The Annals of Statistics
2020, Vol. 48, No. 1, 607–627
https://doi.org/10.1214/19-AOS1870

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

BY NIANSHENG TANG[1], XIAODONG YAN[2] AND XINGQIU ZHAO[3]

[1]*School of Mathematics and Statistics, Yunnan University, nstang@ynu.edu.cn*
[2]*School of Economics, Shandong University, yanxiaodong@sdu.edu.cn*
[3]*Department of Applied Mathematics, Hong Kong Polytechnic University, xingqiu.zhao@polyu.edu.hk*

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.

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) and 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 $-2\log$(empirical likelihood ratio) is a scaled chi-square distribution, where the scale parameter is a function of

Received July 2018; revised February 2019.
MSC2010 subject classifications. Primary 62N02, 62N03; secondary 62F12, 62F05.
Key words and phrases. Censored survival data, penalized generalized empirical likelihood, penalized generalized empirical likelihood ratio test, oracle property, semiparametric efficiency.

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 $-2\log$(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 $-2\log$(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 a Lasso (Tibshirani (1997)), a nonconcave penalized likelihood (Fan and Li (2002)), an adaptive Lasso (Zhang and Lu (2007)), an efficient-adaptive-shrinkage method (Zou (2008)), and the Dantzig selector (Antoniadis, Fryzlewicz and Letué (2010)). In the context of the additive hazards model (Lin and Ying (1994)), Leng and Ma (2007) proposed a weighted L_1 approach and Martinussen and Scheike (2009) considered the ridge, the Lasso, the adaptive Lasso and the Dantzig selector, while Lin and Lv (2013) developed regularization methods using a class of concave penalties. For parameter estimation and variable selection in the AFT model, Wu, Li and Tang (2015) developed an empirical likelihood method using the estimating equation of Tsiatis (1990) and Ritov (1990), and also obtained the central chi-square distributed empirical likelihood ratio.

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) and 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 χ^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 Supplementary Material (Tang, Yan and Zhao (2019)).

2. Methods.

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 $\Delta = I(T \le C)$ 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, Δ). Assume that the joint distribution $\mathcal{F}(t, x)$ of (T, X) is unknown with $(t, x) \in \mathcal{T} \times \mathcal{X} \subset \mathcal{R}^1 \times \mathcal{R}^r$, and $\boldsymbol{\theta} \in \Theta \subset \mathcal{R}^p$ is a p-dimensional parameter vector of interest. Without assuming a specific form of $\mathcal{F}(t, x)$, we are interested in making statistical inference on $\boldsymbol{\theta}$ via k functionally independent estimating functions $g(T, X; \boldsymbol{\theta}) = (g_1(T, X; \boldsymbol{\theta}), \ldots, g_k(T, X; \boldsymbol{\theta}))^T$ satisfying

$$E\{g(T_i, X_i; \boldsymbol{\theta}_0)\} = 0 \tag{2.1}$$

uniquely at some unknown parameter $\boldsymbol{\theta}_0 \in \Theta$, a compact subset of $\mathcal{R}^p$, where $g(\cdot)$ is a $k \times 1$-vector of known moment functions that may be nonlinear in $\boldsymbol{\theta}_0$. First, we present the semiparametric efficiency bound for the estimation of $\boldsymbol{\theta}_0$ 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

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

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

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

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

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

and

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

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

THEOREM 2.1. *Let $\boldsymbol{\theta}_0$ be defined by the moment condition (2.1). Suppose that Condition* (C1) *holds, and $\Gamma(\boldsymbol{\theta})$ has full column rank equal to p and $k \ge p$. Then the asymptotic variance lower bound for all regular estimators of $\boldsymbol{\theta}_0$ is*

$$K(\boldsymbol{\theta}_0) = \left(\Gamma(\boldsymbol{\theta}_0)^T \Sigma(\boldsymbol{\theta}_0)^{-1} \Gamma(\boldsymbol{\theta}_0)\right)^{-1}.$$

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:

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

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

$$E\big(\psi(Z;\boldsymbol{\theta})\big) = Eg(T, X;\boldsymbol{\theta}),$$

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

Clearly, if the true parameter $\boldsymbol{\theta}_0$ satisfies $Eg(T, X;\boldsymbol{\theta}_0) = 0$, then $E\psi(Z;\boldsymbol{\theta}_0) = 0$. In practice, $G(y \mid x) = P(C \geq y \mid x)$ and $\xi(x;\boldsymbol{\theta})$ in (2.3) are unknown. We can estimate $G(y \mid x)$ by the local Kaplan–Meier estimator $G_n(y \mid x)$ (Dabrowska (1989) and He, Wang and Hong (2013)). More specifically,

$$G_n(y \mid x) = \prod_{i=1}^{n}\left\{1 - \frac{B_{ni}(x)}{\sum_{j=1}^{n} I(Y_j \geq Y_i)B_{nj}(x)}\right\}^{I(Y_i \leq y, \delta_i = 0)}, \tag{2.4}$$

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 $\xi(X;\boldsymbol{\theta}) = E\{\frac{\Delta}{G(T|X)}g(T, X;\boldsymbol{\theta}) \mid X\}$, we can estimate $\xi(x;\boldsymbol{\theta})$ by

$$\xi_n(x;\boldsymbol{\theta}) = \sum_{i=1}^{n}\frac{B_{ni}(x)\Delta_i}{G_n(Y_i \mid X_i)}g(Y_i, X_i;\boldsymbol{\theta}). \tag{2.5}$$

The uniform consistency of $G_n(y \mid x)$ and $\xi_n(x;\boldsymbol{\theta})$ with Condition (C2) is presented in the Supplementary Material (Tang, Yan and Zhao (2019)).

Now we propose to approximate $\psi(Z;\boldsymbol{\theta})$ in (2.3) by

$$\psi_n(Z;\boldsymbol{\theta}) = \frac{\Delta}{G_n(Y \mid X)}\{g(Y, X;\boldsymbol{\theta}) - \xi_n(X;\boldsymbol{\theta})\} + \xi_n(X;\boldsymbol{\theta}). \tag{2.6}$$

The price to pay for the approximation is that $\{\psi_n(Z_i;\boldsymbol{\theta})\}_{i=1}^{n}$ 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, that is, 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 $\mathbb{S}$ containing zero. For convenience, let $\rho_j(s) = \partial^j\rho(s)/\partial s^j$ and $\rho_j = \rho_j(0)$ $(j = 0, 1, 2, \ldots)$, 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

$$\ell(\lambda, \boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\rho\big(\lambda^T\psi_n(Z_i;\boldsymbol{\theta})\big) \tag{2.7}$$

and

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

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

$$\widetilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \max_{\lambda\in\widehat{\Lambda}_n(\boldsymbol{\theta})} \ell(\lambda, \boldsymbol{\theta}).$$

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

$$\ell_p(\lambda, \boldsymbol{\theta}) = \ell(\lambda, \boldsymbol{\theta}) + \sum_{j=1}^{p} p_\gamma(|\theta_j|), \tag{2.8}$$

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

$$\begin{aligned}\mathcal{P} = \{&p_\gamma(\cdot) : \varrho_\gamma(s) \text{ is increasing in } s \in [0, \infty), \text{ and the derivative } \varrho'_\gamma(s)\\ &\text{is continuous on } (0, \infty). \text{ In addition, } \varrho'_\gamma(s) \text{ is increasing}\\ &\text{in } \gamma \text{ and } \varrho'_\gamma(0+) \equiv \varrho'(0+) > 0 \text{ is independent of } \gamma\}.\end{aligned} \tag{2.9}$$

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:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \max_{\lambda\in\widehat{\Lambda}_n(\boldsymbol{\theta})} \ell_p(\lambda, \boldsymbol{\theta})$$

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: $\log(T_i) = \beta_0 + X_i^\top \boldsymbol{\beta} + \epsilon_i$ with $E(\epsilon_i \mid X_i) = 0$, 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 $\tilde{X}_i = (1, X_i^\top)^\top$ and $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^\top)^\top$.

