Published in final edited form as:

J Stat Plan Inference. 2019 July ; 201: 71–85. doi:10.1016/j.jspi.2018.12.001.

Penalized Empirical Likelihood for the Sparse Cox Regression Model

Dongliang Wang¹, Tong Tong Wu², Yichuan Zhao³

¹Department of Public Health and Preventive Medicine, SUNY Upstate Medical University

²Department of Biostatistics and Computational Biology, University of Rochester

³Department of Mathematics and Statistics, Georgia State University

Abstract

The current penalized regression methods for selecting predictor variables and estimating the associated regression coefficients in the sparse Cox model are mainly based on partial likelihood. In this paper, a bias-corrected empirical likelihood method is proposed for the sparse Cox model in conjunction with appropriate penalty functions when the dimensionality of data is high. Theoretical properties of the resulting estimator for the large sample are proved. Simulation studies suggest that penalized empirical likelihood works better than partial likelihood in terms of selecting correct predictors without introducing more model errors. The well-known primary biliary cirrhosis data set is used to illustrate the proposed penalized empirical likelihood method.

Keywords

Coordinate descent algorithm; bias-corrected empirical likelihood; high-dimensional data; oracle property; penalized likelihood; sparse proportional hazards model; right censoring; Wilks' theorem

1 Introduction

The proportional hazards models, also known as the Cox models (Cox 1972, 1975), are widely used to identify the risk factors given time to event data. Optimal model selection and estimation procedures are desired to correctly select the true risk factors and yield efficient estimators for regression coefficients. The model selection techniques for linear regression models, such as best-subset selection, stepwise deletion, Akaike information criterion (AIC) and Bayesian information criterion (BIC), can be easily extended for the Cox models. Faraggi and Simon (1998) and Ibrahim *et al.* (1999) discussed Bayesian variable selection and Sauerbrei and Schumacher (1992) proposed a bootstrap procedure for the Cox model. From the vantage point of penalized regression, Tibshirani (1997) and Fan and Li (2002) extended the least absolute shrinkage and selection operator (LASSO) and the smoothly clipped absolute deviation (SCAD) methods, firstly proposed in Tibshirani (1996) and Fan and Li (2001), respectively. Zhang and Lu (2007) investigated the adaptive LASSO method, where the penalty associated with each coefficient is weighted by its magnitude. Both the LASSO and the adaptive LASSO methods have a convex form for the ease of global optimization, but the adaptive LASSO also enjoys the oracle property as the SCAD method.

All the above mentioned penalized likelihoods for the sparse Cox models are defined upon the partial likelihood function. Many other studies were related to penalized likelihood estimations, including Zucker and Karr (1990), Gu (1996) and Goeman (2010). More recently with the motivation to handle ultra high-dimensional data, novel statistical methods for censored data have been developed from a variety of perspectives, such as quantile regression by He *et al.* (2013), accelerated failure time models by Xia *et al.* (2016) and survival impact index by Li *et al.* (2017).

The goal of this paper is to investigate the penalized empirical likelihood method for the Cox model, as a competitor to the penalized partial likelihood. Since empirical likelihood (EL) was proposed by Owen (1988, 2001), empirical likelihood methods have been investigated well from many statistical perspectives. With regards to the model selection, Variyath *et al.* (2009) investigated the properties of the information criteria with the empirical likelihood. Tang and Leng (2010) proposed penalized empirical likelihood methods for the mean vector estimation and linear regression with the divergent number of dimensions, in conjunction with appropriate penalty functions. Leng and Tang (2012) extended their work for more general estimation equations with growing dimensionality. Other work about the impact of growing dimensionality on the inference of empirical likelihood can be found in Hjort *et al.* (2009), Chen and Keilegom (2009), and Chen *et al.* (2009). For right censored data, Zhao and Huang (2007) proposed the bias-corrected empirical likelihood for the accelerated failure time model based on estimated influence functions. Motivated by Zhao and Huang (2007), Wu *et al.* (2015) developed the penalized empirical likelihood method for the censored accelerated failure time model. For the Cox's proportional hazards model, Hou *et al.* (2014) investigated the performance of the penalized empirical likelihood estimator via the bridge penalty. Their empirical likelihood was derived with known baseline hazard function using the method in Qin and Jing (2001), which makes their approach not suitable or applicable in practice. In this paper, we aim to develop the penalized empirical likelihood (PEL) method for sparse Cox model from the following two vantage points: (1) a bias-corrected empirical likelihood, proposed in Sun *et al.* (2009), Zhao and Jinnah (2012) and Zheng and Yu (2013) with a plug-in cumulative baseline hazard function, is evaluated in order to facilitate the application of the method to real data, and; (2) more general requirements regarding the penalty function to maintain the sparsity and asymptotic normality of the coefficient estimators are explored, so that penalty functions such as LASSO and SCAD can be applied with no any concerns.

In the rest of this paper, we study the property of the bias-corrected empirical likelihood in conjunction with appropriate penalty functions for parameter estimation and model selection. In Section 2.1, we define a penalized empirical likelihood after providing some preliminary results about the Cox model. We investigate the asymptotic properties of the penalized empirical likelihood in Section 2.2, and discuss the optimization and tuning procedure in Section 2.3. Results from extensive simulation studies are presented in Section 3 to evaluate the asymptotic properties of the existing methods, the bias-corrected empirical likelihood and penalized empirical likelihood. We illustrate the proposed method by re-analyzing a well-known primary biliary cirrhosis (PBC) data set in Section 3.2, followed by the discussions in Section 4. The proofs of theorems are given in the Appendix.

2 Methodology

We start this section by restating some preliminary results about the Cox model and introducing notations that will be used throughout the paper. Let T , C , and $x = (x_1, 2, \dots, x_p)^\tau$ be the survival time, the censoring time and the associated vector of covariates, respectively. Assume that given x , T and C are independent. Let $Z = \min(T, C)$ be the observed time and $\delta = \mathbb{I}(T < C)$ be the censoring indicator, where $\mathbb{I}(\cdot)$ is the indicator function. The proportional hazards model proposed by Cox (1972) is given as

$$h(t|x) = h_0(t)\exp(x^\tau(t)\beta) \quad (2.1)$$

with the unspecified baseline hazard function $h_0(\cdot)$ and the parameter vector of interest β .

Define $N_i(t) = \delta_i \mathbb{I}(Z_i \leq t)$ and $Y_i(t) = \mathbb{I}(Z_i \geq t)$, $i = 1, \dots, n$, as the counting and at-risk process, respectively. The parameter vector β_0 is commonly estimated by maximizing the log partial likelihood

$$l(\beta) = \sum_{i=1}^n \int_0^1 x_i^\tau(s)\beta - \log\left(\sum_{i=1}^n Y_i(s)\exp(x_i^\tau(s)\beta)\right) dN_i(s). \quad (2.2)$$

For the simplicity, we assume no ties in the observed event times. When ties are present, the log partial likelihood at (2.2) can be readily rewritten; see more details in Breslow (1974). For brevity, x will be used in the rest of this paper to stand for the time-varying covariate $x(t)$.

For any vector a , let $a^{\otimes 0}(t) = 1$, $a^{\otimes 1} = a$ and $a^{\otimes 2} = aa^\tau$. For $r = 0, 1, 2$, let

$$\begin{aligned} \alpha^{(r)}(\beta, t) &= E\left[x^{\otimes r}(t)Y(t)\exp(x^\tau\beta)\right], \\ \omega(\beta, t) &= \frac{\alpha^{(2)}(\beta, t)}{\alpha^{(0)}(\beta, t)} - \left(\frac{\alpha^{(1)}(\beta, t)}{\alpha^{(0)}(\beta, t)}\right)\left(\frac{\alpha^{(1)}(\beta, t)}{\alpha^{(0)}(\beta, t)}\right)^\tau. \end{aligned}$$

The information matrix for the partial likelihood at equation (2.2) can be expressed as

$$I(\beta) = \int_0^1 \omega(\beta, t)\alpha^{(0)}(\beta, t)h_0(t)dt. \quad (2.3)$$

It is well known that under certain regularity conditions such as (A)–(D) in Andersen and Gill (1982),

$$\sqrt{n}(\hat{\beta}_{MLE} - \beta_0) \xrightarrow{d} N(0, I^{-1}(\beta_0)),$$

where $\hat{\beta}_{MLE}$ and β_0 denote the maximum likelihood estimator from equation (2.2) and the true value of β , respectively. Thus an asymptotic $100(1 - \alpha)\%$ confidence region for β_0 can be constructed as follows,

$$R_\alpha = \left\{ \beta : n(\hat{\beta}_{MLE} - \beta)^T \hat{I}(\hat{\beta}_{MLE})(\hat{\beta}_{MLE} - \beta) \leq \chi_p^2(\alpha) \right\},$$

where $\hat{I}(\hat{\beta}_{MLE})$ is an empirical estimator of $I(\beta_0)$ at equation (2.3) by plugging in $\hat{\beta}_{MLE}$ for β and $\chi_p^2(\alpha)$ is the $(1 - \alpha)$ th quantile of the chi-square distribution with p degrees of freedom.

