Penalized empirical likelihood inference for sparse additive hazards regression with a diverging number of covariates

Wang, Shanshan; Xiang, Liming

2016

Wang, S., & Xiang, L. Penalized empirical likelihood inference for sparse additive hazards regression with a diverging number of covariates. Statistics and Computing, in press.

https://hdl.handle.net/10356/83404

https://doi.org/10.1007/s11222-016-9690-x

Penalized Empirical Likelihood Inference for Sparse Additive Hazards Regression with a Diverging number of covariates

Shanshan Wang[1,2] **and Liming Xiang**[2*]

[1] *School of Economics and Management, Beihang University, Beijing 100191, China*

[2] *School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371*

Abstract

High-dimensional sparse modeling with censored survival data is of great practical importance, as exemplified by applications in high-throughput genomic data analysis. In this paper, we propose a class of regularization methods, integrating both the penalized empirical likelihood and pseudoscore approaches, for variable selection and estimation in sparse and high-dimensional additive hazards regression models. When the number of covariates grows with the sample size, we establish asymptotic properties of the resulting estimator and the oracle property of the proposed method. It is shown that the proposed estimator is more efficient than that obtained from the non-concave penalized likelihood approach in the literature. Based on a penalized empirical likelihood ratio statistic, we further develop a nonparametric likelihood approach for testing the linear hypothesis of regression coefficients and constructing confidence regions consequently. Simulation studies are carried out to evaluate the performance of the proposed methodology and also a real data set is analyzed.

Key Words: Penalized empirical likelihood; Empirical likelihood ratio; Oracle property; Smoothly clipped absolute deviation; Survival data; Variable selection.

1 Introduction

A main objective of survival analysis is to assess the relationship of survival time T to a set of p covariates (possibly time-dependent) $\boldsymbol{Z}(t) = (Z_1(t), \cdots, Z_p(t))^T$. A common way of achieving this goal is the hazard regression, which studies how the conditional hazard function of T depends on the covariate $\boldsymbol{Z}(t)$. With technological advances in many scientific areas, a considerable amount of

* Corresponding author. Tel.: +65 65137451 ; Fax: +65 65158213.

E-mail address: lmxiang@ntu.edu.sg .

covariate information is available while observing survival time information on subjects. It is crucial to develop more sophisticated statistical methods to incorporate high-dimensional covariates thus to reduce the modeling bias. In this paper we develop a general framework using the penalized empirical likelihood method for estimation, variable selection and inference in additive hazards regression analysis, where we allow the possibility of a large number of covariates in the sense that p diverges as the sample size increases.

As a useful alternative to the popular regression model for survival data, the Cox proportional hazards (PH) model, the additive hazards model particularly assumes that the hazard function of T conditional on covariates $\boldsymbol{Z}(\cdot)$ takes the form

$$h(t|\boldsymbol{Z}) = h_0(t) + \boldsymbol{\beta}_0^T \boldsymbol{Z}(t), \tag{1}$$

where $h_0(\cdot)$ is an unspecified baseline hazard function and $\boldsymbol{\beta}_0$ is a p-vector of regression coefficients. Unlike the Cox PH model which concerns the relative risk, the additive hazards model pertains to the risk difference. When the absolute change in risk is of primary interest or the proportional hazard assumption is violated, an additive hazard regression model may become a more reasonable choice. For example, Xie et al. (2013) applied the additive hazards regression to analyze data from a study of natural history of human papilloma virus (HPV) infection among human immunodeficiency virus HIV-positive and HIV-negative women, where the excess risk of infecting HPV between the two groups of women was of interest in the study. Due to easy interpretation of the covariate effects, the additive hazards model has received considerable attention in the recent literature (Breslow and Day, 1987, Cox and Oakes, 1984, Lin and Ying, 1994, Ma et al., 2006, Martinussen and Scheike, 2009).

In practice, although a possibly large number of covariates are involved in the initial stage of modeling, such as DAN micro-array, proteinic and SNP data via bio-imaging technology, it is quite likely that not all available covariates are associated with the survival outcome. A promising approach to reduce model complexity is to force less influential variables to have zero impact on the model. This results in sparsity, where many regression coefficients are zeros. Consequently, it is essential to identify important covariates (with nonzero coefficients) and in the meantime to estimate their risk contributions to the survival time. To this end, various variable selection techniques have been extended from classical linear regression to the the Cox PH model, including the best-subset selection, stepwise selection, bootstrap procedures (Sauerbrei and Schumacher, 1992), the least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1997) and adaptive Lasso (Zhang and Lu, 2007), the nonconcave penalized likelihood (Fan and Li, 2002), and the

penalized pseudo-partial likelihood (Cai et al., 2005) in a framework with p growing with n. Beyond the Cox PH model, several variable selection procedures have also been applied to the additive hazards model. For example, Ma and Huang (2005) proposed a simple Lasso approach to select relevant predictors for survival data with multiple covariates in gene selection context; Leng and Ma (2007) explored a path consistent model selector via a weighted Lasso approach; and Martinussen and Scheike (1999, 2009) considered several regularization methods including the Lasso, adaptive Lasso and Dantzig selector. More recently, Lin and Lv (2013) studied regularization methods for simultaneous variable selection and estimation in additive hazards regression model using the non-concave penalized likelihood approach (Fan and Li, 2001).

A valuable alternative method for variable selection is the penalized empirical likelihood (Tang and Leng, 2010). As a nonparametric method of inference, the empirical likelihood method (Owen, 2001) has successful applications in various areas, providing attractive features, such as the improvement of the confidence region, accuracy of coverage and ease of implementation (Chen and Cui, 2006, DiCiccio et al., 1991) in comparison with the classical likelihood-based methods. Chen et al. (2009) and Hjort et al. (2009) studied the empirical likelihood with growing dimensions and established the asymptotic normality for the standardization of -2log-empirical likelihood ratio. By adding a penalty function to the standard empirical likelihood in spirit of the penalized likelihood context (Fan and Li, 2001, Fan and Peng, 2004), Tang and Leng (2010) considered the mean vector and the linear regression models. Leng and Tang (2012) further generalized their results to the growing dimensional general estimating equations. In general, the regularization method via empirical likelihood for model (1) and the corresponding rigorous theory in the "large n, diverging p" framework are undeveloped.

Complementary to the regularization method via the non-concave penalized likelihood approach (Lin and Lv, 2013), the penalized empirical likelihood method has the merits in both efficiency and adaptivity inheriting from a nonparametric likelihood approach. This motivates us to investigate regularization methods through penalized empirical likelihood for model (1) with a diverging number of parameters. In particular, we propose a penalized empirical likelihood procedure, integrating the empirical likelihood together with the pseudoscore method (Lin and Ying, 1994), for variable selection and parameter estimation simultaneously, and show its oracle property when p grows with n. Although the regularization method formulated in Lin and Lv (2013) is applicable in the same "large n, diverging p" framework and also possesses the oracle property, the proposed work leads to more efficient estimates of the nonzero regression coefficients (as shown in Theorem 3).

Furthermore, we show that the penalized empirical likelihood ratio is asymptotically chi-squared distributed under some regularity conditions, facilitating hypothesis tests and confidence regions for nonzero regression coefficients.

The rest of the paper is organized as follows. Section 2 presents the additive hazards model and its existing theoretical results. In Section 3, we propose the empirical likelihood method and its penalized version. Their asymptotic properties are derived, respectively. A penalized empirical likelihood ratio statistic is also provided in this section. In Section 4, we describe the algorithm and discuss selection of tuning parameters. Simulation studies are presented in Section 5 and an application to PBC data is given in Section 6.1 to demonstrate the effectiveness of the proposed method. We conclude with a brief summary and discussions in Section 7. All technical details are relegated to Appendix.

2 Additive Hazards Model

For more precise understanding, we represent the model using counting processes and give a brief review of results in Lin and Ying (1994). Let T be the failure time and C the censoring time. Denote the censored failure time by $X = T \wedge C$ and the failure indicator by $\Delta = I(T \leq C)$, where $I(\cdot)$ is the indicator function. Let $\boldsymbol{Z}(\cdot) = (Z_1(\cdot), \ldots, Z_p(\cdot))^T$ be a vector of predictable covariate processes and assume that T and C are conditionally independent given $\boldsymbol{Z}(\cdot)$. The observed data consist of $(X_i, \Delta_i, \boldsymbol{Z}_i(\cdot)), i = 1, \ldots, n$, which are independent copies of $(X, \Delta, \boldsymbol{Z}(\cdot))$. Assume that the conditional hazard function is given by model (1). We define the observed-failure counting process $N_i(t) = I(X_i \leq t, \Delta_i = 1)$, the at-risk indicator $Y_i(t) = I(X_i \geq t)$, and the counting process martingale

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\{h_0(s) + \boldsymbol{\beta}_0^T \boldsymbol{Z}_i(s)\}ds. \tag{2}$$

We will use $N(t), Y(t)$, and $M(t)$ to denote the generic versions of these processes.

Let $\bar{\boldsymbol{Z}}(t) = \sum_{j=1}^n Y_j(t)\boldsymbol{Z}_j(t)/\sum_{j=1}^n Y_j(t)$. It is easy to show that $\frac{1}{n}\sum_{i=1}^n \int_0^\tau Y_i(t)\{\boldsymbol{Z}(t) - \bar{\boldsymbol{Z}}(t)\}\bar{\boldsymbol{Z}}^T(t)dt = 0$. When the dimension p is fixed, the pseudoscore function is defined by

$$\boldsymbol{U}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\}\{dN_i(t) - Y_i(t)\boldsymbol{\beta}^T \boldsymbol{Z}_i(t)dt\} = \boldsymbol{b}_n - \boldsymbol{D}_n\boldsymbol{\beta}, \tag{3}$$

with $\boldsymbol{b}_n = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\}dN_i(t)$ and

$$\boldsymbol{D}_n = \frac{1}{n}\sum_{i=1}^n \int_0^\tau Y_i(t)\{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\}^{\otimes 2}dt, \tag{4}$$

where τ is the maximum follow-up time of subjects in the study, and $\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^T$ for any vector $\boldsymbol{v}$. Note that the integrals for vectors and matrices are understood component-wise throughout the paper. Then the resulting estimator, denoted by $\hat{\boldsymbol{\beta}}^{\mathrm{LY}}$, takes the explicit form $\hat{\boldsymbol{\beta}}^{\mathrm{LY}} = \boldsymbol{D}_n^{-1}\boldsymbol{b}_n$.

