PENALIZED GENERALIZED EMPIRICAL LIKELIHOOD
WITH A DIVERGING NUMBER OF GENERAL
ESTIMATING EQUATIONS FOR CENSORED DATA

BY NIANSHENG TANG, XIAODONG YAN AND XINGQIU ZHAO*

Yunnan University, Shandong University and The Hong Kong Polytechnic University

This article considers simultaneous variable selection and parameter estimation as well as hypothesis testing in censored survival models where a parametric likelihood is not available. For the problem, we utilize certain growing dimensional general estimating equations and propose a penalized generalized empirical likelihood, where the general estimating equations are constructed based on the semiparametric efficiency bound of estimation with given moment conditions. The proposed penalized generalized empirical likelihood estimators enjoy the oracle properties, and the estimator of any fixed dimensional vector of nonzero parameters achieves the semiparametric efficiency bound asymptotically. Furthermore, we show that the penalized generalized empirical likelihood ratio test statistic has an asymptotic central chi-square distribution. The conditions of local and restricted global optimality of weighted penalized generalized empirical likelihood estimators are also discussed. We present a two-layer iterative algorithm for efficient implementation, and investigate its convergence property. The performance of the proposed methods is demonstrated by extensive simulation studies, and a real data example is provided for illustration.

*Supported in part by the Research Grant Council of Hong Kong (PolyU 153012/18P), the National Natural Science Foundation of China (No. 11771366) and The Hong Kong Polytechnic University.

MSC 2010 subject classifications: Primary 62N02, 62N03; secondary 62F12, 62F05

Keywords and phrases: Censored survival data, penalized generalized empirical likelihood, penalized generalized empirical likelihood ratio test, oracle property, semiparametric efficiency
1. Introduction. Semiparametric regression models are widely used for the analysis of censored data. The commonly used models include the Cox proportional hazards model [Cox (1972)], the additive risk model [Lin and Ying (1994)], and the accelerated failure time model [AFT, Kalbfleisch and Prentice (1980)]. Under these models, the likelihood functions have complicated forms and contain both unknown functions and regression parameters. To avoid estimating unknown functions, a partial likelihood approach was developed for the Cox model [Andersen and Gill (1982)], an estimating equation-based method was designed for the additive risk model [Lin and Ying (1994)], and a rank-based estimation method was proposed for AFT model [Tsiatis (1990); Jin, Lin, Wei and Ying (2003)]. When a parametric likelihood is unspecified, the empirical likelihood (EL) approach is widely used for inference. Qin and Lawless (1994) were the first to study the EL and general estimating equations. Owen (2001) and Chen and Van Keilegom (2009) among others provided a comprehensive review about the attractive advantages and extensive applications of the EL. The EL method has been developed for making inference under the AFT model with a completely unknown error distribution. For example, Li and Wang (2003) constructed the synthetic data empirical likelihood, while Zhou and Li (2008) and Fang, Li, Lu and Qin (2013) used the empirical likelihood method to make inference based on the Buckley-James estimator [Buckley and James (1979)]. However, these methods show that the limiting distribution of \(-2 \log(\text{empirical likelihood ratio})\) is a scaled chi-square distribution, where the scale parameter is a function of the unknown asymptotic variance and must be estimated for making inference. This would increase estimation errors. Zhou (2005) studied the empirical likelihood coupled with the rank-based estimating equation under the AFT model and showed that the \(-2 \log(\text{empirical likelihood ratio})\) converges to a standard chi-square distribution. Recently, using certain influence functions in an estimating equation, He, Liang, Shen and Yang (2016) proposed an empirical likelihood with censored data and concluded that the asymptotic distribution of \(-2 \log(\text{empirical likelihood ratio})\) is a standard chi-square distribution. This method is designed for making inference on linear functionals of the distribution function of the survival time, but it could not be applicable to make inference for semiparametric censored regression models.

High-dimensional sparse modeling with censored survival data is of great practical importance. Several regularization methods originally developed for a linear regression model with a complete response have been adapted to survival models with a censored response. For example, penalized methods have been developed for variable selection in the Cox model, including

To the best of our knowledge, there is no research on the EL approach based on growing dimensional general estimating equations with censored data in the literature. Clearly, this work is challenging due to the presence of censoring, dependence among estimating equations and high correlation among variables. Growing dimensional estimating equations means that the dimension of estimating equations depends on the sample size $n$ and grows to infinity as $n \to \infty$. It is motivated by censored regression. The existing estimating equation-based approach is commonly used for the estimation of covariate effects on survival times in censored linear regression models. In general, increasing the dimension of estimating equations can improve the estimation efficiency. A real data example will be given in the application section. The high dimensionality of parameters means that the dimension of unknown parameters of interest depends on the sample size $n$ and grows to infinity as $n \to \infty$ [Chen, Peng and Qin (2009)]. This situation often occurs in gene expression data and consumer financial history data [Fan and Peng (2004)]. One also needs to deal with such case in the analysis of genomic data sets with censored survival outcome data and nonparametric regression. For example, it is commonly known that the bilirubin predictor has a nonlinear effect on the risk function based on the primary biliary cirrhosis data [Fleming and Harrington (1991); Grambsch, Therneau and Fleming (1995)]. In this case, we can use splines to approximate the nonlinear effect where the dimension of basis functions depends on $n$ so that the dimension of unknown coefficients grows to infinity as $n \to \infty$. This example will be analyzed with more details in the application section. In this article, we develop a penalized generalized empirical likelihood (PGEL) procedure for parameter estimation, variable selection, and hypothesis testing based on growing dimensional general estimating equations with high-dimensional censored survival data. In particular, the proposed PGEL estimator has
the oracle properties [Fan and Li (2001)] and attains the semiparametric efficiency bound with the given estimating equation. The PGEL ratio test statistic follows asymptotically a standard chi-square distribution, which can be used to conduct hypothesis testing and to construct confidence regions of parameters of interest. The unscaled form based on the PGEL avoids estimation of a scale parameter and also simplifies intensive computations for censored data. The new PGEL method can provide a nice framework for statistical inference when a parametric likelihood is not available under a high-dimensional censored survival model.

The rest of this paper is organized as follows. We begin with presenting the semiparametric efficiency bound with the given estimating equation and the construction of general estimating equations, and then propose GEL and PGEL procedures for growing estimating equations in Section 2. We establish the theoretical properties of the resulting estimator in Section 3, and develop a penalized generalized empirical likelihood ratio test which has an asymptotic standard $\chi^2$ distribution in Section 4. Moreover, computing procedures of the new method are provided in Section 5. Simulation and application results are reported in Sections 6 and 7, respectively. Some concluding remarks are made in Section 8. Proofs of theorems are given in the supplement [40].


2.1. General estimating equations. Consider a survival study with right-censored data. Let $T$ and $C$ denote survival and censoring times, respectively. However, we can only observe $Y = \min(T, C)$ and $\Delta = I(T \leq C)$ where $I(\cdot)$ denotes the indicator function. Let $X$ be a $r$-dimensional vector of covariates. Let $\{(X_i, Y_i, \Delta_i) : i = 1, \ldots, n\}$ consist of independent copies of $(X, Y, \Delta)$. Assume that the joint distribution $F(t, x)$ of $(T, X)$ is unknown with $(t, x) \in T \times X \subset \mathbb{R}^1 \times \mathbb{R}^r$, and $\theta \in \Theta \subset \mathbb{R}^p$ is a $p$-dimensional parameter vector of interest. Without assuming a specific form of $F(t, x)$, we are interested in making statistical inference on $\theta$ via $k$ functionally independent estimating functions $g(T, X; \theta) = (g_1(T, X; \theta), \ldots, g_k(T, X; \theta))^T$ satisfying

\begin{equation}
E\{g(T_i, X_i; \theta_0)\} = 0
\end{equation}

uniquely at some unknown parameter $\theta_0 \in \Theta$, a compact subset of $\mathbb{R}^p$, where $g(\cdot)$ is a $k \times 1$-vector of known moment functions that may be nonlinear in $\theta_0$.

First, we present the semiparametric efficiency bound for the estimation of $\theta_0$ implicitly defined by the moment condition (2.1). The following condition is commonly used for right-censored data.
(C1) $T$ and $C$ are conditionally independent given $X$.

Define

$$
(2.2) \quad \xi(X; \theta) = E\{g(T, X; \theta) \mid X\}
$$

to be the conditional expectation of the moment conditions given $X$, and define

$$
V(g(T, X; \theta) \mid X) = E\{g(T, X; \theta)^T g(T, X; \theta) \mid X\} - \xi(X; \theta)^T \xi(X; \theta)
$$

to be the conditional variance of the moment conditions given $X$. In addition, define

$$
G(T \mid X) = P(C > T \mid X), \quad \Gamma(\theta) = E\partial g(T, X; \theta)/\partial \theta^T,
$$

and

$$
\Sigma(\theta) = E\left\{\frac{1}{G(T \mid X)} V(g(T, X; \theta) \mid X) + \xi(X; \theta)\xi(X; \theta)^T\right\}.
$$

In the following, the “regular estimators” are as defined in Newey (1990).

**Theorem 2.1.** Let $\theta_0$ be defined by the moment condition (2.1). Suppose that Condition (C1) holds, and $\Gamma(\theta)$ has full column rank equal to $p$ and $k \geq p$. Then the asymptotic variance lower bound for all regular estimators of $\theta_0$ is

$$
K(\theta_0) = (\Gamma(\theta_0)^T \Sigma(\theta_0)^{-1} \Gamma(\theta_0))^{-1}.
$$

In the proof, we present explicit expressions for the efficient score functions corresponding to the asymptotic variance lower bounds in Theorem 2.1. To attain the semiparametric efficiency bound given in Theorem 2.1, we propose the following estimating function

$$
(2.3) \quad \psi(Z; \theta) = \frac{\Delta}{G(Y \mid X)} \{g(Y, X; \theta) - \xi(X; \theta)\} + \xi(X; \theta),
$$

where $Z = (Y, X, \Delta)$.