2.1 Penalized empirical likelihood for variable selection

Taking the first derivative of the log partial likelihood, the score function $U(\beta)$ can be readily derived as

$$U(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^n \int_0^\infty [x_i - \bar{x}(\beta, t)] dM_i(t), \quad (2.4)$$

where

$$\bar{x}(\beta, t) = \frac{\sum_{i=1}^n Y_i(t) x_i \exp(x_i^T \beta)}{\sum_{i=1}^n Y_i(t) \exp(x_i^T \beta)}, \quad (2.5)$$

$$M_i(\beta, t) = N_i(t) - \int_0^t Y_i(u) \exp(x_i^T \beta) h_0(u) du.$$

It has been well established in Andersen and Gill (1982) that $M_i(\beta_0, t)$ are local martingales and $EU(\beta_0) = 0$. Based on $U(\beta_0)$, an empirical likelihood method has been proposed by Qin and Jing (2001) to make inference for β_0 . However, the resulting likelihood function involves unspecified baseline hazard function $h_0(t)$, which poses severe difficulty for the application in the real data. To overcome the problem, we replace $h_0(u)du$ with $d\hat{H}(\beta_0, t)$,

where

$$\hat{H}(\beta_0, t) = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{i=1}^n Y_i(u) \exp(x_i^T \beta_0)} \quad (2.6)$$

$$= \sum_{z_i \leq t} \frac{\delta_i}{\sum_{i=1}^n Y_j(z_i) \exp(\mathbf{x}_j^\tau \boldsymbol{\beta}_0)}. \quad (2.7)$$

It leads to

$$\widehat{M}_i(\boldsymbol{\beta}_0, t) = N_i(t) - \int_0^t Y_i(u) \exp(\mathbf{x}_i^\tau \boldsymbol{\beta}_0) d\widehat{H}(\boldsymbol{\beta}_0, u).$$

Sun *et al.* (2009) and Zhao and Jinnah (2012) proposed the bias-corrected empirical likelihood for the Cox model based on influence functions using plug-in estimators. We follow the method in Zhao and Jinnah (2012) to develop the penalized empirical likelihood with the bias correction. Towards this end, let

$$\mathbf{W}_{ni}(\boldsymbol{\beta}) = \int_0^\infty [x_i - \bar{\mathbf{x}}(\boldsymbol{\beta}, t)] d\widehat{M}_i(\boldsymbol{\beta}, t). \quad (2.8)$$

It is known in p. 4 of Zhao and Jinnah (2012), the score function satisfies

$$U(\boldsymbol{\beta}_0) = \sum_{i=1}^n \mathbf{W}_{ni}(\boldsymbol{\beta}_0).$$

The bias-corrected empirical likelihood ratio statistic based on $\mathbf{W}_{ni}(\boldsymbol{\beta})$'s can be written as (cf. pp. 4–5 of Zhao and Jinnah (2012) for details)

$$l_e(\boldsymbol{\beta}) = - \sum_{i=1}^n \log(1 + \boldsymbol{\lambda}^\tau \mathbf{W}_{ni}(\boldsymbol{\beta})), \quad (2.9)$$

where $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_p)^\tau$ satisfies

$$n^{-1} \sum_{i=1}^n \frac{\mathbf{W}_{ni}(\boldsymbol{\beta})}{1 + \boldsymbol{\lambda}^\tau \mathbf{W}_{ni}(\boldsymbol{\beta})} = 0.$$

We define the maximum penalized empirical likelihood estimator (MPELE) $\hat{\boldsymbol{\beta}}$ as the solution to maximize the penalized empirical likelihood, which takes a general form as

$$l_p(\boldsymbol{\beta}) = l_e(\boldsymbol{\beta}) - n \sum_{i=1}^p p_\gamma(|\beta_i|), \quad (2.10)$$

where $p_\gamma(|\beta_j|)$ is a penalty function with the tuning parameter γ . Specifically, $p_\gamma(|\beta_j|) = \gamma|\beta_j|$ for the LASSO; $p_\gamma(|\beta_j|) = \gamma^2 - (|\beta_j| - \gamma)^2 I(|\beta_j| < \gamma)$ for the SCAD; and $p_\gamma(|\beta_j|) = \gamma|\beta_j|\omega_j$ for the adaptive LASSO.

2.2 Asymptotic properties

We establish asymptotic properties of the penalized empirical likelihood and the associated MPELE under the framework of counting process. Without loss of generality, we assume that $t \in [0, 1]$. One may extend the results to the interval $[0, \infty)$ as in Andersen and Gill (1982). The covariate vector $\mathbf{x}$ is allowed to be time dependent, denoted by $\mathbf{x}(t)$. Throughout this paper, we assume that the regularity conditions (A)–(D) specified in Andersen and Gill (1982) hold in order to maintain the local asymptotic quadratic property for the partial likelihood.

Let

$$a_n = \max \left\{ p'_{\gamma_n}(|\beta_{j0}|) : \beta_{j0} \neq 0 \right\}, \quad (2.11)$$

$$b_n = \max \left\{ p''_{\gamma_n}(|\beta_{j0}|) : \beta_{j0} \neq 0 \right\}. \quad (2.12)$$

Theorem 2.1. Assume that T and C are conditionally independent given $\mathbf{x}$, conditions (A)–(D) hold and $a_n = O_p(n^{-1/2})$. Then as $b_n \rightarrow 0$, the maximum penalized empirical likelihood estimator $\hat{\beta}$ satisfies $\|\hat{\beta} - \beta_0\| = O_p(n^{-1/2})$, where $\hat{\beta}$ and $\hat{\lambda}$ are the solutions to $\mathcal{Q}_{1n}(\beta, \lambda) = 0$ and $\mathcal{Q}_{2n}(\beta, \lambda) = 0$, where

$$\mathcal{Q}_{1n}(\beta, \lambda) = n^{-1} \sum_{i=1}^n \frac{W_{ni}(\beta)}{1 + \lambda^\tau W_{ni}(\beta)}, \quad (2.13)$$

$$\mathcal{Q}_{2n}(\beta, \lambda) = n^{-1} \sum_{i=1}^n \frac{1}{1 + \lambda^\tau W_{ni}(\beta)} \frac{\partial W_{ni}(\beta)}{\partial \beta} \lambda + \mathbf{b}(\beta), \quad (2.14)$$

with $\mathbf{b}(\beta) = \left(p'_{\gamma_n}(|\beta_1|) \text{sgn}(\beta_1), \dots, p'_{\gamma_n}(|\beta_q|) \text{sgn}(\beta_q) \right)^\tau$.

Next, we show the sparsity and asymptotic normality of $\hat{\beta}$. Without loss of generality, Let $\beta_0 = (\beta_{10}^\tau, \beta_{20}^\tau)^\tau$ denote the true parameter vector, where β_{10} and β_{20} consists of all q nonzero components and the remaining $p - q$ zero components, respectively. Let

$$\begin{aligned}
\Sigma(\beta_0) &= \text{diag}\left(p''_{\gamma_n}(|\beta_{10}|), \dots, p''_{\gamma_n}(|\beta_{p0}|)\right), \\
b(\beta_0) &= \left(p'_{\gamma_n}(|\beta_{10}|)\text{sgn}(\beta_{10}), \dots, p'_{\gamma_n}(|\beta_{p0}|)\text{sgn}(\beta_{p0})\right)^\tau, \\
\Sigma_1(\beta_0) &= \text{diag}\left(p''_{\gamma_n}(|\beta_{10}|), \dots, p''_{\gamma_n}(|\beta_{q0}|)\right), \\
b_1(\beta_0) &= \left(p'_{\gamma_n}(|\beta_{10}|)\text{sgn}(\beta_{10}), \dots, p'_{\gamma_n}(|\beta_{q0}|)\text{sgn}(\beta_{q0})\right)^\tau.
\end{aligned}$$

Theorem 2.2. Assume that

$$\liminf_{n \rightarrow \infty} \liminf_{\theta \rightarrow 0^+} p'_{\gamma_n}(\theta)/\gamma_n > 0, \quad (2.15)$$

$\gamma_n \rightarrow 0, n^{-1/2}\gamma_n^{-1} \rightarrow 0, a_n = O_p(n^{-1/2})$ and the conditions in Theorem 2.1 hold. Then as $b_n \rightarrow 0$, the maximum penalized empirical likelihood estimator $\hat{\beta} = (\hat{\beta}_1^\tau, \hat{\beta}_2^\tau)^\tau$ satisfies:

- (Sparsity of $\hat{\beta}$) $\hat{\beta}_2 = 0$ with probability tending to 1;
- (Asymptotic normality of $\hat{\beta}$)

$$\sqrt{n}(\mathbf{I}_1(\beta_{10}) + \Sigma_1(\beta_0))[\hat{\beta}_1 - \beta_{10} + (\mathbf{I}_1(\beta_{10}) + \Sigma_1(\beta_0))^{-1}b_1(\beta_0)] \xrightarrow{d} N(0, \mathbf{I}_1(\beta_{10})),$$

(2.16)

where $\mathbf{I}_1(\beta_{10})$ is the first $q \times q$ submatrix of the information matrix $\mathbf{I}(\beta_0)$.

- (Asymptotic normality of $\hat{\lambda}$)

$$\sqrt{n}(\hat{\lambda} + \mathbf{V}^{-1}b(\beta_0)) \xrightarrow{d} N(0, \mathbf{U}),$$

where $\mathbf{U} = (\mathbf{I}^{-1}(\beta_0) - \mathbf{V}^{-1})\mathbf{I}^{-1}(\beta_0)(\mathbf{I}^{-1}(\beta_0) - \mathbf{V}^{-1})^\tau$ and $\mathbf{V} = \Sigma(\beta_0) + \mathbf{I}(\beta_0)$.

Theorem 2.3. For testing $H_0: \mathbf{L}\beta_0 = 0$ where $\mathbf{L}$ is a $l \times p$ matrix such that $\mathbf{L}\mathbf{L}^\tau = \mathbf{I}_l$ for a fixed and finite l , the penalized profile empirical likelihood ratio statistic is

$$\phi(\mathbf{L}) = -2 \left(l_p(\hat{\beta}) - \min_{\beta: \mathbf{L}\beta_0 = 0} l_p(\beta) \right). \quad (2.17)$$

Assume the conditions in Theorem 2.2 hold, $a_n = o_p(n^{-1/2})$ and $b_n = o_p(1)$, then as $n \rightarrow \infty$,

$$\phi(\mathbf{L}) \xrightarrow{d} \chi^2_I \text{ under } H_0. \quad (2.18)$$

Therefore a $100(1 - \alpha)\%$ penalized bias-corrected EL confidence region for $\mathbf{L}\boldsymbol{\beta}_0$ can be constructed as follows,

$$V_\alpha = \left\{ \boldsymbol{\beta}: -2 \left(l_p(\hat{\boldsymbol{\beta}}) - \min_{\boldsymbol{\beta}: \mathbf{L}\boldsymbol{\beta}_0 = 0} l_p(\boldsymbol{\beta}) \right) \leq \chi^2_I(\alpha) \right\},$$

where $\chi^2_I(\alpha)$ is the $(1 - \alpha)$ th quantile of the chi-square distribution with I degrees of freedom.

