening (SIS) in Fan and Lv (2008) can be used to reduce the dimensionality from high to a moderate scale that is below the sample size.

4. Computational algorithm and simulation

Firstly, we describe how to approach the optimization problems posed by EL. Due to the nonconvexity, computing EL is nontrivial. Furthermore, PEL computation involving a nondifferentiable penalty is obviously more difficult. We use iterated least squares algorithm of Owen (2001) to obtain the minimizer of EL ratio defined by (6), and use the local quadratic approximation (LQA) algorithm of Fan and Li (2001) to obtain the minimizer of PEL ratio defined by (8). For EL estimator and PEL estimator, we use nested algorithm of Owen (2001) to obtain the minimizer of (6) and (8) through nonlinear optimization, respectively. For the minimizer of (8), assume that θ^0 is an initial value that is close to the minimizer of (8), $\theta_j^{(l)}$ is the lth step estimator of θ_j, if $\theta_j^{(l)}$ is very close to 0, we set $\widehat{\theta}_j^{(l)} = 0$. When $\theta_j^{(l)} \neq 0$, $p_\nu(|\theta_j|)$ can be locally approximated by $p_\nu(|\theta_j^{(l)}|) + \frac{1}{2}\{p'_\nu(|\theta_j^{(l)}|)/|\theta_j^{(l)}|\}\{\theta_j^2 - (\theta_j^{(l)})^2\}$. The procedure is repeated until convergence. Simulation in the next section suggests that this algorithm usually converges given a good initial value.

To choose the penalty parameter ν, we use the following Bayesian information criterion (BIC) motivated by Variyath et al. (2010),

$$BIC(\nu) = -2\breve{l}(\theta_\nu) + df_\nu \log(n),$$

where df_ν is the number of nonzero coefficients.

Next, we conduct simulation studies to illustrate the properties of the EL and PEL inferences for the high-dimensional semiparametric models by using estimating equation framework, and illustrate the usefulness of the EL and PEL by several examples of semiparametric models as follows.

Example 1. We consider the varying coefficient partially linear model:

$$Y_i = X_i^T\theta + Z_i^T u(T_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{11}$$

where Y_i is a response variable, $\mathbf{X}_i = (X_i, Z_i, T_i)$ is the associated covariate, $u(\cdot) = (u_1(\cdot), u_2(\cdot), \ldots, u_{k_1}(\cdot))^T$ is a k_1-dimensional vector of unknown smoothing regression functions, $\theta = (\theta_1, \theta_2, \ldots, \theta_p)^T$ is a p-dimensional vector of unknown regression coefficients and ε_i is an independent random error with mean zero and finite variance, $E(\epsilon_i|X_i, Z_i, T_i) = 0$ almost surely. The estimating function in (1) can be taken by

$$g(\mathbf{X}, H(T), \theta) = (X - (M_3^{-1}M_1^T)^T Z)\{(Y - M_2M_3^{-1}Z) - (X - (M_3^{-1}M_1^T)^T Z)^T\theta\},$$

where $\mathbf{X} = (X, Y, Z)$, $H(t) = (E(XZ^T|T = t), E(YZ^T|T = t), E(ZZ^T|T = t))$, $M_1 = E(XZ^T|T)$, $M_2 = E(YZ^T|T)$ and $M_3 = E(ZZ^T|T)$.

Table 1
Comparison of coverage probability between EL method and GLS method.

p	n	$1-\alpha=0.90$		$1-\alpha=0.95$	
		GLS	EL	GLS	EL
10	200	0.864	0.879	0.921	0.927
	300	0.872	0.884	0.926	0.934
	400	0.883	0.896	0.931	0.945
20	300	0.859	0.868	0.917	0.922
	400	0.867	0.873	0.924	0.931
	500	0.879	0.882	0.928	0.937
30	300	0.842	0.865	0.891	0.913
	400	0.856	0.871	0.895	0.923
	500	0.861	0.889	0.903	0.932

Table 2
Mean (standard deviations) of estimators for parameters of EL and PEL.

p	n	Method	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_5$
10	200	EL	1.09(0.153)	1.92(0.261)	−1.42(0.248)	2.08(0.156)
		Bridge	0.93(0.186)	1.78(0.246)	−1.35(0.203)	1.87(0.195)
		Lasso	0.97(0.174)	1.80(0.248)	−1.47(0.182)	2.06(0.162)
		SCAD	0.99(0.128)	1.82(0.213)	−1.53(0.176)	2.03(0.173)
20	300	EL	1.08(0.210)	1.86(0.234)	−1.38(0.225)	2.11(0.231)
		Bridge	1.12(0.247)	1.72(0.289)	−1.42(0.254)	1.92(0.186)
		Lasso	1.10(0.236)	1.66(0.315)	−1.39(0.247)	1.93(0.193)
		SCAD	0.96(0.179)	1.79(0.232)	−1.43(0.218)	2.02(0.074)
30	400	EL	0.89(0.234)	1.83(0.257)	−1.33(0.273)	2.13(0.241)
		Bridge	0.87(0.287)	1.70(0.248)	−1.26(0.267)	1.75(0.258)
		Lasso	0.83(0.269)	1.74(0.251)	−1.31(0.271)	1.81(0.212)
		SCAD	0.86(0.245)	1.95(0.089)	−1.31(0.263)	1.87(0.187)

Assumptions 6–10 are quite strong and conservative in Section 2. Similar to Li et al. (2012), there are weaker versions of these assumptions for example 1 as follows:

Assumption 6′. $E(\varepsilon|X,Z,T)=0$ almost surely. For some integer $d\geq 4$, $E(\|Z\varepsilon\|^d)<\infty$, $E(\|Z\|^d)<\infty$, $E(\|\varepsilon\|^d)<\infty$;

Assumption 7′. Let $\Upsilon(\theta)=\varepsilon(X-M_3^{-1}M_1^TZ)$, and $\Upsilon(\theta)_t$ be the tth component of $\Upsilon(\theta)$, $t=1,\ldots,r$. For some integer $d\geq 4$, there exists positive constant C_0 such that as $n\to\infty$, $E(\|\Upsilon(\theta)/\sqrt{p}\|^d)<C_0$, $E(\|XZ^T/\sqrt{p}\|^d)<C_0$, $E(\|M_3^{-1}M_1^TZZ^T/\sqrt{p}\|^d)<C_0$, and

$$\frac{1}{r}\sum_{t=1}^{r}E(|\Upsilon(\theta)_t|(\|XZ^T/\sqrt{p}\|^4+\|M_3^{-1}M_1^TZZ^T/\sqrt{p}\|^4))<C_0;$$

Assumption 8′. Positive constant C_0 exists such that as $n\to\infty$,

$$\max_{1\leq t_1,t_2,t_3\leq r}E(\Upsilon(\theta)_{t_1}\Upsilon(\theta)_{t_2}\Upsilon(\theta)_{t_3})^2<C_0;$$

Assumption 9′. Let $\Sigma_1=E(\varepsilon(X-M_3^{-1}M_1^TZ)(X-M_3^{-1}M_1^TZ)^T)$. The eigenvalues of Σ_1 are bounded away from zero and infinity;

Assumption 10′. Assume that $E(\varepsilon^3|X,Z,T)=0$ almost surely. Furthermore, for some integer $d\geq 8$, $p\to\infty$, $p^{3+2/(d-2)}/n\to 0$ and $p/r\to c_0$ $(0<c_0<1)$ as $n\to\infty$.