It follows from (2) and (3) that

$$\boldsymbol{U}(\boldsymbol{\beta}_0) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\}dM_i(t) \tag{5}$$

is a martingale integral. Note that $E\boldsymbol{U}(\boldsymbol{\beta}_0) = \boldsymbol{0}$ and when dimension p is fixed, the random vector $n^{1/2}\boldsymbol{U}(\boldsymbol{\beta}_0)$ converges weakly to a p-dimensional normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$, which can be consistently estimated by $\boldsymbol{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau}\{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\}^{\otimes 2}dN_i(t)$ using the standard counting process arguments (Andersen and Gill, 1982). Furthermore, the random vector $n^{1/2}(\hat{\boldsymbol{\beta}}^{\mathrm{LY}} - \boldsymbol{\beta}_0)$ converges weakly to a p-dimensional normal distribution with mean zero and a covariance matrix which can be consistently estimated by $\boldsymbol{D}_n^{-1}\boldsymbol{\Sigma}_n\boldsymbol{D}_n^{-1}$ with $\boldsymbol{D}_n$ defined in (4).

To derive the penalized empirical likelihood method, some more notation is required. For any vector $\boldsymbol{v}$, recall that $\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^T$. We write $\boldsymbol{v}^{\otimes 0} = 1$ and $\boldsymbol{v}^{\otimes 1} = \boldsymbol{v}$ for notational convenience. Define

$$\begin{aligned}
\boldsymbol{s}^{(k)} &= E\{Y(t)\boldsymbol{Z}(t)^{\otimes k}\}, \quad k = 0, 1, 2, \\
\boldsymbol{e}(t) &= \boldsymbol{s}^{(1)}(t)/\boldsymbol{s}^{(0)}(t), \\
\boldsymbol{D} &= E\Big\{\int_0^{\tau} Y(t)[\boldsymbol{Z}(t) - \boldsymbol{e}(t)]^{\otimes 2}dt\Big\}, \\
\boldsymbol{\Sigma} &= E\Big[\int_0^{\tau}\{\boldsymbol{Z}(t) - \boldsymbol{e}(t)\}^{\otimes 2}dN(t)\Big].
\end{aligned}$$

where $\boldsymbol{D}$ and $\boldsymbol{\Sigma}$ are the theoretical quantities corresponding to the matrices $\boldsymbol{D}_n$ and $\boldsymbol{\Sigma}_n$, respectively. These matrices characterize the covariance structure of the model and will play key roles in the following analysis.

3 Regularization Methodology

In this section, we develop the penalized empirical likelihood for model (1) with a diverging number of parameters, i.e., the dimension p increases to infinite as the sample size $n \to \infty$ at a proper rate, and discuss its theoretic properties. Specifically, we formulate an empirical likelihood method for model (1) first in Section 3.1, then propose the penalized empirical likelihood method and consequently the corresponding penalized empirical likelihood ratio test for a general linear hypothesis of parameters in Section 3.2.

3.1 Empirical Likelihood

Motivated by representation (5), we define

$$\boldsymbol{G}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\}dM_i(\boldsymbol{\beta}, t), \tag{6}$$

where $M_i(\boldsymbol{\beta}, t) = N_i(t) - \int_0^t Y_i(s)\{h_0(s) + \boldsymbol{\beta}^T\boldsymbol{Z}_i(s)\}ds$. It is easy to obtain that $E\boldsymbol{G}(\boldsymbol{\beta}_0) = \mathbf{0}$ since $\boldsymbol{G}(\boldsymbol{\beta}_0) = \boldsymbol{U}(\boldsymbol{\beta}_0)$ is a martingale from the estimating equation (5). Recall that $\frac{1}{n}\sum_{i=1}^n\int_0^\tau Y_i(t)\{\boldsymbol{Z}(t) - \bar{\boldsymbol{Z}}(t)\}\bar{\boldsymbol{Z}}^T(t)dt = 0$. Similar to (3), we have

$$\begin{aligned}
\boldsymbol{G}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\}\{dN_i(t) - Y_i(t)\boldsymbol{\beta}^T\boldsymbol{Z}_i(t)dt - Y_i(t)h_0(t)dt\} \\
&= \boldsymbol{G}(\boldsymbol{\beta}_0) + \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau Y_i(t)\{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\}^{\otimes 2}dt(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) \\
&= \boldsymbol{G}(\boldsymbol{\beta}_0) + \boldsymbol{D}_n(\boldsymbol{\beta}_0 - \boldsymbol{\beta}).
\end{aligned}$$

When the eigenvalues of $\boldsymbol{D}$ deviate from zero and infinity, it follows that $E\boldsymbol{G}(\boldsymbol{\beta}) = \boldsymbol{D}_n(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) \to \boldsymbol{D}(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) = \mathbf{0}$ if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Thus, we can apply the empirical likelihood method (Owen, 2001) based on the estimating equations (6).

Note that the cumulative baseline hazard function $\Lambda_0(t) = \int_0^t h_0(s)ds$ and martingale $M_i(t, \boldsymbol{\beta})$ are unknown. An estimated empirical likelihood can be defined as follows. Let $\boldsymbol{p} = (p_1, \ldots, p_n)$ be a probability vector, namely, $\sum_{i=1}^n p_i = 1$ and $p_i \geq 0$ for all i. For $1 \leq i \leq n$, let

$$\boldsymbol{W}_{ni}(\boldsymbol{\beta}) = \int_0^\tau \{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\}d\hat{M}_i(\boldsymbol{\beta}, t), \tag{7}$$

where $\hat{\boldsymbol{M}}_i(\boldsymbol{\beta}, t) = N_i(t) - \int_0^t Y_i(u)\{\boldsymbol{\beta}^T\boldsymbol{Z}_i(u)du + d\hat{\Lambda}_0(\boldsymbol{\beta}, u)\}$, $\hat{\Lambda}_0(\boldsymbol{\beta}, t) = \int_0^t \frac{\sum_{i=1}^n\{dN_i(u) - Y_i(u)\boldsymbol{\beta}^T\boldsymbol{Z}_i(u)du\}}{\sum_{i=1}^n Y_i(u)}$ being the Aalen-Breslow type estimator for $\Lambda_0(t)$ when evaluated at $\boldsymbol{\beta}_0$. Then, $\boldsymbol{W}_{ni}(\boldsymbol{\beta}_0)$ can be treated as an estimated version of the zero-mean martingale $\int_0^\tau\{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\}dM_i(t)$. Similar to the definition of the conventional empirical likelihood for mean (Owen, 1990), we therefore define the following estimated empirical likelihood function for $\boldsymbol{\beta}$ with zero mean restriction on $W_{ni}(\boldsymbol{\beta})$:

$$L(\boldsymbol{\beta}) = \sup\Big\{\prod_{i=1}^n p_i : 0 \leq p_i \leq 1, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i\boldsymbol{W}_{ni}(\boldsymbol{\beta}) = 0\Big\}. \tag{8}$$

The empirical likelihood estimate of $\boldsymbol{\beta}$ is obtained by maximizing $\prod_{i=1}^n p_i$ subject to $0 \leq p_i \leq 1, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i W_{ni}(\beta) = 0$, which is equivalent to maximizing $\sum_{i=1}^n \log p_i$ subject to the same conditions. Applying the Lagrange Multiplier method, it is easy to obtain that $p_i = n^{-1}\{1 + \boldsymbol{\nu}^T\boldsymbol{W}_{ni}\}^{-1}$, where $\boldsymbol{\nu}$ is an $p \times 1$ vector of Lagrange Multipliers satisfying the equation

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{W}_{ni}(\boldsymbol{\beta})}{1 + \boldsymbol{\nu}^T W_{ni}(\boldsymbol{\beta})} = 0. \tag{9}$$

Note that $\prod_{i=1}^n p_i$, subject to $\sum_{i=1}^n p_i = 1$, attains its maximum n^{-n} at $p_i = n^{-1}$. Thus the empirical likelihood ratio for $\boldsymbol{\beta}$ is given by $R(\boldsymbol{\beta}) = \prod_{i=1}^n (np_i) = \prod_{i=1}^n \{1 + \boldsymbol{\nu}^T \boldsymbol{W}_{ni}(\boldsymbol{\beta})\}^{-1}$. Consequently, the empirical log-likelihood ratio can be formulated as

$$l(\boldsymbol{\beta}) = -2 \log R(\boldsymbol{\beta}) = 2 \sum_{i=1}^n \log \Big\{1 + \boldsymbol{\nu}^T \boldsymbol{W}_{ni}(\boldsymbol{\beta})\Big\}, \tag{10}$$

where $\boldsymbol{\nu}$ is the solution of (9). Hence, the empirical likelihood estimator of $\boldsymbol{\beta}$ is obtained by minimizing $l(\boldsymbol{\beta})$, i.e., $\hat{\boldsymbol{\beta}}_E = \arg\min_\beta l(\boldsymbol{\beta})$.

In comparison with Lin and Ying's estimator $\hat{\boldsymbol{\beta}}^{\mathrm{LY}}$ based on the pseudo-likelihood method, the estimator $\hat{\boldsymbol{\beta}}_E$ is derived from the empirical likelihood method and has no explicit expression. However, both estimators are equivalent when the dimension p is fixed. In fact, we note that $l(\boldsymbol{\beta}) \geq 0$. Since the dimension of $\boldsymbol{\beta}$ is equal to the number of estimating functions $\boldsymbol{W}_{ni}(\boldsymbol{\beta})$, it is easy to see that $l(\boldsymbol{\beta})$ attains its minimum 0 at $\tilde{\boldsymbol{\beta}}$, where $\tilde{\boldsymbol{\beta}}$ is the root of the estimating equations $\sum_{i=1}^n \boldsymbol{W}_{ni}(\boldsymbol{\beta}) = 0$. Thus, $\hat{\boldsymbol{\beta}}_E = \arg\min_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) = \tilde{\boldsymbol{\beta}}$. Next, recall the definition of $\boldsymbol{W}_{ni}(\boldsymbol{\beta})$ and the facts that $\bar{\boldsymbol{Z}}(t) = \sum_{i=1}^n \boldsymbol{Z}_i(t) / \sum_{i=1}^n Y_i(t)$ and $\sum_{i=1}^n \{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\} Y_i(t) = 0$. We have

$$\begin{aligned}
\sum_{i=1}^n \boldsymbol{W}_{ni}(\boldsymbol{\beta}) &= \sum_{i=1}^n \int_0^\tau \{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\} d\hat{M}_i(\boldsymbol{\beta}, t) \\
&= \sum_{i=1}^n \int_0^\tau \{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\} d\{N_i(t) - Y_i(t) \boldsymbol{\beta}^T \boldsymbol{Z}_i(t) dt\} - \sum_{i=1}^n \int_0^\tau \{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\} Y_i(t) d\hat{\Lambda}_0(\boldsymbol{\beta}, t) \\
&= U(\boldsymbol{\beta}) - \int_0^\tau \Big[\sum_{i=1}^n \{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\} Y_i(t)\Big] d\hat{\Lambda}_0(\beta, t) = U(\boldsymbol{\beta}).
\end{aligned}$$

Thus, solving equations $\sum_{i=1}^n \boldsymbol{W}_{ni}(\boldsymbol{\beta}) = 0$ is equivalent to solving equations $U(\boldsymbol{\beta}) = 0$. This implies that $\hat{\boldsymbol{\beta}}_E$ reduces to $\hat{\boldsymbol{\beta}}^{\mathrm{LY}}$ in the fixed-dimensional cases.