2.3 Optimization using nested coordinate descent

To optimize the penalized bias-corrected empirical likelihood at equation (2.10), we adopt the nested coordinate descent algorithm proposed by Tang and Wu (2014), in conjunction with the majorization-minimization (MM) algorithm (Hunter and Lange, 2004; Lange, 2004; Wu and Lange, 2010). The nested coordinate descent algorithm is employed to ensure effective and stable computational performance in solving EL optimization problems. It is a two-layer nested algorithm that is constructed by an inner layer of evaluating the EL at a given parameter and an outer layer of optimizing the EL with respect to the parameter.

A common practice is to use the pseudo-logarithm function $\log_\star(\cdot)$ of the bias-corrected empirical likelihood at (2.9) to ensure the closure property required for the convergence (Owen 2001). The pseudo-logarithm function is the second degree Taylor approximation of the logarithm at the point $1/n$ and is twice differentiable. In the inner layer with a fixed value of $\boldsymbol{\beta}$, one can readily optimize the bias-corrected empirical likelihood $l_\star(\boldsymbol{\beta})$ at equation (2.9) with respect to the Lagrange multiplier $\boldsymbol{\lambda}$. Suppose $\boldsymbol{\lambda}$ is started from its initial value $\boldsymbol{\lambda}^{(0)}$. Given the other coordinates fixed, the $(m + 1)$ th Newton's update of λ_k , $k = 1, \dots, p$, is given by

$$\hat{\lambda}_k^{(m+1)} = \hat{\lambda}_k^{(m)} - \frac{\sum_{i=1}^n \log'_\star(t_i^{(m)}) \cdot W_k(\boldsymbol{\beta})}{\sum_{i=1}^n \log''_\star(t_i^{(m)}) \cdot \{W_k(\boldsymbol{\beta})\}^2}, \quad (2.19)$$

where $t_i^{(m)} = 1 + (\hat{\lambda}^{(m)})^\tau \mathbf{W}_{ni}(\boldsymbol{\beta})$ and $W_k(\boldsymbol{\beta})$ is the k th component of $\mathbf{W}_{ni}(\boldsymbol{\beta})$. All the p components of $\boldsymbol{\lambda}$ get updated by turn. This procedure is repeated until the convergence. The coordinate descent procedure (2.19) can be viewed as a univariate sequential optimization. Its convergence rate and stability are studied in optimization literature; see for example in Gillett *et al.* (1981).

In the outer layer at a given $\boldsymbol{\lambda}$, one can optimize the penalized empirical likelihood at equation (2.10) with respect to $\boldsymbol{\beta}$. The coordinate algorithm updates β_j , $j = 1, \dots, p$, by

minimizing $\tilde{\ell}_\star(\beta, \lambda)$ w.r.t. β_j with other $\beta_i, i \neq j$ fixed. The $(m+1)$ th iterative Newton's update of β_j is given by

$$\hat{\beta}_j^{(m+1)} = \hat{\beta}_j^{(m)} - \frac{\sum_{i=1}^n \log'_\star(s_i^{(m)}) \left[\lambda^\tau \frac{\partial W_{ni}(\beta^{(m)})}{\partial \beta_j} \right]}{\sum_{i=1}^n \left\{ \log''_\star(s_i^{(m)}) \left[\lambda^\tau \frac{\partial W_{ni}(\beta^{(m)})}{\partial \beta_j} \right]^2 + \log'_\star(s_i^{(m)}) \left[\lambda^\tau \frac{\partial^2 W_{ni}(\beta^{(m)})}{\partial \beta_j^2} \right] \right\}}, \quad (2.20)$$

where $s_i^{(m)} = 1 + \lambda^\tau W_{ni}(\beta^{(m)})$ and $\hat{\theta}_j^{(0)}$ is an initial value. Note that in (2.20) λ actually depends on $\hat{\beta}^{(m)}$ by the definition (2.9). This implies that upon updating one component β_j , λ needs to be updated. Compared to a joint updating algorithm that updates all components in β , the update (2.20) is advantageous in its stability because coordinate-wise updates of components in β can effectively avoid drastic changes in $\ell(\beta)$ and λ .

As we have shown above, both layers of optimizations are solved by the cyclic version of coordinate descent. This iterative optimization algorithm can be modified using the MM algorithm with the parameter updated by

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} - D^{-1} \nabla \ell_\star(\hat{\beta}^{(m)}),$$

where the matrix D can be considered as an approximated Hessian matrix and satisfies $D \geq d^2 \ell_\star(\beta)$, i.e., $D - d^2 \ell_\star(\beta)$ is positive definite. This modification can further simplify the optimizations when D is well chosen. The readers are referred to Tang and Wu (2014) for more details.

To select the appropriate tuning parameter γ , we propose to minimize the Bayesian Information Criterion (BIC), defined as

$$BIC(\gamma) = -2l_p(\beta_\gamma) + d_\gamma \log n, \quad (2.21)$$

where d_γ is the number of non-zero coefficient estimate $\hat{\beta}$ for a fixed γ . Other BIC-type criteria, such as the BICC proposed by Wang *et al.* (2009) and EBIC proposed by Chen and Chen (2008), can also be used for tuning. It has been shown by Wang *et al.* (2007) and Wang and Leng (2007) in the least square regression setting that the true model can be consistently identified by SCAD and adaptive LASSO methods with the tuning parameter determined by

a BIC-type criterion. Variyath *et al.* (2010) developed a BIC approach to the model selection problem with empirical likelihood for the generalized linear models.

3 Simulation Studies and Applications

3.1 Simulation studies

In this section, we evaluate the performance of the penalized bias-corrected empirical likelihood in the Cox model in model selection and parameter estimation and compare it with the penalized partial likelihood method, for the purpose of investigating the potential advantages of replacing partial likelihood with the bias-corrected empirical likelihood in the context of the penalized Cox model. The SCAD penalty in Fan and Li (2001) is used as the penalty function in (2.10) for direct comparison between the bias-corrected empirical likelihood and partial likelihood.

The predictor $x_i, i = 1, 2, \dots, n$ is generated from a multivariate normal distribution with mean zero and a compound symmetry variance-covariance matrix with $\sigma = 1$ and ρ , that is,

$$\text{Cov}(x_{ij}, x_{ik}) = \begin{cases} 1 & \text{for } j = k \\ \rho & \text{for } j \neq k \end{cases}$$

for $i = 1, \dots, n, j = 1, \dots, p$, and $k = 1, \dots, p$. Both independent ($\rho = 0$) and correlated cases ($\rho = 0.5$ and 0.9) are considered. The true hazard function is defined in equation (2.1) with $h_0(t) = 1$ and $\beta_0 = (3, 1.5, 0, 0, 2, 0, \dots, 0)^T$. In other words, true survival time T_i is generated by

$$T_i = -\frac{\log(U)}{x_i^T \beta_0},$$

where U is uniformly distributed between 0 and 1. The censoring time is generated from a uniform distribution from 1 to 10. The average censoring rate is 35%, estimated by simulations.

In Tables 1 and 2 we report the results for different combinations of (n, p, ρ) . Our penalized empirical likelihood method is combined with three information criteria – BIC, BICC, and EBIC – for selecting the tuning parameter τ . The results are compared with the oracle Cox model (i.e., containing only the true predictors) and two R packages for the penalized Cox models, *penalized*() and *ncvreg* (Breheny and Huang 2011), which use the LASSO and SCAD penalties, respectively. The column of N_{nonzero} reports the average total number of selected predictors and the column of N_{true} reports the average total number of selected true predictors from (x_1, x_2, x_5) . The results show that both the PELs and *ncvreg* methods select all three true predictors. The performance of both BIC and BICC in conjunction with the PEL is comparable, while the selection from the EBIC is usually more accurate. As compared to the penalized Cox model from an R package *ncvreg*, the PEL with EBIC penalty almost uniformly leads to less false positives, which are defined as those β 's with zero coefficient but wrongly included in the model. All methods tend to select more false positive predictors, when the predictors are correlated.

It is worthy mentioning that for the simulation study, a fixed range of lambda values for Lagrange multiplier is applied for grid searching. Occasionally the penalized empirical likelihood is maximized at the boundary of the range, leading to all zero coefficient estimates and the results are not included in Tables 1 and 2. To evaluate the overall performance in terms of parameter estimation, the last column in Tables 1 and 2 reports the model error (ME) defined by $ME = \|\hat{\beta} - \beta\|_2^2$. Standard errors are reported in the parentheses.

The model error of the oracle model, which only uses the non-zero predictors, is also reported in the last column of each setting. It is uniformly true that all penalized methods are superior to the un-penalized method. The penalized empirical likelihood methods have model errors slightly larger than those from the oracle estimation and as small as half of those from the penalized Cox method from *penalized*. The penalized empirical likelihood tends to be superior to *ncvreg*, except for the cases where *ncvreg* has smaller model errors than the oracle method (which is a bit surprising).

3.2 Primary Biliary Cirrhosis (PBC) Data Analysis

For the purpose of illustration, we re-analyzed the well-known PBC data set, taken from Appendix D of Fleming and Harrington (1991). Dickson *et al.* (1989) originally published the data set for the Mayo Clinic trial of the liver PBC conducted between January 1974 and May 1984, comparing the drug D-penicillamine (DPCA) with a placebo. The data set was re-analyzed by Tibshirani (1997) using LASSO and by Zhang and Lu (2007) using SCAD and adaptive LASSO respectively. Previous analyses included 276 out of 312 randomized subjects with no missing values for all variables, in order to assess the contribution of 17 risk factors to the survival time: *drug* (1 = D-penicillamine, 2 = placebo); *Age* (in years); *Sex* (0=male, 1=female); *Ascites* - Presence of ascites (0=no, 1=yes); *Hepato* - Presence of hepatomegaly (0=no, 1=yes); *Spiders* - Presence of spiders (0=no, 1=yes); *Oedema* - Presence of oedema (0=no, 0.5 = yes but responded to diuretic treatment, 1=yes, did not respond to treatment); *Bili* - Serum bilirubin (in mg/dl); *Chol* - Serum cholesterol (in mg/dl); *Albumin* (in g/dl); *Copper* - Urine copper (in kg/day); *Alk phos* - Alkaline phosphatase (in U/litre); *SGOT* (in U/ml); *Trig* - Triglycerides (in mg/dl); *Platelet* - Platelet count (in count/ml³ of blood divided by 1000), *PT*-Prothrombine time (in seconds); *Stage* - Histologic state of disease (graded 1, 2, 3 or 4).