(i) We take $g(T, X; \boldsymbol{\theta}) = \tilde{X}\{\log(T) - \tilde{X}^\top\boldsymbol{\theta}\}$ 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, that is, $k = p = r + 1$.

(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:

$$g_{js}(T, X; \theta) = B_s(x_j)\{\log(T) - \tilde{X}^\top \boldsymbol{\theta}\}, \quad s = 1, \ldots, q_n.$$

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

EXAMPLE 2 (Censored partially linear model). Suppose that the survival time T_i follows the model

$$H(T_i) = \beta_0 + U_i^\top \boldsymbol{\beta} + \sum_{j=1}^{r_2} \sum_{s=1}^{q_n} \gamma_{js} B_s(v_{ij}) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $E(\epsilon_i \mid X_i) = 0$, $X_i = (U_i^\top, V_i^\top)^\top$ with $U_i = (u_{i,1}, \ldots, u_{i,r_1})^\top$ and $V_i = (v_{i,1}, \ldots, v_{i,r_2})^\top$, 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 $\tilde{U} = (1, U^\top)^\top$, $\boldsymbol{\theta}_1 = (\beta_0, \boldsymbol{\beta}^\top)^\top$, $\boldsymbol{\gamma}_j = (\gamma_{j,1}, \ldots, \gamma_{j,q_n})^\top$, $\boldsymbol{\theta}_2 = (\boldsymbol{\gamma}_j^\top, j = 1, \ldots, r_2)^\top$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$, and $\mathbf{B}(\cdot) = (B_1(\cdot), \ldots, B_{q_n}(\cdot))^\top$. We take

$$g_u(T, X; \boldsymbol{\theta}) = \tilde{U}\left\{H(T) - \tilde{U}^\top \boldsymbol{\theta}_1 - \sum_{j=1}^{r_2} \mathbf{B}_n(v_j)^\top \boldsymbol{\gamma}_j\right\}$$

and

$$g_{vl}(T, X; \boldsymbol{\theta}) = \mathbf{B}_n(v_l)\left\{H(T) - \tilde{X}^\top \boldsymbol{\theta}_1 - \sum_{j=1}^{r_2} \mathbf{B}_n(v_j)^\top \boldsymbol{\gamma}_j\right\}, \quad l = 1, \ldots, r_2.$$

Set $g(T, X; \boldsymbol{\theta}) = (g_u(T, X; \boldsymbol{\theta})^\top, g_{vj}(T, X; \boldsymbol{\theta})^\top, j = 1, \ldots, r_2)^\top$. 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 $\mathbb{J} = \{j : \theta_{0j} \neq 0\}$ be the index set of nonzero components of the true parameter vector $\boldsymbol{\theta}_0$, and denote the cardinality of $\mathbb{J}$ as $q = |\mathbb{J}|$, which is unknown. Assume $q \leq k$. Without loss of generality, we write $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)^T$, where $\boldsymbol{\theta}_1 \in \mathcal{R}^q$ and $\boldsymbol{\theta}_2 \in \mathcal{R}^{p-q}$ correspond to the nonzero and zero components of $\boldsymbol{\theta}$, respectively, which implies that the true parameter vector $\boldsymbol{\theta}_0$ has the following form $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{10}^T, \mathbf{0}^T)^T$. The corresponding decomposition of $\widehat{\boldsymbol{\theta}}$ can be written as $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\theta}}_1^T, \widehat{\boldsymbol{\theta}}_2^T)^T$. 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

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

$$\bar{\psi}_n(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^{n} \psi_n(Z_i; \boldsymbol{\theta}), \qquad \Gamma_1(\boldsymbol{\theta}) = E\partial g(T, X; \boldsymbol{\theta})/\partial \boldsymbol{\theta}_1^T,$$

and $\mathcal{K}(\boldsymbol{\theta}) = (\Gamma_1(\boldsymbol{\theta})^T \Sigma(\boldsymbol{\theta})^{-1}\Gamma_1(\boldsymbol{\theta}))^{-1}$. Let Ω, Σ, Γ and Γ_1 denote $\Omega(\boldsymbol{\theta}_0)$, $\Sigma(\boldsymbol{\theta}_0)$, $\Gamma(\boldsymbol{\theta}_0)$ and $\Gamma_1(\boldsymbol{\theta}_0)$, respectively.

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

(C3) (i) $\max_{1\le j\le k}(\sup_{\boldsymbol{\theta}\in\Theta} E|g_j(T, X; \boldsymbol{\theta})|^\nu) < \infty$ with $\nu > 2$;
(ii) $B^{-1} \le \inf_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\min}\{\Omega(\boldsymbol{\theta})\} \le \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\max}\{\Omega(\boldsymbol{\theta})\} \le B$ for $B > 1$;
(iii) $B_1^{-1} \le \inf_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\min}\{\Gamma^\top(\boldsymbol{\theta})\Gamma(\boldsymbol{\theta})\} \le \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\max}\{\Gamma^\top(\boldsymbol{\theta})\Gamma(\boldsymbol{\theta})\} \le B_1$ for $B_1 > 1$.

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

$$\inf_{\boldsymbol{\theta}\in\Theta:\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\ge\varepsilon} \|Eg(T, X; \boldsymbol{\theta})\| \ge \pi_1(k)\pi_2(\varepsilon) > 0,$$

where $\liminf_{k\to\infty} \pi_1(k) > 0$;
(ii) $\sup_{\boldsymbol{\theta}\in\Theta} \|\bar{g}(\boldsymbol{\theta}) - Eg(T, X; \boldsymbol{\theta})\| = o_p\{\pi_1(k)\}$ where $\bar{g}(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n g(T_i, X_i; \boldsymbol{\theta})$.

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

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* $k = o(n^{1/2-1/\nu}\sqrt{h^r/\log(n)})$, *then there is a strict local maximizer* $\widehat{\boldsymbol{\theta}}$ *of the PGEL likelihood* $\ell_p(\lambda, \boldsymbol{\theta})$ *such that* $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(\sqrt{k\log n/(nh^r)})$.

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

(C6) (i) $\max_{1\le l,j\le k} \sup_{\boldsymbol{\theta}\in\Theta} E|g_l(T, X; \boldsymbol{\theta})g_j(T, X; \boldsymbol{\theta})|^2 < H_1$ for a constant H_1;
(ii) $\max_{1\le r\le k, 1\le j\le p} \sup_{\boldsymbol{\theta}\in\Theta} E|\partial g_r(T, X; \boldsymbol{\theta})/\partial\theta_j|^2 \le H_2$ for a constant H_2;
(iii) $\max_{1\le r\le k, 1\le j,l\le p} \sup_{\boldsymbol{\theta}\in\Theta} E|\partial^2 g_r(T, X; \boldsymbol{\theta})/\partial\theta_j\partial\theta_l|^2 \le H_3$ for a constant H_3.

(C7) (i) As $n\to\infty$, $\liminf_{\gamma\to 0}\liminf_{s\to 0^+} \varrho'_\gamma(s) > 0$ and the tuning parameter γ satisfies $\gamma/\sqrt{k\log n/(nh^r)} \to \infty$;
(ii) The derivative of the penalty function $p_\gamma(t)$ satisfies
$\gamma \max_{j\in\mathbb{J}} \varrho'_\gamma(|\theta_{0j}|) = o(\frac{1}{\sqrt{nq}})$ and $\sup_{\boldsymbol{\theta}\in\Theta} \max_{j\in\mathbb{J}} p''_\gamma(|\theta_j|) = o(\sqrt{h^r/(k\log n)})$.