The first term of the proposed estimating function is the adjusted version with inverse probability weighting. The inverse probability weighting method is actually popular in the literature of time-to-event analysis. Among others, Li and Wang (2003) constructed a synthetic response with the inverse probability weighting, Zhou et al. (2006) considered the empirical likelihood inference based on the inverse probability weighted estimating equations.
proposed by Lin (2003), and He et al. (2016) utilized the inverse probability weighting to construct the estimating functions for the empirical likelihood inference with censored data.

Direct calculations yield

\[ E(\psi(Z; \theta)) = Eg(T; X; \theta), \]
\[ \text{Var}(\psi(Z; \theta)) = E\left\{ \frac{1}{G(T; X)} V(g(T; X; \theta) | X) + \xi(X; \theta) \xi(X; \theta)^T \right\} = \Sigma(\theta). \]

Clearly, if the true parameter \( \theta_0 \) satisfies \( Eg(T; X; \theta_0) = 0 \), then \( E\psi(Z; \theta_0) = 0 \). In practice, \( G(y; x) = P(C \geq y|x) \) and \( \xi(x; \theta) \) in (2.3) are unknown. We can estimate \( G(y; x) \) by the local Kaplan-Meier estimator \( G_n(y; x) \) [Dabrowska (1989); He, Wang and Hong (2013)]. More specifically,

\[ G_n(y; x) = \prod_{i=1}^{n} \left( 1 - \frac{B_{ni}(x)}{\sum_{j=1}^{n} I(Y_j \geq Y_i)B_{nj}(x)} \right) I(Y_i \leq y, \delta_i = 0), \]

(2.4)

where \( B_{nj}(x) = K(\frac{x-X_j}{h})/\sum_{i=1}^{n} K(\frac{x-X_i}{h}) \), \( j = 1, \ldots, n \), are the Nadaraya-Watson weights, \( h \) is the bandwidth, and \( K(\cdot) \) is a probability density function. Since \( \xi(X; \theta) = E\{ \frac{\Delta}{G(T; X)} g(T; X; \theta) | X \} \), we can estimate \( \xi(x; \theta) \) by

\[ \xi_n(x; \theta) = \sum_{i=1}^{n} \frac{B_{ni}(x) \Delta_i}{G_n(Y_i; X_i)} g(Y_i, X_i; \theta). \]

(2.5)

The uniform consistency of \( G_n(y; x) \) and \( \xi_n(x; \theta) \) with Condition (C2) is presented in the supplement [40].

Now we propose to approximate \( \psi(Z; \theta) \) in (2.3) by

\[ \psi_n(Z; \theta) = \frac{\Delta}{G_n(Y; X)} \{ g(Y, X; \theta) - \xi_n(X; \theta) \} + \xi_n(X; \theta). \]

(2.6)

The price to pay for the approximation is that \( \{ \psi_n(Z_i; \theta) \}_{i=1}^{n} \) are not independent which complicates the ensuing analysis due to censoring.

2.2. Generalized empirical likelihood. To investigate the parameter estimation under the constructed general estimating equations (2.6), we present a more general alternative to GMM, i.e., the generalized empirical likelihood (GEL) [Newey and Smith (2004)]. The GEL can be described as a function of a general concave function \( \rho(s) \), whose domain is an open interval \( S \) containing zero. For convenience, let \( \rho_j(s) = \partial^j \rho(s)/\partial s^j \) and \( \rho_j = \rho_j(0) \).
where \( \rho_1 \neq 0, \rho_2 < 0 \) for concavity. Similar to Newey and Smith (2004), it is normalized such that \( \rho_1 = \rho_2 = -1 \) and \( \rho(0) = 0 \). Define

\[
(2.7) \quad \ell(\lambda, \theta) = n^{-1} \sum_{i=1}^{n} \rho(\lambda^T \psi_n(Z_i; \theta))
\]

and

\[
\hat{\Lambda}_n(\theta) = \{ \lambda : \lambda^T \psi_n(Z_i; \theta) \in \mathbb{S}, i = 1, \ldots, n \}.
\]

The GEL estimator is the solution to a saddle point problem [Chang, Chen and Chen (2015)]

\[
\tilde{\theta} = \arg \min_{\theta \in \Theta} \max_{\lambda \in \hat{\Lambda}_n(\theta)} \ell(\lambda, \theta).
\]

The GEL includes the empirical likelihood of Owen (2001) with \( \rho(s) = \log(1 - s) \) and \( \mathbb{S} = (-\infty, 1) \), the exponentially tilting (ET) likelihood estimator of Kitamura and Stutzer (1997) with \( \rho(s) = 1 - \exp(s) \), the continuous updating (CU) GMM of Hansen, Heaton and Yaron (1996) with a quadratic \( \rho(s) = \frac{1}{2} - \frac{(1+s)^2}{2} \) as special cases.

2.3. Penalized generalized empirical likelihood. Under the sparsity assumption on a \( p \)-dimensional parameter we need to identify the zero components and estimate the nonzero parameters. To this end, we consider the following penalized generalized empirical likelihood function

\[
(2.8) \quad \ell_p(\lambda, \theta) = \ell(\lambda, \theta) + \sum_{j=1}^{p} p_\gamma(|\theta_j|),
\]

where \( p_\gamma(s), s \geq 0 \) is a penalty function and its amount of penalty depends on the regularization parameter \( \gamma \) controlling the trade-off between the bias and the model complexity. For convenience, we rewrite the penalty function as \( p_\gamma(\cdot) = \gamma \varphi_\gamma(\cdot) \) and \( \varphi_\gamma(\cdot) \) as \( \varphi(\cdot) \) when it is free of \( \gamma \). Here, \( p_\gamma(s) \) is taken to be the penalties as defined in the following family of functions [Lv and Fan (2009); Fan and Lv (2011)]:

\[
\mathcal{P} = \{ p_\gamma(\cdot) : \varphi_\gamma(s) \text{ is increasing in } s \in [0, \infty), \text{ and the derivative } \varphi'_\gamma(s) \text{ is continuous on } (0, \infty). \text{ In addition, } \varphi'_\gamma(s) \text{ is increasing in } \gamma \text{ and } \varphi'_\gamma(0+) \equiv \varphi'(0+) > 0 \text{ is independent of } \gamma. \}
\]

The commonly used penalties in the family \( \mathcal{P} \) include \( L_1 \) (Lasso) penalty [Tibshirani (1996)], the SCAD penalty [Fan and Li (2001)], and MCP [Zhang (2010)].
In the next section, we study the large sample properties of the following penalized generalized empirical likelihood (PGEL) estimator
\[
\hat{\theta} = \arg\min_{\theta \in \Theta} \max_{\lambda \in \hat{\Lambda}_n(\theta)} \ell_p(\lambda, \theta)
\]
under some regularity conditions.

2.4. Examples. We illustrate the moment conditions through two examples: the accelerated failure time model and the censored partially linear model.

Example 1. (Accelerated Failure Time Model). Consider the accelerated failure time model for the survival time \( T_i \): \( \log(T_i) = \beta_0 + X_i^\top \beta + \epsilon_i \) with \( E(\epsilon_i | X_i) = 0 \), where the covariates \( X_i = (x_{i1}, \ldots, x_{ir})^\top \). The observed data consist of \( \{(X_i, Y_i, \Delta_i) : i = 1, \ldots, n\} \). Let \( \bar{X}_i = (1, X_i^\top)^\top \) and \( \theta = (\beta_0, \beta)^\top \).

(i) We take \( g(T, X; \theta) = \bar{X}\{\log(T) - \bar{X}^\top \theta\} \) and construct the influence function through (2.3) such that the estimators can attain the semiparametric efficiency bound asymptotically. Here, the dimension of the moment functions is the same as the dimension of unknown parameters, i.e. \( k = p = r + 1 \).

(ii) We consider increasing the dimension of the moment functions to improve the estimation efficiency. If covariate \( x_j \) is continuous, we take a \( q_n \)-dimensional vector of known B-spline basis functions \( \{B_s, s = 1, \ldots, q_n\} \) that can approximate any smooth functions of \( x_j \), and then construct over-identified restrictions by using these spline basis functions. For this, we can take additional moment functions as follows:
\[
g_{js}(T, X; \theta) = B_s(x_j)\{\log(T) - \bar{X}^\top \theta\}, \quad s = 1, \ldots, q_n.
\]

Correspondingly, we construct the estimating functions through (2.3) based on the over-identified moment restrictions.

Example 2. (Censored Partially Linear Model). Suppose that the survival time \( T_i \) follows the model:
\[
H(T_i) = \beta_0 + U_i^\top \beta + \sum_{j=1}^{r^2} \sum_{s=1}^{q_n} \gamma_{js} B_s(v_{ij}) + \epsilon_i, \quad i = 1, \ldots, n
\]
where \( E(\epsilon_i | X_i) = 0, X_i = (U_i^\top, V_i^\top)^\top \) with \( U_i = (u_{i1}, \ldots, u_{ir1})^\top \) and \( V_i = (v_{i1}, \ldots, v_{ir2})^\top \), and \( H(\cdot) \) is a known transformation function. The observed data consist of \( \{(X_i, Y_i, \Delta_i) : i = 1, \ldots, n\} \). Here \( r = r_1 + r_2 \) and \( p = 1 + r_1 + r_2 q_n \).
(i) Let $\hat{U} = (1, U^T)^T$, $\theta_1 = (\beta_0, \beta^T)^T$, $\gamma_j = (\gamma_{j,1}, \ldots, \gamma_{j,q_n})^T$, $\theta_2 = (\gamma_j^T, j = 1, \ldots, r_2)^T$, $\theta = (\theta_1^T, \theta_2^T)$, and $B(\cdot) = (B_1(\cdot), \ldots, B_{q_n}(\cdot))^T$. We take

$$g_u(T, X; \theta) = \hat{U} \left\{ H(T) - \hat{U}^T \theta_1 - \sum_{j=1}^{r_2} B_n(v_j)^T \gamma_j \right\}$$

and

$$g_v(T, X; \theta) = B_n(v_l) \left\{ H(T) - \tilde{X}^T \theta_1 - \sum_{j=1}^{r_2} B_n(v_j)^T \gamma_j \right\}, \quad l = 1, \ldots, r_2.$$

Set $g(T, X; \theta) = (g_u(T, X; \theta)^T, g_v(T, X; \theta)^T, j = 1, \ldots, r_2)^T$. Then we can construct the estimating function through (2.3).