We re-analyze the data set using the proposed penalized empirical likelihood method. Missing covariates are imputed with the mean values and all 312 subjects are included in the analysis. The results are shown in Table 3, together with the results from the unpenalized Cox model and the penalized Cox model with LASSO, using the R package *penalized*. The seven out of eight predictors selected from both PEL-BIC and PEL-BICC are the same as those selected by LASSO, SCAD and Adaptive LASSO in Tibshirani (1997) and Zhang and Lu (2007), namely, *Oedema*, *Bili*, *Albumin*, *Copper*, *SGOT*, *PT* and *Stage*. It is worthy noting that the predictor *drug* is also selected using our methods, which has not been selected previously (*Age* is selected instead). This selection of *Drug* is in line with the evidence from a systemic meta analysis in Gong *et al.* (2006) that D-penicillamine compared with placebo/no intervention tends to increase mortality. The PEL-EBIC method only selects four predictors, including *Bili*, *Albumin*, *Copper*, and *Stage*, which is consistent with the observation from our simulation study that the PEL-EBIC method appears to select fewer

number of predictors. The four predictors selected from the PEL-EBIC method are the only four predictors from both the LASSO in Table 1 of Tibshirani (1997) and the Adaptive LASSO in Table 5 of Zhang and Lu (2007) with the absolute value of the Z-score greater than 1.96.

4 Discussions

In this paper, we evaluate the performance of the bias-corrected empirical likelihood for variable selection and parameter estimation of the sparse Cox model in conjunction with commonly used penalty functions, as an alternative to the partial likelihood. The resulting parameter estimators share desirable asymptotic properties with the maximized penalized partial likelihood estimators, and demonstrate superior empirical performance, particularly from the perspective of model selection.

The proposed bias-corrected empirical likelihood method is developed under the framework of the sparse Cox regression model, and thus its performance relies on the proportional hazards assumption. The method should not be considered unless the Cox model is deemed as appropriate for real data. In case that the proportional hazards assumption is violated, one may use model-free selection approaches, such as those recently proposed by He *et al.* (2013) and Li *et al.* (2016). The current work is limited to dealing with a fixed number of parameters and further efforts may be taken to extend the methods when the dimensionality diverges. One potential way is to generalize the general multivariate model that was firstly proposed by Chen *et al.* (2009) and applied by Tang and Leng (2010) for linear regression with the diverging dimensionality.

Acknowledgment

We are grateful to the Editor, the Associate Editor and the Reviewers for their constructive comments and insightful suggestions, which lead to significant improvement of the manuscript. Wu's research was partly supported by NIH grants 2R01HLI9058-03A1, 4P30AI078498-09, and 4R01MH075017-09. Zhao was supported by NSA Grant (H98230-12-1-0209) and NSF Grant (DMS-1613176).

A Appendix: Proofs of Theorems

A.1 Proof of Theorem 2.1.

Let

$$\bar{W}_n(\beta) = n^{-1} \sum_{i=1}^n W_{ni}(\beta)$$

$$\hat{S}(\beta) = n^{-1} \sum_{i=1}^n W_{ni}(\beta) W_{ni}^T(\beta).$$

Lemma A.1. *Under regularity conditions (A)-(D), then*

$$\sqrt{n}\bar{W}_n(\beta_0) \xrightarrow{d} \mathbf{N}(0, \mathbf{I}(\beta_0)), \quad (\text{A-1})$$

$$\hat{S}(\beta_0) \xrightarrow{P} \mathbf{I}(\beta_0). \quad (\text{A-2})$$

Proof. Let $\tilde{W}_{ni}(\beta) = \int_0^\infty [x_i - \bar{x}(\beta, t)] d\hat{M}_i(\beta, t)$. By the definition of $\hat{M}_i(\beta, t)$, we have

$$\begin{aligned} \sum_{i=1}^n W_{ni}(\beta_0) &= \sum_{i=1}^n \int_0^\infty [x_i - \bar{x}(\beta_0, t)] d\hat{M}_i(\beta_0, t) \quad (\text{A-3}) \\ &= \sum_{i=1}^n \int_0^\infty [x_i - \bar{x}(\beta_0, t)] dN_i(t) \\ &= \sum_{i=1}^n \tilde{W}_{ni}(\beta_0). \end{aligned}$$

Thus from Theorem 3.2 in Andersen and Gill (1982) one has equation (A-1).

Let $W_i(\beta) = \int_0^\infty \left[x_i - \frac{\alpha^{(1)}(\beta, t)}{\alpha^{(0)}(\beta, t)} \right] dM_i(\beta, t)$. By the law of large number,

$$n^{-1} \sum_{i=1}^n W_i(\beta_0) W_i^\tau(\beta_0) \xrightarrow{P} \mathbf{I}(\beta_0).$$

By Lemma 8.2 in Zheng and Yu (2013), we have

$$n^{-1} \sum_{i=1}^n W_i(\beta_0) W_i^\tau(\beta_0) = n^{-1} \sum_{i=1}^n W_{ni}(\beta_0) W_{ni}^\tau(\beta_0) + o_p(1),$$

which leads to equation (A-2). $\square$

Let $\alpha_n = n^{-1/2} + a_n$. When $a_n = O_p(n^{-1/2})$, $\alpha_n = n^{-1/2} + a_n = O_p(n^{-1/2})$. The next lemma gives a bound for the magnitude of $\lambda(\beta)$.

Lemma A.2. Assume the conditions in Theorem 2.1 hold, then as $n \rightarrow \infty$,

$$\|\lambda(\beta)\| = O_p(n^{-1/2}), \text{ for all } \|\beta - \beta_0\| \leq n^{-1/2}.$$

Proof. As $\|\beta - \beta_0\| = n^{-1/2}$, one has

$$\|W_{ni}(\beta) - W_{ni}(\beta_0)\| \leq n^{-1/2}.$$

Combining this with Lemma A.1, it leads to

$$\bar{W}_n(\beta) = O_p(n^{-1/2}), \quad (\text{A-4})$$

$$\hat{S}(\beta) = I(\beta_0) + o_p(1). \quad (\text{A-5})$$

Let $u = \lambda(\beta)/\|\lambda(\beta)\|$, $Z_n^* = \max_{1 \leq i \leq n} \|W_{ni}(\beta)\|$ and

$$\tilde{S}(\beta) = n^{-1} \sum_{i=1}^n W_{ni}(\beta) W_{ni}^T(\beta) / (1 + \lambda^T(\beta) W_{ni}(\beta)).$$

From (2.13) it is ready to show that $\|\lambda(\beta)\| u^T \tilde{S}(\beta) u = u^T \bar{W}_n(\beta)$. Thus,

$$\begin{aligned} \|\lambda(\beta)\| u^T \hat{S}(\beta) u &\leq \|\lambda(\beta)\| u^T \tilde{S}(\beta) u (1 + \|\lambda(\beta)\| Z_n^*) \\ &\leq u^T \bar{W}_n(\beta) (1 + \|\lambda(\beta)\| Z_n^*). \end{aligned}$$

As $E\|W_{ni}(\beta)\|^2 < \infty$ when $\|\beta - \beta_0\| = n^{-1/2}$, one has $Z_n^* = o_p(n^{1/2})$. Combining with equations (A-4) and (A-5), it leads to

$$\|\lambda(\beta)\| (I(\beta_0) + o_p(1)) = O_p(n^{-1/2}),$$

and hence $\|\lambda\| = O_p(n^{-1/2})$. $\square$

Proof of Theorem 2.1. Let $\beta = \beta_0 + u a_n$, where $\|u\| = C$ and $a_n = n^{-1/2} + a_n = O_p(n^{-1/2})$. It follows from equation (2.13) that,

$$\begin{aligned} 0 &= n^{-1} \sum_{i=1}^n \frac{W_{ni}(\beta)}{1 + \lambda^T(\beta) W_{ni}(\beta)} \\ &= n^{-1} \sum_{i=1}^n W_{ni}(\beta) \left(1 - \lambda^T(\beta) W_{ni}(\beta) + \frac{[\lambda^T(\beta) W_{ni}(\beta)]^2}{1 + \lambda^T(\beta) W_{ni}(\beta)} \right) \\ &= n^{-1} \sum_{i=1}^n W_{ni}(\beta) - \hat{S}(\beta) \lambda(\beta) + n^{-1} \sum_{i=1}^n \frac{W_{ni}(\beta) [\lambda^T(\beta) W_{ni}(\beta)]^2}{1 + \lambda^T(\beta) W_{ni}(\beta)}. \end{aligned} \quad (\text{A-6})$$

From Lemma A.2, one has the final term in (A-6) is bounded by $o_p(n^{-1/2})$. Thus

$$\lambda(\beta) = I^{-1}(\beta_0)n^{-1} \sum_{i=1}^n W_{ni}(\beta) + o_p(n^{-1/2}). \quad (\text{A-7})$$

By this and Taylor series expansion, one has

$$\begin{aligned} -2l_e(\beta) &= 2 \sum_{i=1}^n \log(1 + \lambda^\tau(\beta) W_{ni}(\beta)) \\ &= n \left(n^{-1} \sum_{i=1}^n W_{ni}^\tau(\beta) \right) I^{-1}(\beta_0) \left(n^{-1} \sum_{i=1}^n W_{ni}(\beta) \right) + o_p(1) \\ &= n \left(n^{-1} \sum_{i=1}^n W_{ni}^\tau(\beta_0) + n^{-1} \sum_{i=1}^n (W'_{ni}(\beta_0))^\tau (\beta - \beta_0) \right) I^{-1}(\beta_0) \\ &\quad \left(n^{-1} \sum_{i=1}^n W_{ni}(\beta_0) + n^{-1} \sum_{i=1}^n (W'_{ni}(\beta_0))^\tau (\beta - \beta_0) \right) + o_p(1). \end{aligned} \quad (\text{A-8})$$

Similarly,

$$-2l_e(\beta_0) = n \left(n^{-1} \sum_{i=1}^n W_{ni}^\tau(\beta_0) \right) I^{-1}(\beta_0) \left(n^{-1} \sum_{i=1}^n W_{ni}(\beta_0) \right) + o_p(1).$$

Thus, $-2(l_e(\beta) - l_e(\beta_0))$ is bounded by $nO_p(C^2\alpha_n^2)$.