(C8) For $l = 1, \ldots, k$ and $j = 1, \ldots, q$, let

$$\mathcal{A}_l = \begin{pmatrix} E\big[g_l(T, X; \boldsymbol{\theta})g(T, X; \boldsymbol{\theta})g^T(T, X; \boldsymbol{\theta})\big] & E\Big[g(T, X; \boldsymbol{\theta})\dfrac{\partial g_l(T, X; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1^T}\Big] \\ E\Big[\dfrac{\partial g_l(T, X; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1} g^T(T, X; \boldsymbol{\theta})\Big] & E\Big[\dfrac{\partial^2 g_l(T, X; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1\partial \boldsymbol{\theta}_1^T}\Big] \end{pmatrix}$$

and

$$\mathbb{A}_j = \begin{pmatrix} E\Big[\dfrac{\partial g(T, X; \boldsymbol{\theta})}{\partial\theta_{1j}} g^T(T, X; \boldsymbol{\theta})\Big] & E\Big[\dfrac{\partial^2 g(T, X; \boldsymbol{\theta})}{\partial\theta_{1j}\partial \boldsymbol{\theta}_1^T}\Big] \\ E\Big[\dfrac{\partial^2 g^T(T, X; \boldsymbol{\theta})}{\partial\theta_{1j}\partial \boldsymbol{\theta}_1}\Big] & \mathbf{0} \end{pmatrix}.$$

The eigenvalues of the matrixes $\mathcal{A}_l$'s and $\mathbb{A}_j$'s are bounded by

$$B_3^{-1} \le \min_{1\le l\le k} \inf_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\min}(\mathcal{A}_l) \le \max_{1\le l\le k} \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\max}(\mathcal{A}_l) \le B_3 \quad \text{and}$$

$$B_4^{-1} \le \min_{1\le j\le q} \inf_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\min}(\mathbb{A}_j) \le \max_{1\le j\le q} \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\max}(\mathbb{A}_j) \le B_4$$

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, $\gamma = \gamma \max_{j\in\mathbb{J}} \varrho'_\gamma(|\theta_{0j}|) = o(\frac{1}{\sqrt{nq}})$ is incompatible with $\gamma \gg \sqrt{k\log n/(nh^r)}$, which suggests that the PGEL estimator of $\widehat{\boldsymbol{\theta}}_1$ 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*) $\widehat{\boldsymbol{\theta}}_2 = 0$ *with probability tending to one*.
(ii) (*Asymptotic normality*) *For a fixed* q, $\sqrt{n}\mathcal{K}^{-1/2}(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}) \to \mathcal{N}(0, I_q)$ *in distribution when* $k^5 = o(nh^{2r}/\log^2 n)$, *where* $\mathcal{K} = \mathcal{K}(\boldsymbol{\theta}_{10})$.

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 $\widehat{\boldsymbol{\theta}}_1$ of the true nonzero parameter $\boldsymbol{\theta}_{10}$ achieves the semiparametric efficiency bound asymptotically.

For the case of completed data with no censoring, we replace ψ_n with g and get the following results under some regularity conditions:

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

(C7*) (i) As $n \to \infty$, $\liminf_{\gamma\to 0} \liminf_{s\to 0^+} \varrho'_\gamma(s) > 0$ and the tuning parameter γ satisfies $\gamma/\sqrt{k/n} \to \infty$.
(ii) The derivative of the penalty function $p_\gamma(t)$ satisfies $\gamma \max_{j\in\mathbb{J}} \varrho'_\gamma(|\theta_{0j}|) = o(1/\sqrt{nq})$ and $\sup_{\boldsymbol{\theta}\in\Theta} \max_{j\in\mathbb{J}} p''_\gamma(|\theta_j|) = o(1/\sqrt{k})$.

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* $\widehat{\boldsymbol{\theta}}$ *of the PGEL likelihood* $\ell_p(\lambda, \boldsymbol{\theta})$ *such that* $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(\sqrt{k/n})$.

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*) $\widehat{\boldsymbol{\theta}}_2 = 0$ *with probability tending to one*.

(ii) (*Asymptotic normality*) $\sqrt{n}A_n\mathcal{K}^{-1/2}(\widehat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}) \to \mathcal{N}(0, V)$ *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* $\mathcal{K} = \mathcal{K}(\boldsymbol{\theta}_{10})$.

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 $q \le k$, 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 $\boldsymbol{\theta}_1$, nonzero elements of parameters, via our proposed PGEL. Now we consider the following hypothesis $H_0 : B_n\boldsymbol{\theta}_1 = 0$ versus $H_1 : B_n\boldsymbol{\theta}_1 \neq 0$, 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, $H_{0j} : \theta_{1j} = 0$, $j \in \{1, \ldots, q\}$ which can be used to construct the confidence region for θ_{1j}; $H_{0j} : \theta_{1j} + \theta_{1l} = 0$, $j, l \in \{1, \ldots, q\}$ 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: $B_n\boldsymbol{\theta}_1 = 0$:

$$\widehat{\ell}_p(B_n) = -2n\Big\{\ell(\widetilde{\boldsymbol{\theta}}) - \min_{\boldsymbol{\theta}\in\Theta: B_n\boldsymbol{\theta}_1 = 0} \ell_p(\boldsymbol{\theta})\Big\},$$

where $\ell(\boldsymbol{\theta}) = \ell(\lambda(\boldsymbol{\theta}), \boldsymbol{\theta})$ with $\lambda(\boldsymbol{\theta}) = \arg\max_{\lambda\in\widehat{\Lambda}_n(\boldsymbol{\theta})}\{\sum_{i=1}^n \rho(\lambda^T\psi_n(Z_i;\boldsymbol{\theta}))\}$, and $\ell_p(\boldsymbol{\theta}) = \ell_p(\lambda(\boldsymbol{\theta}), \boldsymbol{\theta})$. One may ask why we take $\ell(\widetilde{\boldsymbol{\theta}})$ rather than $\ell_p(\widehat{\boldsymbol{\theta}})$. This is because the test statistic $\ell_p(\widehat{\boldsymbol{\theta}}) - \min_{\boldsymbol{\theta}\in\Theta: B_n\boldsymbol{\theta}_1=0}\ell_p(\boldsymbol{\theta})$ 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* $\widehat{\ell}_p(B_n) \to \chi_d^2$ *in distribution, where* χ_d^2 *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 $B_n\boldsymbol{\theta}_1$ is

$$R_\alpha = \Big\{\boldsymbol{\phi} : -2n\Big\{\ell(\widetilde{\boldsymbol{\theta}}) - \min_{B_n\boldsymbol{\theta}_1 = \boldsymbol{\phi}} \ell_p(\boldsymbol{\theta})\Big\} \le \chi_d^2(1-\alpha)\Big\}, \tag{4.1}$$

where $\chi_d^2(1-\alpha)$ 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 $v_i(\lambda,\boldsymbol{\theta}) = \lambda^T\psi_n(Z_i;\boldsymbol{\theta})$, $w_{ij}(\lambda,\boldsymbol{\theta}) = \lambda^T\partial_{\theta_j}\psi_n(Z_i;\boldsymbol{\theta})$, $\pi_{ij}(\lambda,\boldsymbol{\theta}) = \lambda^T\partial^2_{\theta_j}\psi_n(Z_i;\boldsymbol{\theta})$, $w_{iA}(\lambda,\boldsymbol{\theta}) = \{\lambda^T\partial_{\theta_j}\psi_n(Z_i;\boldsymbol{\theta}) : j\in A\}^T$, $\pi_{iA}(\lambda,\boldsymbol{\theta}) = \{\lambda^T\partial^2_{\boldsymbol{\theta}_{l,\theta_j}}\psi_n(Z_i;\boldsymbol{\theta}) : l, j\in A\}^T$, and if $A=\{1,\ldots,p\}$, $w_i(\lambda,\boldsymbol{\theta}) = w_{iA}(\lambda,\boldsymbol{\theta})$ and $\pi_i(\lambda,\boldsymbol{\theta}) = \pi_{iA}(\lambda,\boldsymbol{\theta})$.