In this simulation, the covariate T_i is uniformly distributed on $[0,1]$, the covariate X_i is a p-dimensional normal distribution random vector with mean zero and covariance matrix (σ_{ij}) with $\sigma_{ij}=0.2^{|i-j|}$, the nonparametric component $u(t)=(u_1(t),u_2(t))^T$ with $k_1=2$ in which $Z_{i1}=1$ and $Z_{i2}\sim N(0,1)$. The model errors ε_i are generated from $N(0,1)$ and the coefficient functions are given as $u_1(t)=\sin\pi t$, and $u_1(t)=2t(1-t)$. Furthermore, we use the Epanechnikov kernel function $K(t)=\frac{3}{4}(1-t^2)_+$, and use the cross-validation method to select the optimal bandwidth h satisfying Assumption 2. Let $\theta=(1,2,-1.5,0,2,0,\ldots,0)^T$ in the model (11). We consider $p=10$, 20 or 30, and $n=200$, 300, 400 or 500, respectively. Such simulation is repeated 1000 times, and the results are summarized in Tables 1–4.

Table 3
Variable selection results for various methods of PEL.

p	n	Method	Average number of zeros coefficients	
			Correct	Incorrect
10	200	Bridge	4.13[69%]	1.86
		Lasso	4.08[68%]	1.91
		SCAD	4.97[83%]	1.12
20	300	Bridge	12.89[81%]	3.24
		Lasso	13.93[87%]	2.08
		SCAD	14.87[93%]	1.35
30	400	Bridge	22.75[88%]	3.37
		Lasso	22.81[88%]	3.28
		SCAD	24.65[95%]	1.13

Table 4
The empirical frequency (%) that a given value of θ_1 does not fall in the 95% confidence interval constructed by (10). The truth is $\theta_1 = 1$.

p	n	0.8	0.9	1	1.1	1.2
10	200	76.1	36.9	6.7	38.2	79.6
	300	92.4	43.6	5.8	41.5	93.1
20	300	78.6	47.1	7.2	48.3	76.4
	400	94.3	54.5	5.6	55.7	93.9
30	400	82.5	41.8	6.1	43.8	83.5
	500	96.2	58.4	5.4	56.4	95.3

Table 1, when the nominal level is 0.90 or 0.95, shows the coverage probability of confidence regions for θ based on the method of EL and the general least-squares (GLS), respectively. The GLS estimator is

$$\hat{\theta}_{GLS} = \left(\sum_{i=1}^{n} \hat{X}_i \hat{X}_i^T\right)^{-1} \left(\sum_{i=1}^{n} \hat{X}_i \hat{Y}_i\right),$$

where $\hat{X}_i = X_i - (M_{3i}^{-1} M_{1i}^T)^T Z_i$, $\hat{Y}_i = Y_i - M_{2i} M_{3i}^{-1} Z_i$, $M_{1i} = \hat{E}(X_i Z_i^T | T_i)$, $M_{2i} = \hat{E}(Y_i Z_i^T | T_i)$ and $M_{3i} = \hat{E}(Z_i Z_i^T | T_i)$. Similar to Lam and Fan (2008), the GLS estimator has the asymptotic normality as

$$\sqrt{n} B_n B \Sigma^{-1/2} (\hat{\theta}_{GLS} - \theta) \xrightarrow{L} N(0, G)$$

with the estimators $\hat{B} = \frac{1}{n} \sum_{i=1}^{n} \hat{X}_i \hat{X}_i^T$ and $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \hat{X}_i \hat{X}_i^T$, where $\varepsilon_i = \hat{Y}_i - X_i^T \hat{\theta}$. Therefore, the confidence region based on the GLS method is

$$\mathcal{I}_\alpha(GLS) = \{\theta : n(\theta - \hat{\theta}_{GLS})^T \hat{\Sigma}^{-1/2} \hat{B}^T B_n^T G^T B_n \hat{B} \hat{\Sigma}^{-1/2} (\hat{\theta}_{GLS} - \theta) \le \chi^2_{p,(1-\alpha)}\}.$$

From Table 1, at each nominal level and each p, the coverage probability for the method of EL increases as the sample size n increases, and the coverage probability appears to be close to the nominal levels especially with moderate sample size. Furthermore, we find that the confidence regions based on EL consistently have better coverage probability than those based on GLS.

The mean and standard deviations of the EL estimators and the PEL estimators of non-zero parameters $(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_5)$ are summarized in Table 2. From Table 2, we can see that all estimators are close to the true values of the parameters, and the standard deviations of estimators of SCAD-PEL are slightly smaller than Bridge-PEL and Lasso-PEL. Table 3 summaries results of variable selection by PEL. It can be found that the PEL with SCAD penalty gives smaller number of average false estimated zeros than Lasso and Bridge, and the average number of zero components is closer to $p - 4$ than Lasso and Bridge. This demonstrates the good performance of the proposed PEL with SCAD penalty approach in variable selection. From Table 4, we can see that the coverage probability for θ_1 based the PEL method increases as the sample size n increases, and the coverage probability appears to be close to the nominal levels especially with moderate sample size. This shows that the proposed test based on PEL has a good power for testing the null hypothesis.

Example 2. Consider the partially linear errors-in-variable model:

$$\begin{cases} Y_i = X_i^T \theta + u(T_i) + \varepsilon_i, \\ Z_i = X_i + \vartheta_i, \quad i = 1, 2, \ldots, n, \end{cases} \tag{12}$$

where Y_i is a response variable, T_i are the associated covariates, $X_i \in R^p$ are the unobserved latent covariates which are measured in an error-prone way, Z_i is the observed surrogate of X_i, θ is an unknown parameter in R^p, $u(\cdot)$ is an unknown function, ε_i and ϑ_i are random error and p-variate measurement error, respectively, satisfying $E\{(\varepsilon_i, \vartheta_i^T)^T\} = 0$ and

Table 5
Comparison of coverage probability between EL method and ANS method.

p	n	$1-\alpha=0.90$		$1-\alpha=0.95$	
		ANS	EL	ANS	EL
20	500	0.848	0.854	0.895	0.903
	700	0.875	0.872	0.904	0.917
	900	0.881	0.885	0.932	0.936
30	800	0.845	0.849	0.871	0.873
	900	0.862	0.867	0.896	0.908
	1000	0.879	0.882	0.918	0.929

Table 6
Mean (standard deviations) of estimators for parameters of EL and SCAD-PEL.

p	n	Method	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_4$	$\hat{\theta}_5$
20	500	EL	2.15(0.272)	−1.67(0.241)	1.34(0.256)	2.06(0.132)
		SCAD	1.76(0.236)	−1.58(0.164)	0.93(0.225)	1.98(0.123)
30	1000	EL	2.18(0.253)	−1.75(0.279)	1.22(0.293)	2.15(0.298)
		SCAD	1.87(0.214)	−1.72(0.262)	1.26(0.281)	1.83(0.274)