By Taylor series expansion, one has

$$\begin{aligned} &\sum_{i=1}^q p_{\gamma_n}(|\beta_{i0} + u_i \alpha_n|) - \sum_{i=1}^q p_{\gamma_n}(|\beta_{i0}|) \\ &= \sum_{i=1}^q p'_{\gamma_n}(|\beta_{i0}|) \text{sgn}(\beta_{i0}) u_i \alpha_n + \sum_{i=1}^q p''_{\gamma_n}(|\beta_i^*|) u_i^2 \alpha_n^2, \end{aligned} \quad (\text{A-9})$$

where β^* is on the line segment between β and β_0 . By Cauchy-Schwarz inequality, we further have

$$\sum_{i=1}^q p'_{\gamma_n}(|\beta_{i0}|) \text{sgn}(\beta_{i0}) u_i \alpha_n + \sum_{i=1}^q p''_{\gamma_n}(|\beta_i^*|) u_i^2 \alpha_n^2 \leq \sqrt{q} \alpha_n a_n \|u\| + \alpha_n^2 b_n \|u\|^2. \quad (\text{A-10})$$

Thus $\sum_{i=1}^q p_{\gamma_n}(|\beta_{i0} + u_i \alpha_n|) - \sum_{i=1}^q p_{\gamma_n}(|\beta_{i0}|)$ is bounded by $\sqrt{q} C \alpha_n^2$ as $b_n \rightarrow 0$ combining (A-8) with (A-10), one has for any fixed ϵ , there always exists a large constant C such that

$$P \left\{ \max_{\|u\|=C} l_p(\beta_0 + u\alpha_n) < l_p(\beta_0) \right\} \geq 1 - \epsilon.$$

From the continuity of $l_p(\beta)$ when $\|\beta - \beta_0\| \rightarrow 0$, one has Theorem 2.1. $\square$

A.2 Proof of Theorem 2.2.

To prove the sparsity of $\hat{\beta}$ is equivalent to proving that when $\beta - \beta_0 = O_p(n^{-1/2})$, with probability tending to 1 as $n \rightarrow \infty$, $l_p(\beta)/\beta_j$ and β_j have different signs, for $j = q+1, \dots, p$. Towards this end,

$$\begin{aligned} \frac{\partial l_p(\beta)}{\partial \beta_j} &= \frac{\partial l_e(\beta)}{\partial \beta_j} - np'_{\gamma_n}(\beta_j) \text{sgn}(\beta_j) \\ &= - \sum_{i=1}^n \frac{\lambda(\beta) W'_{ni}(\beta)}{1 + \lambda^\tau(\beta) W_{ni}(\beta)} - np'_{\gamma_n}(\beta_j) \text{sgn}(\beta_j) \\ &= - \sum_{i=1}^n \frac{\lambda(\beta_0) W'_{ni}(\beta_0)}{1 + \lambda^\tau(\beta_0) W_{ni}(\beta_0)} + nO_p(n^{-1/2}) - np'_{\gamma_n}(\beta_j) \text{sgn}(\beta_j). \end{aligned}$$

Note that

$$\begin{aligned} - \sum_{i=1}^n \frac{\lambda(\beta_0) W'_{ni}(\beta_0)}{1 + \lambda^\tau(\beta_0) W_{ni}(\beta_0)} &\leq - \sum_{i=1}^n \frac{\lambda(\beta_0) W'_{ni}(\beta_0)}{1 + \lambda^\tau(\beta_0) \max_{1 \leq i \leq n} W_{ni}(\beta_0)} \quad (\text{A-11}) \\ &= -O_p(n^{-1/2}) W'_{ni}(\beta_0) \\ &= O_p(n^{1/2}). \end{aligned}$$

Thus, it follows that

$$\begin{aligned} \frac{\partial l_p(\beta)}{\partial \beta_j} &= -np'_{\gamma_n}(\beta_j) \text{sgn}(\beta_j) + O_p(n^{1/2}) \\ &= -np'_{\gamma_n}(\beta_j) \text{sgn}(\beta_j) + O_p(n^{1/2}) \\ &= n\gamma_n \left(-\gamma_n^{-1} p'_{\gamma_n}(\beta_j) \text{sgn}(\beta_j) + O_p(n^{-1/2} \gamma_n^{-1}) \right). \end{aligned}$$

Since $\liminf_{n \rightarrow \infty} \liminf_{\theta \rightarrow 0^+} p'_{\gamma_n}(\theta)/\gamma_n > 0$, the sign of $l_p(\beta)/\beta_j$ is completely determined by β_j ,

which leads to the sparsity of $\hat{\beta}$.

Now we prove the asymptotic normality of $\hat{\beta}_1$ and $\hat{\lambda}$. Taking derivative of $Q_{1n}(\beta, \lambda)$ at equation (2.13) and $Q_{2n}(\beta, \lambda)$ at equation (2.14) at $\beta = \beta_0$ and $\lambda = 0$, one has

$$\begin{aligned}\frac{\partial Q_{1n}(\beta_0, 0)}{\partial \beta} &= \frac{1}{n} \sum_{i=1}^n w'_{ni}(\beta_0), & \frac{\partial Q_{1n}(\beta_0, 0)}{\partial \lambda'} &= -\frac{1}{n} \sum_{i=1}^n w_{ni}^{\tau}(\beta_0) w_{ni}(\beta_0), \\ \frac{\partial Q_{2n}(\beta_0, 0)}{\partial \beta} &= \Sigma(\beta_0), & \frac{\partial Q_{2n}(\beta_0, 0)}{\partial \lambda} &= \frac{1}{n} \sum_{i=1}^n w'_{ni}{}^{\tau}(\beta_0).\end{aligned}$$

Expanding $Q_{1n}(\hat{\beta}, \hat{\lambda})$ and $Q_{2n}(\hat{\beta}, \hat{\lambda})$ at $(\hat{\beta} = \beta_0, \lambda = 0)$, one has

$$\begin{aligned}0 &= Q_{1n}(\hat{\beta}, \hat{\lambda}) \\ &= Q_{1n}(\beta_0, 0) + \frac{\partial Q_{1n}(\beta_0, 0)}{\partial \beta}(\hat{\beta} - \beta_0) + \frac{\partial Q_{1n}(\beta_0, 0)}{\partial \lambda'}\hat{\lambda} + o_p(c_n)\end{aligned}\tag{A-12}$$

$$\begin{aligned}0 &= Q_{2n}(\hat{\beta}, \lambda), \\ &= Q_{2n}(\beta_0, 0) + \frac{\partial Q_{2n}(\beta_0, 0)}{\partial \beta}(\hat{\beta} - \beta_0) + \frac{\partial Q_{2n}(\beta_0, 0)}{\partial \lambda'}\hat{\lambda} + o_p(c_n),\end{aligned}\tag{A-13}$$

where $c_n = \|\hat{\beta} - \beta_0\| + \|\hat{\lambda}\|$. From Theorem 2.1, $c_n = O_p(n^{-1/2})$. Solving equations (A-12) and (A-13), it leads to

$$\begin{pmatrix} \hat{\lambda} \\ \hat{\beta} - \beta_0 \end{pmatrix} = A_n^{-1} \begin{pmatrix} -Q_{1n}(\beta_0, 0) + o_p(n^{-1/2}) \\ -Q_{2n}(\beta_0, 0) + o_p(n^{-1/2}) \end{pmatrix}, \tag{A-14}$$

where

$$A_n = \begin{pmatrix} \frac{\partial Q_{1n}(\beta_0, 0)}{\partial \lambda'} & \frac{\partial Q_{1n}(\beta_0, 0)}{\partial \beta} \\ \frac{\partial Q_{2n}(\beta_0, 0)}{\partial \lambda'} & \frac{\partial Q_{2n}(\beta_0, 0)}{\partial \beta} \end{pmatrix}.$$

By the law of large number, one has $A_n \xrightarrow{P} A$ where,

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix} = \begin{pmatrix} -\mathbf{E}\mathbf{W}_{ni}(\boldsymbol{\beta}_0)\mathbf{W}_{ni}^{\tau}(\boldsymbol{\beta}_0) & \mathbf{E}\mathbf{W}_{ni}^{\tau}(\boldsymbol{\beta}_0) \\ \mathbf{E}\mathbf{W}_{ni}'^{\tau}(\boldsymbol{\beta}_0) & \boldsymbol{\Sigma}(\boldsymbol{\beta}_0) \end{pmatrix} \\ &= \begin{pmatrix} -\mathbf{I}(\boldsymbol{\beta}_0) & -\mathbf{I}(\boldsymbol{\beta}_0) \\ -\mathbf{I}^{\tau}(\boldsymbol{\beta}_0) & \boldsymbol{\Sigma}(\boldsymbol{\beta}_0) \end{pmatrix}. \end{aligned}$$

It follows that

$$\sqrt{n}\mathbf{V}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 + \mathbf{V}^{-1}\mathbf{Q}_{2n}(\boldsymbol{\beta}_0, 0)) = \sqrt{n}\mathbf{Q}_{1n}(\boldsymbol{\beta}_0, 0) + o_p(1),$$

where $\mathbf{V} = \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12} = \boldsymbol{\Sigma}(\boldsymbol{\beta}_0) + \mathbf{I}(\boldsymbol{\beta}_0)$, $\mathbf{Q}_{2n}(\boldsymbol{\beta}_0, 0) = \mathbf{b}(\boldsymbol{\beta}_0)$ and

$$\sqrt{n}\mathbf{Q}_{1n}(\boldsymbol{\beta}_0, 0) = \sqrt{n}\bar{\mathbf{W}}_{ni}(\boldsymbol{\beta}_0) \xrightarrow{d} \mathbf{N}(0, \mathbf{I}(\boldsymbol{\beta}_0)) \text{ by Lemma A.1.}$$