Nevertheless, it is worth noting that $\hat{\boldsymbol{\beta}}_E$ is obtained from the empirical likelihood method under a high-dimensional framework by allowing p to tend to ∞ as the sample size increases. Hence the derivation of its subsequent asymptotic theory is much more challenging. In addition, unlike the fixed dimensional cases (Chen et al., 2009, Hjort et al., 2009), the magnitude of $\|\boldsymbol{\nu}\|$ is no longer $O_p(n^{-1/2})$ in the empirical likelihood with diverging p. To develop an asymptotic expansion for (10), $\boldsymbol{\nu}^T \boldsymbol{W}_{ni}(\boldsymbol{\beta})$ has to be stochastically and uniformly small. Let $\boldsymbol{W}_i(\boldsymbol{\beta}) = \int_0^\tau \{\boldsymbol{Z}_i(t) - \boldsymbol{e}(t)\} dM_i(\boldsymbol{\beta}, t)$ and $\boldsymbol{W}(\boldsymbol{\beta}) = E(\boldsymbol{W}_i(\boldsymbol{\beta}))$. We define that $a_n = (p/n)^{1/2}$ and $E_n = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq C a_n\}$ being a neighborhood of $\boldsymbol{\beta}_0$ for some constant $C > 0$. Moreover, we make the following assumptions.

Assumption 1. Assume that $0 \leq \tau \leq \infty$ and $P(X \geq \tau) > 0$, which implies $P(T \geq \tau) > 0$ and $P(C \geq \tau) > 0$.

Assumption 2. $\sup_{t \in [0,\tau]} |\hat{\boldsymbol{s}}^{(1)}(t) - \boldsymbol{s}^{(1)}(t)| = O_p(1)$, where $\hat{\boldsymbol{s}}^{(1)}(t) = n^{-1} \sum_{i=1}^n Y_i(t) \boldsymbol{Z}_i(t)$.

Assumption 3. Let $\boldsymbol{\Sigma}(\boldsymbol{\beta}) = E\{\boldsymbol{W}_1(\boldsymbol{\beta}) - \boldsymbol{W}(\boldsymbol{\beta})\}^{\otimes 2}$. Let $\gamma_j\{\boldsymbol{A}\}$ denote the jth ordered eigenvalue of matrix $\boldsymbol{A}$. There exists b and B such that the eigenvalues of $\boldsymbol{\Sigma}(\boldsymbol{\beta})$ satisfy $0 < b \leq \gamma_1\{\boldsymbol{\Sigma}(\boldsymbol{\beta})\} \leq \cdots \leq \gamma_p\{\boldsymbol{\Sigma}(\boldsymbol{\beta})\} \leq B < \infty$ for all $\boldsymbol{\beta} \in E_n$ when n is large. Especially, $\boldsymbol{\Sigma}(\boldsymbol{\beta}_0) = \boldsymbol{\Sigma}$. Matrix $\boldsymbol{D}$ also satisfies $0 < b \leq \gamma_1\{\boldsymbol{D}\} \leq \cdots \leq \gamma_p\{\boldsymbol{D}\} \leq B < \infty$.

Assumption 4. Assume that, for each component of $\boldsymbol{W}_i(\boldsymbol{\beta}_0)$,

$$
\begin{aligned}
E\Big| \int_0^\tau \{Z_{ij}(t) - e_j(t)\}\{dM_i(\boldsymbol{\beta}_0, t)\}\Big|^q &< \infty, \\
E\Big| \int_0^\tau Y_i(t)\{Z_{ij}(t) - e_j(t)\}\frac{\sum_{j=1}^n dM_j(\boldsymbol{\beta}_0, t)}{\sum_{j=1}^n Y_j(t)}\Big|^q &< \infty, \\
E\Big| \int_0^\tau \{d\tilde{N}_i(t) - Y_i(t)\boldsymbol{\beta}_0^T\{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\}dt\}\Big|^q &< \infty, \\
E\Big| \int_0^\tau e_j(t)\{d\tilde{N}_i(t) - Y_i(t)\boldsymbol{\beta}_0^T\{\boldsymbol{Z}_i(t) - \bar{\boldsymbol{Z}}(t)\}dt\}\Big|^q &< \infty,
\end{aligned}
$$

for some $q \geq 10$ when n is large, where $d\tilde{N}_i(t) = dN_i(t) - \frac{Y_i(t)\sum_{j=1}^n dN_j(t)}{\sum_{j=1}^n Y_j(t)}$.

Assumption 5. $p \to \infty$ and $p^5/n \to 0$.

Remark 1. Assumption 1 is needed for technical reasons such that $\inf_{t\in[0,\tau]}\sum_{i=1}^n Y_i(t) \xrightarrow{p} \infty$ as $n \to \infty$, which simply guarantees that the number of individuals at risk at each time point becomes larger. Assumption 2 is necessary for $\bar{\boldsymbol{Z}}(t)$ to estimate $\boldsymbol{e}(t)$ well, while Assumption 3 is the regularized condition imposed in the high-dimensional context. Assumption 4 is needed to ensure the existence and consistency of the minimizer of (10) and to control the tail probability behavior of the estimating function. Assumption 5 requires that $p = o(n^{1/5})$, which is also the case in Leng and Tang (2012) and Fan and Peng (2004). Compared with the growth rates of p obtained in some early studies (e.g., Hjort et al., 2009 and Chen et al., 2009) on empirical likelihood for high-dimensional data, Assumption 5 leads to a relatively conservative rate due to lack of particular structural information available for the proposed estimating equations, data subject to censoring and hence providing incomplete information and an unknown baseline hazard function estimated nonparametrically in the estimation procedure. In fact, based on the proofs of the following Theorems 1-2, Assumption 5 can be relaxed to $p^2 = o_p(n^{1/2-1/q})$ for some $q \geq 10$. In other words, $p^5/n \to 0$ when $q \leq 10$ and nearly $p^4/n \to 0$ as q increases.

Given the assumptions above, we can show that the proposed estimator $\hat{\boldsymbol{\beta}}_E$ is consistent and asymptotic normal when dimension p is diverging. Our main theoretical results are summarized in the following two theorems. Rigorous proofs are provided in the appendix.

Theorem 1. *Suppose that Assumptions 1-5 hold. The minimizer $\hat{\boldsymbol{\beta}}_E$ of (10) satisfies (a) $\hat{\boldsymbol{\beta}}_E \to \boldsymbol{\beta}_0$*

in probability as $n \to \infty$, *and (b)* $\|\hat{\boldsymbol{\beta}}_E - \boldsymbol{\beta}_0\| = O_p(a_n)$.

Theorem 2. *Suppose that Assumptions 1-5 hold.* $\sqrt{n}\boldsymbol{A}_n\boldsymbol{\Psi}^{-1/2}\{\hat{\boldsymbol{\beta}}_E - \boldsymbol{\beta}_0\} \to N_q(0, \boldsymbol{G})$ *in distribution as* $n \to \infty$, *where* $\boldsymbol{A}_n$ *is a* $q \times p$ *matrix such that* $\boldsymbol{A}_n\boldsymbol{A}_n^T \to \boldsymbol{G}$, $\boldsymbol{G}$ *is a* $q \times q$ *nonnegative symmetric matrix with fixed* q *and* $\boldsymbol{\Psi} = \boldsymbol{D}^{-1}\boldsymbol{\Sigma}\boldsymbol{D}^{-1}$.

In the fixed-dimensional setting, the asymptotic covariance of $\hat{\boldsymbol{\beta}}_E$ is obtained from Theorem 2 as $\boldsymbol{\Psi} = \boldsymbol{D}^{-1}\boldsymbol{\Sigma}\boldsymbol{D}^{-1}$, which is actually the same as the asymptotic covariance of the pseudoscore estimator $\hat{\boldsymbol{\beta}}^{LY}$ given by Lin and Ying (1994). Theorem 2 provides an further extension of this result to the additive hazards model with diverging number of variables. The empirical likelihood estimation developed in this section results in that none of the estimated coefficients is exactly zero, leaving all covariates in the final model. So it is unable to select important variables. When the sparsity exists, a more efficient estimator is desirable.

3.2 Penalized Empirical Likelihood

To balance modeling biases and achieve variable selection in the high-dimensional setting, Lin and Lv (2013) defined the regularized estimator as the solution of a penalized least squares-type loss function. In this section, we propose to use a penalized empirical likelihood for model (1) by complementing (10) with a penalty function $p_\lambda(|\beta_j|)$. In particular, the penalized log-empirical likelihood function is defined as

$$l_p(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log\left\{1 + \boldsymbol{\nu}^T\boldsymbol{W}_{ni}(\boldsymbol{\beta})\right\} + n\sum_{j=1}^{p} p_\lambda(|\beta_j|). \tag{11}$$

where $p_\lambda(\cdot)$ is a given nonnegative penalty function with λ as a tuning parameter, which in general can be chosen by a data-driven criterion. The penalized empirical likelihood estimator, denoted by $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\lambda)$, minimizes (11) for given λ.