Table 7
Variable selection results for various methods of PEL.

p	n	Method	Average number of zeros coefficients	
			Correct	Incorrect
20	500	Bridge	13.72[85%]	2.34
		Lasso	12.41[78%]	3.65
		SCAD	14.16[89%]	1.27
30	800	Bridge	23.24[89%]	2.58
		Lasso	21.69[83%]	4.31
		SCAD	24.73[95%]	1.46

$Cov(\varepsilon_i, \vartheta_i) = \mathrm{diag}(\sigma^2, \Sigma_p)$ with unknown σ^2 and known $\Sigma_p > 0$. In this section, it is assumed that ϑ_i is independent of $\mathbf{X}_i$, where $\mathbf{X}_i = (X_i, Y_i, T_i)$. The estimating function in (1) of partially linear errors-in-variable model can be taken by

$$g(\mathbf{X}, H(T), \theta) = (Z - E(Z|T))\{(Y - E(Y|T)) - (Z - E(Z|T))^T\theta\} + \Sigma_p\theta,$$

where $\mathbf{X} = (X, Y, Z), H(t) = (E(Z|T = t), E(Y|T = t))$. $E\{g(\mathbf{X}, h(T), \theta)\} = 0$ when θ is the true value of parameter.

Assumptions 6–8 in Section 2 can be relaxed for Example 2.

Assumption 6′. There exist $v_1(\mathbf{X}, T)$ and C_2 such that

$$\frac{\partial g(\mathbf{X}, H(T), \theta)}{\partial \theta_l} < v_1(X, T), E\{v_1^2(\mathbf{X}, T)\} \le C_2 < \infty \quad (l = 1, 2, \ldots, p);$$

Assumption 7′. $E(|\varepsilon|^4) < \infty, E(\|Z\|^4) < \infty, E(\|\vartheta\|^4) < \infty$;

Assumption 8′. Let ϑ_{ij} be the jth component of ϑ_i, $i = 1, \ldots, n, j = 1, \ldots, p$. Positive integer d exists such that as $|j_1 - j_2| > d, cov(\vartheta_{ij_1}, \vartheta_{ij_2}) = 0, 1 \le j_1, j_2 \le p$.

In this simulation, kernel function and bandwidth are similar to Example 1. The covariate T_i is uniformly distributed on $[0, 1], X_i$ is a p-dimensional normal distribution random vector with mean zero and covariance matrix (σ_{ij}) with $\sigma_{ij} = 0.5^{|i-j|}$, measurement error ϑ_i is from p-dimensional normal distribution random vector with mean zero and $\{\Sigma_p\}_{ij} = 0.3^{|i-j|}$, and $u(t) = \sin(2\pi t)$. Let $\theta = (2, -1.5, 0, 1, 2, 0, \ldots, 0)^T$ in the model (12), and consider $p = 20$ or 30, and $n = 500, 700, \ldots, 1000$, respectively. Such simulation is repeated 1000 times, and the results of simulation are summarized in Tables 5–8.

In order to compare the confidence region constructed by EL method with the confidence region formed by asymptotic normality of the parameter, similar to Chen et al. (2009), we can give the moment estimator of θ as

$$\hat{\theta}_M = \left[\sum_{i=1}^{n}(Z_i - \hat{E}(Z_i|T_i))(Z_i - \hat{E}(Z_i|T_i))^T - n\Sigma_p\right]^{-1} \sum_{i=1}^{n}(Z_i - \hat{E}(Z_i|T_i))(Y_i - \hat{E}(Y_i|T_i)),$$

Table 8
The empirical frequency (%) that a given value of θ_1 does not fall in the 95% confidence interval constructed by (10). The truth is $\theta_1 = 2$.

p	n	1.8	1.9	2	2.1	2.2
20	500	76.2	43.5	8.4	46.3	79.9
	800	92.5	57.6	6.1	58.9	93.4
30	800	78.9	52.8	9.3	54.7	81.2
	1000	94.7	61.2	5.9	59.5	93.8

Table 9
Affymetrix probe IDs and the SCAD-PEL estimators of 6 nonzero coefficients.

Probe IDs	SCAD-PEL estimators
31536_at	−0.0935
36131_at	−0.0506
37761_at	0.1104
39837_s_at	−0.1191
40718_at	−0.0669
754_s_at	0.0916

and

$$\sqrt{n}^{-1/2}(\theta - \hat{\theta}_M) \xrightarrow{L} N(0, \Sigma_X^{-1}\Omega\Sigma_X^{-1})$$

with the estimators

$$\hat{\Omega} = n^{-1}\sum_{i=1}^{n}\{(Z_i - \hat{E}(Z_i|T_i))((Y_i - \hat{E}(Y_i|T_i)) - (Z_i - \hat{E}(Z_i|T_i))^T\hat{\theta}_M) + \Sigma_p\hat{\theta}_M\}^{\otimes 2}$$

and

$$\hat{\Sigma}_X = \frac{1}{n}\sum_{i=1}^{n}\{(Z_i - \hat{E}(Z_i|T_i))(Z_i - \hat{E}(Z_i|T_i))^T - n\Sigma_p\},$$

where $A^{\otimes 2} = AA^T$. In this case we can give the confidence region formed from asymptotic normality, that is,

$$\mathcal{I}_\alpha(ASN) = \{\theta : \{n(\theta - \hat{\theta}_M)^T\hat{\Sigma}_X\hat{\Omega}^{-1}\hat{\Sigma}_X(\theta - \hat{\theta}_M)\} \le \chi^2_{p,(1-\alpha)}\}.$$

Table 5 reports empirical coverage of two types of confidence regions. We choose nominal levels 90% and 95% respectively. From the results, we can see that the EL confidence region has slightly higher coverage probability than the asymptotic normal confidence region, and the coverage probability tends to the nominal level as the sample size increases. From Table 6, we can see that the standard deviations of the estimators of SCAD-PEL are slightly smaller than others, and the mean of the estimators is more accurate. Results of variable selection by PEL are shown in Table 7. It can be seen that the PEL with the SCAD, Lasso and Bridge penalties effectively reduces model complexity, and that the SCAD can do significantly better than the Lasso and Bridge in terms of estimation accuracy and model complexity. Similar to Table 4, Table 8 also shows that the proposed test based on PEL has a good power for testing the null hypothesis.

5. Real data application

We further illustrate our proposed method by applying the partially linear logistic regression model to cancer classification. The acute lymphoblastic leukemia (ALL) data, available from http://www.bioconductor.org/, contains microarray expressions for 128 patients with either T-cell or B-cell type leukemia. It has been analyzed by Dudoit et al. (2008) and Chen and Qin (2010). We consider a subset of the ALL data representing 79 samples from patients with B-cell acute lymphoblastic leukemia. Of particular interest is the classification of 37 samples with the BCR/ABL and 42 samples with NEG. Because many of the genes represented by the 12,625 probesets on the array are not expressed, we perform preliminary screening for gene-filtering, and retain only those genes for which: (1) at least 75% of the subjects have a measured intensity of at least 100, and (2) the coefficient of variation of the intensities across samples is between 0.7 and 10. The filtered data set contains expression measures on 2396 genes, for 79 patients. We are interested in a general model where age can interact with gene expression levels. This is the partially linear logistic regression model coming in, with age acting as the index variable. In the 79 samples, 3 samples have missing age information and are removed, so the final sample size is 76.