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

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

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

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

For simplicity, we assume that $W_{jj}(\boldsymbol{\theta}) > 0$ and adopt its absolute value if $W_{jj}(\boldsymbol{\theta}) < 0$ for all j. The main difficulty in implementing the nonlinear optimization procedure to minimize $\widetilde{\ell}_p(\boldsymbol{\theta})$ 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 θ_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 $\partial^2_{\theta_j}p_\gamma(|\theta_j|) = p'_\gamma(|\theta_j^{(m-1)}|)/|\theta_j^{(m-1)}|$. 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 λ by maximizing the concave function $\ell(\lambda,\boldsymbol{\theta})$ with respect to λ for given $\boldsymbol{\theta}$. The outer layer also uses coordinate descent algorithm to search for optimizer $\widehat{\boldsymbol{\theta}}$ 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 λ 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 $\boldsymbol{\theta}$. We update θ_j $(j=1,\ldots,p)$ at a given λ and other fixed θ_l's $(l\neq j)$. Specifically, we obtain the $(m+1)$th update for θ_j by

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

where $\boldsymbol{\theta}^{(m)} = (\theta_1^{(m)}, \ldots, \theta_p^{(m)})^T$. 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 $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$.

5.2. *Convergence analysis.* Denote $\boldsymbol{\theta}^{(m)}$ as the mth 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 $\varrho_\gamma(s)$ at $\mathbf{a} = (a_1, \ldots, a_q)^T \in \mathcal{R}^q$, that is,

$$\kappa(\varrho_\gamma; \mathbf{a}) = \max_{1\le j\le q} \lim_{\epsilon\to 0+} \sup_{|a_j|-\epsilon<s_1<s_2<|a_j|+\epsilon} -\frac{\varrho'_\gamma(s_2) - \varrho'_\gamma(s_1)}{s_2 - s_1}, \tag{5.2}$$

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 $\kappa(\varrho_\gamma; \mathbf{a})$ and $\kappa(p_\gamma)$. For the penalty functions Lasso, SCAD and MCP, we have $\kappa(p_\gamma) = 0$, $\frac{1}{a-1}$ and $\frac{1}{a}$, respectively. The concavity of $p_\gamma(\cdot)$ ensures the nonconvex of $\widetilde{\ell}_p(\boldsymbol{\theta})$.

Let $\oplus$ denote the Hadamard (entrywise) product, and $\operatorname{diag}(A)$ be the diagonal elements of matrix A. Write

$$\begin{aligned}
\mathcal{B}_{\widehat{\mathbb{J}}}(\widehat{\lambda}, \widehat{\boldsymbol{\theta}}) &= \frac{1}{n}\sum_{i=1}^n \rho_1\big[v_i(\widehat{\lambda}, \widehat{\boldsymbol{\theta}})\big] w_{i\widehat{\mathbb{J}}}(\widehat{\lambda}, \widehat{\boldsymbol{\theta}}), \mathcal{B}_{\widehat{\mathbb{J}}^c}(\widehat{\lambda}, \widehat{\boldsymbol{\theta}}) \\
&= \frac{1}{n}\sum_{i=1}^n \rho_1\big(v_i(\widehat{\lambda}, \widehat{\boldsymbol{\theta}})\big) w_{i\widehat{\mathbb{J}}^c}(\widehat{\lambda}, \widehat{\boldsymbol{\theta}}), \\
W_{\widehat{\mathbb{J}}}(\widehat{\lambda}, \widehat{\boldsymbol{\theta}}) &= \frac{1}{n}\sum_{i=1}^n \rho_2\big[v_i(\widehat{\lambda}, \widehat{\boldsymbol{\theta}})\big] w_{i\widehat{\mathbb{J}}}(\widehat{\lambda}, \widehat{\boldsymbol{\theta}}) w_{i\widehat{\mathbb{J}}}(\widehat{\lambda}, \widehat{\boldsymbol{\theta}})^T \\
&\quad + \frac{1}{n}\sum_{i=1}^n \rho_1\big[v_i(\widehat{\lambda}, \widehat{\boldsymbol{\theta}})\big] \pi_{i\widehat{\mathbb{J}}}(\widehat{\lambda}, \widehat{\boldsymbol{\theta}}).
\end{aligned}$$

The following theorem gives a sufficient and necessary condition on the strict local minimizer of $\widetilde{\ell}_p$ and its restricted global optimality.

THEOREM 5.1.

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

$$\mathcal{B}_{\widehat{\mathbb{J}}}(\widehat{\lambda}, \widehat{\boldsymbol{\theta}}) + \gamma\varrho'_\gamma\big(|\widehat{\boldsymbol{\theta}}_{\widehat{\mathbb{J}}}|\big) \oplus \operatorname{sgn}(\widehat{\boldsymbol{\theta}}_{\widehat{\mathbb{J}}}) \oplus \operatorname{diag}\big(W_{\widehat{\mathbb{J}}}(\widehat{\lambda}, \widehat{\boldsymbol{\theta}})\big) = 0, \tag{5.3}$$

$$\big\|\mathcal{B}_{\widehat{\mathbb{J}}^c}(\widehat{\lambda}, \widehat{\boldsymbol{\theta}})\big\|_\infty < \gamma\varrho'_\gamma\big(|0_+|\big) \min_{j\in\widehat{\mathbb{J}}^c} W_{jj}(\widehat{\boldsymbol{\theta}}), \tag{5.4}$$

$$\mathbb{E}_{\min}\big(W_{\widehat{\mathbb{J}}}(\widehat{\lambda}, \widehat{\boldsymbol{\theta}})\big) > \gamma\kappa(\varrho_\gamma; \widehat{\boldsymbol{\theta}}_{\widehat{\mathbb{J}}}) \max_{j\in\widehat{\mathbb{J}}} W_{jj}(\widehat{\boldsymbol{\theta}}), \tag{5.5}$$

$$\widehat{\lambda} = \arg\max_{\lambda\in\widehat{\Lambda}_n(\widehat{\boldsymbol{\theta}})} 1/n \sum_{i=1}^n \rho\big[v_i(\lambda, \widehat{\boldsymbol{\theta}})\big]. \tag{5.6}$$

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

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 $\widehat{\boldsymbol{\theta}}$ is a strict local minimizer of (5.1) when constrained on the $\|\widehat{\boldsymbol{\theta}}\|_0$-dimensional subspace of $\{\boldsymbol{\theta}\in\mathcal{R}^p:\boldsymbol{\theta}^c=0\}$ of $\mathcal{R}^p$, where $\boldsymbol{\theta}^c$ denotes the subvector of $\boldsymbol{\theta}$ formed by the components in the complement of $\operatorname{supp}(\widehat{\boldsymbol{\theta}})$. Condition (5.4) makes sure that the sparse vector is indeed a strict local minimizer of (5.1) on the whole space $\mathcal{R}^p$. 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 $\mathbb{J}$ 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 $\widehat{\boldsymbol{\theta}}$ on the union of all $|\mathbb{J}|$-dimensional coordinate subspaces of $\mathcal{R}^p$.

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 $\{\boldsymbol{\theta}^{(m)}\}$ 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 $\{\boldsymbol{\theta}^{(m)}\}$ is a local minimizer of $\widetilde{\ell}_p(\boldsymbol{\theta})$. 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 $\{\boldsymbol{\theta}^{(m)}\}$ eventually drops in a compact neighborhood Θ of $\boldsymbol{\theta}^*$ such that $\boldsymbol{\theta}^*$ is the unique local minimizer of $\widetilde{\ell}_p(\boldsymbol{\theta})$ in Θ, then the sequence $\{\boldsymbol{\theta}^{(m)}\}$ converges to $\boldsymbol{\theta}^*$. (iii) If the sequence $\{\boldsymbol{\theta}^{(m)}\}$ generated by the two-layer iterative sparsity estimator algorithm eventually lies in $\mathcal{E}_{\mathbb{J}}$, and $\mathbb{E}_{\min}(W_{\mathbb{J}})>\kappa(p_\gamma)\max_{j\in\mathbb{J}}W_{jj}$, then $\widetilde{\ell}_p(\boldsymbol{\theta})$ has a unique global minimizer $\boldsymbol{\theta}^*$ in $\mathcal{E}_{\mathbb{J}}$ and the sequence $\{\boldsymbol{\theta}^{(m)}\}$ converges to $\boldsymbol{\theta}^*$.