To solve the optimization problems of (10) or (11), 0 needs to be in the convex hull formed by $\{\boldsymbol{W}_{ni}(\boldsymbol{\beta})\}_{i=1}^n$. When $l_p(\boldsymbol{\beta})$ in (11) is non-convex, we will consider a local minimizer, as is common in the literature. Without the penalty term, $\hat{\boldsymbol{\beta}}$ reduces to the empirical likelihood estimate $\hat{\boldsymbol{\beta}}_E$ or the pseudo-score estimator $\hat{\boldsymbol{\beta}}^{LY}$ when dimension p is fixed. When the dimensionality is high, however, some form of regularization is needed to guard against over-fitting, and the performance of the regularized estimator depends on the choice of the penalty function.

There have been many penalty functions proposed in the literature for variable selection, such as the Lasso penalty (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD) penalty (Fan

and Li, 2001), elastic-net penalty (Zou and Hastie, 2005), adaptive L_1 (Zou, 2006), and minimax concave penalty (Zhang, 2010). We use the SCAD penalty in our method though other penalties can also be used. Specifically, the first derivative of SCAD penalty satisfies

$$p'_\lambda(|\boldsymbol{\beta}|) = \lambda \, \mathrm{sgn}(\boldsymbol{\beta})\Big\{I(|\boldsymbol{\beta}| \le \lambda) + \frac{(a\lambda - |\boldsymbol{\beta}|)_+}{(a-1)\lambda} I(|\boldsymbol{\beta}| > \lambda)\Big\}, \text{ for some } a > 2,$$

and $(s)_+ = s$ for $s > 0$ and 0 otherwise. The SCAD penalty is non-convex, leading to a non-convex optimization. For the non-convex SCAD penalized optimization, Fan and Li (2001) proposed the local quadratic approximation, while Zou and Li (2008) proposed the local linear approximation. Here, whenever necessary we use the local quadratic approximation together with the nested optimization procedure (Owen, 2001) to solve the SCAD penalized empirical likelihood optimization.

We now state the main asymptotic findings for the penalized empirical likelihood estimator $\hat{\boldsymbol{\beta}}$. Some additional notations and conditions are needed. Let $\mathcal{A} = \{j : \beta_{0j} \neq 0\}$ be the set of nonzero components of the true parameter vector $\boldsymbol{\beta}_0$ and denote the cardinality of $\mathcal{A}$ as $|\mathcal{A}| = d$ which is unknown. Here we allow d to grow at the same rate as p when $n \to \infty$ without imposing any specific restriction. Without loss of generality, let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, where $\boldsymbol{\beta}_1 \in \mathbf{R}^d$ and $\boldsymbol{\beta}_2 \in \mathbf{R}^{p-d}$. We assume that the last $p-d$ components of the true parameter $\boldsymbol{\beta}_0$ are zeros; that is, $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, \mathbf{0}^T)^T$, where $\boldsymbol{\beta}_{10}$ contains non-zero components. Correspondingly, we partition the asymptotic covariance matrix $\boldsymbol{\Psi}$ as $\begin{pmatrix} \boldsymbol{\Psi}_{11} & \boldsymbol{\Psi}_{12} \\ \boldsymbol{\Psi}_{21} & \boldsymbol{\Psi}_{22} \end{pmatrix}$ with diagonal submatrices $\boldsymbol{\Psi}_{11}$ and $\boldsymbol{\Psi}_{22}$ of dimensions $d \times d$ and $(p-d) \times (p-d)$, respectively. In addition, the following two conditions are imposed on the penalty function $p_\lambda(\cdot)$. It is important to note that the quantities p and λ depend on the sample size n, and we have suppressed this dependency for notational simplicity.

Condition 1. As $n \to \infty$, the tuning parameter λ satisfies $\lambda \to 0$ and $\lambda(n/p)^{1/2} \to \infty$. The nonzero components satisfy $\min_{j \in \mathcal{A}} |\beta_{0j}|/\lambda \to \infty$.

Condition 2. The function $p_\lambda(\cdot)$ satisfies $\max_{j \in \mathcal{A}} p'_\lambda(|\beta_{0j}|) = o_p((np)^{-1/2})$ and $\max_{j \in \mathcal{A}} p''_\lambda(|\beta_{0j}|) = o_p(p^{-1/2})$.

Remark 2. Condition 1 states that the weakest signal should dominate the penalty parameter λ, which is routinely made to ensure the recovery of signals. Condition 2 is used in controlling the impact of penalty on the nonzero component. For the SCAD penalty (Fan and Li, 2001), Condition 2 is satisfied, because $\max_{j \in \mathcal{A}} p'_\lambda(|\beta_{0j}|) = 0$ for n large enough given Condition 1.

The following theorem establishes the theoretical properties of the estimator $\hat{\boldsymbol{\beta}}$. The proof is given in the Appendix.

Theorem 3. *Suppose that Assumptions 1-5 and Conditions 1-2 hold. When $n \to \infty$, the penalized empirical estimator $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ satisfies*

(i) Sparsity: $\hat{\boldsymbol{\beta}}_2 = 0$ with probability tending to 1;

(ii) Asymptotic Normality: $\sqrt{n}\boldsymbol{A}_n I_{\mathcal{A}}^{-1/2}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{10}) \to N_q(0, \boldsymbol{G})$ in distribution, where $I_{\mathcal{A}} = \boldsymbol{\Psi}_{11} - \boldsymbol{\Psi}_{12}\boldsymbol{\Psi}_{22}^{-1}\boldsymbol{\Psi}_{21}$, $\boldsymbol{A}_n$ is a $q \times d$ matrix such that $\boldsymbol{A}_n\boldsymbol{A}_n^T \to \boldsymbol{G}$, and $\boldsymbol{G}$ is a $q \times q$ nonnegative symmetric matrix with fixed q.

Theorem 3 shows that the penalized empirical likelihood estimator $\hat{\boldsymbol{\beta}}$ enjoys the oracle property as defined in Fan and Peng (2004). That is, the penalized empirical likelihood approach is consistent in model selection and performs as efficient as if the true sparse model was known.

Remark 3. The oracle property formulated in Theorem 2 of Lin and Lv (2013) is also applicable in the "large n, diverging p" framework of our study. In fact, their regularized estimator, denoted by $\hat{\boldsymbol{\beta}}^{LL} = (\hat{\boldsymbol{\beta}}_1^{LL}, \hat{\boldsymbol{\beta}}_2^{LL})$, satisfies that $\hat{\boldsymbol{\beta}}_2^{LL} = \mathbf{0}$ with probability tending to one and $n^{1/2}\boldsymbol{A}_n\boldsymbol{\Psi}_{11}^{-1/2}(\hat{\boldsymbol{\beta}}_1^{LL} - \boldsymbol{\beta}_{10}) \to N_d(0, \boldsymbol{G})$ in distribution. We note that $\hat{\boldsymbol{\beta}}_1^{LL}$ coincides with the corresponding subvector $\hat{\boldsymbol{\beta}}_{E1}$ of $\hat{\boldsymbol{\beta}}_E$, whose asymptotic distribution is characterized through result that $n^{1/2}A_n\Psi_{11}^{-1/2}(\hat{\boldsymbol{\beta}}_{E1} - \boldsymbol{\beta}_{10}) \to N_d(0, \boldsymbol{G})$ in distribution from Theorem 2. Based on result (ii) in Theorem 3, the proposed penalized empirical likelihood method therefore results in a more efficient estimator of the nonzero component, $\hat{\boldsymbol{\beta}}_1$, than $\hat{\boldsymbol{\beta}}_1^{LL}$ or $\hat{\boldsymbol{\beta}}_{E1}$. As shown in the proof of Theorem 3 in the supplementary material, this efficiency gain is due to the reduction of the effective dimension of $\boldsymbol{\beta}$ via penalization.

A remarkable advantage of empirical likelihood lies in testing hypotheses and constructing confidence regions for $\boldsymbol{\beta}$. The penalized empirical likelihood inherits this advantage in model (1) with a diverging number of parameters. To understand this more clearly, we consider the problem of testing linear hypotheses:

$$H_0 : \boldsymbol{A}\boldsymbol{\beta}_{10} = \mathbf{0} \text{ vs } \ H_1 : \boldsymbol{A}\boldsymbol{\beta}_{10} \neq \mathbf{0}, \tag{12}$$

where $\boldsymbol{A}$ is a $q \times d$ matrix satisfying $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}_q$ with fixed q. Note that $d = dim(\boldsymbol{\beta_{01}})$ is possible to be very big, while q is fixed and often small. For example, the most common hypothesis is for the significance of the individual component of β_{10}, i.e., $q = 1$. For cases with diverging q, further work is required to extend the results in Theorem 3 as discussed in Hjort et al. (2008). That is beyond the scope of the current paper. Thus, we consider (12) with $q \ll d$ only. It includes hypotheses for individual and multiple components of $\boldsymbol{\beta}_{10}$ as special cases. A similar type of hypothesis testing was studied by Fan and Peng (2004) in a parametric likelihood framework, and Tang and Leng (2010)

and Leng and Tang (2012) in an empirical likelihood framework. Based on the penalized empirical likelihood formulation, a penalized empirical likelihood ratio test statistic can be obtained by

$$PELR = -2\Big\{l_p(\hat{\boldsymbol{\beta}}) - \min_{\boldsymbol{\beta},\boldsymbol{A}\boldsymbol{\beta}_1=0} l_p(\boldsymbol{\beta})\Big\}. \tag{13}$$

The following theorem states the asymptotic null distribution of the test statistic in (13).

Theorem 4. *Suppose that Assumptions 1-5 and Conditions 1-2 hold. Then, under the null hypothesis H_0, $PELR \to \chi^2_q$ in distribution as $n \to \infty$, where χ^2_q is the chi-squared distribution with q degrees of freedom.*

Theorem 4 extends the result of Leng and Tang (2012) to high-dimensional survival data. It shows that the Wilks' phenomenon holds for the penalized empirical likelihood approach proposed in this study. This theorem provides a convenient way for testing hypotheses and constructing data-oriented confidence regions, avoiding to estimate the asymptotic covariance matrix of the estimated parameters. As a direct consequence of Theorem 4, a $(1-\alpha)100\%$ confidence region for $\boldsymbol{A}\boldsymbol{\beta}_1$ can be constructed as

$$C_\alpha = \Big\{\boldsymbol{v} : -2\{l_p(\hat{\boldsymbol{\beta}}) - \min_{\boldsymbol{\beta},\boldsymbol{A}\boldsymbol{\beta}_1=\boldsymbol{v}} l_p(\boldsymbol{\beta})\} \le \chi^2_{q,1-\alpha}\Big\},$$

where $\chi^2_{q,1-\alpha}$ is the $1-\alpha$ quantile of the χ^2_q distribution.