(ii) Similarly, we can increase the dimension of the moment restrictions for each continuous component of covariate vector $U$ to improve the estimation efficiency.

3. Asymptotic properties of the PGEL estimator. We first introduce some notation. Let $\mathcal{J} = \{ j : \theta_{0,j} \neq 0 \}$ be the index set of nonzero components of the true parameter vector $\theta_0$, and denote the cardinality of $\mathcal{J}$ as $q = |\mathcal{J}|$, which is unknown. Assume $q \leq k$. Without loss of generality, we write $\theta = (\theta_1^T, \theta_2^T)^T$, where $\theta_1 \in \mathbb{R}^q$ and $\theta_2 \in \mathbb{R}^{n-q}$ correspond to the nonzero and zero components of $\theta$, respectively, which implies that the true parameter vector $\theta_0$ has the following form $\theta_0 = (\theta_{1,0}^T, 0^T)^T$. The corresponding decomposition of $\hat{\theta}$ can be written as $\hat{\theta} = (\hat{\theta}_1^T, \hat{\theta}_2^T)^T$. Let $||A||$ denote its Frobenius norm, $\mathbb{E}(A)$ be its eigenvalues, $\mathbb{E}_{\min}(A)$ and $\mathbb{E}_{\max}(A)$ denote its minimum and maximum eigenvalues, respectively. "w.p.a.1" denotes "with probability approaching one". Define

$$\Omega(\theta) = E\{g(T, X; \theta)^T g(T, X; \theta)\}, \quad \psi(\theta) = n^{-1} \sum_{i=1}^{n} \psi(Z_i; \theta),$$

$$\tilde{\psi}_n(\theta) = n^{-1} \sum_{i=1}^{n} \psi_n(Z_i; \theta), \quad \Gamma(\theta) = E \partial_g(T, X; \theta) / \partial \theta_1^T,$$

and $K(\theta) = (\Gamma(\theta)^T \Sigma(\theta)^{-1} \Gamma(\theta))^{-1}$. Let $\Omega$, $\Sigma$, $\Gamma$ and $\Gamma_1$ denote $\Omega(\theta_0)$, $\Sigma(\theta_0)$, $\Gamma(\theta_0)$ and $\Gamma_1(\theta_0)$, respectively.

To establish the consistency of the PGEL estimator, we need the following conditions.

(C3) (i) $\max_{1 \leq j \leq k} \left\{ \sup_{\theta \in \Theta} E |g_j(T, X; \theta)|^\nu \right\} < \infty$ with $\nu > 2$;
(ii) $B^{-1} \leq \inf_{\theta \in \Theta} \mathbb{E}_{\min}\{\Omega(\theta)\} \leq \sup_{\theta \in \Theta} \mathbb{E}_{\max}\{\Omega(\theta)\} \leq B$ for $B > 1$;

(iii) $B^{-1} \leq \inf_{\theta \in \Theta} \mathbb{E}_{\min}\{\Gamma(\theta)\} \leq \sup_{\theta \in \Theta} \mathbb{E}_{\max}\{\Gamma(\theta)\} \leq B_1$ for $B_1 > 1$.

(C4) (i) There are positive functions $\pi_1(k), \pi_2(\varepsilon)$ such that for any $\varepsilon$,

\[
\inf_{\theta \in \Theta: \|\theta - \theta_0\| \geq \varepsilon} \|Eg(T, X; \theta)\| \geq \pi_1(k)\pi_2(\varepsilon) > 0,
\]

where $\lim \inf_{k \to \infty} \pi_1(k) > 0$;

(ii) $\sup_{\theta \in \Theta} \|\tilde{g}(\theta) - Eg(T, X; \theta)\| = o_p\{\pi_1(k)\}$ where

\[
\tilde{g}(\theta) = n^{-1} \sum_{i=1}^n g(T_i, X_i; \theta).
\]

(C5) The penalty function $p_\gamma(t)$ and the tuning parameter $\gamma$ satisfy that $p_\gamma(0) = 0$, and $\max_{j \in J} p_\gamma(|\theta_{0j}|) \leq B_2 k \log n/(nh^r q)$.

Condition (C3) (i) is utilized to control the tail probability behavior of the influence function, which is the same as that required by Chang, Chen and Chen (2015). Conditions (C3) (ii) and (iii) allows for bounding the eigenvalues of the matrixes. Condition (C4) (i) is the population identification condition for the case of the diverging parameter space, and Condition (C4) (ii) is the uniform convergence; the detailed interpretation can be found in Chang, Chen and Chen (2015). Condition (C5) on penalty function is a technical condition for controlling the impact of the penalty on the nonzero components and deriving the consistency of the PGEL estimators.

**Theorem 3.1. (Consistency)** Suppose that Conditions (C1)–(C5) hold. If $k = o(n^{1/2-1/u} \sqrt{h^r / \log(n)})$, then there is a strict local maximizer $\hat{\theta}$ of the PGEL likelihood $\ell_p(\lambda, \theta)$ such that $\|\theta - \theta_0\| = O_p(\sqrt{k \log n/(nh^r q)})$.

To establish the oracle property of the proposed PGEL estimator, we need more conditions.

(C6) (i) $\max_{1 \leq i \leq k} \sup_{\theta \in \Theta} E|g_{r}(T, X; \theta)g_j(T, X; \theta)|^2 < H_1$ for a constant $H_1$;

(ii) $\max_{1 \leq i \leq k, 1 \leq j \leq p} \sup_{\theta \in \Theta} E|\partial g_r(T, X; \theta)/\partial \theta_j|^2 < H_2$ for a constant $H_2$;

(iii) $\max_{1 \leq i \leq k, 1 \leq j \leq l} \sup_{\theta \in \Theta} E|\partial^2 g_r(T, X; \theta)/\partial \theta_j \partial \theta_l|^2 < H_3$ for a constant $H_3$.

(C7) (i) As $n \to \infty$, $\lim \inf_{r \to 0} \lim \inf_{s \to 0} \varphi_\gamma'(s) > 0$ and the tuning parameter $\gamma$ satisfies $\gamma/\sqrt{k \log n/(nh^r q)} \to \infty$;

(ii) The derivative of the penalty function $p_\gamma(t)$ satisfies

$\gamma \max_{j \in J} \varphi_\gamma'(|\theta_{0j}|) = o(\frac{1}{\sqrt{m}})$ and

$\sup_{\theta \in \Theta} \max_{j \in J} p_\gamma''(|\theta_{0j}|) = o(\sqrt{h^r/(k \log n)})$.  


(C8) For $l = 1, \ldots, k$ and $j = 1, \ldots, q$, let
\[
A_l = \begin{pmatrix}
E[g_l(T, X; \theta)g(T, X; \theta)g^T(T, X; \theta)] & E[g(T, X; \theta)\frac{\partial g(T, X; \theta)}{\partial \theta_i^1}] \\
E[\frac{\partial g(T, X; \theta)}{\partial \theta_i^1}g^T(T, X; \theta)] & E[\frac{\partial^2 g(T, X; \theta)}{\partial \theta_i^1 \partial \theta_i^1}] 
\end{pmatrix}
\]
and
\[
A_j = \begin{pmatrix}
E[\frac{\partial g(T, X; \theta)}{\partial \theta_i^1}g^T(T, X; \theta)] & E[\frac{\partial^2 g(T, X; \theta)}{\partial \theta_i^1 \partial \theta_i^1}] \\
E[\frac{\partial^2 g(T, X; \theta)}{\partial \theta_i^1 \partial \theta_i^1}] & 0 
\end{pmatrix}.
\]

The eigenvalues of the matrixes $A_l$’s and $A_j$’s are bounded by $B_3^{-1} \leq \min_{1 \leq l \leq k} E_{\min} (A_l) \leq \max_{1 \leq l \leq k} E_{\max} (A_l) \leq B_3$ and $B_4^{-1} \leq \min_{1 \leq j \leq q} E_{\min} (A_j) \leq \max_{1 \leq j \leq q} E_{\max} (A_j) \leq B_4$ for constants $B_3 > 1$ and $B_4 > 1$.

Condition (C6) is a standard assumption for the asymptotic normality of the GEL-based estimator. Condition (C7) (i) combined with the penalty function defined in (2.9), is needed to ensure the sparsity property of the PGEL estimator because of the singularity of the penalty function at the origin; Condition (C7) (ii) is designed to reduce the impact of the penalty function on the nonzero parameter estimators. The similar assumptions about the penalty were also required by Chang, Chen and Chen (2015). Condition (C7) holds for the commonly used penalty functions such as the SCAD penalty, the MCP and the hard-threshold penalty. However, for $L_1$ penalty, $\gamma = \gamma \max_{j \in J} g_j'(\theta_j) = o\left(\frac{1}{\sqrt{nq}}\right)$ is incompatible with $\gamma \gg \sqrt{k \log n/(nh^r)}$, which suggests that the PGEL estimator of $\hat{\theta}_1$ with $L_1$ penalty generally cannot achieve the oracle property given in Theorem 3.2 below. Condition (C8) is required to control the order of the remaining terms in the high-order Taylor expansion of the objective function through bounding the eigenvalues of the related matrixes.

**Theorem 3.2.** (Oracle Property) Suppose Conditions (C1)–(C7) hold and $k = o[n^{1/2-1/\nu} \sqrt{h^r / \log n}]$. As $n \to \infty$, we have the following conclusions:

(i) (Sparsity) $\hat{\theta}_2 = 0$ with probability tending to one.

(ii) (Asymptotic Normality) For a fixed $q$, $\sqrt{n}K^{-1/2}(\hat{\theta}_1 - \theta_0) \to N(0, I_q)$ in distribution when $k^3 = o(nh^{2r} / \log^2 n)$, where $K = K(\theta_0)$.

In addition, if Condition (C8) holds, then the asymptotic normality still holds when $k^3 = o(nh^{2r} / \log^2 n)$.
Remark 1. It follows from Theorems 2.1 and 3.2 that the proposed PGEL estimator $\hat{\theta}_1$ of the true nonzero parameter $\theta_{10}$ achieves the semiparametric efficiency bound asymptotically.