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 γ. Then we consider the following BIC criterion:

$$\mathrm{BIC}(\gamma)=2n\ell(\widehat{\boldsymbol{\theta}}_\gamma)+C_n\log(n)\mathrm{df}_\gamma,$$

where $\widehat{\boldsymbol{\theta}}_\gamma$ is the PGEL estimator of $\boldsymbol{\theta}$ depending on the penalty parameter γ, df_γ is the number of nonzero components in $\boldsymbol{\theta}$ 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 $\xi_n(x;\boldsymbol{\theta})$. For this, we assume that the survival time follows a general index model

$$\Pr(T\le t\mid X)=\Pr\bigl(T\le t\mid\alpha^\top X\bigr),$$

where the α 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 $\widehat{\alpha}$. Correspondingly, all the estimators involving the kernel smoothing need to be modified, that is,

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

and

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

where $\tilde{B}_{nj}(x) = K(\frac{\widehat{\alpha}^\top x - \widehat{\alpha}^\top X_j}{h}) / \{\sum_{i=1}^{n} K(\frac{\widehat{\alpha}^\top x - \widehat{\alpha}^\top X_i}{h})\}$, $j = 1, \ldots, n$, 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

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

where $\widetilde{X}_i = (\widetilde{x}_{i1}, \ldots, \widetilde{x}_{ip})^T$ was generated from a multivariate normal distribution with zero mean and covariance matrix $\mathcal{D} = \{d_{jl}\}$ with $d_{jl} = \sigma^{|j-l|}$, and ϵ_i was taken from the standard normal distribution $\mathcal{N}(0,1)$. The censoring time C_i was generated from a uniform distribution Unif$(0, \kappa \exp(|\widetilde{x}_{i1} - \widetilde{x}_{i2}|))$, where κ was chosen to achieve censoring rates of 20% and 40%. Set $\boldsymbol{\theta}_0 = (3, 1.5, 0, 0, 2, 0, \ldots, 0)$ 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 $(8(3n)^{1/5.1} - 14)$. We choose the bandwidth $h = \widehat{\sigma}_X n^{-1/5}$, where $\widehat{\sigma}_X$ 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 overidentified general estimating equations, we introduce an instrumental variable $U_i = (u_{i1}, \ldots, u_{ip})^T$ generated from the following models:

$$\text{MODEL C}: \quad u_{ij} \overset{\text{iid}}{\sim} \widetilde{x}_{ij} + \mathcal{N}(0,1), \quad j = 1, \ldots, p, \epsilon \sim \mathcal{N}(0,1),$$

$$\text{MODEL M}: \quad u_{ij} \overset{\text{iid}}{\sim} \widetilde{x}_{ij} + \mathcal{N}(1,1), \quad j = 1, \ldots, p, \epsilon \sim \mathcal{N}(0.5,1).$$

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

$$g(T, X; \boldsymbol{\theta}) = \big(\widetilde{x}_1\big(\log(T) - \widetilde{X}^T\boldsymbol{\theta}\big), \ldots, \widetilde{x}_p\big(\log(T) - \widetilde{X}^T\boldsymbol{\theta}\big), u_1\big(\log(T) - \widetilde{X}^T\boldsymbol{\theta}\big), \ldots, u_p\big(\log(T) - \widetilde{X}^T\boldsymbol{\theta}\big)\big)^T.$$

Tables 1 and 2 report the RMSE values of nonzero components in $\boldsymbol{\theta}_0$ 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

TABLE 1

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

		PEL			PET			PCU		
(n, p)		Oracle	SCAD	Lasso	Oracle	SCAD	Lasso	Oracle	SCAD	Lasso
(50, 7)	$\widehat{\theta}_1$	0.119	0.160	0.208	0.114	0.150	0.201	0.125	0.161	0.219
	$\widehat{\theta}_2$	0.105	0.140	0.186	0.102	0.136	0.183	0.106	0.143	0.203
	$\widehat{\theta}_5$	0.112	0.143	0.195	0.109	0.147	0.192	0.115	0.148	0.205
	T	4	3.74	2.2	4	3.78	3.08	4	3.63	3.02
	F	0	0.02	0.00	0	0	0	0	0.04	0.00
(100, 10)	$\widehat{\theta}_1$	0.110	0.121	0.155	0.113	0.125	0.160	0.115	0.123	0.166
	$\widehat{\theta}_2$	0.090	0.116	0.145	0.091	0.118	0.152	0.095	0.119	0.155
	$\widehat{\theta}_5$	0.106	0.117	0.150	0.107	0.120	0.151	0.110	0.124	0.155
	T	7	6.81	5.99	7	6.81	5.90	7	6.78	5.90
	F	0	0	0	0	0	0	0	0	0
(200, 14)	$\widehat{\theta}_1$	0.081	0.090	0.118	0.084	0.091	0.120	0.088	0.093	0.124
	$\widehat{\theta}_2$	0.064	0.067	0.115	0.068	0.070	0.119	0.073	0.078	0.120
	$\widehat{\theta}_5$	0.068	0.072	0.116	0.072	0.077	70.117	0.076	0.080	0.121
	T	11	11.00	10.02	11	11.00	10.08	11	10.97	10.08
	F	0	0	0	0	0	0	0	0	0

"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

		PEL			PET			PCU		
(n, p)		Oracle	SCAD	Lasso	Oracle	SCAD	Lasso	Oracle	SCAD	Lasso
(50, 7)	$\widehat{\theta}_1$	0.130	0.168	0.221	0.126	0.166	0.214	0.132	0.172	0.226
	$\widehat{\theta}_2$	0.118	0.150	0.197	0.109	0.145	0.196	0.115	0.152	0.212
	$\widehat{\theta}_5$	0.121	0.157	0.207	0.120	0.156	0.201	0.123	0.159	0.217
	T	4	3.44	2.01	4	3.50	2.10	4	3.44	2.03
	F	0	0.08	0.00	0	0	0	0	0.09	0.00
(100, 10)	$\widehat{\theta}_1$	0.120	0.129	0.163	0.121	0.131	0.165	0.123	0.132	0.175
	$\widehat{\theta}_2$	0.095	0.121	0.154	0.096	0.125	0.160	0.103	0.128	0.165
	$\widehat{\theta}_5$	0.112	0.124	0.160	0.115	0.127	0.161	0.119	0.131	0.165
	T	7	6.42	4.89	7	6.40	4.93	7	6.36	4.90
	F	0	0	0	0	0	0	0	0	0
(200, 14)	$\widehat{\theta}_1$	0.092	0.098	0.130	0.095	0.100	0.132	0.101	0.105	0.131
	$\widehat{\theta}_2$	0.074	0.076	0.122	0.078	0.082	0.120	0.080	0.083	0.124
	$\widehat{\theta}_5$	0.079	0.083	0.128	0.082	0.086	0.128	0.087	0.092	0.130
	T	11	10.72	9.57	11	10.60	9.42	11	10.56	9.39
	F	0	0	0	0	0	0	0	0	0

"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.