4 Implementation

Due to nonconvexity, computing the empirical likelihood is nontrivial. The penalized empirical likelihood computation involving a non-differentiable penalty tends to be more challenging, especially when performing hypothesis testing.

We adopt a nested optimization procedure as given in Owen (2001) to minimize (11), and iterate between solving for $\boldsymbol{\nu}$ and $\boldsymbol{\beta}$. When $\boldsymbol{\nu}$ is fixed, we use the local quadratic approximation to the penalty following Fan and Li (2001). Suppose that we have the current kth iterative estimate $\hat{\boldsymbol{\beta}}^{(k)}$ of $\boldsymbol{\beta}$. If $\hat{\beta}_j^{(k)}$ is near zero, namely, $|\hat{\beta}_j^{(k)}| \le \epsilon$ with ϵ is the a pre-specified tolerance, then set $\hat{\beta}_j = 0$. Otherwise, approximate $p_\lambda(|\beta_j|)$ by $p_\lambda(|\hat{\beta}_j^{(k)}|) + \frac{p'_\lambda(|\hat{\beta}_j^{(k)}|)}{2|\hat{\beta}_j^{(k)}|}\{\beta_j^2 - \hat{\beta}_j^{(k)}\}$. With the aid of the local quadratic approximation, we then minimize the penalized empirical likelihood (11) through the nested optimization. Removing the irrelevant constants, it consequently follows that

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \arg\min_{\boldsymbol{\beta}} \Big\{ \sum_{i=1}^{n} \log\{1 + \boldsymbol{\nu}^{(k)T}\boldsymbol{W}_{ni}(\boldsymbol{\beta})\} + n\sum_{j=1}^{p} \frac{p'_\lambda(|\hat{\beta}_j^{(k)}|)}{2|\hat{\beta}_j^{(k)}|}\beta_j^2 \Big\}. \tag{14}$$

We plug $\hat{\boldsymbol{\beta}}^{(k+1)}$ into Eq (9) and obtain the updated $\boldsymbol{\nu}^{(k+1)}$ through solving the equation

$$\sum_{i=1}^{n} \frac{\boldsymbol{W}_{ni}(\hat{\boldsymbol{\beta}}^{(k+1)})}{1+\boldsymbol{\nu}^T \boldsymbol{W}_{ni}(\hat{\boldsymbol{\beta}}^{(k+1)})} = 0. \tag{15}$$

For a given value of λ, we use $\hat{\boldsymbol{\beta}}^{LY}$ as an initial value for $\boldsymbol{\beta}_0$, and $\mathbf{0}$ for the Lagrange multiplier coefficient $\boldsymbol{\nu}$. Taking into account of the derivations above, the estimation algorithm is proposed as follows:

- Step 1. Set $\hat{\boldsymbol{\nu}}^{(0)} = \mathbf{0}, \hat{\boldsymbol{\beta}}^{(0)} = \hat{\boldsymbol{\beta}}^{LY}$ and $k = 0$.
- Step 2. For fixed $\hat{\boldsymbol{\nu}}^{(k)}$, update $\hat{\boldsymbol{\beta}}^{(k+1)}$ using Eq (14).
- Step 3. For fixed $\hat{\boldsymbol{\beta}}^{(k+1)}$, update $\hat{\boldsymbol{\nu}}^{(k+1)}$ using Eq (15).
- Step 4. Set $k = k+1$; repeat steps 2-3 until the algorithm converges and denote the resulting estimates by $\hat{\boldsymbol{\beta}}_\lambda$ and $\hat{\boldsymbol{\nu}}_\lambda$.

Our experience in both the simulations and real data examples suggests that this algorithm usually converges with the initial values given above. Note that $\hat{\boldsymbol{\nu}}_\lambda$ depends on $\hat{\boldsymbol{\beta}}_\lambda$, then we obtain

$$l_p(\hat{\boldsymbol{\beta}}_\lambda) = \sum_{i=1}^{n} \log\left\{1+\hat{\boldsymbol{\nu}}_\lambda^T W_{ni}(\hat{\boldsymbol{\beta}}_\lambda)\right\} + n\sum_{j=1}^{p} p_\lambda(\hat{\boldsymbol{\beta}}_\lambda).$$

The tuning parameter λ can be chosen through a data-driven algorithm. We determine the value of λ based on the Bayesian information criterion (BIC) (Schwarz, 1978). Specifically, the optimal λ is selected through minimizing the following BIC-type criterion

$$\mathrm{BIC}(\lambda) = -2l_p(\hat{\boldsymbol{\beta}}_\lambda) + C_n\{\log n\} df_\lambda, \tag{16}$$

where df_λ is the number of nonzero coefficients in $\hat{\boldsymbol{\beta}}_\lambda$, and C_n is a scaling factor diverging to infinity at a slow rate as $p \to \infty$. When p is fixed, we can simply take $C_n = 1$ as for the usual BIC. Otherwise, $C_n = \max\{\log\log p, 1\}$ seems to be a good choice. A rigorous proof of the consistency of this BIC for penalized empirical likelihood is much more challenging than for the linear regression model considered in Wang et al. (2009), and merits further investigation. Our simulation studies demonstrate that the BIC-type criterion defined in (16) often selects the tuning parameter satisfactorily and identifies the true model consistently.

5 Simulation Study

We generated data from the model $h(t|\boldsymbol{Z}) = 1 + \boldsymbol{\beta}_0^T\boldsymbol{Z}$, where $\boldsymbol{Z}$ has a multivariate normal distribution with mean zero and covariance matrix $(\rho^{|i-j|})_{i,j=1}^p$ and subject to $\boldsymbol{\beta}_0^T\boldsymbol{Z} > -1$ and $\boldsymbol{\beta}_0 = (\boldsymbol{v}^T, \ldots, \boldsymbol{v}^T, 0, \ldots, 0)^T$ with the pattern $\boldsymbol{v} = (2, 0, -2)^T$ repeated k times. We set $\rho = 0.1$ and 0.5, and $k = 2$ and 3 so that the sparsity dimension $d = 4$ and 6, respectively. We consider the sample size $n = 100$, 200 and 400 with the dimension $p = 9, 15$ and 21. The censoring time C has a uniform distribution $U(0, c_0)$, where c_0 is chosen to obtain a censoring rate about 10% and 20%, respectively. The BIC-type criterion defined in (16) is used to estimate the optimal tuning parameter λ in the SCAD penalty. Figure 1 shows the solution paths for regression coefficients with an excerpt $(0, 0.4185)$ of possible λ values and the BIC scores in the case with $n = 400$, $\rho = 0.1$ and $k = 2$. We observe that the four variables corresponding to the nonzero coefficients are selected correctly using the proposed method and the BIC-type criterion is able to identify the true model consistently. All simulations are done with R codes and available upon request.

[Figure 1]

We evaluate the performance of the resulting estimators by five measures. The first measure quantify prediction performance, namely, PE $= \|\boldsymbol{Z}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2$, which is the L_2 prediction error in the excess risk and computed from an independent test sample of size 500. For estimation accuracy, we report the L_2-loss $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = \{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\}^{1/2}$ and L_1-loss $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_1 = \sum_{j=1}^p |\hat{\beta}_j - \beta_{0j}|$. The other two measures pertain to model selection consistency: #C and #IC refer to the number of correctly estimated zero coefficients and the number of incorrectly excluded variables, respectively. The means and standard deviations of each measure over 200 replicates are summarized in Tables 1 and 2 for sparsity dimension $d = 4$ and 6, respectively. As sample size n increases, the number of correctly selected variables approaches the actual number of nonzero coefficients $p-d$ ($p-4$ in Table 1 and $p - 6$ in Table 2) in most cases. This tendency becomes more obvious for a large censoring rate or large correlation ρ among variables. The number of incorrectly excluded variables is near 0 throughout all the simulated cases. It confirms the selection consistency results in Theorem 3.

[Tables 1 and 2]

To further demonstrate the shrinkage effect of the proposed method over different parameter settings in the simulation, we show the boxplots of the estimated coefficient $\hat{\boldsymbol{\beta}}$ in Figures 2 and 3

for sparsity levels $d = 4$ and $d = 6$, respectively. It is seen that our method performs very well in isolating the non-zero coefficients and shrinking the others considerably in all settings.

[Figures 2 and 3]

Inspired by a referee's suggestion, we further estimate the parameter using Lin and Ying's pseudo-likelihood estimation. The corresponding results are summarized graphically in Figure 4 based on 500 simulations with different sparsity dimensions when $n = 100$ and $n = 200$. It can be seen that $\hat{\boldsymbol{\beta}}^{LY}$ performs worse than the proposed PEL estimator in shrinking the parameters even when the sample size is 100.

[Figure 4]

We next examine the performance of the penalized empirical likelihood ratio test given in Theorem 4. The null hypothesis is specified as $H_0 : \beta_{01} = 2$, where β_{01} is the first component of β_0. Using a nominal level $\alpha = 0.05$, we compute the empirical percentage of rejecting $H_0 : \beta_{01} = 2$ for different values of the truth $\beta_{01} = 2, 2.10, 2.20, \ldots, 2.50$, respectively. The empirical size and power results are reported in Table 3. The size of the test is found to be close to the nominal level as the sample size increases. The power increases as either the sample size increases or the true value of β_{01} deviates far from the null hypothesis $H_0 : \beta_{01} = 2$. In addition, we show the QQplots of the penalized empirical likelihood ratio test statistic against the nominal χ^2_1 distribution under H_0 when sample size $n = 400$ and sparsity dimension $d = 4$ in Figure 5, which verifies the asymptotic distribution result in Theorem 4.

[Table 3]

[Figure 5]

Following a referee's suggestion, we also have conducted further simulations for comparison of different penalization methods. As an alternative of the SCAD method, we consider the adaptive Lasso penalty in the proposed penalized empirical likelihood estimation. The weights used in the adaptive Lasso are specified by $1/|\beta_i|^{0.5}$ (Zhang and Lu, 2010), where β_i is the non-penalized EL estimates of β_i. Results of parameter estimation and variable selection based on the adaptive Lasso penalty are incorporated in Tables 1-3. It can be seen that the PEL-adaptive Lasso performs very similar to the proposed PEL-SCAD. In addition, we have carried out simulations for smaller signal

levels, e.g., $\beta_0 = (\nu^T, \cdots, \nu^T, 0, \cdots, 0)^T$ with $\nu = (1.5, 0, -1.5)^T$ and $(0.8, 0, -0.8)^T$ repeated k times, while the settings for the other variables are kept unchanged. Our simulation results show that in general the proposed method performs well in parameter estimation and variable selection for small signals though the estimated standard derivations of PE, L_2-loss, L_1-loss and $\#C$ are slightly greater than their corresponding counterparts in Tables 1 and 2. Results of the size and power of the PELR test for $H_0 : \beta_{01} = 0.8$ are found very similar to that those for $H_0 : \beta_{01} = 2$ given in Table 3.