Since our current implementation dealing with $p = 2396$ is computationally infeasible, similar to Fan and Lv (2008), we use a sure independence screening (SIS) as a preprocessing step. By using SIS, we retain only the top 15 probesets that show the largest marginal effect on responses. By using our PEL approach with BIC for tuning parameter selection, we identify 6 probesets with nonzero coefficients. Table 9 shows the results of variable selection by SCAD-PEL and provides Affymetrix probe IDs and the SCAD-PEL estimators of 6 nonzero coefficients.

Acknowledgment

The second author's project was supported by Open Fund of Innovation Platform in Hunan province colleges and universities, 13k030.

Appendix

Throughout the Appendix, C will denote a generic positive constant that may be different in different uses. Let $\gamma_n = O_p((p/n)^{1/2})$, $D_n = \{\theta \mid \|\theta - \theta_0\| \le C\gamma_n\}$, $g(\theta) = g(\mathbf{X}, H(T), \theta)$, $g_i(\theta) = g(\mathbf{X}_i, H(T_i), \theta)$ and $\hat{g}_i(\theta) = g(\mathbf{X}_i, \hat{H}(T_i), \theta)$.

Lemma 1 (*See Wann and Zheng, 1997*)**.** *Under Assumptions 2–3, we have*

(1) $E\{\mathcal{W}_{ni}(T_l)\}^2 \le C(n^2h)$, *for* $i \ne l$;
(2) $E\{\mathcal{W}_{ni}(T_l)\}^4 \le C(n^4h)$, *for* $i \ne l$;
(3) $E\{\mathcal{W}_{ni}(t)\}^2 \le C(n^2h)$, *for* $i = 1, 2, \ldots, n$.

Lemma 2. *If in Assumptions 1–4, then we have*

$$\|\hat{H}_l(T_i) - H_l(T_i)\| = O_p(n^{-1/3}), \quad i = 1, 2, \ldots, n;\ l = 1, 2, \ldots, k.$$

Proof of Lemma 2. Because of $\hat{H}(t) = \sum_{j=1}^n \mathcal{W}_{nj}(T_i)\varphi(X_j)$, for $i = 1, 2, \ldots, n$, we have

$$\begin{aligned}\hat{H}_l(T_i) - H_l(T_i) &= \sum_{j=1}^n \mathcal{W}_{nj}(T_i)\varphi_l(X_j) - H_l(T_i)\\ &= \sum_{j=1}^n \mathcal{W}_{nj}(T_i)\varphi_l(X_j) - H_l(T_j) + H_l(T_j) - H_l(T_i) = R_{1i} + R_{2i},\end{aligned}$$

where $R_{1i} = \sum_{j=1}^n \mathcal{W}_{nj}(T_i)(\varphi_l(X_j) - H_l(T_j))$, $R_{2i} = \sum_{j=1}^n \mathcal{W}_{nj}(T_i)(H_l(T_j) - H_l(T_i))$. By Assumption 4, Lemma 1 and the property of the kernel estimator, we have

$$\begin{aligned}E\|R_{1i}\|^2 &= E\left\|\sum_{j=1}^n \mathcal{W}_{nj}(T_i)(\varphi_l(X_j) - H_l(T_j))\right\|^2\\ &\le C\sum_{j=1}^n \mathcal{W}_{nj}^2(T_i)E\{(\varphi_l(X_j) - E(\varphi_l(X_j)|T_j))^2|T_i, X_j\} = O_p((nh)^{-1}),\end{aligned}$$

and by Assumptions 1, 3 and Lemma 1(1), we have

$$\begin{aligned}E\|R_{2i}\|^2 &= E\left\|\sum_{j=1}^n \mathcal{W}_{nj}(T_i)(H_l(T_j) - H_l(T_i))\right\|^2\\ &\le E\left\{\sum_{j=1}^n \mathcal{W}_{nj}(T_i)|T_j - T_i|\right\}^2 \le h^2\sum_{j=1}^n E\left\{\mathcal{W}_{nj}^2(T_i)\left|\frac{T_j - T_i}{h}\right|^2\right\}\\ &= h^2\sum_{j=1}^n E\left\{\mathcal{W}_{nj}^2(T_i)\left|\frac{T_j - T_i}{h}\right|^2 I\left(\left|\frac{T_j - T_i}{h}\right| \le \rho\right)\right\}\\ &\quad + h^2\sum_{j=1}^n E\left\{\mathcal{W}_{nj}^2(T_i)\left|\frac{T_j - T_i}{h}\right|^2 I\left(\left|\frac{T_j - T_i}{h}\right| > \rho\right)\right\}\\ &\le nh^2\rho^2 E\{\mathcal{W}_{nj}^2(T_i)\} = O_p(n^{-4/3}).\end{aligned}$$

It implies that

$$\|\hat{H}_l(T_i) - H_l(T_i)\|^2 = O_p((nh)^{-1}) + O_p(n^{-4/3}) = O_p(n^{-2/3}).$$

Thus, Lemma 2 hold, and the proof of Lemma 2 is completed. □

Lemma 3. *Suppose* $E\{\frac{\partial g_i(\theta)}{\partial H}|T=t\}=0$, *and under Assumptions 1–5 and 9, we have*

(1) $\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{g}_i(\theta)=\frac{1}{\sqrt{n}}\sum_{i=1}^n g_i(\theta)+o_p(1)$;
(2) $\frac{1}{n}\sum_{i=1}^n \hat{g}_i(\theta)\hat{g}_i^T(\theta)=\frac{1}{n}\sum_{i=1}^n g_i(\theta)g_i^T(\theta)+o_p(1)$.