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 exam 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 (Li and Wang (2003)) and the Buckley–James method (Fang et al. (2013)) as follows:

$$\psi_n(Z_i;\boldsymbol{\theta}) = \frac{\Delta_i}{G_n(Y_i \mid X_i)}\{g(Y_i, X_i;\boldsymbol{\theta}) - \xi_n(X_i;\boldsymbol{\theta})\} + \xi_n(X_i;\boldsymbol{\theta}),$$

$$\psi_n^{\mathrm{SD}}(Z_i;\boldsymbol{\theta}) = \frac{\Delta_i g(Y_i, X_i;\boldsymbol{\theta})}{G_n(Y_i)},$$

$$\psi_n^{\mathrm{BJ}}(Z_i;\boldsymbol{\theta}) = \Delta_i g(Y_i, X_i;\boldsymbol{\theta}) + (1-\Delta_i)X_i \frac{\int_{\log(Y_i)-\widetilde{X}_i^T\boldsymbol{\theta}}^{\infty} t\, dF_n(t \mid \boldsymbol{\theta})}{1 - F_n(\log(Y_i) - \widetilde{X}_i^T\boldsymbol{\theta} \mid \boldsymbol{\theta})},$$

where $F_n(t \mid \boldsymbol{\theta})$ denotes the Kaplan–Meier estimator of distribution function of ϵ. Setting different vector forms of B_n in (4.1) leads to a confidence set for θ_1, θ_2 and θ_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

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 θ_1, θ_2 and θ_5 over 200 simulated data sets with $\sigma = 0.7$ under Model C

			$\alpha = 10\%$			$\alpha = 5\%$		
Par.	CR	(n, p)	Proposed	S-D	B–J	Proposed	S-D	B–J
θ_1	20%	(50, 7)	88.8	87.5	88.1	92.8	91.8	91.9
		(100, 10)	89.5	88.1	88.3	94.1	93.0	93.1
		(200, 14)	89.7	88.6	89.0	94.7	93.5	93.6
	40%	(50, 7)	87.1	85.7	85.9	91.9	89.5	89.8
		(100, 10)	88.0	86.4	86.8	93.0	90.6	91.1
		(200, 14)	89.2	87.0	87.3	93.9	92.3	92.6
θ_2	20%	(50, 7)	88.4	86.5	87.1	92.5	90.1	91.2
		(100, 10)	88.9	86.9	87.6	93.7	91.7	92.5
		(200, 14)	89.7	87.5	88.3	94.6	92.5	93.1
	40%	(50, 7)	87.5	85.2	86.2	91.2	89.2	89.8
		(100, 10)	87.9	85.9	86.8	92.4	90.1	91.1
		(200, 14)	89.2	86.6	87.4	93.7	91.3	92.2
θ_5	20%	(50, 7)	88.9	87.2	88.0	92.8	91.4	91.9
		(100, 10)	89.8	87.8	88.2	94.1	91.8	92.6
		(200, 14)	89.9	88.2	88.7	94.8	92.5	93.3
	40%	(50, 7)	87.4	85.1	85.7	91.9	89.0	89.5
		(100, 10)	88.2	85.8	86.6	92.9	90.3	90.8
		(200, 14)	89.1	86.5	87.2	94.2	91.8	92.1

Note: "Par." stands for parameter; "CR" denotes censoring rate; "S-D" stands for the Synthetic-data method (Li and Wang (2003)) and "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

	PET				PEL				PCU			
	$\sigma = 0.3$		$\sigma = 0.7$		$\sigma = 0.3$		$\sigma = 0.7$		$\sigma = 0.3$		$\sigma = 0.7$	
	BIAS	SSD	BIAS	SSD	BIAS	SSD	BIAS	SSD	BIAS	SSD	BIAS	SSD
$\widehat{\theta}_1$	0.058	0.114	0.056	0.104	0.750	0.840	0.642	1.220	0.579	0.442	0.534	0.318
$\widehat{\theta}_2$	0.133	0.165	0.120	0.158	0.673	1.236	0.534	1.223	0.610	0.536	0.603	0.316
$\widehat{\theta}_3$	0.000	0.008	0.001	0.005	0.241	0.832	0.224	0.522	0.211	0.180	0.202	0.176
$\widehat{\theta}_4$	0.001	0.007	0.000	0.003	0.233	0.633	0.223	0.422	0.224	0.160	0.212	0.150
$\widehat{\theta}_5$	0.121	0.196	0.125	0.193	0.850	1.033	0.643	1.216	0.746	0.336	0.741	0.318
$\widehat{\theta}_6$	0.001	0.004	0.000	0.001	0.240	0.532	0.224	0.422	0.250	0.178	0.230	0.189
$\widehat{\theta}_7$	0.001	0.005	0.000	0.003	0.234	0.630	0.214	0.520	0.234	0.190	0.221	0.168
T	3.62		3.71		1.52		1.70		1.56		1.80	
F	0		0		1.10		0.92		1.30		1.02	

function ρ. 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) and Grambsch, Therneau and Fleming (1995)). The data set, which can be downloaded from *R* package survival, consists of 424 observations and 19 variables including censored survival time (T), censoring indicator (Δ), 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 $\Delta_i = I(T_i \le C_i)$ 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

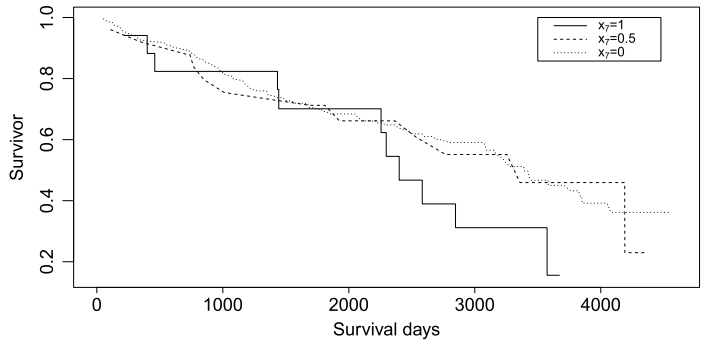

FIG. 1. *The Kaplan–Meier estimates of survival functions based on three groups in the presence of edema* ($x_7 = 0, 0.5, 1$) *with the PBC data.*

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 $\{0, 0, 0, 1/3, 2/3, 1, 1, 1, 1\}$ and $q_n = 5$.

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

Case I. Consider the moment functions:

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

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

$$j = 1, \ldots, 16,$$

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

$$l = 1, \ldots, 5.$$

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

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

$$l = 1, \ldots, 5,$$

and

$$g_{27+l}(T, X, \boldsymbol{\theta}) = B_l(x_{11}) \left\{ \log(T) - \beta_0 - \sum_{j=1}^{16} x_j\beta_j - \sum_{s=1}^{q_n} \gamma_s B_s(x_{17}) \right\},$$

$$l = 1, \ldots, 5.$$

TABLE 5

Estimated coefficients for the PBC data by various methods. The numbers in parentheses are p-values of the estimated nonzero coefficients