For the case of completed data with no censoring, we replace $\psi_n$ with $g$ and get the following results under some regularity conditions.

(C5*) The penalty function $p_\gamma(t)$ and the tuning parameter $\gamma$ satisfy that $p_\gamma(0) = 0$, and $\max_{j \in J} p_\gamma(|\theta_{0j}|) \leq B_2 k/(nq)$.

(C7*) (i) As $n \to \infty$, $\lim \inf_{n \to 0} \lim \inf_{s \to 0^+} \varphi_\gamma(s) > 0$ and the tuning parameter $\gamma$ satisfies $\gamma/\sqrt{k/n} \to \infty$.

(ii) The derivative of the penalty function $p_\gamma(t)$ satisfies $\gamma \max_{j \in J} \varphi_\gamma'(|\theta_{0j}|) = o(1/\sqrt{nq})$ and $\sup_{\theta \in \Theta} \max_{j \in J} p''_\gamma(|\theta_j|) = o(1/\sqrt{k})$.

Theorem 3.3. (Consistency) Suppose that Conditions (C3)–(C4) and (C5*) hold. If $k = o(n^{1/2-1/\nu})$, then there is a strict local maximizer $\hat{\theta}$ of the PGEL likelihood $\ell_p(\lambda, \theta)$ such that $\|\hat{\theta} - \theta_{10}\| = O_p(\sqrt{k/n})$.

Theorem 3.4. (Oracle Property) Suppose Conditions (C3)–(C4), (C5*), (C6) and (C7*) hold and $k = o(n^{1/2-1/\nu})$. As $n \to \infty$, we have the following conclusions:

(i) (Sparsity) $\hat{\theta}_2 = 0$ with probability tending to one.

(ii) (Asymptotic Normality) $\sqrt{n} A_n K^{-1/2}(\hat{\theta}_1 - \theta_{10}) \to \mathcal{N}(0, V)$ in distribution when $k^5 = o(n)$, where $A_n$ is a $d \times q$ matrix such that $A_n A_n^T \to V$, $V$ is a $d \times d$ nonnegative symmetric matrix with the fixed $d$, and $K = K(\theta_{10})$.

In addition, if Condition (C8) holds, then the asymptotic normality still holds when $k^3 = o(n)$.

The proofs of Theorems 3.3 and 3.4 are omitted since they are the simplified versions of Theorems 3.1 and 3.2.

Remark 2. Taking $\rho(s) = \log(1 + s)$, the proposed PGEL reduces to the PEL considered in Leng and Tang (2012). Under weaker conditions, we obtain the same conclusion as those in Leng and Tang (2012). In particular, we remove the assumption $p/k \to \kappa \in (0, 1)$ required by Leng and Tang (2012). Under the sparsity assumption, we only need $q \leq k$, while $p$ can be larger than $k$. Furthermore, the required condition $k^5 = o(n)$ by Leng and Tang (2012) can be relaxed to $k^3 = o(n)$. 
Remark 3. Theorem 3.4 also relaxes the diverging rate $k^3 p^2 = o(n)$ required by Chang, Chen and Chen (2015) as $k^3 = o(n)$. It is easy to see that the proposed PGEL estimator of nonzero parameters can attain the semiparametric efficiency bound asymptotically.

4. Penalized generalized empirical likelihood ratio test. In this section, we present a unified framework for testing hypothesis and constructing confidence regions of $\theta_1$, nonzero elements of parameters, via our proposed PGEL. Now we consider the following hypothesis $H_0 : B_n \theta_1 = 0$ versus $H_1 : B_n \theta_1 \neq 0$, where $B_n$ is a $d \times q$ matrix such that $B_n B_n^T = I_d$ for a fixed $d$, and $I_d$ is a $d \times d$ identity matrix. Such hypothesis includes many hypotheses as its special cases, for example, $H_0 j : \theta_1 j = 0$, $j \in \{1, \ldots, q\}$ which can be used to construct the confidence region for $\theta_1 j$; $H_0 j : \theta_1 j + \theta_1 l = 0$, $j, l \in \{1, \ldots, q\}$ which can be used to test the linear relation between covariates under a linear regression model. Next we propose a PGEL ratio test statistic for testing $H_0 : B_n \theta_1 = 0$:

$$\hat{\ell}_p(B_n) = -2n \{ \ell(\hat{\theta}) - \min_{\theta \in \Theta : B_n \theta_1 = 0} \ell_p(\theta) \},$$

where $\ell(\theta) = \ell(\lambda(\theta), \hat{\theta})$ with $\lambda(\theta) = \arg \max_{\lambda \in \hat{\Lambda}_n(\theta)} \{ \sum_{i=1}^n \rho(\lambda^T \psi_n(Z_i; \theta)) \}$, and $\ell_p(\theta) = \ell_p(\lambda(\theta), \theta)$. One may ask why we take $\ell(\hat{\theta})$ rather than $\ell_p(\hat{\theta})$. This is because the test statistic $\hat{\ell}_p(\hat{\theta}) - \min_{\theta \in \Theta : B_n \theta_1 = 0} \ell_p(\theta)$ only makes sense when the restrictions are imposed on the nonzero coefficients. On the other hand, if the restrictions with the matrix $B_n$ are on the zero coefficients, then the likelihood ratio test equals zero with probability approaching 1, implying the test degenerates.

Theorem 4.1. Suppose that Conditions (C1)–(C8) hold. Under the null hypothesis, we have $\hat{\ell}_p(B_n) \to \chi^2_d$ in distribution, where $\chi^2_d$ denotes the chi-square distribution with $d$ degrees of freedom.

Theorem 4.1 indicates that the well-known Wilks’ theorem holds for the proposed PGEL method. Using the PGEL ratio test, an approximate $100(1 - \alpha)$% confidence region for $B_n \theta_1$ is

$$R_\alpha = \left\{ \phi : -2n \{ \ell(\hat{\theta}) - \min_{B_n \theta_1 = \phi} \ell_p(\theta) \} \leq \chi^2_d(1 - \alpha) \right\},$$

where $\chi^2_d(1 - \alpha)$ is the $(1 - \alpha)$-quantile of the standard chi-square distribution with $d$ degrees of freedom.
5. Computation. In this section, we present a nonlinear optimization procedure for the implementation of minimizing the PGEL, the convergence properties, and the selection of tuning parameters. In the following, let $v_i(\lambda, \theta) = \lambda^T \psi_n(Z_i; \theta)$, $w_{ij}(\lambda, \theta) = \lambda^T \partial_{\theta_j} \psi_n(Z_i; \theta)$, $\pi_{ij}(\lambda, \theta) = \lambda^T \partial_{\theta_j}^2 \psi_n(Z_i; \theta)$, $w_{iA}(\lambda, \theta) = \{\lambda^T \partial_{\theta_j} \psi(Z_i; \theta) : j \in A\}^T$, $\pi_{iA}(\lambda, \theta) = \{\lambda^T \partial_{\theta_j}^2 \psi(Z_i; \theta) : j \in A\}^T$, if $A = \{1, \cdots, p\}$, $w_i(\lambda, \theta) = w_{iA}(\lambda, \theta)$ and $\pi_i(\lambda, \theta) = \pi_{iA}(\lambda, \theta)$.

5.1. Algorithm. To balance the regularization strengths on different components of $\theta$, we minimize the weighted version of the objective function,

$$
(5.1) \quad \tilde{\ell}_p(\theta) = \ell(\theta) + \sum_{j=1}^{p} W_{jj}(\theta)p_\gamma(|\theta_j|),
$$

where $\ell(\theta) = \ell(\lambda(\theta), \theta)$ with $\lambda(\theta) = \arg \max_{\lambda \in \hat{\Lambda}_n(\theta)} \{\sum_{i=1}^{n} \rho(\lambda^T \psi_n(Z_i; \theta))\}$, $W_{jj}(\theta)$ is the $j$th diagonal element of $W(\theta)$ and

$$
W(\theta) = \partial^2_\theta \ell(\theta) = n^{-1} \sum_{i=1}^{n} \rho_2\{v_i(\lambda(\theta), \theta)\}w_i(\lambda(\theta), \theta)w_i(\lambda(\theta), \theta)^T + n^{-1} \sum_{i=1}^{n} \rho_1\{v_i(\lambda(\theta), \theta)\}\pi_i(\lambda(\theta), \theta).
$$

For simplicity, we assume that $W_{jj}(\theta) > 0$ and adopt its absolute value if $W_{jj}(\theta) < 0$ for all $j$. The main difficulty in implementing the nonlinear optimization procedure to minimize $\tilde{\ell}_p(\theta)$ given in Equation (5.1) is the involved nonconcave penalty function $p_\gamma(|\theta_j|)$. For tackling this issue, we conduct the local quadratic approximation of the penalty function [Fan and Li (2001)] at $\theta_j^{(m-1)}$, the iterative value of $\theta_j$ at the $(m - 1)$th step, that is, $p_\gamma(|\theta_j|) \approx p_\gamma(|\theta_j^{(m-1)}|) + \frac{1}{2} \{p'_\gamma(|\theta_j^{(m-1)}|)/|\theta_j^{(m-1)}|\}\{\theta_j^2 - (\theta_j^{(m-1)})^2\}$. Therefore, the first and second derivatives are approximated by $\partial_{\theta_j} p_\gamma(|\theta_j|) = \{p'_\gamma(|\theta_j^{(m-1)}|)/|\theta_j^{(m-1)}|\}\theta_j$ and $\partial^2_\theta p_\gamma(|\theta_j|) = p''_\gamma(|\theta_j^{(m-1)}|)/|\theta_j^{(m-1)}|$. Motivated by Chang, Tang and Wu (2017), we address the computational challenge with the high-dimensionality through a modified two-layer iterative algorithm. The inner layer searches the optimal $\lambda$ by maximizing the concave function $\ell(\lambda, \theta)$ with respect to $\lambda$ for given $\theta$. The outer layer also uses coordinate descent algorithm to search for optimizer $\hat{\theta}$ by cycling through and updating each of the coordinates.