Proof of Lemma 3. By Taylor's expansion, we have

$$\begin{aligned}\hat{g}_i(\theta)&=g_i(\theta)+\sum_{l=1}^k\frac{\partial g_i(\theta)}{\partial H_l}(\hat{H}_l(T_i)-H_l(T_i))+\xi_i\\&=g_i(\theta)+\sum_{l=1}^k\frac{\partial g_i(\theta)}{\partial H_l}\left\{\sum_{j=1}^n \mathcal{W}_{nj}(T_i)\varphi_l(X_j)-H_l(T_i)\right\}+\xi_i\\&=g_i(\theta)+\sum_{l=1}^k\frac{\partial g_i(\theta)}{\partial H_l}\left\{\sum_{j=1}^n \mathcal{W}_{nj}(T_i)(\varphi_l(X_j)-H_l(T_j))\right\}\\&\quad+\sum_{l=1}^k\frac{\partial g_i(\theta)}{\partial H_l}\left\{\sum_{j=1}^n \mathcal{W}_{nj}(T_i)(H_l(T_j)-H_l(T_i))\right\}+\xi_i,\end{aligned}$$

where $\xi_i=O_p(\|\hat{H}(T_i)-H(T_i)\|^2)$. So,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\hat{g}_i(\theta)=\frac{1}{\sqrt{n}}\sum_{i=1}^n g_i(\theta)+\frac{1}{\sqrt{n}}\sum_{i=1}^n\xi_i+A_{n1}+A_{n2},\tag{13}$$

where

$$\begin{aligned}A_{n1}&=\sum_{l=1}^k A_{n1,l}=\sum_{l=1}^k\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{\partial g_i(\theta)}{\partial H_l}\sum_{j=1}^n\mathcal{W}_{nj}(T_i)(H_l(T_j)-H_l(T_i)),\\A_{n2}&=\sum_{l=1}^k A_{n2,l}=\sum_{l=1}^k\frac{1}{\sqrt{n}}\sum_{i=1}^n\frac{\partial g_i(\theta)}{\partial H_l}\sum_{j=1}^n\mathcal{W}_{nj}(T_i)(\varphi_l(X_j)-H_l(T_j)).\end{aligned}$$

From Lemma 2, we know that, as $n\to\infty$,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\xi_i=\sqrt{n}\times(O_p(n^{-1/3}))^2=O_p(n^{-1/6})=o_p(1).\tag{14}$$

By Assumptions 1–2, Assumption 5 and Lemma 1(1), we have

$$\begin{aligned}E\|A_{n1}\|^2&\le\sum_{l=1}^k E\|A_{n1,l}\|^2=\frac{1}{n}\sum_{l=1}^k E\left\|\sum_{i=1}^n\frac{\partial g_i(\theta)}{\partial H_l}\sum_{j=1}^n\mathcal{W}_{nj}(T_i)(H_l(T_j)-H_l(T_i))\right\|^2\\&=\frac{1}{n}\sum_{l=1}^k\sum_{i=1}^n E\left\{E\left\{\left\|\frac{\partial g_i(\theta)}{\partial H_l}\right\|^2\middle|T_i\right\}\left\{\sum_{j=1}^n\mathcal{W}_{nj}(T_i)(H_l(T_j)-H_l(T_i))\right\}^2\right\}\\&\le Ch^2r\sum_{l=1}^k\sum_{i=1}^n\sum_{j=1}^n E\left\{\mathcal{W}_{nj}^2(T_i)\left|\frac{T_j-T_i}{h}\right|^2\right\}=O_p(krh).\end{aligned}$$

Similar to the proof of Lemma 3 in Wang et al. (2013), we can show that

$$E(A_{n2,l}A_{n2,l}^T)=\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n E\left\{\left(\frac{\partial g_i(\theta)}{\partial H_l}\right)\left(\frac{\partial g_i(\theta)}{\partial H_l}\right)^T\mathcal{W}_{nj}^2(\varphi_l(X_j)-h_l(T_j))^2\right\}=O_p\left(\frac{r^2}{nh}\right).$$

So it means that $E\|A_{n2}\|^2=O_p(\frac{kr^2}{nh})$. By Assumptions 2 and 9, we have $E\|A_{n1}\|^2=o_p(1)$ and $E\|A_{n2}\|^2=o_p(1)$, therefore,

$$A_{n1}=o_p(1);\qquad A_{n2}=o_p(1).\tag{15}$$

By (13)–(15), we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\hat{g}_i(\theta)=\frac{1}{\sqrt{n}}\sum_{i=1}^n g_i(\theta)+o_p(1),$$

and the first part of Lemma 3 is proved. Next, we begin to proof the second part of Lemma 3. According to (13), we have

$$\begin{aligned}\sum_{i=1}^{n}\hat{g}_i(\theta)\hat{g}_i^T(\theta) &= \sum_{i=1}^{n}r_{i1}r_{i1}^T + 2\sum_{i=1}^{n}r_{i1}r_{i2}^T + 2\sum_{i=1}^{n}r_{i1}r_{i3}^T + \sum_{i=1}^{n}r_{i2}r_{i2}^T + 2\sum_{i=1}^{n}r_{i2}r_{i3}^T\\ &\quad + \sum_{i=1}^{n}r_{i3}r_{i3}^T + 2\sum_{i=1}^{n}r_{i1}\xi_i^T + 2\sum_{i=1}^{n}r_{i2}\xi_i^T + 2\sum_{i=1}^{n}r_{i3}\xi_i^T + \sum_{i=1}^{n}\xi_i\xi_i^T\\ &= R_1 + 2R_2 + 2R_3 + R_4 + 2R_5 + R_6 + 2R_7 + 2R_8 + 2R_9 + R_{10},\end{aligned}$$

where $\xi_i = O_p(\|\hat{H}(T_i) - H(T_i)\|^2)$, $r_{i1} = g_i$, $r_{i2} = \sum_{l=1}^{k}\frac{\partial g_i(\theta)}{\partial H_l}\sum_{j=1}^{n}\mathcal{W}_{nj}(T_i)(\varphi_l(X_j) - H_l(T_j))$, $r_{i3} = \sum_{l=1}^{k}\frac{\partial g_i(\theta)}{\partial H_l}\sum_{j=1}^{n}\mathcal{W}_{nj}(T_i)(H_l(T_j) - H_l(T_i))$. Let $R_{2,st}$ be the (s,t) element in R_2, and $r_{ij,s}$ be the sth component of $r_{ij}, j = 1, 2$. By the Cauchy–Schwarz inequality, it implies that $|R_{2,st}| \le (\sum_{i=1}^{n}r_{i1,s}^2)^{1/2}(\sum_{i=1}^{n}r_{i2,t}^2)^{1/2}$. According to Assumption 2, Assumption 5 and the proof of Lemma 2, we know that $r_{i1,s} = o_p(n^{1/\alpha})$, and

$$r_{i2,t}^2 \le \sum_{l=1}^{k}\left\|\frac{\partial g_i(\theta)_t}{\partial H_l}\right\|^2\left\|\sum_{j=1}^{n}\mathcal{W}_{nj}(T_i)(\varphi_l(X_j) - H_l(T_j))\right\|^2 = O_p\left(\frac{k}{nh}\right),$$

where $g_i(\theta)_t$ is the tth component of $g_i(\theta)$, $t = 1, 2\ldots, r$. Therefore,

$$|R_{2,st}| \le n \times o_p(n^{1/\alpha}) \times O_p\left(\sqrt{\frac{k}{nh}}\right) = o_p(n^{5/6}),$$

and $\|\frac{1}{n}R_2\| = \frac{r}{n} \times o_p(n^{5/6}) = o_p(1)$. Similarly, $\|\frac{1}{n}R_i\| = o_p(1)$, for $i = 3, \ldots, 10$. This implies

$$\frac{1}{n}\sum_{i=1}^{n}\hat{g}_i(\theta)\hat{g}_i^T(\theta) = \frac{1}{n}\sum_{i=1}^{n}g_i(\theta)g_i^T(\theta) + o_p(1).$$

Thus, the second part of Lemma 3 holds, and the proof is completed. □

Proof of Theorem 1. By Lemma 3, we can show that the nonparametric estimation does not affect the asymptotic result of the EL ratio, and the proof is similar to that of Theorem 1 and Theorem 2 of Leng and Tang (2012), hence is omitted. □

Lemma 4. *Under Assumptions 1–10, we have, with probability tending to* 1, $\lambda_{\theta_0} = \mathrm{argmax}_{\lambda\in\hat{\Lambda}_n(\theta_0)}\tilde{l}(\lambda,\theta_0)$ *exists,* $\|\lambda_{\theta_0}\| = O_p(\gamma_n)$, *and* $\|\lambda_{\hat{\theta}}\| = O_p(\gamma_n)$.