As $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^{\tau}, 0)^{\tau}$, one has by Slutsky's Theorem that

$$\sqrt{n}(\mathbf{I}_1(\boldsymbol{\beta}_{10}) + \boldsymbol{\Sigma}_1(\boldsymbol{\beta}_0))[\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10} + (\mathbf{I}_1(\boldsymbol{\beta}_{10}) + \boldsymbol{\Sigma}_1(\boldsymbol{\beta}_0))^{-1}\mathbf{b}_1(\boldsymbol{\beta}_0)] \xrightarrow{d} \mathbf{N}(0, \mathbf{I}_1(\boldsymbol{\beta}_{10})). \quad (\text{A-15})$$

Similarly, from equation (A-14) one has

$$\sqrt{n}(\hat{\boldsymbol{\lambda}} + \mathbf{V}^{-1}\mathbf{Q}_{2n}(\boldsymbol{\beta}_0, 0)) = \sqrt{n}(\mathbf{I}^{-1}(\boldsymbol{\beta}_0) - \mathbf{V}^{-1})\mathbf{Q}_{1n}(\boldsymbol{\beta}_0, 0) + o_p(1), \quad (\text{A-16})$$

and hence

$$\sqrt{n}(\hat{\boldsymbol{\lambda}} + \mathbf{V}^{-1}\mathbf{b}(\boldsymbol{\beta}_0)) \xrightarrow{d} \mathbf{N}(0, \mathbf{U}),$$

where $\mathbf{U} = (\mathbf{I}^{-1}(\boldsymbol{\beta}_0) - \mathbf{V}^{-1})\mathbf{I}^{-1}(\boldsymbol{\beta}_0)(\mathbf{I}^{-1}(\boldsymbol{\beta}_0) - \mathbf{V}^{-1})^{\tau}$. This completes the proof of Theorem 2.2.

A.3 Proof of Theorem 2.3.

Let $\hat{Y}_i = \hat{\boldsymbol{\lambda}}^{\tau}\mathbf{W}_{ni}(\hat{\boldsymbol{\beta}})$. By Taylor series expansion,

$$-2l_p(\hat{\boldsymbol{\beta}}) = 2 \sum_{i=1}^n \hat{Y}_i - \sum_{i=1}^n \hat{Y}_i^2 + \sum_{i=1}^n o_p(\hat{Y}_i^3) + nO_p\left(\sum_{i=1}^p p_{\gamma_n}(\boldsymbol{\beta}_{i0}) + \sum_{i=1}^p p'_{\gamma_n}(\boldsymbol{\beta}_{i0})(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_{i0})\right).$$

Note that by Lemma A.2 and Theorem 2.2, one has

$$\sum_{i=1}^n O_p(\hat{Y}_i^3) \leq \|\hat{\lambda}\|^3 \sum_{i=1}^n O_p(w_{ni}^3(\hat{\beta})) = o_p(1).$$

As $a_n = o_p(n^{-1/2})$ and $\|\hat{\beta} - \beta_0\| = O(n^{-1/2})$, we also have

$nO_p\left(\sum_{i=1}^p p_{\gamma_n}(\beta_{i0}) + \sum_{i=1}^p p'_{\gamma_n}(\beta_{i0})(\hat{\beta}_i - \beta_{i0})\right) = o_p(1)$. Thus one has

$$-2l_p(\hat{\beta}) = 2 \sum_{i=1}^n \hat{Y}_i - \sum_{i=1}^n \hat{Y}_i^2 + o_p(1). \quad (\text{A-17})$$

Since $a_n = o_p(n^{-1/2})$ and $b_n = o_p(1)$, one has $V = I(\beta_0) + o_p(1)$ and

$\hat{\lambda} = (I^{-1}(\beta_0) - V^{-1})Q_{1n}(\beta_0, 0) - V^{-1}b = o_p(n^{-1/2})$. Let β^* lie on the line segment between $\hat{\beta}$ and β_0 .

$$\begin{aligned} \sum_{i=1}^n \hat{Y}_i &= \sum_{i=1}^n \hat{\lambda} w_{ni}(\hat{\beta}) \\ &= \hat{\lambda}' \sum_{i=1}^n w_{ni}(\beta_0) + \frac{\partial w_{ni}(\beta^*)}{\partial \beta} (\hat{\beta} - \beta_0) \\ &= o_p(1). \end{aligned}$$

Similarly one has $\sum_{i=1}^n \hat{Y}_i^2 = o_p(1)$. Thus one has

$$-2l_p(\hat{\beta}) = o_p(1). \quad (\text{A-18})$$

Next we need to derive the maximum penalized empirical likelihood estimator $\tilde{\beta}$ under H_0 . According to the standard Lagrange multiplier method, the estimator is the maximizer of a new objective function

$$\tilde{l}_p(\beta) = l_e(\beta) - n \sum_{i=1}^p p_{\gamma}(|\beta_i|) - \nu^T L \beta_0, \quad (\text{A-19})$$

where ν is the additional Lagrange multiplier vector. Thus the minimizer $(\tilde{\beta}, \tilde{\lambda}, \tilde{\nu})$ satisfies

$$\tilde{Q}_{1n}(\beta, \lambda, \nu) = n^{-1} \sum_{i=1}^n \frac{W_{ni}(\beta)}{1 + \lambda^T W_{ni}(\beta)} = 0, \quad (\text{A-20})$$

$$\tilde{\mathcal{Q}}_{2n}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = n^{-1} \sum_{i=1}^n \frac{1}{1 + \boldsymbol{\lambda}^\tau \mathbf{W}_{ni}(\boldsymbol{\beta})} \frac{\partial \mathbf{W}_{ni}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \boldsymbol{\lambda} + \mathbf{b}(\boldsymbol{\beta}) + \mathbf{L}^\tau \boldsymbol{\nu} = 0, \quad (\text{A-21})$$

$$\tilde{\mathcal{Q}}_{3n}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \mathbf{L}^\tau \boldsymbol{\beta} = 0. \quad (\text{A-22})$$

Taking derivative of $\tilde{\mathcal{Q}}_{kn}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\nu})$, $k = 1, 2, 3$ at equations (A-20), (A-21) and (A-22) at $(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = (\boldsymbol{\beta}_0, 0, 0)$, one has

$$\begin{aligned} \mathbf{B}_n &= \begin{pmatrix} \frac{\partial \mathcal{Q}_{1n}(\boldsymbol{\beta}_0, 0, 0)}{\partial \boldsymbol{\lambda}'} & \frac{\partial \mathcal{Q}_{1n}(\boldsymbol{\beta}_0, 0, 0)}{\partial \boldsymbol{\beta}} & \frac{\partial \mathcal{Q}_{1n}(\boldsymbol{\beta}_0, 0, 0)}{\partial \boldsymbol{\nu}'} \\ \frac{\partial \mathcal{Q}_{2n}(\boldsymbol{\beta}_0, 0, 0)}{\partial \boldsymbol{\lambda}} & \frac{\partial \mathcal{Q}_{2n}(\boldsymbol{\beta}_0, 0, 0)}{\partial \boldsymbol{\beta}} & \frac{\partial \mathcal{Q}_{2n}(\boldsymbol{\beta}_0, 0, 0)}{\partial \boldsymbol{\nu}} \\ \frac{\partial \mathcal{Q}_{3n}(\boldsymbol{\beta}_0, 0, 0)}{\partial \boldsymbol{\lambda}} & \frac{\partial \mathcal{Q}_{3n}(\boldsymbol{\beta}_0, 0, 0)}{\partial \boldsymbol{\beta}} & \frac{\partial \mathcal{Q}_{3n}(\boldsymbol{\beta}_0, 0, 0)}{\partial \boldsymbol{\nu}} \end{pmatrix} \\ &= \begin{pmatrix} -\frac{1}{n} \sum_{i=1}^n \mathbf{W}_{ni}^\tau(\boldsymbol{\beta}_0) \mathbf{W}_{ni}(\boldsymbol{\beta}_0) & \frac{1}{n} \sum_{i=1}^n \mathbf{W}_{ni}'(\boldsymbol{\beta}_0) & 0 \\ \frac{1}{n} \sum_{i=1}^n \mathbf{W}_{ni}^{\tau}(\boldsymbol{\beta}) & \boldsymbol{\Sigma}(\boldsymbol{\beta}_0) & \mathbf{L} \\ 0 & \mathbf{L}^\tau & 0 \end{pmatrix}. \end{aligned}$$

Expanding $\tilde{\mathcal{Q}}_{kn}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$, $k = 1, 2, 3$ at $(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = (\boldsymbol{\beta}_0, 0, 0)$, one has

$$\begin{aligned} 0 &= \tilde{\mathcal{Q}}_{kn}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) \\ &= \tilde{\mathcal{Q}}_{kn}(\boldsymbol{\beta}_0, 0, 0) + \frac{\partial \tilde{\mathcal{Q}}_{kn}(\boldsymbol{\beta}_0, 0, 0)}{\partial \boldsymbol{\beta}} (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \frac{\partial \tilde{\mathcal{Q}}_{kn}(\boldsymbol{\beta}_0, 0, 0)}{\partial \boldsymbol{\lambda}'} \tilde{\boldsymbol{\lambda}} \\ &\quad + \frac{\partial \tilde{\mathcal{Q}}_{kn}(\boldsymbol{\beta}_0, 0, 0)}{\partial \boldsymbol{\nu}'} \tilde{\boldsymbol{\nu}} + o_p(d_n), \end{aligned} \quad (\text{A-23})$$

where $d_n = \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| + \|\tilde{\boldsymbol{\lambda}}\| + \|\tilde{\boldsymbol{\nu}}\|$. From Lemma A.2 one has that $\tilde{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = O_p(n^{-1/2})$ for $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < n^{-1/2}$. It follows from Theorem 2.1 that there exists a local minimizer $\tilde{\boldsymbol{\beta}}$ of $\tilde{l}_p(\boldsymbol{\beta})$ within the interior of the ball $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < Cn^{-1/2}$. From $\tilde{\mathcal{Q}}_{1n}(\boldsymbol{\beta}_0, 0, 0) = n^{-1} \sum_{i=1}^n \mathbf{W}_{ni}(\boldsymbol{\beta}_0) = 0$, one has $d_n = O_p(n^{-1/2})$.