Specifically, in the inner layer, we adopt the strategy of Owen (2001) and generate a proper step size to update $\lambda$ repeatedly until convergence. Note that the objective function needs to be checked to get optimized in each step. If not, the step size continues to be halved until the objective function
gets driven in the right direction. The iterative updating procedure is stable because this layer only involves maximizing a concave function.

The outer layer of the algorithm utilizes the coordinate descent algorithm to optimize the objective function with respect to \( \theta \). We update \( \theta_j \) \((j = 1, \ldots, p)\) at a given \( \lambda \) and other fixed \( \theta_l \)'s \((l \neq j)\). Specifically, we obtain the \((m + 1)\)th update for \( \theta_j \) by

\[
\theta_j^{(m+1)} = \theta_j^{(m)} - \left\{ \sum_{i=1}^{n} \rho_1(v_i(\lambda, \theta^{(m)})) w_{ij}(\lambda, \theta^{(m)}) + nW_{jj}(\theta^{(m)}) \partial_{\theta_j} p_\gamma(|\theta_j^{(m)}|) \right\} \\
\times \left[ \sum_{i=1}^{n} \left\{ \rho_1(v_i(\lambda, \theta^{(m)})) \pi_{ij}(\lambda, \theta^{(m)}) + \rho_2(v_i(\lambda, \theta^{(m)})) w_{ij}(\lambda, \theta^{(m)})^2 \right\} \\
+ nW_{jj}(\theta^{(m)}) \partial_{\theta_j}^2 p_\gamma(|\theta_j^{(m)}|) \right]^{-1}
\]

where \( \theta^{(m)} = (\theta_1^{(m)}, \ldots, \theta_p^{(m)})^T \). During the nonlinear optimization procedure, we set \( \theta_j^{(m)} \) as zero whenever it is less than a threshold level in an iteration. Repeating the two-layer nonlinear optimization procedures until convergence yields the PGEL estimate \( \hat{\theta} \) of \( \theta \).

### 5.2. Convergence analysis

Denote \( \theta^{(m)} \) as the \( m \)th result updated by the two-layer iterative sparsity estimate of our method. We control the convexity of the whole optimization problem (2.8) through introducing the “local concavity” of the penalty function \( \rho_\gamma(s) \) at \( a = (a_1, \ldots, a_q)^T \in \mathcal{R}^q \), i.e.,

\[
\kappa(\rho_\gamma; a) = \max_{1 \leq j \leq q} \lim_{\epsilon \to 0^+} \max_{|a_j| - \epsilon < s_2 < a_j < |a_j| + \epsilon} \max_{0 < s_1 < s_2 < \infty} \frac{\rho'_\gamma(s_2) - \rho'_\gamma(s_1)}{s_2 - s_1},
\]

and define the maximum concavity of the penalty function \( p_\gamma(\cdot) \) by

\[
\kappa(p_\gamma) = \sup_{0 < s_1 < s_2 < \infty} \left\{ -\frac{p'_{\gamma}(s_2) - p'_{\gamma}(s_1)}{s_2 - s_1} \right\}.
\]

One can refer to Lv and Fan (2009) and Fan and Lv (2011) for the detailed interpretation of \( \kappa(\rho_\gamma; a) \) and \( \kappa(p_\gamma) \). For the penalty functions Lasso, SCAD and MCP, we have \( \kappa(p_\gamma) = 0 \), \( \frac{1}{\alpha - 1} \), and \( \frac{1}{\gamma} \), respectively. The concavity of \( p_\gamma(\cdot) \) ensures the nonconvex of \( \ell_p(\theta) \).
Let $\oplus$ denote the Hadamard (entrywise) product, and $\text{diag}(A)$ be the diagonal elements of matrix $A$. Write

$$B_{\mathcal{J}}(\hat{\lambda}, \hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \rho_1(v_i(\hat{\lambda}, \hat{\theta})) w_{\mathcal{J}}(\hat{\lambda}, \hat{\theta}),$$

$$= \frac{1}{n} \sum_{i=1}^{n} \rho_1(v_i(\hat{\lambda}, \hat{\theta})) w_{\mathcal{J}}(\hat{\lambda}, \hat{\theta}),$$

$$W_{\mathcal{J}}(\hat{\lambda}, \hat{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \rho_2(v_i(\hat{\lambda}, \hat{\theta})) w_{\mathcal{J}}(\hat{\lambda}, \hat{\theta}) w_{\mathcal{J}}(\hat{\lambda}, \hat{\theta})^T$$

$$+ \frac{1}{n} \sum_{i=1}^{n} \rho_1(v_i(\hat{\lambda}, \hat{\theta})) \pi_{\mathcal{J}}(\hat{\lambda}, \hat{\theta}).$$

The following theorem gives a sufficient and necessary condition on the strict local minimizer of $\ell_p$ and its restricted global optimality.

**Theorem 5.1.**

(i) (Characterization of PGEL). $\hat{\theta} \in \mathbb{R}^p$ is a strict local minimizer of the objective function in (5.1) if and only if the following conditions hold:

(5.3) $B_{\mathcal{J}}(\hat{\lambda}, \hat{\theta}) + \gamma \rho'(\hat{\theta}_{\mathcal{J}}) \oplus \text{sgn}(\hat{\theta}_{\mathcal{J}}) \oplus \text{diag}(W_{\mathcal{J}}(\hat{\lambda}, \hat{\theta})) = 0,$

(5.4) $\|B_{\mathcal{J}}(\hat{\lambda}, \hat{\theta})\|_{\infty} < \gamma \rho'(\|\cdot\|) \min_{j \in \mathcal{J}^c} W_{\mathcal{J}j}(\hat{\theta}),$

(5.5) $\mathbb{E}_{\min}(W_{\mathcal{J}}(\hat{\lambda}, \hat{\theta})) > \gamma \kappa(\rho, \hat{\theta}_{\mathcal{J}}) \max_{j \in \mathcal{J}} W_{\mathcal{J}j}(\hat{\theta}),$

(5.6) $\hat{\lambda} = \arg \max_{\hat{\lambda} \in \Lambda_n(\hat{\theta})} 1/n \sum_{i=1}^{n} \rho(v_i(\hat{\lambda}, \hat{\theta})).$

(ii) (Restricted global optimality). Let $\hat{\theta}$ be a local minimizer of $\ell_p(\theta)$. For any subset of $\{1, \cdots, p\}$, denote the $|\mathcal{J}|$-dimensional subspace $\{\theta \in \mathbb{R}^p : \theta_j = 0, j \in \mathcal{J}^c\}$ by $\mathcal{E}_\mathcal{J}$. If $\hat{\theta}$ lies in $\mathcal{E}_\mathcal{J}$ and $\mathbb{E}_{\min}(W_{\mathcal{J}}(\hat{\lambda}, \hat{\theta})) > \kappa(p, \rho) \max_{j \in \mathcal{J}} W_{\mathcal{J}j}(\hat{\theta})$, then $\hat{\theta}$ is a global minimizer of $\ell_p(\theta)$ in $\mathcal{E}_\mathcal{J}$.

Note that if we relax the strict inequalities in (5.3)-(5.6) by the nonstrict inequalities in the necessity of Theorem 5.1 (i), the strict local minimizer becomes the local minimizer. Conditions (5.3) and (5.5) imply that $\hat{\theta}$ is a strict local minimizer of (5.1) when constrained on the $\|\hat{\theta}\|_0$-dimensional subspace of $\{\theta \in \mathbb{R}^p : \theta^c = 0\}$ of $\mathbb{R}^p$, where $\theta^c$ denotes the subvector of $\theta$ formed by the components in the complement of $\text{supp}(\hat{\theta})$. Condition (5.4)
makes sure that the sparse vector is indeed a strict local minimizer of (5.1) on the whole space \( \mathbb{R}^p \). Condition (5.6) ensures the achievement of a saddle point.

The condition for gaining global optimality in Theorem 5.1 (ii) is trivially satisfied for the \( L_1 \)-penalty. For the SCAD and MCP, the condition can be satisfied with some \( J \) if the correlation among covariates is not too strong and the concavity of the penalty function is not too large. Following Fan and Lv (2011), under some mild regularity conditions we can further establish the global optimality of \( \hat{\theta} \) on the union of all \(|J|\)-dimensional coordinate subspaces of \( \mathbb{R}^p \).

Based on the Karush-Kuhn-Tucker (KKT) conditions in Theorem 5.1, we can explore the convergence of the two-layer iterative sparsity estimator. Assuming the sequence \( \{\theta^{(m)}\} \) resulted from the two-layer iterative sparsity estimate of our method is bounded, we have the following conclusions: (i) If the penalty function \( p_\gamma(\cdot) \) satisfies \( \kappa(p_\gamma) < 1 \), then every cluster point of \( \{\theta^{(m)}\} \) is a local minimizer of \( \tilde{\ell}_p(\theta) \). Note that the condition \( \kappa(p_\gamma) < 1 \) is always satisfied for the \( L_1 \)-penalty, SCAD \( (a > 2) \), and MCP \( (a > 1) \). (ii) If the sequence \( \{\theta^{(m)}\} \) eventually drops in a compact neighborhood \( \Theta \) of \( \theta^* \) such that \( \theta^* \) is the unique local minimizer of \( \tilde{\ell}_p(\theta) \) in \( \Theta \), then the sequence \( \{\theta^{(m)}\} \) converges to \( \theta^* \). (iii) If the sequence \( \{\theta^{(m)}\} \) generated by the two-layer iterative sparsity estimator algorithm eventually lies in \( \mathcal{E}_J \), and \( \mathbb{E}_{\min}(W_{\bar{J}}) > \kappa(p_\gamma) \max_{j \in J} W_{jj} \), then \( \tilde{\ell}_p(\theta) \) has a unique global minimizer \( \theta^* \) in \( \mathcal{E}_J \) and the sequence \( \{\theta^{(m)}\} \) converges to \( \theta^* \).