Proof of Lemma 4. By Lemma 3, the proof is similar to Lemma 2 and Lemma 4 of Leng and Tang (2012), hence is omitted. □

Proof of Theorem 2. Assumption 8 implies that

$$\max_{1\le i\le n}\sup_{\theta\in\Theta_\theta}\|g_i(\theta)\| = o_p(n^{1/\alpha}r^{1/2}), \tag{16}$$

then, by Lemmas 3 and 4, we have

$$\begin{aligned}\max_{1\le i\le n}|\lambda^T\hat{g}_i(\theta_0)| &= \max_{1\le i\le n}|\lambda^T g_i(\theta)| + o_p(1) \le \|\lambda\|\max_{1\le i\le n}\|g_i(\theta)\| + o_p(1)\\ &= o_p(\gamma_n n^{1/\alpha}r^{1/2}) = o_p(1).\end{aligned}$$

We expand the equation of (5),

$$\begin{aligned}0 &= \frac{1}{n}\sum_{i=1}^{n}\hat{g}_i(\theta_0) - \frac{1}{n}\sum_{i=1}^{n}\hat{g}_i(\theta_0)\hat{g}_i^T(\theta_0)\lambda + R_{n1}\\ &= \frac{1}{n}\sum_{i=1}^{n}g_i(\theta_0) - \frac{1}{n}\sum_{i=1}^{n}g_i(\theta_0)g_i^T(\theta_0)\lambda + R_{n2},\end{aligned} \tag{17}$$

where $R_{n1} = \frac{1}{n}\sum_{i=1}^{n}\hat{g}_i(\theta_0)\frac{(\lambda^T\hat{g}_i(\theta_0))^2}{(1+\xi_i)^3}$, $R_{n2} = \frac{1}{n}\sum_{i=1}^{n}g_i(\theta_0)\frac{(\lambda^T g_i(\theta_0))^2}{(1+\xi_i)^3} + o_p(1)$, and $|\xi_i| \le \max_{1\le i\le n}|\lambda^T\hat{g}_i(\theta_0)|$. As $\max_{1\le i\le n}|\lambda^T\hat{g}_i(\theta_0)| = o_p(1)$, $\max_{1\le i\le n}|\xi_i| = o_p(1)$ as well. According to (16), (17) and Lemma 4,

$$\|R_{nl}\| \le \|\lambda\|^2\max_{1\le i\le n}\|g_i(\theta)\|^3 + o_p(1) = o_p(\gamma_n^2 n^{3/\alpha}r^{3/2}) = o_p(1), \quad l = 1, 2. \tag{18}$$

Let $\bar{g}_n = \frac{1}{n}\sum_{i=1}^{n}g_i(\theta_0)$, $S_n = \frac{1}{n}\sum_{i=1}^{n}g_i(\theta_0)g_i^T(\theta_0)$, it follows from (18) that

$$\lambda = S_n^{-1}\bar{g}_n + S_n^{-1}R_{n2}. \tag{19}$$

By Taylor's expansion and Lemma 3, we have

$$\begin{aligned}\tilde{l}(\theta_0) &= 2\sum_{i=1}^{n}\lambda^T g_i(\theta_0) - \sum_{i=1}^{n}(\lambda^T g_i(\theta_0))^2 + \frac{2}{3}\sum_{i=1}^{n}(\lambda^T g_i(\theta_0))^3(1+\eta_i)^{-4} + o_p(1)\\ &= n\bar{g}_n^T S_n^{-1}\bar{g}_n - nR_{n2}^T S_n^{-1}R_{n2} + \frac{2}{3}\sum_{i=1}^{n}\frac{(\lambda^T g_i(\theta_0))^3}{(1+\eta_i)^4} + o_p(1),\end{aligned}$$

where $|\eta_i| \le \max_{1\le i\le n}|\lambda^T g_i(\theta_0)|$. By Assumptions 8–10, as $n\to\infty$, $S_n \to \Sigma(\theta_0)$ in probability, so, by (19), $\|nR_{n2}^T S_n^{-1}R_{n2}\| = o_p(\sqrt{p})$, and

$$\begin{aligned}\left|\frac{2}{3}\sum_{i=1}^{n}\frac{(\lambda^T g_i(\theta_0))^3}{(1+\eta_i)^4}\right| &\le \frac{2}{3}\|\lambda\|^3\sum_{i=1}^{n}\|(g_i(\theta_0))\|^3\{1+o_p(1)\}\\ &= O_p(p^3 n^{3/\alpha-1/2}) = o_p(\sqrt{p}).\end{aligned}$$

Similar to Lemma 5 and Lemma 6 of Chen et al. (2009), we can show that $(2p)^{-\frac{1}{2}}\{n\bar{g}_n^T\Sigma(\theta_0)^{-1}\bar{g}_n - p\} \xrightarrow{L} N(0,1)$, and $n\bar{g}_n^T(\Sigma(\theta_0)^{-1} - S_n^{-1})\bar{g}_n = o_p(\sqrt{p})$. Hence, $(2p)^{-\frac{1}{2}}(\tilde{l}(\theta_0) - p) \xrightarrow{L} N(0,1)$, and the proof of Theorem 2 is completed. □

Proof of Theorem 3. According to Newey and Smith (2004), $\breve{l}(\theta)$ given by (8) has a minimizer in D_n. Considering $\theta \in D_n$, by (16) and Lemma 2, $\max_{1\le i\le n}\|\lambda^T g(\theta)\| = o_p(1)$. By using Lemma 3 and Taylor's expansion, we have

$$\begin{aligned}\frac{1}{n}\frac{\partial\breve{l}(\theta)}{\partial\theta_j} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\lambda^T\partial\hat{g}_i(\theta)/\partial\theta_j}{1+\lambda^T\hat{g}_i(\theta)} + P_\nu'(|\theta_j|)\mathrm{sign}(\theta_j)\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{\lambda^T\partial g_i(\theta)/\partial\theta_j}{1+\lambda^T g_i(\theta)} + o_p(1) + P_\nu'(|\theta_j|)\mathrm{sign}(\theta_j)\\ &= \frac{1}{n}\sum_{i=1}^{n}\lambda^T\left\{\frac{\partial g_i(\theta_0)}{\partial\theta_j} + \frac{\partial^2 g_i(\theta_0)}{\partial\theta_j\partial\theta^T}(\theta-\theta_0)\right\} + o_p(1) + P_\nu'(|\theta_j|)\mathrm{sign}(\theta_j)\\ &= A_1 + A_2 + P_\nu'(|\theta_j|)\mathrm{sign}(\theta_j) + o_p(1).\end{aligned}$$