6 Practical Examples

For illustration purpose, we apply the proposed method to analyze data sets from two practical examples in this section.

6.1 Analysis of the PBC Data

Data of 424 patients suffering from primary biliary cirrhosis of the liver (PBC) were collected by the Mayo Clinic between January 1974 and May 1984. Seventeen covariates (clinical and laboratory measurements) were collected for 312 randomized patients, while the other 112 subjects did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. We discarded observations with missing values in the covariates, leaving 276 valid observations. More details about the PBC data can be found in Fleming and Harrington (2011). Several authors investigated this dataset. Among these, Tibshirani (1997) illustrated the PBC data under the Cox model assumption, while Ma and Huang (2005) and Leng and Ma (2007) studied it under the additive hazards model assumption. Leng and Ma (2007) observed that the additive hazards model was better fitted than the Cox model in terms of the area under curve (AUC) criterion in the time-dependent receiver operating characteristic (ROC) approach.

We standardize all the covariates so that they are more comparable in the penalization scheme. Due to the fact that the values of $\hat{\boldsymbol{\beta}}$ are small as shown in Table 4, we use the data-dependent weighted SCAD penalty (Leng and Ma, 2007) $\hat{w}_j p_\lambda(|\beta_j|)$ and the non-negative weight $\hat{w}_j = 1/|\hat{\beta}_j^{LY}|$ with $\hat{\boldsymbol{\beta}}^{LY}$ being Lin and Ying (1994)'s estimator. The tuning parameter λ is chosen by the BIC-type criterion in (16). We present the solution paths for an excerpt of λ together with the BIC scores in Figure 6. The plot of the BIC scores shows that the score $BIC(\lambda)$ is well defined as a function of the tuning parameter and there exists a well-separated and unique minimizer. In this

study, it requires a very big value of λ to let all $\hat{\beta}_j, j = 1, \cdots, 17$ shrink towards zero. To verify whether variables with non-zero coefficients are selected correctly and the other coefficients shrink towards zero when the optimal λ is used, we only illustrate the solution paths of $\hat{\beta}_j$ in an excerpt of the whole λ grid in Figure 6, where the selected model is indicated by the broken vertical line. It is shown that the BIC-type criterion selects a model with 9 non-zero coefficients.

[Figure 6]

The estimation results are summarized in Table 4. For the null hypothesis $H_0 : \beta_j = 0$, the corresponding empirical likelihood ratio (ELR) and penalized empirical likelihood ratio (PELR) test statistics together with their p-values are also provided in Table 4.

[Table 4]

We summarize the main findings from Table 4 as follows:

1) 17 estimated coefficients of the full additive hazards model obtained from the empirical likelihood (EL) method are the same as those from Lin and Ying (1994)'s estimator $\hat{\beta}^{LY}$.

2) The significance of each covariate is the same when using both the normal approximation (NA) and EL methods at the nominal level $\alpha = 0.10$.

3) The proposed penalized empirical likelihood with the weighted SCAD penalty (PEL-SCAD) leads to a final model with 9 predictors. This finding is consistent with the results produced by both the mLasso2 in Leng and Ma (2007) and the Lasso under the Cox model in Zhang and Lu (2007) in terms of signs of estimated nonzero coefficients. Note that the variable spiders was particularly selected by Martinussen and Scheike (2009) using the Lasso, adaptive Lasso and Dantzig estimator under the same model as ours. This difference may be caused by the sensitivity of the cross-validation parameters in their work. Compared to 8 covariates selected by their mLasso1, the PEL-SCAD includes the additional covariate "protime", which is shown to be marginally significant in the model.

4) P-values corresponding to covariates selected by the PEL-SCAD are consistent with those from the full additive hazards model. This verifies that the proposed variable selection does not change significant effects of the main covariates much.

6.2 Analysis of the NKI Breast Cancer data

We now apply the proposed method to another practical setting, breastCancerNKI data, collected in a breast cancer study initiated by the Netherlands Cancer Institute (Nederlands Kanker Instituut in Dutch, NKI). This data set includes 337 patients aged from 26 to 62 years old with stage I , II or III disease (van de Vijver et al., 2002, van't Veer et al., 2002). In addition to 24481 gene expression measurements, the clinical information of each patient are considered in the study, including tumor size (size), age at diagnosis (age), estrogen receptor status by IHC (er), histological grade (grade), lymph nodal status (node) and chemotherapy or hormonal therapy treatment (treatment). The survival outcome of our interest is time for distant metastasis-free survival (DMFS) defined as the duration from the initial diagnosis to the time at which a distant metastasis was detected. Both clinical features and gene expression profiles have been used recently to predict patient survival outcome and assist treatment decisions for breast cancers. However, when combining both clinical and genomic covariates together, none of genomic covariates would be selected by the Lasso or Adaptive Lasso selector (Martinussen and Scheike, 2009). Hence we perform analysis separately for them.

To examine the potential effects of clinical features, we fit the additive hazards model to the data of n=319 patients who have valid times for DMFS and clinical covariates, and apply the proposed PEL procedure for estimation and variable selection using the Adaptive Lasso and SCAD penalties, respectively. We choose adaptive weights as $1/|\beta_i|^{0.5}$ in the Adaptive Lasso, where β_i is the non-penalized EL estimates of β_i. The BIC-type criterion is applied to select the tuning parameter, giving the optimal $\lambda = 0.06657$ for both the Adaptive Lasso and SCAD methods. Table 5 lists the PEL estimates using the two types of penalties. As a confirmation of the selected model, we also report the estimation results and corresponding p-values obtained from the non-penalized EL method in the table. Table 5 clearly shows that there are only two variables, age and historical grade, selected in the model, and they are coincident with the significant covariates indicated by p-values (less than level 0.05) under the EL estimation. Findings of van't Veer et al. (2002) and van de Vijver et al. (2002) suggest that histological grade and age at diagnosis are two of the strongest predictors for metastases. Interestingly, chemotherapy or hormonal therapy is not selected in the model though it is known a treatment to reduce the risk of distant metastases. In fact, the Early Breast Cancer Trialists' Collaborative Group (1998) showed that $70 - 80\%$ of patients receiving this treatment would have survived without it.

[Table 5]

For the genomic covariates, there are a huge number of gene expression measurements compared with the sample size. Note that each gene expression profile has at least one missing record for all samples, and sample indices corresponding to these missing records are different for different genes. Thus we cannot perform a screen in this case. For simplicity, we consider the first 20 gene expression profiles in the NKI data, corresponding to 312 samples. The optimal tuning parameter λ is chosen as 0.0552 for the Adaptive Lasso and 0.06651 for the SCAD by the BIC method. The resulting estimates of β are given in Table 6. It is seen that the PEL with two different penalties leads to slightly different results in selecting variables. Out of 20 genes, both methods select 4 genes: $Contig26811, Contig368292, Contig42854$ and $Contig8376\_RC$, while the PEL-Adaptive Lasso can select one more gene $Contig40179\_RC$. However, further studies are required to confirm the prognostic impacts of these selected genes.

[Table 6]

7 Discussions

We study the problem of variable selection, parameter estimation and inference for sparse additive hazards regression in the asymptotic framework with a growing number of regression coefficients as the sample size $n \to \infty$ at a proper rate. To this end, we have proposed a penalized empirical likelihood method, shown the asymptotic properties of the resulting estimator and further developed a penalized empirical likelihood ratio test for testing a general linear hypothesis for the parameters. The proposed method can be implemented through a local quadratic approximation to the penalty and a nonlinear optimization procedure. Our numerical results demonstrate that the proposed penalized empirical likelihood method performs satisfactorily in drawing inference for the additive hazards regression model in the "large n, diverging p" framework. When "large p, small n" is encountered, some preliminary method such as the sure independent screening (Fan and Lv, 2008, Fan et al., 2010) has to be employed to reduce the dimensionality before applying the proposed penalized empirical likelihood estimator.

In the context of empirical likelihood, the growth rate of p has been proposed in some existing studies in the literature. Hjort et al. (2009) proved that the asymptotic normality of the empirical likelihood ratio with the normalizing constants p and $(2p)^{1/2}$ holds if $p^{3+6/(q-2)}/n \to 0$, after imposing conditions on boundedness of the eigenvalues of covariance matrix and boundedness of

the q-moment of each component for some $q \geq 4$. Meanwhile, Chen et al. (2009) improved their rate to $p = o(n^{1/2-1/(8k)})$ within a specific model framework of high-dimensional data. More restrictive growth typically suffices for certain quadratic approximations associated with Wilks theorems. When empirical likelihood is studied in a broader framework of general estimating equations, Leng and Tang (2012) established the Wilks theorems when $p^5/n \to 0$, and they argued that the rate $p = o(n^{1/5})$ could not be weakened. In the current work, there is no particularly structural information available for the estimating equations, data are subject to censoring and hence provide incomplete information, and an additional unknown baseline hazard function has to be estimated nonparametrically in the estimation procedure. These pose great challenges in establishing the theoretical results. Consequently, stronger regularity conditions are required and the bounds in the stochastic analysis are relatively conservative. Based on the proofs of Theorems 1-2, we have relaxed the rate $p = o(n^{1/5})$ in Assumption 5 to $p^2 = o_p(n^{1/2-1/q})$ for some $q \geq 10$.

Variable selection via empirical likelihood in regression is generally a difficult problem, and is even more challenging in the survival model context. As pointed out in Section 4, the implementation of penalized empirical likelihood in our study is nontrivial and time-consuming. In our experience, especially when dimension p is high, its convergence and stability may be problematic because the objective function is nonconvex and the Hessian matrix is very likely to be singular in the iterative procedure. Owing to the semiparametric nature of survival models and the presence of censoring, the performance of variable selection and hypothesis testing is affected by several factors, such as the covariates, baseline hazard, and censoring rate. These factors call for a need to increase the sample size to make reliable inference.