5.3. Selection of tuning parameters. To implement the proposed PGEL procedure, it is necessary to find a data-driven approach to select the penalty parameter \( \gamma \). Then we consider the following BIC criterion:

\[
\text{BIC}(\gamma) = 2n\ell(\hat{\theta}_\gamma) + C_n \log(n)\text{df}_\gamma,
\]

where \( \hat{\theta}_\gamma \) is the PGEL estimator of \( \theta \) depending on the penalty parameter \( \gamma \), \( \text{df}_\gamma \) is the number of nonzero components in \( \theta \) representing the “degrees of freedom” of the estimated estimating equations, and \( C_n \) is a scaling factor diverging to infinity at a slow rate for \( k \to \infty \). When \( k \) is fixed, we set \( C_n = 1 \); otherwise we take \( C_n = \max\{\log \log k, 1\} \) [Tang and Leng (2010)]. Although the proposed BIC shows good performance, a rigorous proof of the consistency of the BIC for the PGEL objective function merits further theoretical investigation.

5.4. Dimension reduction. When the dimension \( r \) of covariate vector is large, to handle a curse of dimensionality, instead of using all components
of $X$ in the kernel smoothing, we can apply some dimension reduction techniques for ensuring the properties of $G_n(y \mid x)$ and $\xi_n(x; \theta)$. For this, we assume that the survival time follows a general index model

$$
\Pr(T \leq t \mid X) = \Pr(T \leq t \mid \alpha^T X),
$$

where the $\alpha$ is a $r \times m$ index regression coefficient matrix with $m < r$. Then we adopt the counting process-based dimension reduction strategy developed by Sun, Zhu, Wang and Zeng (2019) using the semiparametric inverse regression approach and employ R Package “orthoDr” (Zhu et al., 2018) for computation to obtain the estimator $\hat{\alpha}$. Correspondingly, all the estimators involving the kernel smoothing need to be modified, i.e.,

$$
G_n(y \mid x) = \prod_{i=1}^{n} \left\{ 1 - \frac{\tilde{B}_{ni}(x)}{\sum_{j=1}^{n} I(Y_j \geq Y_i) \tilde{B}_{nj}(x)} \right\} I(Y_i \leq y, \delta_i = 0),
$$

and

$$
\xi_n(x; \theta) = \sum_{i=1}^{n} \frac{\tilde{B}_{ni}(x) \Delta_i}{G_n(Y_i \mid X_i)} g(Y_i, X_i; \theta),
$$

where $\tilde{B}_{nj}(x) = K(\tilde{\alpha}^T x - \tilde{\alpha}^T X_j) / \{ \sum_{i=1}^{n} K(\tilde{\alpha}^T x - \tilde{\alpha}^T X_i) \}$, $j = 1, \ldots, n$, are the Nadaraya-Watson weights, $h$ is the bandwidth, and $K(\cdot)$ is a probability density function.

6. Simulation studies. We conducted simulation studies to evaluate the finite-sample performance of the new method. To apply the proposed PGEL procedure, we took $\rho(s) = \log(1 - s)$, $1 - \exp(s)$, $1/2 - (1 + s)^2/2$, denoted by PEL, PET, PCU, respectively, based on two penalties SCAD and Lasso. The goal of the study is to examine the parameter estimation and sparsity recovery of the proposed method and the performance of the PGEL ratio test. For the purpose, we considered the accelerated failure time model

$$
\log(T_i) = \tilde{X}_i^T \theta + \epsilon_i, \quad i = 1, \ldots, n,
$$

where $\tilde{X}_i = (\tilde{x}_{i1}, \ldots, \tilde{x}_{ip})^T$ was generated from a multivariate normal distribution with zero mean and covariance matrix $D = \{d_{jl}\}$ with $d_{jl} = \sigma^{j-l}$, and $\epsilon_i$ was taken from the standard normal distribution $\mathcal{N}(0, 1)$. The censoring time $C_i$ was generated from a uniform distribution $\text{Unif}(0, \kappa \exp(|\tilde{x}_{i1} - \tilde{x}_{i2}|))$, where $\kappa$ was chosen to achieve censoring rates of 20% and 40%. Set $\theta_0 = (3, 1.5, 0, 0.2, 0, \ldots, 0)$ including three nonzero components and $p - 3$ zero components. For dimensionality $p$ and sample size $n$,
we took \((n, p) = (50, 7), (100, 10), (200, 14)\), where \(p\) is the integer part of \((8(3n)^{1.5} - 14)\). We choose the bandwidth \(h = \hat{\sigma}_X n^{-1/5}\), where \(\hat{\sigma}_X\) is the estimated standard deviation of \(X\) in the sample and adopt standard multivariate normal distribution as the kernel density function \(K(\cdot)\). For each setting, 200 repetitions were conducted to investigate the accuracy of our proposed estimates in terms of the root mean square error (RMSE) and the performance of our proposed variable selection method.

To illustrate our proposed approach to over-identified general estimating equations, we introduce an instrumental variable \(U_i = (u_{i1}, \ldots, u_{ip})^T\) generated from the following models

**Model C**: \(u_{ij} \sim i.i.d. \tilde{x}_{ij} + N(0, 1), j = 1, \ldots, p, \epsilon \sim N(0, 0.1)\)

**Model M**: \(u_{ij} \sim i.i.d. \tilde{x}_{ij} + N(1, 1), j = 1, \ldots, p, \epsilon \sim N(0, 0.5)\).

We set \(X = (\tilde{X}^T, U^T)\) and utilize influence functions \(g(T, X; \theta)\):

\[
g(T, X; \theta) = (\tilde{x}_1 (\log(T) - \tilde{X}^T \theta), \ldots, \tilde{x}_p (\log(T) - \tilde{X}^T \theta), u_1 (\log(T) - \tilde{X}^T \theta), \ldots, u_p (\log(T) - \tilde{X}^T \theta))^T.
\]

Tables 1 and 2 report the RMSE values of nonzero components in \(\theta_0\) and the average numbers of zero coefficients that are correctly and incorrectly identified using three PGEL methods with the SCAD and Lasso penalties under two censoring rates, respectively. The corresponding oracle estimates are also included in the tables for comparison. It can be seen from the tables that the estimation and selection results by the proposed PGEL methods with the SCAD are close to oracle results, whilst PGEL with the Lasso yields biased estimates due to the large penalty on nonzero parameters, indicating that our simulation results are consistent with those given in Theorem 3.2.

To examine the behavior of the proposed PGEL ratio test in Theorem 4.1, we constructed confidence regions using the PEL method with the SCAD penalty based on three types of estimating functions. These include the proposed, the synthetic-data method \cite{Li and Wang (2003)} and the Buckley-James method \cite{Fang et al. (2013)} as follows:

\[
\psi_n(Z_i; \theta) = \frac{\Delta_i}{G_n(Y_i | X_i)} \{ g(Y_i, X_i; \theta) - \xi_n(X_i; \theta) \} + \xi_n(X_i; \theta),
\]

\[
\psi_n^{SD}(Z_i; \theta) = \frac{\Delta_i g(Y_i, X_i; \theta)}{G_n(Y_i)},
\]

\[
\psi_n^{BJ}(Z_i; \theta) = \Delta_i g(Y_i, X_i; \theta) + (1 - \Delta_i) X_i \int_{\log(Y_i) - \tilde{X}_i^T \theta}^{\infty} \frac{t \, dF_n(t | \theta)}{1 - F_n(\log(Y_i) - \tilde{X}_i^T \theta | \theta)}.
\]
where \( F_n(t | \theta) \) denotes the Kaplan-Meier estimator of distribution function of \( \epsilon \). Setting different vector forms of \( B_n \) in (4.1) leads to a confidence set for \( \theta_1, \theta_2 \) and \( \theta_5 \), at the 1-\( \alpha \) level. Table 3 summarizes the coverage percentages of the confidence intervals estimated by three methods based on 200 replicates. The simulation results indicate that the proposed method outperforms the methods in Li and Wang (2003) and Fang et al. (2013) in each setting.

The results in Tables 1 and 2 show that PEL, PET and PCU under Model C have similar performance. This confirms the conclusion made in Theorem 3.2 with different concave function \( \rho \). Next we turn to investigate the performance of the PEL, PET and PCU procedures in misspecified estimating equations (i.e., Model M). Table 4 includes the estimated bias (BIAS) given by the average of the estimates minus the true value, the sample standard deviation (SSD) of the estimates, and the average numbers of zero coefficients that are correctly and incorrectly identified under \( (n, p) = (50, 7) \) from 200 independent runs. From the table, we observe that the PET estimates of parameters are robust to misspecified estimating equations in terms of BIAS, while the PEL and PCU estimates of parameters are sensitive to misspecified estimating equations in the sense that their corresponding biases deviate from zero. Additionally, the PET variable selection procedure performs better than the PEL and PCU in the sense that the average number of correctly identify nonzero components by the PET method is quite close to the true number 3 of nonzero components even though estimating equations are misspecified.

7. An application. In this section, we analyzed an example taken from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver to illustrate our proposed PGEL method. The PBC data have been analyzed by many authors [e.g., Fleming and Harrington (1991); Grambsch, Therneau and Fleming (1995)]. The data set, which can be downloaded from \( R \) package \textit{survival}, consists of 424 observations and 19 variables including censored survival time \( (T) \), censoring indicator \( (\Delta) \), and 17 covariates \( X = (x_1, \ldots, x_{17})^T \). The 17 covariates include treatment code \( (x_1) \), age in years \( (x_2) \), sex \( (x_3) \), presence of ascites \( (x_4) \), presence of hepatomegaly \( (x_5) \), presence of spiders \( (x_6) \), presence of edema \( (x_7) \), serum cholesterol \( (x_8) \), log(albumin) \( (x_9) \), urine copper \( (x_{10}) \), alkaline phosphatase \( (x_{11}) \), SGOT \( (x_{12}) \), triglycerides \( (x_{13}) \), platelet count \( (x_{14}) \), log(prothrombin time) \( (x_{15}) \), histologic stage of disease \( (x_{16}) \), and log(serum bilirubin)(\( x_{17} \)). Here, the log-transformation of the three covariates was made according to the analysis in Fleming and Harrington (1991) and Grambsch, Therneau and Fleming.
(1995); a transformation was further made for each covariate such that all covariates took values in $[0, 1]$ for convenience.

We only considered $n = 276$ observations through deleting ones with some missing covariates. The observed data consist of $\{Z_i = (Y_i, X_i, \Delta_i) : i = 1, \ldots, 276\}$, where $Y_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$ with $C_i$ being the censoring time for patient $i$.