By Assumption 6, we can show that

$$\begin{aligned}\max_{j\notin\mathcal{A}}(|A_1|) &= \max_{j\notin\mathcal{A}}\frac{1}{n}\left|\sum_{i=1}^{n}\lambda^T\left\{E\left(\frac{\partial g(\theta_0)}{\partial\theta_j}\right) + \left(\frac{\partial g_i(\theta_0)}{\partial\theta_j} - E\frac{\partial g(\theta_0)}{\partial\theta_j}\right)\right\}\right|\\ &\le \max_{j\notin\mathcal{A}}\left|\lambda^T E\left(\frac{\partial g(\theta_0)}{\partial\theta_j}\right)\right| + \frac{1}{n}\left\|\sum_{i=1}^{n}\lambda^T\left(\frac{\partial g_i(\theta_0)}{\partial\theta_j} - E\frac{\partial g(\theta_0)}{\partial\theta_j}\right)\right\| = o_p(1),\end{aligned}$$

and

$$\begin{aligned}\max_{j\notin\mathcal{A}}(|A_2|) &= \max_{j\notin\mathcal{A}}\frac{1}{n}\left|\sum_{i=1}^{n}\lambda^T\left\{\left[E\left(\frac{\partial^2 g(\theta_0)}{\partial\theta_j\partial\theta^T}\right) + \left(\frac{\partial^2 g_i(\theta_0)}{\partial\theta_j\partial\theta^T} - E\frac{\partial^2 g(\theta_0)}{\partial\theta_j\partial\theta^T}\right)\right](\theta-\theta_0)\right\}\right|\\ &= o_p(1).\end{aligned}$$

According to $|\theta_j|_{\{j\notin\mathcal{A}\}} \le \gamma_n$ and Assumption 11, we have that $P_\nu'(|\theta_j|)_{\{j\notin\mathcal{A}\}} = \nu$ and $P_\nu'(|\theta_j|)\mathrm{sign}(\theta_j)_{\{j\notin\mathcal{A}\}} = \nu\mathrm{sign}(\theta_j)_{\{j\notin\mathcal{A}\}}$. Therefore, as $n\to\infty$, the sign of θ_j dominates $\frac{\partial\breve{l}(\theta)}{\partial\theta_j}$ for $j\notin\mathcal{A}$, with probability tending to 1. It implies that $\breve{\theta}^{(2)} = 0$ with probability tending to 1, and the first part of Theorem 3 is proved.

Next we show the second part of Theorem 3. We consider constrained minimization of (8) subject to $D_2\theta = 0$. For $k = 0$, this type of estimator is studied in Leng and Tang (2012). By the Lagrange multiplier method, this is equivalent to minimizing a new objective function

$$\breve{l}(\theta,\lambda,\mu) = \frac{1}{n}\sum_{i=1}^{n}\log(1+\lambda^T\hat{g}_i(\theta)) + \sum_{j=1}^{p}p_\nu(|\theta_j|) + \mu^T D_2\theta, \tag{20}$$

where $\mu\in\mathbb{R}^{(p-s)}$ is another Lagrange multiplier. By Lemma 3, minimizing (20) is equivalent to minimizing the following function

$$\breve{l}(\theta, \lambda, \mu) = \frac{1}{n}\sum_{i=1}^{n}\log(1+\lambda^T g_i(\theta)) + \sum_{j=1}^{p} p_\nu(|\theta_j|) + \mu^T D_2\theta. \tag{21}$$

The rest of the proof of this theorem is similar to that of Theorem 3 in Leng and Tang (2012), hence is omitted. □

Proof of Theorem 4. Firstly, we present the asymptotic expansion of $\breve{l}(\breve{\theta})$ where $\breve{\theta}$ is the minimizer of (8). Let $\hat{y}_i = \hat{\lambda}^T\hat{g}_i(\breve{\theta})$ and $y_i = \hat{\lambda}^T g_i(\breve{\theta})$, for $i = 1, 2, \ldots, n$. It implies that $\max_{1\le i\le n}|y_i| = o_p(1)$ by Lemma 4 and (16). By Taylor's expansion and Lemma 3, we have

$$\begin{aligned}\breve{l}(\breve{\theta}) &= \sum_{i=1}^{n}\hat{y}_i - \sum_{i=1}^{n}\frac{\hat{y}_i^2}{2} + \sum_{i=1}^{n}\frac{\hat{y}_i^3}{3(1+\zeta_i)^4 + o_p(1)}\\ &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\frac{y_i^2}{2} + \sum_{i=1}^{n}\frac{y_i^3}{3(1+\zeta_i)^4 + o_p(1)} + o_p(1),\end{aligned} \tag{22}$$

where $|\zeta_i| \le |y_i|$. Let $Q_{1n}(\theta, \lambda) = n^{-1}\sum_{i=1}^{n}\frac{g_i(\theta)}{1+\lambda^T g_i(\theta)}$ and $\Sigma = \Sigma(\theta_0)$. According to (19)–(21), we can show that an expansion for $\hat{\lambda}$ is given by

$$\hat{\lambda} = \{\Sigma^{-1} + \Sigma^{-1}G(V - VD_2^T(D_2VD_2^T)^{-1}D_2V)G^T\Sigma^{-1}\}(Q_{1n}(\theta_0, 0) + o_p(1)).$$

Let $\bar{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\theta)$, from (22), we can gain the expansion of the PEL ratio as follows

$$2\breve{l}(\breve{\theta}) = n\bar{g}_n(\theta_0)^T D_2^T(D_2\Sigma D_2^T)^{-1}D_2\bar{g}_n(\theta_0) + o_p(1). \tag{23}$$

Under the null hypothesis, since $L_nL_n^T = I_q$, there exists $\widetilde{D}_2$ such that $\widetilde{D}_2\theta = 0$ and $\widetilde{D}_2\widetilde{D}_2^T = I_{p-d+q}$. By Lemma 3, we establish that under the null hypothesis, the estimator of θ can be obtained by minimizing

$$\breve{l}(\theta, \lambda, \mu) = \frac{1}{n}\sum_{i=1}^{n}\log(1+\lambda^T g_i(\theta)) + \sum_{j=1}^{p} p_\nu(|\theta_j|) + \mu^T\tilde{D}_2\theta. \tag{24}$$

Denote the minimizer of (24) by $(\breve{\theta}, \breve{\lambda}, \breve{\mu})$. By the proof of the first part in Theorem 1, we have, with probability tending to 1, $\breve{\theta}^{(2)} = 0$. Therefore, by Assumption 12, with probability tending to 1, $n\{\sum_{j=1}^{p} p_\nu(|\hat{\theta}_{p,j}|) - \sum_{j=1}^{p} p_\nu(|\breve{\theta}_j|)\} = 0$. By replacing D_2 in (22) by $\widetilde{D}_2$ and (23), we establish that

$$2\breve{l}(\breve{\theta})_{L_n\theta=0} = 2\breve{l}(\breve{\theta}) = n\bar{g}_n(\theta_0)^T\tilde{D}_2^T(\tilde{D}_2\Sigma\tilde{D}_2^T)^{-1}\tilde{D}_2\bar{g}_n(\theta_0) + o_p(1). \tag{25}$$