Acknowledgement

We are grateful to an Associate Editor and two anonymous referees for their very constructive comments and suggestions which helped improve the paper greatly. This research is supported partly by the Singapore Ministry of Education Academic Research Fund Tier 1 (RG30/12), Tier 2 (MOE2013-T2-2-118) and the National Natural Science Foundation of China (Grant No. 71420107025).

References

P. K. Andersen and R. D. Gill. Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

N. E. Breslow and N. E. Day. *Statistical methods in cancer research*, volume 2. International Agency for Research on Cancer Lyon, 1987.

J. Cai, J. Fan, R. Li, and H. Zhou. Variable selection for multivariate failure time data. *Biometrika*, 92(2):303–316, 2005.

S. X. Chen and H. Cui. On Bartlett correction of empirical likelihood in the presence of nuisance parameters. *Biometrika*, 93(1):215–220, 2006.

S. X. Chen, L. Peng, and Y. L. Qin. Effects of data dimension on empirical likelihood. *Biometrika*, 96(3):711–722, 2009.

D. R. Cox and D. Oakes. *Analysis of survival data*, volume 21. CRC Press, 1984.

T. DiCiccio, P. Hall, and J. Romano. Empirical likelihood is Bartlett-correctable. *The Annals of Statistics*, 19(2):1053–1061, 1991.

Early Breast Cancer Trialists' Collaborative Group. Polychemotherapy for early breast cancer: an overview of the randomised trials. *Lancet*, 352(1):930–942, 1998.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

J. Fan and R. Li. Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, 30(1):74–99, 2002.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961, 2004.

J. Fan, Y. Feng, and Y. Wu. High-dimensional variable selection for Cox's proportional hazards model. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, pages 70–86. Institute of Mathematical Statistics, 2010.

T. R. Fleming and D. P. Harrington. *Counting processes and survival analysis*, volume 169. John Wiley & Sons, 2011.

N. L. Hjort, I. W. McKeague, and I. Van Keilegom. Extending the scope of empirical likelihood. *The Annals of Statistics*, 37(3):1079–1111, 2009.

C. Leng and S. Ma. Path consistent model selection in additive risk model via Lasso. *Statistics in medicine*, 26(20):3753–3770, 2007.

C. Leng and C. Y. Tang. Penalized empirical likelihood and growing dimensional general estimating equations. *Biometrika*, 99(3):703–716, 2012.

D. Lin and Z. Ying. Semiparametric analysis of the additive risk model. *Biometrika*, 81(1):61–71, 1994.

W. Lin and J. Lv. High-dimensional sparse additive hazards regression. *Journal of the American Statistical Association*, 108(501):247–264, 2013.

S. Ma and J. Huang. Lasso method for additive risk models with high dimensional covariates. Technical report, Department of Statistics and Actuarial Science, University of Iowa, Iowa, 2005.

S. Ma, M. R. Kosorok, and J. P. Fine. Additive risk models for survival data with high-dimensional covariates. *Biometrics*, 62(1):202–210, 2006.

T. Martinussen and T. H. Scheike. A semiparametric additive regression model for longitudinal data. *Biometrika*, 86(3):691–702, 1999.

T. Martinussen and T. H. Scheike. Covariate selection for the semiparametric additive risk model. *Scandinavian Journal of Statistics*, 36(4):602–619, 2009.

A. B. Owen. *Empirical Likelihood*. CRC Press, 2001.

W. Sauerbrei and M. Schumacher. A bootstrap resampling procedure for model building: application to the Cox regression model. *Statistics in medicine*, 11(16):2093–2109, 1992.

G. E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

C. Y. Tang and C. Leng. Penalized high-dimensional empirical likelihood. *Biometrika*, 97(4): 905–920, 2010.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

R. Tibshirani. The Lasso method for variable selection in the Cox model. *Statistics in medicine*, 16(4):385–395, 1997.

M. J. van de Vijver, Y. He, L. J. vant Veer, H. Dai, A. M. Hart, D. W. Voskuil, G. J. Schreiber, H. L. Peterse, C. Roberts, M. J. Marton, , M. Parrish, D. Atsma, A. T. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.

L. J. van't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(1):530–536, 2002.

H. Wang, B. Li, and C. Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3): 671–683, 2009.

X. Xie, H. D. Strickler, and X. Xue. Additive hazard regression models: an application to the natural history of human papillomavirus. *Computational and mathematical methods in medicine*, 2013, 2013.

C. H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

H. H. Zhang and W. Lu. Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 94(3): 691–703, 2007.

H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of statistics*, 36(4):1509–1533, 2008.

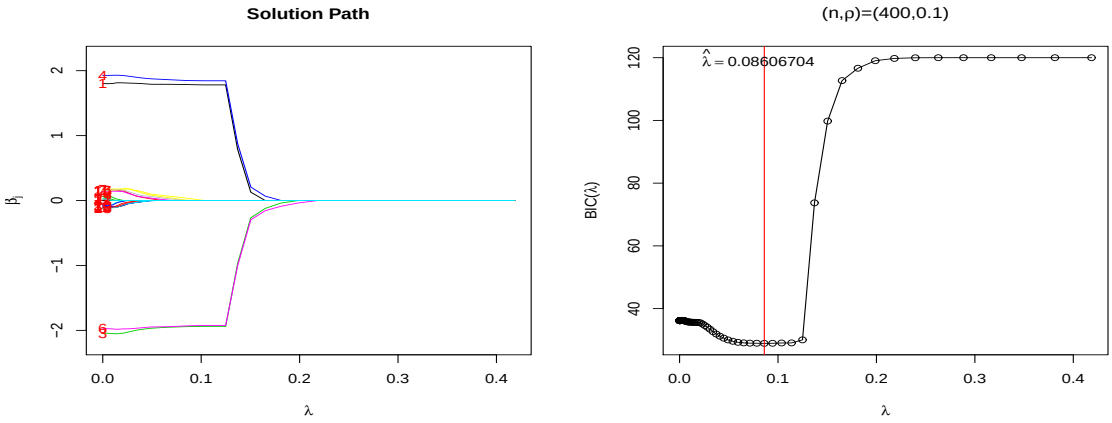

Figure 1: SCAD coefficient paths and the corresponding BIC(λ) for simulated data when $n = 400$, $p = 21$, $d = 4$, $k = 2$ and $\rho = 0.1$.

Table 1: Simulation results for different methods with sample size $n = 100, 200$ and 400, sparsity level $d = 4$, and censoring rate about 10% and 20%, respectively. Values shown are means and standard deviations of each performance measure over 200 replicates.

		PE		L_2-loss		L_1-loss		#C		#IC	
(n, p, ρ)	Method	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
		Censoring rate: 10%									
(100,9,0.1)	SCAD	7.409	4.308	0.824	0.499	1.531	0.993	4.765	0.567	0.000	0.000
	ALASSO	7.409	4.307	0.823	0.499	1.531	0.992	4.765	0.569	0.000	0.000
(100,9,0.5)	SCAD	5.992	3.595	0.789	0.512	1.477	1.036	4.865	0.409	0.000	0.000
	ALASSO	5.991	3.596	0.790	0.512	1.477	1.036	4.866	0.409	0.000	0.000
(200,15,0.1)	SCAD	7.518	4.603	0.584	0.373	1.096	0.748	10.575	0.937	0.000	0.000
	ALASSO	7.517	4.603	0.585	0.371	1.095	0.746	10.576	0.936	0.000	0.000
(200,15,0.5)	SCAD	5.867	3.181	0.550	0.338	1.035	0.681	10.590	0.643	0.000	0.000
	ALASSO	5.869	3.182	0.551	0.340	1.035	0.681	10.589	0.644	0.000	0.000
(400,21,0.1)	SCAD	6.938	4.077	0.379	0.228	0.707	0.464	16.590	0.710	0.000	0.000
	ALASSO	6.938	4.079	0.380	0.228	0.708	0.464	16.588	0.711	0.000	0.000
(400,21,0.5)	SCAD	5.772	3.267	0.383	0.245	0.734	0.501	16.260	0.931	0.000	0.000
	ALASSO	5.770	3.268	0.382	0.247	0.733	0.500	16.259	0.932	0.000	0.000
		Censoring rate: 20%									
(100,9,0.1)	SCAD	9.582	5.879	1.050	0.665	1.942	1.351	4.875	0.332	0.005	0.071
	ALASSO	9.581	5.877	1.048	0.665	1.941	1.352	4.875	0.332	0.000	0.000
(100,9,0.5)	SCAD	7.510	4.002	0.976	0.585	1.802	1.188	4.840	0.430	0.000	0.000
	ALASSO	7.510	4.002	0.975	0.587	1.801	1.188	4.840	0.430	0.000	0.000
(200,15,0.1)	SCAD	8.579	5.375	0.665	0.430	1.228	0.891	10.885	0.377	0.000	0.000
	ALASSO	8.578	5.374	0.665	0.430	1.229	0.891	10.883	0.376	0.000	0.000
(200,15,0.5)	SCAD	7.489	4.148	0.672	0.415	1.252	0.867	10.785	0.480	0.000	0.000
	ALASSO	7.489	4.149	0.671	0.417	1.252	0.867	10.785	0.479	0.000	0.000
(400,21,0.1)	SCAD	7.846	4.270	0.426	0.242	0.781	0.499	16.910	0.304	0.000	0.000
	ALASSO	7.846	4.269	0.427	0.242	0.780	0.500	16.909	0.305	0.000	0.000
(400,21,0.5)	SCAD	6.624	3.790	0.436	0.279	0.810	0.568	16.835	0.411	0.000	0.000
	ALASSO	6.624	3.791	0.436	0.279	0.810	0.569	16.835	0.411	0.000	0.000

Table 2: Simulation results for different methods with sample size $n = 100, 200$ and 400, sparsity level $d = 6$, and censoring rate 10% and 20%, respectively. Values shown are means and standard deviations of each performance measure over 200 replicates.