	Parameter	PEL		PET		PCU	
		SCAD	MCP	SCAD	MCP	SCAD	MCP
Case I	β_1	0	0	0	0	0	0
	β_2	−1.35 (0.00)	−1.30 (0.00)	−1.47 (0.00)	−1.42 (0.00)	−1.62 (0.00)	−1.57 (0.00)
	β_3	0.03 (0.08)	0	0.07 (0.03)	0.09 (0.03)	0.12 (0.01)	0.15 (0.01)
	β_4	−0.32 (0.00)	−0.27 (0.00)	−0.43 (0.00)	−0.38 (0.00)	−0.31 (0.00)	−0.28 (0.00)
	β_5	0	0	0	0	0	0
	β_6	0	0	0	0	0	0
	β_7	−0.86 (0.00)	−0.92 (0.00)	−1.06 (0.00)	−1.10 (0.00)	−0.96 (0.00)	−1.02 (0.00)
	β_8	0	0	0	0	0	0
	β_9	0.47 (0.00)	0.42 (0.00)	0.64 (0.00)	0.62 (0.00)	0.50 (0.00)	0.44 (0.00)
	β_{10}	−0.76 (0.00)	−0.73 (0.00)	−0.93 (0.00)	−0.87 (0.00)	−0.69 (0.00)	−0.67 (0.00)
	β_{11}	0.05 (0.12)	0	0.11 (0.22)	0.08 (0.30)	0.18 (0.10)	0
	β_{12}	−1.43 (0.00)	−1.36 (0.00)	−1.65 (0.00)	−1.58 (0.00)	−1.55 (0.00)	−1.47 (0.00)
	β_{13}	0.18 (0.00)	0.16 (0.00)	0.26 (0.00)	0.20 (0.00)	0.28 (0.00)	0.25 (0.00)
	β_{14}	0	0	0	0	0	0
	β_{15}	−0.80 (0.00)	−0.83 (0.00)	−0.95 (0.00)	−0.98 (0.00)	−0.91 (0.00)	−0.93 (0.00)
	β_{16}	−0.82 (0.00)	−0.85 (0.00)	−0.86 (0.00)	−0.90 (0.00)	−0.97 (0.00)	−0.99 (0.00)
	γ_1	1.73 (0.00)	1.75 (0.00)	1.83 (0.00)	1.89 (0.00)	1.86 (0.00)	1.90 (0.00)
	γ_2	0.81 (0.00)	0.82 (0.00)	0.71 (0.00)	0.77 (0.00)	0.90 (0.00)	0.93 (0.00)
	γ_3	0	0	0	0	0	0
	γ_4	0	0	0	0	0	0
	γ_5	−0.14 (0.00)	−0.18 (0.00)	−0.33 (0.00)	−0.38 (0.00)	−0.23 (0.00)	−0.24 (0.00)
Case II	β_1	0	0	0	0	0	0
	β_2	−1.17 (0.00)	−1.13 (0.00)	−1.32 (0.00)	−1.28 (0.00)	−1.25 (0.00)	−1.18 (0.00)
	β_3	0.19 (0.00)	0.17 (0.00)	0.29 (0.00)	0.26 (0.00)	0.25 (0.00)	0.21 (0.00)
	β_4	−0.32 (0.00)	−0.28 (0.00)	−0.35 (0.00)	−0.27 (0.00)	−0.44 (0.00)	−0.37 (0.00)
	β_5	0	0	0	0	0	0
	β_6	0	0	0	0	0	0
	β_7	−0.42 (0.00)	−0.47 (0.00)	−0.71 (0.00)	−0.78 (0.00)	−0.60 (0.00)	−0.64 (0.00)
	β_8	0	0	0	0	0	0
	β_9	0.48 (0.00)	0.43 (0.00)	0.81 (0.00)	0.77 (0.00)	0.67 (0.00)	0.63 (0.00)
	β_{10}	−0.79 (0.00)	−0.75 (0.00)	−0.84 (0.00)	−0.75 (0.00)	−0.62 (0.00)	−0.57 (0.00)
	β_{11}	0	0	0	0	0	0
	β_{12}	−1.45 (0.00)	−1.36 (0.00)	−1.52 (0.00)	−1.46 (0.00)	−1.76 (0.00)	−1.55 (0.00)
	β_{13}	0.27 (0.00)	0.24 (0.00)	0.42 (0.00)	0.39 (0.00)	0.35 (0.00)	0.33 (0.00)
	β_{14}	0	0	0	0	0	0
	β_{15}	−0.89 (0.00)	−0.96 (0.00)	−1.01 (0.00)	−1.09 (0.00)	−1.02 (0.00)	−1.06 (0.00)
	β_{16}	−0.84 (0.00)	−0.90 (0.00)	−0.98 (0.00)	−1.01 (0.00)	−0.87 (0.00)	−0.92 (0.00)
	γ_1	1.61 (0.00)	1.65 (0.00)	1.74 (0.00)	1.80 (0.00)	1.62 (0.00)	1.70 (0.00)
	γ_2	0.87 (0.00)	0.91 (0.00)	0.91 (0.00)	0.98 (0.00)	0.89 (0.00)	0.93 (0.00)
	γ_3	−0.39 (0.00)	−0.34 (0.00)	−0.57 (0.00)	−0.50 (0.00)	−0.48 (0.00)	−0.42 (0.00)
	γ_4	0	0	0	0	0	0
	γ_5	−0.20 (0.00)	−0.23 (0.00)	−0.52 (0.00)	−0.55 (0.00)	−0.30 (0.00)	−0.33 (0.00)

For the estimation of G_n and ξ_n, we utilized the dimension reduction method given in Section 5.4. and calculated the estimate of a 17×2 projection matrix α. For each case, we used the three PGEL methods with the SCAD and MCP penalties to estimate $\boldsymbol{\theta}$ and to select important covariates with the initial value of $\boldsymbol{\theta}$ taken from its GEL estimate without penalties. Estimates of nonzero regression coefficients in $\boldsymbol{\theta}$ 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}: $\theta_{1j} = 0$, that is, $H_{0j} : B_{nj}\boldsymbol{\theta}_1 = 0$, where B_{nj} is a $1 \times \widehat{q}$ vector with jth component being 1 and others being 0's, $j = 1, \ldots, \widehat{q}$, and $\widehat{q}$ 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 $k^3 p^2 = o(n)$ 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 $E(g(T, X; \boldsymbol{\theta}_0)) = 0$ and the semiparametric efficiency with this condition, while the Buckley–James method is based on the imputation approach. If we use the idea of the Buckley–James method, $g(T, X; \boldsymbol{\theta})$ should be estimated by its conditional expectation $E\{g(T, X; \boldsymbol{\theta}) \mid Y, X, \Delta\} = \Delta g(T, X; \boldsymbol{\theta}) + (1 - \Delta)E\{g(T, X; \boldsymbol{\theta}) \mid Y, X, \Delta\}$.

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

$$\ell(\widetilde{\lambda}, \widetilde{\boldsymbol{\theta}}) \le \inf_{\boldsymbol{\theta}\in\Theta} \sup_{\lambda\in\widehat{\Lambda}_n(\boldsymbol{\theta})} \ell(\lambda, \boldsymbol{\theta})) + o_p\big(n^{-\tau}\big),$$

where $\ell(\lambda, \boldsymbol{\theta})$ is as defined in (2.7), τ is nonnegative, and

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

Similarly, for nonsmooth moment functions, the PGEL estimator is defined to satisfy

$$\ell_p(\widehat{\lambda}, \widehat{\boldsymbol{\theta}}) \le \inf_{\boldsymbol{\theta}\in\Theta} \sup_{\lambda\in\widehat{\Lambda}_n(\boldsymbol{\theta})} \ell_p(\lambda, \boldsymbol{\theta}) + o_p\big(n^{-\tau}\big),$$

where $\ell_p(\lambda, \boldsymbol{\theta})$ is as defined in (2.8), and $\widehat{\lambda} = \arg\max_{\lambda\in\widehat{\Lambda}_n(\widehat{\boldsymbol{\theta}})} \ell_p(\lambda, \widehat{\boldsymbol{\theta}})$.

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 nonsmooth moment functions. Further research is needed to investigate the theoretical properties of $\widehat{\boldsymbol{\theta}}$ with nonsmooth 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.

The first author was supported by the National Natural Science Foundation of China (Grant No. 11671349) and the Key Projects of the National Natural Science Foundation of China (Grant No. 11731101).

The third author was supported by the Research Grant Council of Hong Kong (15301218) and the National Natural Science Foundation of China (No. 11771366), and Hong Kong Polytechnic University.

SUPPLEMENTARY MATERIAL

Supplement to "Penalized generalized empirical likelihood with a diverging number of general estimating equations for censored data" (DOI: 10.1214/19-AOS1870SUPP; .pdf). The supplement (Tang, Yan and Zhao (2019)) contains all technical proofs of Theorems 2.1, 3.1, 3.2, 4.1 and 5.1.

REFERENCES

ANDERSEN, P. K. and GILL, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10** 1100–1120. MR0673646

ANTONIADIS, A., FRYZLEWICZ, P. and LETUÉ, F. (2010). The Dantzig selector in Cox's proportional hazards model. *Scand. J. Stat.* **37** 531–552. MR2779635 https://doi.org/10.1111/j.1467-9469.2009.00685.x

BUCKLEY, J. and JAMES, I. (1979). Linear regression with censored data. *Biometrika* **66** 429–436.