Combining Eqs. (23)–(25), we have

$$\breve{l}(L_n) = n\bar{g}_n(\theta_0)^T\Sigma^{-1/2}(A_1 - A_2)\Sigma^{-1/2}\bar{g}_n(\theta_0) + o_p(1),$$

where

$$P_1 = \Sigma^{-1/2}GV\widetilde{D}_2^T(\widetilde{D}_2V\widetilde{D}_2^T)^{-1}\widetilde{D}_2VG^T\Sigma^{-1/2},$$
$$P_2 = \Sigma^{-1/2}GVD_2^T(D_2VD_2^T)^{-1}D_2VG^T\Sigma^{-1/2}.$$

As $A_1 - A_2$ is an idempotent matrix of rank q, $A_1 - A_2$ can be written as $A_n^TA_n$, A_n is a $q \times p$ matrix such that $A_nA_n^T = I_q$. By Lindeberg and Feller central limit theorem, we have that $\sqrt{n}A_n\Sigma^{-1/2}\bar{g}_n(\theta_0) \xrightarrow{L} N(0, I_q)$. Then $n\bar{g}_n(\theta_0)^T\Sigma^{-1/2}(A_1 - A_2)\Sigma^{-1/2}\bar{g}_n(\theta_0) \xrightarrow{L} \chi_q^2$ and the proof of Theorem 4 is completed. □

References

Bai, Z., Aranadasa, H., 1996. Effect of high dimension: By an example of a two sample problem. Statist. Sinica 6, 311–329.
Bertail, P., 2006. Empirical likelihood in some semiparametric models. Bernoulli 12, 299–331.
Breiman, L., 1996. Heuristics of instability and stabilization in model selection. Ann. Statist. 24, 2350–2383.
Chen, S., Peng, L., Qin, Y., 2009. Effects of data dimension on empirical likelihood. Biometrika 96, 712–722.
Chen, S.X., Qin, Y.L., 2010. A two-sample test for high-dimensional data with applications to gene-set testing. Ann. Statist. 38, 808–835.
Chen, Q.H., Zhong, P.S., Cui, H.J., 2009. Empirical likelihood for mixed-effects error-in-variables model. Acta Math. Appl. Sin. Engl. Ser. 25, 561–578.
Diciccio, T., Hall, P., Romano, J., 1991. Bartlett adjustment for empirical likelihood. Ann. Statist. 19, 1053–1061.
Donoho, D., Johnstone, I., 1994a. Ideal spatial adaptation by wavelet shrinkage. Biometrika 81, 425–455.
Du, P., Ma, S., Liang, H., 2010. Penalized variable selection procedure for cox models with semiparametric relative risk. Ann. Statist. 38, 2092–2117.
Dudoit, S., Keles, S., Vanderlaan, M., 2008. Multiple tests of association with biological annotation metadata. Inst. Math. Stat. Collect. 2, 153–218.
Engle, R., Granger, C., Rice, J., Weiss, A., 1986. Semiparametric estimates of the relation between weather and electricity sales. J. Amer. Statist. Assoc. 81, 310–320.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360.

Fan, J.Q., Lv, J.C., 2008. Sure independence screening for ultrahigh dimensional feature space. J. Roy. Statist. Soc. B 70, 894–911.
Fan, J.Q., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. Ann. Statist. 32, 928–961.
Härdle, W., Liang, H., Gao, J., 2000. Partically Linear Models. Pysica-Verlag, Heidelberg.
Hjort, H., Mckeague, I., Keilegom, I.V., 2009. Extending the scope of empirical likelihood. Ann. Statist. 37, 1079–1111.
Hoerl, A., Kennard, R., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.
Ichimura, H., 1993. Semiparametric least square (sls) and weigted sls estimation of single-index models. J. Econometrics 58, 71–120.
Lahiri, S., Mukhopadhyay, S., 2012. A penalized empirical likelihood method in high dimensions. Ann. Statist. 40, 2511–2540.
Lam, C., Fan, J.Q., 2008. Profile-Kernel likelihood inference with diverging number of parameters. Ann. Statist. 36, 2232–2260.
Ledoit, O., Wolf, M., 2002. Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size. Ann. Statist. 30, 1801-1102.
Leng, C., Tang, C., 2012. Penalized empirical likelihood and growing dimensional general estimating equations. Biometrika 99, 706–716.
Li, G., Lin, L., Zhu, L., 2012. Empirical likelihood for a varying coefficient partially linear model with diverging number of parameters. J. Multivariate Anal. 105, 85–111.
Newey, W., Smith, R., 2004. Higher order properties of gmm and generalized empirical likelihood estimators. Econometrica 72, 219–255.
Owen, A., 1988. Empirical likelihood ratio confidence intervals for a single function. Biometrika 75, 237–249.
Owen, A., 1990. Empirical likelihood ratio confidence regions. Ann. Statist. 18, 90–120.
Owen, A., 1991. Empirical likelihood for linear models. Ann. Statist. 19, 1725–1747.
Owen, A., 2001. Empirical Likelihood. Chapman and Hall, CRC, London.
Qin, J., Lewless, J., 1994. Empirical likelihood and general estimating equations. Ann. Statist. 22, 300–325.
Shi, J., Lau, T.S., 2000. Empirical likelihood for partially linear models. J. Multivariate Anal. 72, 132–148.
Tang, C., Leng, C., 2010. Penalized high-dimensional empirical likelihood. Biometrika 97, 905–920.
Tang, X., Li, B., Lian, H., 2013. Empirical likelihood for partially linear proportional hazards models with growing dimensions. J. Multivariate Anal. 121, 22–32.
Tibshirani, R., 1997. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B 58, 267–288.
Variyath, A., Chen, J., Abrahamc, B., 2010. Empirical likelihood based variable selection. J. Statist. Plann. Inference 140, 971–981.
Wang, S., Cui, H., Li, R., 2013. Empirical likelihood inference for semi-parametric estimating equations. Sci. China(Math.) 06, 1247–1262.
Wann, Q., Zheng, Z., 1997. Asymptotic properties for the semiparametric regression model with randomly sensored data. Sci. China Ser. A 40, 945–957.
Xue, L., Zhu, L., 2006. Empirical likelihood for single-index models. J. Multivariate Anal. 97, 1295–1312.
Xue, L., Zhu, L., 2007a. Empirical likelihood for a varying coefficient model with longitudinal data. J. Amer. Statist. Assoc. 102, 642–654.
Xue, L., Zhu, L., 2007b. Empirical likelihood semiparametric regression analysis for longitudinal data. Biometrika 94, 921–937.
Zhang, H., Lu, W., 2007. Adaptive Lasso for cox proportional hazards model. Biometrika 94, 691–703.