Solving equations (A-23), it leads to

$$\begin{pmatrix} \tilde{\lambda} \\ \tilde{\beta} - \beta_0 \\ \tilde{\nu} \end{pmatrix} = \mathbf{B}_n^{-1} \begin{pmatrix} -\tilde{\mathbf{Q}}_{1n}(\beta_0, 0, 0) + o_p(n^{-1/2}) \\ o_p(n^{-1/2}) \\ o_p(n^{-1/2}) \end{pmatrix}. \quad (\text{A-24})$$

Note that $b_n = o_p(1)$, one has $\Sigma(\beta_0) = o_p(1)$. By the law of large number, one has $\mathbf{B}_n \xrightarrow{P} \mathbf{B}$, where

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix}, \quad (\text{A-25})$$

with

$$\begin{aligned} \mathbf{B}_{11} &= -\mathbf{I}(\beta_0), \quad \mathbf{B}_{12} = (-\mathbf{I}(\beta_0), \mathbf{0}), \\ \mathbf{B}_{12} &= \mathbf{B}_{21}^\tau, \quad \mathbf{B}_{22} = \begin{pmatrix} \mathbf{0} & \mathbf{L} \\ \mathbf{L}^\tau & \mathbf{0} \end{pmatrix}. \end{aligned}$$

Inverting matrix $\mathbf{B}$ by blocks, one has

$$\tilde{\beta} - \beta_0 = \mathbf{I}^{-1}(\beta_0)(\mathbf{I} - \mathbf{P}\mathbf{I}^{-1}(\beta_0))\tilde{\mathbf{Q}}_{1n}(\beta_0, 0, 0) + o_p(n^{-1/2}), \quad (\text{A-26})$$

where $\mathbf{P} = \mathbf{L}^\tau(\mathbf{L}\mathbf{I}^{-1}(\beta_0)\mathbf{L}^\tau)^{-1}\mathbf{L}$. Furthermore,

$$\begin{aligned} n^{-1} \sum_{i=1}^n \mathbf{W}_{ni}(\tilde{\beta}) &= n^{-1} \sum_{i=1}^n \mathbf{W}_{ni}(\beta_0) + n^{-1} \sum_{i=1}^n \mathbf{W}_{ni}'(\beta_0)(\tilde{\beta} - \beta_0) + o_p(n^{-1/2}) \quad (\text{A-27}) \\ &= \tilde{\mathbf{Q}}_{1n}(\beta_0, 0, 0) - \mathbf{I}(\beta_0)(\tilde{\beta} - \beta_0) + o_p(n^{-1/2}) \\ &= \mathbf{P}\mathbf{I}^{-1}(\beta_0)\tilde{\mathbf{Q}}_{1n}(\beta_0, 0, 0) + o_p(n^{-1/2}). \end{aligned}$$

Similar to the derivation for $\lambda(\beta)$ at equation (A-7), it is ready to show from equation (A-20) that

$$\tilde{\lambda} = \mathbf{I}^{-1}(\beta_0)n^{-1} \sum_{i=1}^n \mathbf{W}_{ni}(\tilde{\beta}) + o_p(n^{-1/2}). \quad (\text{A-28})$$

Let $\tilde{Y}_i = \tilde{\lambda}^\tau \mathbf{W}_{ni}(\tilde{\beta})$. Similar to the derivation for $l_p(\hat{\beta})$ at (A-17), one has

$$\begin{aligned}
-2l_p(\tilde{\beta}) &= 2 \min_{\beta: L\beta_0=0} l_p(\beta) \\
&= 2 \sum_{i=1}^n \tilde{Y}_i - \sum_{i=1}^n \tilde{Y}_i^2 + o_p(1) \\
&= n \left(n^{-1} \sum_{i=1}^n W_{ni}^{\tau}(\tilde{\beta}) \right) I^{-1}(\beta_0) \left(n^{-1} \sum_{i=1}^n W_{ni}(\tilde{\beta}) \right) + o_p(1).
\end{aligned} \tag{A-29}$$

Substituting the expansion of $n^{-1} \sum_{i=1}^n W_{ni}(\tilde{\beta})$ at equation (A-27) into $-2l_p(\tilde{\beta})$, one has

$$-2l_p(\tilde{\beta}) = H^{\tau} (L I^{-1}(\beta_0) L^{\tau})^{-1} H + o_p(1), \tag{A-30}$$

where $H = \sqrt{n} L I^{-1}(\beta_0) \tilde{Q}_{1n}(\beta_0, 0, 0)$. By Lemma A.1, one has $\tilde{Q}_{1n}(\beta_0, 0, 0) \xrightarrow{d} N(0, I(\beta_0))$. Thus

$$\begin{aligned}
H &\xrightarrow{d} N(0, L I^{-1}(\beta_0) L^{\tau}) \text{ and} \\
-2l_p(\tilde{\beta}) &\xrightarrow{d} \chi_L^2,
\end{aligned} \tag{A-31}$$

since $LL^{\tau} = I_L$. Combining equation (A-18) with equation (A-31), one has

$$\phi(L) = -2 \left(l_p(\hat{\beta}) - \min_{\beta: L\beta_0=0} l_p(\beta) \right) \xrightarrow{d} \chi_L^2,$$

which completes the proof of Theorem 2.3.

References

- Andersen PK, Gill RD (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist* 10, 1100–1120.
- Breheny P and Huang J (2011). Coordinate decent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics* 51, 232–253.
- Breslow N (1974). Covariance analysis of censored survival data. *Biometrics* 30, 89–99. [PubMed: 4813387]
- Chen J, Chen Z (2008). Extended Bayesian information criterion for model selection with large model space. *Biometrika* 95, 759–771.
- Chen SX, Peng L, Qin YL (2009). Effects of data dimension on empirical likelihood. *Biometrika* 96, 711–722.
- Chen SX, Van Keilegom I (2009). A review on empirical likelihood methods for regression (with discussions). *Test* 18, 415–447.
- Cox DR (1972). Regression model and life tables (with discussion). *J R Stat Soc B* 34, 187–220.
- Cox DR (1975). Partial likelihood. *Biometrika* 62, 269–276.
- Dickson E, Grambsch P, Fleming T, Fisher L, Langworthy A (1989). Prognosis in primary biliary cirrhosis: model for decision making. *Hepatology* 10, 1–7. [PubMed: 2737595]

- Fan J, Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc* 96, 1348–1360.
- Fan J, Li R (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics* 30, 74–99.
- Faraggi D, Simon R (1998). Bayesian variable selection method for censored survival data. *Biometrics* 54, 1475–1485. [PubMed: 9883546]
- Fleming TR, Harrington DP (2013). *Counting Processes and Survival Analysis*, 2nd edition, John Wiley & Sons, Hoboken, NJ.
- Gill PE and Murray W and Wright MH (1981). *Practical Optimization*, New York, Academic Press.
- Goeman JJ (2010). L_1 penalized estimation in the Cox proportional hazards model. *Biometrical Journal* 52(1), 70–84. [PubMed: 19937997]
- Gong Y, Klingenberg SL, Gluud C (2004). D-penicillamine for primary biliary cirrhosis. *Cochrane Database of Systematic Reviews* 4, Art. No.: CD004789.
- Gu C (1996). Penalized likelihood hazard estimation: a general procedure. *Statistica Sinica* 6, 861–876.
- He X, Wang L, Hong H (2013). Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. *The Annals of Statistics*, 41(1), 342–369.
- Hjort NL, McKeague IW, Van Keilegom I (2009). Extending the scope of empirical likelihood. *Ann. Statist.* 37(3), 1079–1111.
- Hoerl AE, Kennard RW (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67.
- Hou W, Song L, Hou X, Wang X (2014). Penalized empirical likelihood via bridge estimator in Cox's proportional hazard model. *Communications in Statistics –Theory and Methods* 43(2), 426–440
- Hunter DR, Lange K (2004). A tutorial on MM algorithms. *Am. Statistician* 58, 30–37.
- Ibrahim JG, Chen MH, Maceachern SN (1999). Bayesian variable selection for proportional hazards models. *Can. J. Statist.* 27, 701–717.
- Lange K (2004). *Optimization*, Springer-Verlag, New York.
- Leng C, Tang CY (2012). Penalized empirical likelihood and growing dimensional general estimating equations. *Biometrika* 99, 703–716.
- Li J, Zheng Q, Peng L, Huang Z (2016). Survival impact index and ultrahighdimensional model-free screening with survival outcomes. *Biometrics*, 72(4), 1145–1154. [PubMed: 26910137]
- Owen AB (1988). Empirical likelihood ratio confidences for single functional. *Biometrika* 75, 237–249.
- Owen AB (2001). *Empirical Likelihood*, Chapman and Hall/CRC press, New York.
- Qin GS, Jing BY (2001). Empirical likelihood for Cox regression model under random censorship. *Commun. Statist. Simul. Computat* 30, 79–90.
- Sauerbrei W, Schumacher M (1992). A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statist. Med* 11, 2093–2109.
- Sun Y, Sundaram R, Zhao Y (2009). Empirical likelihood inference for the Cox models with time-dependent coefficients via local partial likelihood, *Scandinavian Journal of Statistics* 36, 444–462.
- Tang CY, Leng CL (2010). Penalized high-dimensional empirical likelihood. *Biometrika* 97, 905–920.
- Tang C, Wu T (2014). Nested coordinate descent algorithms for empirical likelihood. *Journal of Statistical Computation and Simulation* 84(9), 1917–1930.
- Tibshirani R (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B* 58(1), 267–288.
- Tibshirani R (1997). The lasso method for variable selection in the Cox model. *Statist. Med* 16, 385–395.
- Variyath AM, Chen JH, Abrahamc B (2010). Empirical likelihood based variable selection. *J. Statist. Plann. Infer* 140, 971–981.
- Wu TT, Lange K (2010). The MM alternative to EM. *Stat. Sci* 25, 492–505.
- Wang H, Leng C (2007). Unified LASSO estimation by least square approximation. *Journal of the American Statistical Association* 102(479), 1039–1048.