Before fitting an AFT model to the PBC data, we obtained the Kaplan-Meier estimates of the survival functions based on the three groups in the presence of edema $(0, 0.5, 1)$, as shown in Figure 1. It can be seen that three survival curves are overlapping. This implies that the Cox model may not be suitable for the PBC data. In addition, log(serum bilirubin) still has a nonlinear effect on the survival time [Cao, Huang, Liu and Zhao (2016)]. Thus, we considered the following AFT model for analyzing the PBC data:

$$\log(T_i) = \beta_0 + \sum_{j=1}^{16} x_{i,j} \beta_j + \sum_{s=1}^{q_n} \gamma_s B_s(x_{i,17}) + \varepsilon_i,$$

where $\{B_s, s = 1, \ldots, q_n\}$ are cubic B-spline basis functions on $[0, 1]$ with the knots $\{0, 0, 0, 1/3, 2/3, 1, 1, 1\}$ and $q_n = 5$.

Let $\theta = (\beta_0, \beta_1, \ldots, \beta_{16}, \gamma_1, \ldots, \gamma_5)^T$. We applied the proposed approach to the PBC data through increasing the dimension $k$ of moment function $g(T, X; \theta)$ for parameter estimation and variable selection.

**Case I.** Consider the moment functions:

$$g_1(T, X, \theta) = \log(T) - \beta_0 - \sum_{j=1}^{16} x_{j} \beta_j - \sum_{s=1}^{q_n} \gamma_s B_s(x_{17}),$$

$$g_{j+1}(T, X, \theta) = x_{j} \left\{ \log(T) - \beta_0 - \sum_{j=1}^{16} x_{j} \beta_j - \sum_{s=1}^{q_n} \gamma_s B_s(x_{17}) \right\}, j = 1, \ldots, 16,$$

$$g_{17+l}(T, X, \theta) = B_l(x_{17}) \left\{ \log(T) - \beta_0 - \sum_{j=1}^{16} x_{j} \beta_j - \sum_{s=1}^{q_n} \gamma_s B_s(x_{17}) \right\}, l = 1, \ldots, 5.$$

**Case II.** Increase the dimension of the moment functions $g$ by taking

$$g_{22+l}(T, X, \theta) = B_l(x_3) \left\{ \log(T) - \beta_0 - \sum_{j=1}^{16} x_{j} \beta_j - \sum_{s=1}^{q_n} \gamma_s B_s(x_{17}) \right\}, l = 1, \ldots, 5.$$

and

$$g_{27+l}(T, X, \theta) = B_l(x_{11}) \left\{ \log(T) - \beta_0 - \sum_{j=1}^{16} x_{j} \beta_j - \sum_{s=1}^{q_n} \gamma_s B_s(x_{17}) \right\}, l = 1, \ldots, 5.$$
For the estimation of $G_n$ and $\xi_n$, we utilized the dimension reduction method given in Section 5.4. and calculated the estimate of a $17 \times 2$ projection matrix $\alpha$. For each case we used the three PGEL methods with the SCAD and MCP penalties to estimate $\theta$ and to select important covariates with the initial value of $\theta$ taken from its GEL estimate without penalties. Estimates of nonzero regression coefficients in $\theta$ identified by our proposed PGEL methods are presented in Table 5. Furthermore, we conducted tests for covariate effects. To this end, we considered the null hypothesis $H_{0j}$: $\theta_{1j} = 0$, i.e., $H_{0j} : B_{nj} \theta_1 = 0$, where $B_{nj}$ is a $1 \times \hat{q}$ vector with $j$th component being 1 and others being 0's, $j = 1, \ldots, \hat{q}$, and $\hat{q}$ is the number of nonzero estimates. Using the asymptotic distribution of the proposed PGEL ratio test statistics in Theorem 4.1, we obtained the p-values for all cases, as shown in Table 5.

From the estimation results in Case I, we note that some nonzero coefficients are not statistically significant. However, from the estimation results in Case II with increasing the dimension of moment functions, we find that all regression coefficients of unimportant covariates are estimated as zero, while all nonzero coefficients are statistically significant. This finding demonstrates that increasing the dimension of estimating functions can improve estimation efficiency.

8. Concluding remarks. When a parametric likelihood is unspecified for censored survival models, we develop a penalized generalized empirical likelihood approach for simultaneous variable selection and parameter estimation problems. The oracle properties of the proposed PGEL estimator are established, and the estimators of nonzero parameters attains the semiparametric efficiency bound asymptotically. Also the proposed PGEL ratio is asymptotically distributed as the standard chi-square distribution. In particular, for complete data, the condition $k/p \to \kappa \in (0, 1)$ required by Leng and Tang (2012) is removed, and also the condition $k^5 = o(n)$ or $k^3p^2 = o(n)$ required by Leng and Tang (2012) and Chang, Chen and Chen (2015) is relaxed to $k^3 = o(n)$ for deriving the oracle properties of estimators.

Note that for censored data, general estimating functions involve two unknown functions: conditional survival function of censoring variable $C$ given covariate $X$ and conditional mean of moment function $g$ given $X$. The convergence rates of their nonparametric estimators given in Section 2 depend on the dimension $r$ of covariate vector so that the PGEL estimator converges slowly when $r$ is large in the presence of censoring. For this situation, we can utilize some dimension reduction methods for ensuring the properties
of $G_n$ [e.g., Sun et al. (2019)]. Another possible solution is to first reduce the dimension of covariate vector by using the well-developed model-free screening procedures [e.g., He, Wang and Hong (2013)] and then the proposed approach can be used. Some other solutions deserve further research.

Note that the proposed new approach is different from the Buckley-James method. The estimating function is constructed based on the general moment condition $E(g(T, X; \theta_0)) = 0$ and the semiparametric efficiency with this condition, while the Buck-James method is based on the imputation approach. If we use the idea of the Buckley-James method, $g(T, X; \theta)$ should be estimated by its conditional expectation $E\{g(T, X; \theta)|Y, X, \Delta\} = \Delta g(T, X; \theta) + (1 - \Delta)E\{g(T, X; \theta)|Y, X, \Delta\}$.

Also note that differentiability of moment functions is required. For non-smooth moment functions $g$ with respect to $\theta$, the definitions of the GEL and the PGEL estimators should be modified. Following Parente and Smith (2011), the GEL estimator is defined to satisfy

$$\ell(\lambda, \theta) \leq \inf_{\theta \in \Theta} \sup_{\lambda \in \hat{\Lambda}_n(\theta)} \ell(\lambda, \theta) + o_p(n^{-\tau}),$$

where $\ell(\lambda, \theta)$ is as defined in (2.7), $\tau$ is nonnegative, and

$$\lambda = \arg \max_{\lambda \in \hat{\Lambda}_n(\theta)} \ell(\lambda, \theta).$$

Similarly, for non-smooth moment functions, the PGEL estimator is defined to satisfy

$$\ell_p(\lambda, \theta) \leq \inf_{\theta \in \Theta} \sup_{\lambda \in \hat{\Lambda}_n(\theta)} \ell_p(\lambda, \theta) + o_p(n^{-\tau}),$$

where $\ell_p(\lambda, \theta)$ is as defined in (2.8), and

$$\lambda = \arg \max_{\lambda \in \hat{\Lambda}_n(\theta)} \ell_p(\lambda, \theta).$$

Note that the proofs of the theoretical results in this article rely on differentiability of moment functions, and these cannot be straightforwardly extended to the case of non-smooth moment functions. Further research is needed to investigate the theoretical properties of $\theta$ with non-smooth moment functions for censored data.

Acknowledgments. The authors would like to thank the Editor, the Associate Editor and the two reviewers for their constructive and insightful comments and suggestions that greatly improved the paper.

SUPPLEMENTARY MATERIAL

Supplement A: Supplement to “Penalized generalized empirical likelihood with a diverging number of general estimating equations
for censored data” 


REFERENCES


Table 1. Simulation results of the PGEL estimates in AFT model with $\sigma = 0.7$ and a censoring rate of 20% under Model C.

<table>
<thead>
<tr>
<th>(n, p)</th>
<th>PEL</th>
<th>PET</th>
<th>PCU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\theta_1$</td>
<td>$\theta_2$</td>
<td>$\theta_5$</td>
</tr>
<tr>
<td>(50, 7)</td>
<td>0.119</td>
<td>0.160</td>
<td>0.208</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>3.74</td>
<td>2.2</td>
</tr>
<tr>
<td>(100, 10)</td>
<td>0.110</td>
<td>0.121</td>
<td>0.155</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>6.81</td>
<td>5.99</td>
</tr>
<tr>
<td>(200, 14)</td>
<td>0.081</td>
<td>0.090</td>
<td>0.118</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>11.00</td>
<td>10.02</td>
</tr>
</tbody>
</table>

“T” represents the average number of correctly estimated zero coefficients, “F” denotes the average number of incorrectly estimated zero coefficients. Other values are root mean square errors. PEL: penalized empirical likelihood; PET: penalized exponentially tilted likelihood; PCU: penalized continuous updating method.
Table 2. Simulation results of the PGEL estimates in AFT model with $\sigma = 0.7$ and a censoring rate of 40% under Model C.

<table>
<thead>
<tr>
<th>$(n, p)$</th>
<th>PEL</th>
<th>PET</th>
<th>PCU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oracle SCAD</td>
<td>Lasso</td>
<td>Oracle SCAD</td>
</tr>
<tr>
<td>$(50, 7)$</td>
<td>$\hat{\theta}_1$</td>
<td>0.130 0.168 0.221</td>
<td>$\hat{\theta}_2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$(100, 10)$</td>
<td>$\hat{\theta}_1$</td>
<td>0.120 0.129 0.163</td>
<td>$\hat{\theta}_2$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$(200, 14)$</td>
<td>$\hat{\theta}_1$</td>
<td>0.092 0.098 0.130</td>
<td>$\hat{\theta}_2$</td>
</tr>
</tbody>
</table>

“T” represents the average number of correctly estimated zero coefficients, “F” denotes the average number of incorrectly estimated zero coefficients. Other values are root mean square errors. PEL: penalized empirical likelihood; PET: penalized exponentially tilted likelihood; PCU: penalized continuous updating method.
Table 3. Coverage probabilities (%) of estimated confidence intervals from the PEL ratio test statistic with the SCAD penalty based on three estimating functions from the proposed method, the Synthetic-data method, and Buckley-James method, for $\theta_1$, $\theta_2$ and $\theta_5$ over 200 simulated data sets with $\sigma = 0.7$ under Model C.