CAO, Y., HUANG, J., LIU, Y. and ZHAO, X. (2016). Sieve estimation of Cox models with latent structures. *Biometrics* **72** 1086–1097. MR3591593 https://doi.org/10.1111/biom.12529

CHANG, J., CHEN, S. X. and CHEN, X. (2015). High dimensional generalized empirical likelihood for moment restrictions with dependent data. *J. Econometrics* **185** 283–304. MR3300347 https://doi.org/10.1016/j.jeconom.2014.10.011

CHANG, J., TANG, C. Y. and WU, T. T. (2017). A new scope of penalized empirical likelihood with high-dimensional estimating equations. Preprint. Available at arXiv:1704.00566.

CHEN, S. X., PENG, L. and QIN, Y.-L. (2009). Effects of data dimension on empirical likelihood. *Biometrika* **96** 711–722. MR2538767 https://doi.org/10.1093/biomet/asp037

CHEN, S. X. and VAN KEILEGOM, I. (2009). A review on empirical likelihood methods for regression. *TEST* **18** 415–447. MR2566404 https://doi.org/10.1007/s11749-009-0159-5

COX, D. R. (1972). Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B* **34** 187–220. MR0341758

DABROWSKA, D. M. (1989). Uniform consistency of the kernel conditional Kaplan–Meier estimate. *Ann. Statist.* **17** 1157–1167. MR1015143 https://doi.org/10.1214/aos/1176347261

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

FAN, J. and LI, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Ann. Statist.* **30** 74–99. MR1892656 https://doi.org/10.1214/aos/1015362185

FAN, J. and LV, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory* **57** 5467–5484. MR2849368 https://doi.org/10.1109/TIT.2011.2158486

FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194 https://doi.org/10.1214/009053604000000256

FANG, K.-T., LI, G., LU, X. and QIN, H. (2013). An empirical likelihood method for semiparametric linear regression with right censored data. *Comput. Math. Methods Med.* Art. ID 469373. MR3032121 https://doi.org/10.1155/2013/469373

FLEMING, T. R. and HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR1100924

GRAMBSCH, P. M., THERNEAU, T. M. and FLEMING, T. R. (1995). Diagnostic plots to reveal functional form for covariates in multiplicative intensity models. *Biometrics* **51** 1469–1482.

HANSEN, L. P., HEATON, J. and YARON, A. (1996). Finite-sample properties of some alternative GMM estimators. *J. Bus. Econom. Statist.* **14** 262–280.

HE, X., WANG, L. and HONG, H. G. (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *Ann. Statist.* **41** 342–369. MR3059421 https://doi.org/10.1214/13-AOS1087

HE, S., LIANG, W., SHEN, J. and YANG, G. (2016). Empirical likelihood for right censored lifetime data. *J. Amer. Statist. Assoc.* **111** 646–655. MR3538694 https://doi.org/10.1080/01621459.2015.1024058

JIN, Z., LIN, D. Y., WEI, L. J. and YING, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90** 341–353. MR1986651 https://doi.org/10.1093/biomet/90.2.341

KALBFLEISCH, J. D. and PRENTICE, R. L. (1980). *The Statistical Analysis of Failure Time Data*. *Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0570114

KITAMURA, Y. and STUTZER, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica* **65** 861–874. MR1458431 https://doi.org/10.2307/2171942

LENG, C. and MA, S. (2007). Path consistent model selection in additive risk model via Lasso. *Stat. Med.* **26** 3753–3770. MR2395831 https://doi.org/10.1002/sim.2834

LENG, C. and TANG, C. Y. (2012). Penalized empirical likelihood and growing dimensional general estimating equations. *Biometrika* **99** 703–716. MR2966779 https://doi.org/10.1093/biomet/ass014

LI, G. and WANG, Q.-H. (2003). Empirical likelihood regression analysis for right censored data. *Statist. Sinica* **13** 51–68. MR1963919

LIN, D. Y. (2003). Regression analysis of incomplete medical cost data. *Stat. Med.* **15** 1181–1200.

LIN, W. and LV, J. (2013). High-dimensional sparse additive hazards regression. *J. Amer. Statist. Assoc.* **108** 247–264. MR3174617 https://doi.org/10.1080/01621459.2012.746068

LIN, D. Y. and YING, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81** 61–71. MR1279656 https://doi.org/10.1093/biomet/81.1.61

LV, J. and FAN, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37** 3498–3528. MR2549567 https://doi.org/10.1214/09-AOS683

MARTINUSSEN, T. and SCHEIKE, T. H. (2009). Covariate selection for the semiparametric additive risk model. *Scand. J. Stat.* **36** 602–619. MR2572578 https://doi.org/10.1111/j.1467-9469.2009.00650.x

NEWEY, W. (1990). Semiparametric efficiency bounds. *J. Appl. Econometrics* **5** 99–135.

NEWEY, W. K. and SMITH, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* **72** 219–255. MR2031017 https://doi.org/10.1111/j.1468-0262.2004.00482.x

OWEN, A. B. (2001). *Empirical Likelihood*. CRC Press, New York.

PARENTE, P. M. D. C. and SMITH, R. J. (2011). GEL methods for nonsmooth moment indicators. *Econometric Theory* **27** 74–113. MR2771012 https://doi.org/10.1017/S0266466610000137

QIN, J. and LAWLESS, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22** 300–325. MR1272085 https://doi.org/10.1214/aos/1176325370

RITOV, Y. (1990). Estimation in a linear regression model with censored data. *Ann. Statist.* **18** 303–328. MR1041395 https://doi.org/10.1214/aos/1176347502

SUN, Q., ZHU, R., WANG, T. and ZENG, D. (2019). Counting process-based dimension reduction methods for censored outcomes. *Biometrika* **106** 181–196. MR3912390 https://doi.org/10.1093/biomet/asy064

TANG, C. Y. and LENG, C. (2010). Penalized high-dimensional empirical likelihood. *Biometrika* **97** 905–919. MR2746160 https://doi.org/10.1093/biomet/asq057

TANG, N., YAN, X. and ZHAO, X. (2019). Supplement to "Penalized generalized empirical likelihood with a diverging number of general estimating equations for censored data." https://doi.org/10.1214/19-AOS1870SUPP.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TIBSHIRANI, R. (1997). The lasso method for variable selection in the Cox model. *Stat. Med.* **16** 385–395.

TSIATIS, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.* **18** 303–328.

WU, T. T., LI, G. and TANG, C. (2015). Empirical likelihood for censored linear regression and variable selection. *Scand. J. Stat.* **42** 798–812. MR3391693 https://doi.org/10.1111/sjos.12137

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

ZHANG, H. H. and LU, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika* **94** 691–703. MR2410017 https://doi.org/10.1093/biomet/asm037

ZHOU, M. (2005). Empirical likelihood analysis of the rank estimator for the censored accelerated failure time model. *Biometrika* **92** 492–498. MR2201374 https://doi.org/10.1093/biomet/92.2.492

ZHOU, M. and LI, G. (2008). Empirical likelihood analysis of the Buckley–James estimator. *J. Multivariate Anal.* **99** 649–664. MR2406076 https://doi.org/10.1016/j.jmva.2007.02.007

ZHOU, X. H., QIN, G. S., LIN, H. Z. and LI, G. (2006). Inferences in censored cost regression models with empirical likelihood. *Statist. Sinica* **16** 1213–1232. MR2327487

ZHU, R., ZHANG, J., ZHAO, R., XU, P., ZHOU, W. and ZHANG, X. (2018). orthoDr: Semiparametric dimension reduction via orthogonality constrained optimization. Available at arXiv:1811.11733.

ZOU, H. (2008). A note on path-based variable selection in the penalized proportional hazards model. *Biometrika* **95** 241–247. MR2409726 https://doi.org/10.1093/biomet/asm083