- Wang H, Li B, Leng C (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Stat. Soc., Ser. B* 71, 671–683.
- Wang H, Li R, Tsai C (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 94, 553–568. [PubMed: 19343105]
- Wu T, Li G, Tang C (2015). Empirical likelihood for censored linear regression and variable selection. *Scandinavian Journal of Statistics* 42(3), 798–812.
- Xia X, Jiang B, Li J, Zhang W (2016). Low-dimensional confounder adjustment and high-dimensional penalized estimation for survival analysis. *Lifetime Data Analysis*, 22(4), 547–569. [PubMed: 26463818]
- Zhang H, Lu W (2007). Adaptive lasso for Cox’s proportional hazards model. *Biometrika* 94(3), 691–703.
- Zhao Y, Huang Y (2007). Test-based interval estimation under the accelerated failure time model, *Communications in Statistics-Simulation and Computation* 36, 593–605.
- Zhao Y, Jinnah A (2012). Inference for Cox’s regression models via adjusted empirical likelihood. *Computational Statistics* 27(1), 1–12.
- Zheng M, Yu W (2013). Empirical likelihood method for multivariate Cox regression. *Comput. Stat* 28, 1241–1267.
- Zucker DM; Karr AF. (1990) Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *Ann. Statist* 18(1), 329–353.

Table 1:

Comparison of the proposed penalized empirical likelihood (PEL) and R packages *penalized* and *ncvreg* with $p = 10$ based on 100 replicates (Part A).

(n, p)	ρ	Method	N_{nonzero}	N_{true}	ME
(100, 10)	0	Oracle	NA	3 (0)	0.255 (0.039)
		<i>penalized</i>	7.19 (0.14)	3 (0)	0.502 (0.037)
		<i>ncvreg</i>	3.17 (0.06)	3 (0)	0.225 (0.036)
		PEL-BIC	3.5 (0.1)	3 (0)	0.308 (0.033)
		PEL-BICC	3.5 (0.1)	3 (0)	0.308 (0.033)
		PEL-EBIC	3.1 (0.04)	3 (0)	0.269 (0.031)
	0.5	Oracle	NA	3 (0)	0.310 (0.043)
		<i>penalized</i>	6.06 (0.17)	3 (0)	0.656 (0.043)
		<i>ncvreg</i>	3.25 (0.06)	3 (0)	0.331 (0.050)
		PEL-BIC	3.47 (0.08)	3 (0)	0.456 (0.048)
		PEL-BICC	3.47 (0.08)	3 (0)	0.456 (0.048)
		PEL-EBIC	3.14 (0.05)	3 (0)	0.407 (0.046)
	0.9	Oracle	NA	3 (0)	0.695 (0.052)
		<i>penalized</i>	5.44 (0.13)	2.99 (0.01)	1.442 (0.092)
		<i>ncvreg</i>	3.19 (0.06)	2.93 (0.03)	1.311 (0.145)
		PEL-BIC	3.5 (0.1)	3 (0)	0.932 (0.108)
		PEL-BICC	3.5 (0.1)	3 (0)	0.932 (0.108)
		PEL-EBIC	3.42 (0.1)	3 (0)	0.892 (0.105)
(200, 10)	0	Oracle	NA	3 (0)	0.089 (0.01)
		<i>penalized</i>	7.44 (0.15)	3 (0)	0.232 (0.017)
		<i>ncvreg</i>	3.05 (0.03)	3 (0)	0.124 (0.017)
		PEL-BIC	3.04 (0.02)	3 (0)	0.113 (0.014)
		PEL-BICC	3.04 (0.02)	3 (0)	0.113 (0.014)
		PEL-EBIC	3 (0)	3 (0)	0.113 (0.015)
	0.5	Oracle	NA	3 (0)	0.129 (0.015)
		<i>penalized</i>	6.88 (0.16)	3 (0)	0.285 (0.019)
		<i>ncvreg</i>	3.07 (0.03)	3 (0)	0.143 (0.022)
		PEL-BIC	3.07 (0.03)	3 (0)	0.16 (0.017)
		PEL-BICC	3.07 (0.03)	3 (0)	0.16 (0.017)
		PEL-EBIC	3.04 (0.02)	3 (0)	0.158 (0.017)
	0.9	Oracle	NA	3 (0)	0.314 (0.027)
		<i>penalized</i>	5.56 (0.13)	3 (0)	0.680 (0.038)
		<i>ncvreg</i>	3.11 (0.04)	3 (0)	0.399 (0.042)
		PEL-BIC	3.18 (0.06)	3 (0)	0.440 (0.044)
		PEL-BICC	3.18 (0.06)	3 (0)	0.440 (0.044)

(n,p)	ρ	Method	N_{nonzero}	N_{true}	ME
		PEL-EBIC	3.01 (0.01)	3 (0)	0.343 (0.029)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2:

Comparison of the proposed penalized empirical likelihood (PEL) and R packages *penalized* and *ncvreg* with $p = 20$ based on 100 replicates (Part B).

(n, p)	ρ	Method	N_{nonzero}	N_{true}	ME
(200, 0)	0	Oracle	NA	3 (0)	0.088 (0.015)
		<i>penalized</i>	11.66 (0.26)	3 (0)	0.386 (0.023)
		<i>ncvreg</i>	3.06 (0.03)	3 (0)	0.082 (0.008)
		PEL-BIC	3.73 (0.11)	3 (0)	0.162 (0.020)
		PEL-BICC	3.62 (0.11)	3 (0)	0.162 (0.020)
		PEL-EBIC	3.08 (0.04)	3 (0)	0.157 (0.021)
	0.5	Oracle	NA	3 (0)	0.143 (0.022)
		<i>penalized</i>	9.67 (0.27)	3 (0)	0.521 (0.032)
		<i>ncvreg</i>	3.1 (0.04)	3 (0)	0.118 (0.021)
		PEL-BIC	3.38 (0.08)	3 (0)	0.204 (0.022)
		PEL-BICC	3.33 (0.07)	3 (0)	0.204 (0.022)
		PEL-EBIC	3.04 (0.02)	3 (0)	0.189 (0.021)
	0.9	Oracle	NA	3 (0)	0.289 (0.023)
		<i>penalized</i>	7.71 (0.21)	3 (0)	0.867 (0.047)
		<i>ncvreg</i>	3.47 (0.08)	3 (0)	0.719 (0.077)
		PEL-BIC	3.61 (0.11)	3 (0)	0.780 (0.076)
		PEL-BICC	3.53 (0.1)	3 (0)	0.746 (0.074)
		PEL-EBIC	3.09 (0.03)	3 (0)	0.480 (0.042)
(500, 20)	0	Oracle	NA	3 (0)	0.036 (0.005)
		<i>penalized</i>	11.77 (0.29)	3 (0)	0.164 (0.011)
		<i>ncvreg</i>	3.02 (0.01)	3 (0)	0.037 (0.005)
		PEL-BIC	3.07 (0.03)	3 (0)	0.047 (0.006)
		PEL-BICC	3.02 (0.01)	3 (0)	0.047 (0.006)
		PEL-EBIC	3 (0)	3 (0)	0.047 (0.006)
	0.5	Oracle	NA	3 (0)	0.041 (0.004)
		<i>penalized</i>	10.57 (0.26)	3 (0)	0.185 (0.013)
		<i>ncvreg</i>	3.04 (0.02)	3 (0)	0.043 (0.004)
		PEL-BIC	3 (0)	3 (0)	0.049 (0.005)
		PEL-BICC	3 (0)	3 (0)	0.049 (0.005)
		PEL-EBIC	3 (0)	3 (0)	0.049 (0.005)
	0.9	Oracle	NA	3 (0)	0.099 (0.009)
		<i>penalized</i>	8.13 (0.24)	3 (0)	0.372 (0.022)
		<i>ncvreg</i>	3.23 (0.05)	3 (0)	0.173 (0.022)
		PEL-BIC	3.41 (0.07)	3 (0)	0.237 (0.027)
		PEL-BICC	3.28 (0.06)	3 (0)	0.212 (0.027)

(n,p)	ρ	Method	N_{nonzero}	N_{true}	ME
		PEL-EBIC	3.04 (0.02)	3 (0)	0.148 (0.019)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3:

PBC example.

Covariate	Unpenalized Cox	Penalized Cox	PEL-BIC	PEL-BICC	PEL-EBIC
<i>Drug</i>	−0.355	−0.140	−0.298	−0.298	0
<i>Age</i>	0.109	0.014	0	0	0
<i>Sex</i>	−0.110	0	0	0	0
<i>Ascites</i>	0.071	0.053	0	0	0
<i>Hepato</i>	0.108	0.051	0	0	0
<i>Spiders</i>	−0.027	0	0	0	0
<i>Oedema</i>	0.242	0.147	0.190	0.190	0
<i>Bili</i>	0.419	0.365	0.422	0.422	0.659
<i>Chol</i>	0.097	0	0	0	0
<i>Albumin</i>	−0.320	−0.226	−0.329	−0.329	−0.440
<i>Copper</i>	0.274	0.247	0.255	0.255	0.075
<i>Alk_phos</i>	−0.120	0	0	0	0
<i>SGOT</i>	0.215	0.091	0.187	0.187	0
<i>Trig</i>	−0.144	0	0	0	0
<i>Platelet</i>	0.103	0	0	0	0
<i>PT</i>	0.229	0.130	0.245	0.245	0
<i>Stage</i>	0.391	0.261	0.427	0.427	0.448