<table>
<thead>
<tr>
<th>Par.</th>
<th>CR</th>
<th>$(n, p)$</th>
<th>$\alpha = 10%$</th>
<th>$\alpha = 5%$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Proposed S-D B-J</td>
<td>Proposed S-D B-J</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>20%</td>
<td>(50,7)</td>
<td>88.8 87.5 88.1</td>
<td>92.8 91.8 91.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(100,10)</td>
<td>89.5 88.1 88.3</td>
<td>94.1 93.0 93.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(200,14)</td>
<td>89.7 88.6 89.0</td>
<td>94.7 93.5 93.6</td>
</tr>
<tr>
<td>40%</td>
<td>(50,7)</td>
<td>87.1 85.7 85.9</td>
<td>91.9 89.5 89.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(100,10)</td>
<td>88.0 86.4 86.8</td>
<td>93.0 90.6 91.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(200,14)</td>
<td>89.2 87.0 87.3</td>
<td>93.9 92.3 92.6</td>
<td></td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>20%</td>
<td>(50,7)</td>
<td>88.4 86.5 87.1</td>
<td>92.5 90.1 91.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(100,10)</td>
<td>88.9 86.9 87.6</td>
<td>93.7 91.7 92.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(200,14)</td>
<td>89.7 87.5 88.3</td>
<td>94.6 92.5 93.1</td>
</tr>
<tr>
<td>40%</td>
<td>(50,7)</td>
<td>87.5 85.2 86.2</td>
<td>91.2 89.2 89.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(100,10)</td>
<td>87.9 85.9 86.8</td>
<td>92.4 90.1 91.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(200,14)</td>
<td>89.2 86.6 87.4</td>
<td>93.7 91.3 92.2</td>
<td></td>
</tr>
<tr>
<td>$\theta_5$</td>
<td>20%</td>
<td>(50,7)</td>
<td>88.9 87.2 88.0</td>
<td>92.8 91.4 91.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(100,10)</td>
<td>89.8 87.8 88.2</td>
<td>94.1 91.8 92.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(200,14)</td>
<td>89.9 88.2 88.7</td>
<td>94.8 92.5 93.3</td>
</tr>
<tr>
<td>40%</td>
<td>(50,7)</td>
<td>87.4 85.1 85.7</td>
<td>91.9 89.0 89.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(100,10)</td>
<td>88.2 85.8 86.6</td>
<td>92.9 90.3 90.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(200,14)</td>
<td>89.1 86.5 87.2</td>
<td>94.2 91.8 92.1</td>
<td></td>
</tr>
</tbody>
</table>

Note: “Par.” stands for parameter; “CR” denotes censoring rate; “S-D” stands for the Synthetic-data method [Li and Wang (2003)]; “B-J” stands for Buckley-James method [Fang et al. (2013)].
Table 4. Performance of the three PGEL estimates with SCAD penalty under Model M with $n = 50$ and $p = 7$ based on 200 replicates.

<table>
<thead>
<tr>
<th></th>
<th>PET BIAS</th>
<th>PET SSD</th>
<th>PEL BIAS</th>
<th>PEL SSD</th>
<th>PCU BIAS</th>
<th>PCU SSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\theta}_1$</td>
<td>0.058</td>
<td>0.114</td>
<td>0.056</td>
<td>0.104</td>
<td>0.750</td>
<td>0.840</td>
</tr>
<tr>
<td>$\hat{\theta}_2$</td>
<td>0.133</td>
<td>0.165</td>
<td>0.120</td>
<td>0.158</td>
<td>0.673</td>
<td>1.236</td>
</tr>
<tr>
<td>$\hat{\theta}_3$</td>
<td>0.000</td>
<td>0.008</td>
<td>0.001</td>
<td>0.005</td>
<td>0.241</td>
<td>0.832</td>
</tr>
<tr>
<td>$\hat{\theta}_4$</td>
<td>0.001</td>
<td>0.007</td>
<td>0.000</td>
<td>0.003</td>
<td>0.233</td>
<td>0.633</td>
</tr>
<tr>
<td>$\hat{\theta}_5$</td>
<td>0.121</td>
<td>0.196</td>
<td>0.125</td>
<td>0.193</td>
<td>0.850</td>
<td>1.033</td>
</tr>
<tr>
<td>$\hat{\theta}_6$</td>
<td>0.001</td>
<td>0.004</td>
<td>0.000</td>
<td>0.001</td>
<td>0.240</td>
<td>0.532</td>
</tr>
<tr>
<td>$\hat{\theta}_7$</td>
<td>0.001</td>
<td>0.005</td>
<td>0.000</td>
<td>0.003</td>
<td>0.234</td>
<td>0.630</td>
</tr>
<tr>
<td>$\hat{\theta}_T$</td>
<td>3.62</td>
<td>3.71</td>
<td>1.52</td>
<td>1.70</td>
<td>1.56</td>
<td>1.80</td>
</tr>
</tbody>
</table>

**Fig 1.** The Kaplan-Meier estimates of survival functions based on three groups in the presence of edema ($x_7 = 0, 0.5, 1$) with the PBC data.
Table 5. Estimated coefficients for the PBC data by various methods. The numbers in parentheses are p-values of the estimated nonzero coefficients.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Case I</th>
<th>SCAD</th>
<th>PEL</th>
<th>SCAD</th>
<th>MCP</th>
<th>SCAD</th>
<th>PEL</th>
<th>SCAD</th>
<th>MCP</th>
<th>SCAD</th>
<th>PEL</th>
<th>SCAD</th>
<th>PEL</th>
<th>SCAD</th>
<th>PEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-1.35(0.00)</td>
<td>-1.30(0.00)</td>
<td></td>
<td>-1.47(0.00)</td>
<td>-1.42(0.00)</td>
<td></td>
<td>-1.62(0.00)</td>
<td>-1.57(0.00)</td>
<td></td>
<td>-1.62(0.00)</td>
<td>-1.57(0.00)</td>
<td></td>
<td>-1.62(0.00)</td>
<td>-1.57(0.00)</td>
<td></td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.03(0.08)</td>
<td>0</td>
<td>0</td>
<td>0.07(0.03)</td>
<td>0.09(0.03)</td>
<td></td>
<td>0.12(0.01)</td>
<td>0.15(0.01)</td>
<td></td>
<td>0.12(0.01)</td>
<td>0.15(0.01)</td>
<td></td>
<td>0.12(0.01)</td>
<td>0.15(0.01)</td>
<td></td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>-0.32(0.00)</td>
<td>-0.27(0.00)</td>
<td></td>
<td>-0.43(0.00)</td>
<td>-0.38(0.00)</td>
<td></td>
<td>-0.31(0.00)</td>
<td>-0.28(0.00)</td>
<td></td>
<td>-0.31(0.00)</td>
<td>-0.28(0.00)</td>
<td></td>
<td>-0.31(0.00)</td>
<td>-0.28(0.00)</td>
<td></td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>-0.86(0.00)</td>
<td>-0.92(0.00)</td>
<td></td>
<td>-1.06(0.00)</td>
<td>-1.10(0.00)</td>
<td></td>
<td>-0.96(0.00)</td>
<td>-1.02(0.00)</td>
<td></td>
<td>-0.96(0.00)</td>
<td>-1.02(0.00)</td>
<td></td>
<td>-0.96(0.00)</td>
<td>-1.02(0.00)</td>
<td></td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\beta_9$</td>
<td>0.47(0.00)</td>
<td>0.42(0.00)</td>
<td></td>
<td>0.64(0.00)</td>
<td>0.62(0.00)</td>
<td></td>
<td>0.59(0.00)</td>
<td>0.64(0.00)</td>
<td></td>
<td>0.59(0.00)</td>
<td>0.64(0.00)</td>
<td></td>
<td>0.59(0.00)</td>
<td>0.64(0.00)</td>
<td></td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>-0.76(0.00)</td>
<td>-0.73(0.00)</td>
<td></td>
<td>-0.93(0.00)</td>
<td>-0.87(0.00)</td>
<td></td>
<td>-0.99(0.00)</td>
<td>-0.93(0.00)</td>
<td></td>
<td>-0.99(0.00)</td>
<td>-0.93(0.00)</td>
<td></td>
<td>-0.99(0.00)</td>
<td>-0.93(0.00)</td>
<td></td>
</tr>
<tr>
<td>$\gamma_1$</td>
<td>1.73(0.00)</td>
<td>1.75(0.00)</td>
<td></td>
<td>1.83(0.00)</td>
<td>1.89(0.00)</td>
<td></td>
<td>1.86(0.00)</td>
<td>1.90(0.00)</td>
<td></td>
<td>1.86(0.00)</td>
<td>1.90(0.00)</td>
<td></td>
<td>1.86(0.00)</td>
<td>1.90(0.00)</td>
<td></td>
</tr>
<tr>
<td>$\gamma_2$</td>
<td>0.81(0.00)</td>
<td>0.82(0.00)</td>
<td></td>
<td>0.71(0.00)</td>
<td>0.77(0.00)</td>
<td></td>
<td>0.90(0.00)</td>
<td>0.93(0.00)</td>
<td></td>
<td>0.90(0.00)</td>
<td>0.93(0.00)</td>
<td></td>
<td>0.90(0.00)</td>
<td>0.93(0.00)</td>
<td></td>
</tr>
<tr>
<td>$\gamma_3$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\gamma_4$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

School of Mathematics and Statistics  
Yunnan University  
Kunming, China  
E-mail: nstang@ynu.edu.cn

School of Mathematics and Statistics  
Shandong University  
Jinan, China  
E-mail: xiaodong.yan@polyu.edu.hk

Department of Applied Mathematics  
The Hong Kong Polytechnic University  
Hong Kong, China  
E-mail: xingqiu.zhao@polyu.edu.hk