		PE		L_2-loss		L_1-loss		#C		#IC	
(n, p, ρ)	Method	mean	sd	mean	sd	mean	sd	mean	sd	mean	sd
		Censoring rate: 10%									
(100,9,0.1)	SCAD	10.500	5.778	1.155	0.656	2.552	1.605	2.830	0.427	0.010	0.100
	ALASSO	10.502	5.780	1.155	0.655	2.551	1.606	2.828	0.427	0.010	0.100
(100,9,0.5)	SCAD	8.203	4.524	1.127	0.679	2.537	1.691	2.860	0.402	0.000	0.000
	ALASSO	8.202	4.524	1.129	0.680	2.537	1.691	2.859	0.403	0.000	0.000
(200,15,0.1)	SCAD	9.642	5.734	0.739	0.456	1.644	1.143	8.850	0.372	0.000	0.000
	ALASSO	9.642	5.735	0.738	0.456	1.644	1.144	8.851	0.372	0.000	0.000
(200,15,0.5)	SCAD	7.500	3.827	0.711	0.420	1.593	1.036	8.825	0.535	0.000	0.000
	ALASSO	7.501	3.825	0.712	0.419	1.592	1.036	8.824	0.534	0.000	0.000
(400,21,0.1)	SCAD	9.218	4.825	0.500	0.273	1.109	0.677	14.825	0.464	0.000	0.000
	ALASSO	7.091	3.734	0.476	0.291	1.068	0.715	14.643	0.672	0.000	0.000
(400,21,0.5)	SCAD	7.090	3.733	0.476	0.291	1.068	0.715	14.645	0.672	0.000	0.000
	ALASSO	9.216	4.825	0.499	0.273	1.108	0.677	14.826	0.463	0.000	0.000
		Censoring rate: 20%									
(100,9,0.1)	SCAD	11.269	5.549	1.223	0.623	2.637	1.482	2.830	0.402	0.015	0.122
	ALASSO	11.268	5.548	1.224	0.622	2.636	1.480	2.831	0.402	0.015	0.122
(100,9,0.5)	SCAD	8.985	4.946	1.208	0.734	2.660	1.809	2.845	0.376	0.035	0.380
	ALASSO	8.983	4.945	1.207	0.733	2.660	1.809	2.845	0.377	0.035	0.380
(200,15,0.1)	SCAD	11.119	5.394	0.857	0.438	1.878	1.084	8.820	0.468	0.000	0.000
	ALASSO	11.118	5.395	0.859	0.437	1.878	1.085	8.820	0.470	0.000	0.000
(200,15,0.5)	SCAD	9.628	5.089	0.907	0.537	2.017	1.347	8.850	0.509	0.000	0.000
	ALASSO	9.630	5.088	0.906	0.537	2.018	1.347	8.852	0.508	0.000	0.000
(400,21,0.1)	SCAD	11.608	5.997	0.627	0.337	1.394	0.845	14.865	0.343	0.000	0.000
	ALASSO	11.610	5.998	0.626	0.337	1.394	0.844	14.865	0.344	0.000	0.000
(400,21,0.5)	SCAD	8.330	3.499	0.542	0.269	1.199	0.685	14.880	0.383	0.000	0.000
	ALASSO	8.329	3.497	0.543	0.270	1.200	0.685	14.880	0.383	0.000	0.000

Table 3: Empirical percentages of rejecting $H_0 : \beta_{01} = 2$ using the PELR for different values of the truth over 500 replicates. The nominal level is 5% and the censoring rate is about 15%.

Case		n	p	Method	2	2.10	2.20	2.30	2.40	2.50
$d = 4$	$\rho = 0.1$	100	9	SCAD	0.128	0.163	0.216	0.280	0.371	0.465
				ALASSO	0.126	0.164	0.214	0.280	0.371	0.466
		200	15	SCAD	0.089	0.202	0.281	0.375	0.470	0.558
				ALASSO	0.090	0.201	0.281	0.376	0.470	0.558
		400	21	SCAD	0.057	0.376	0.461	0.591	0.697	0.779
				ALASSO	0.057	0.376	0.460	0.591	0.698	0.778
	$\rho = 0.5$	100	9	SCAD	0.112	0.203	0.253	0.330	0.406	0.472
				ALASSO	0.111	0.203	0.254	0.330	0.406	0.471
		200	15	SCAD	0.074	0.218	0.298	0.405	0.526	0.588
				ALASSO	0.075	0.218	0.298	0.407	0.527	0.589
		400	21	SCAD	0.064	0.353	0.413	0.526	0.654	0.716
				ALASSO	0.063	0.355	0.412	0.525	0.655	0.716
$d = 6$	$\rho = 0.1$	100	9	SCAD	0.122	0.168	0.209	0.284	0.362	0.462
				ALASSO	0.123	0.168	0.208	0.285	0.362	0.461
		200	15	SCAD	0.095	0.199	0.278	0.369	0.472	0.564
				ALASSO	0.095	0.198	0.279	0.369	0.473	0.565
		400	21	SCAD	0.058	0.381	0.460	0.585	0.699	0.772
				ALASSO	0.059	0.380	0.461	0.587	0.699	0.773
	$\rho = 0.5$	100	9	SCAD	0.107	0.197	0.262	0.323	0.413	0.470
				ALASSO	0.108	0.196	0.262	0.323	0.414	0.468
		200	15	SCAD	0.078	0.227	0.296	0.401	0.532	0.588
				ALASSO	0.079	0.227	0.296	0.400	0.531	0.587
		400	21	SCAD	0.057	0.357	0.419	0.522	0.662	0.721
				ALASSO	0.057	0.357	0.418	0.521	0.661	0.721

Table 4: Estimation results for the PBC data: the PEL-SCAD refers to the penalized empirical likelihood results for the additive hazards model with SCAD penalty; the NA refers to the normal approximation method; the EL refers to the empirical likelihood method.

	Full model						PEL-SCAD		
	Coefficient	NA			EL		Coefficient		
Variable	($\times 10^4$)	SE($\times 10^4$)	Z	P-values	ELR	P-values	($\times 10^4$)	PELR	P-values
trt	−0.041	0.197	−0.208	0.836	0.047	0.828	0.000	0.000	1.000
age	0.640	0.224	2.856	0.004	6.517	0.011	0.466	5.097	0.024
sex	−0.142	0.289	−0.490	0.624	0.326	0.568	0.000	0.000	1.000
ascites	1.697	0.880	1.928	0.054	4.065	0.044	1.201	3.820	0.051
hepato	−0.068	0.259	−0.264	0.792	0.082	0.775	0.000	0.000	1.000
spiders	0.224	0.270	0.829	0.407	0.695	0.404	0.000	0.000	1.000
edema	1.088	0.553	1.969	0.049	3.883	0.049	1.160	5.669	0.017
bili	2.319	0.802	2.890	0.004	23.517	0.000	2.324	46.265	0.000
chol	−0.194	0.427	−0.456	0.649	0.317	0.573	0.000	0.000	1.000
albumin	−0.533	0.306	−1.743	0.081	4.158	0.041	−0.443	3.434	0.064
copper	0.749	0.455	1.647	0.099	3.446	0.063	0.455	2.130	0.104
alk.phos	−0.040	0.232	−0.173	0.863	0.028	0.867	0.000	0.000	1.000
ast	0.387	0.275	1.404	0.160	2.603	0.107	0.209	0.998	0.318
trig	−0.239	0.304	−0.786	0.432	0.525	0.469	0.000	0.000	1.000
platelet	0.038	0.215	0.179	0.858	0.037	0.847	0.000	0.000	1.000
protime	0.258	0.256	1.008	0.313	1.142	0.285	0.190	0.834	0.361
stage	0.305	0.206	1.482	0.138	2.149	0.143	0.279	2.665	0.103

Table 5: Estimation results for the breastCancerNKI data with clinical variables: the EL refers to the empirical likelihood method for the additive hazards model; the PEL refers to the penalized empirical likelihood method with the Adaptive Lasso and SCAD penalties, respective.

	EL				PEL	
Variables	Estimate($\times 10^4$)	SD($\times 10^4$)	Z value	P-values	Adaptive Lasso($\times 10^4$)	SCAD($\times 10^4$)
size	0.2460	0.1594	1.5437	0.1227	0.0000	0.0000
age	−0.0979	0.0279	−3.5034	0.0005	−0.0546	−0.0773
er	−0.4551	0.4200	−1.0836	0.2785	0.0000	0.0000
grade	0.6875	0.1638	4.1962	0.0000	0.5039	0.6730
node	0.2818	0.4734	0.5953	0.5517	0.0000	0.0000
treatment	−0.4050	0.3257	−1.2435	0.2137	0.0000	0.0000

Table 6: Estimation results for the breastCancerNKI data with the first 20 gene expression: the EL refers to the empirical likelihood method for the additive hazards model; the PEL refers to the penalized empirical likelihood method with the Adaptive Lasso and SCAD penalties, respective.

	EL				PEL	
Variables	Estimate($\times 10^4$)	SD($\times 10^4$)	Z value	P-values	Adap Lasso($\times 10^4$)	SCAD($\times 10^4$)
Contig45645_*RC*	−0.2592	0.3529	−0.7343	0.4628	0.0000	0.0000
Contig44916_*RC*	1.3837	0.6965	1.9865	0.0470	0.0000	0.0000
D25272	−0.9518	0.4588	−2.0746	0.0380	0.0000	0.0000
J00129	0.2345	0.6938	0.3380	0.7353	0.0000	0.0000
Contig29982_*RC*	−0.2165	0.5936	−0.3648	0.7152	0.0000	0.0000
Contig26811	−1.9436	0.7945	−2.4463	0.0144	−0.8335	−0.6420
D25274	2.0232	1.1060	1.8294	0.0673	0.0000	0.0000
Contig36292	−2.1230	1.6943	−1.2530	0.2102	−0.5768	−0.1574
Contig42854	−1.7521	0.5558	−3.1523	0.0016	−0.3616	−0.2026
Contig34839	7.0514	3.0440	2.3165	0.0205	0.0000	0.0000
Contig8376_*RC*	−5.7248	2.2147	−2.5849	0.0097	−0.0968	0.0000
Contig42014_*RC*	−2.3733	1.3449	−1.7647	0.0776	0.0000	0.0000
D49958	−0.9698	0.9078	−1.0683	0.2854	0.0000	0.0000
Contig25622_*RC*	−0.1543	1.6776	−0.0920	0.9267	0.0000	0.0000
Contig13475_*RC*	2.2758	1.3635	1.6690	0.0951	0.0000	0.0000
Contig40179_*RC*	2.7907	1.1517	2.4231	0.0154	0.3545	0.0404
Contig27915_*RC*	−1.0698	0.6304	−1.6972	0.0897	0.0000	0.0000
Contig44682_*RC*	−2.8791	3.4190	−0.8421	0.3997	0.0000	0.0000
Contig35934_*RC*	−0.5069	1.6563	−0.3061	0.7596	0.0000	0.0000
Contig29373_*RC*	5.9787	3.0411	1.9659	0.0493	0.0000	0.0000

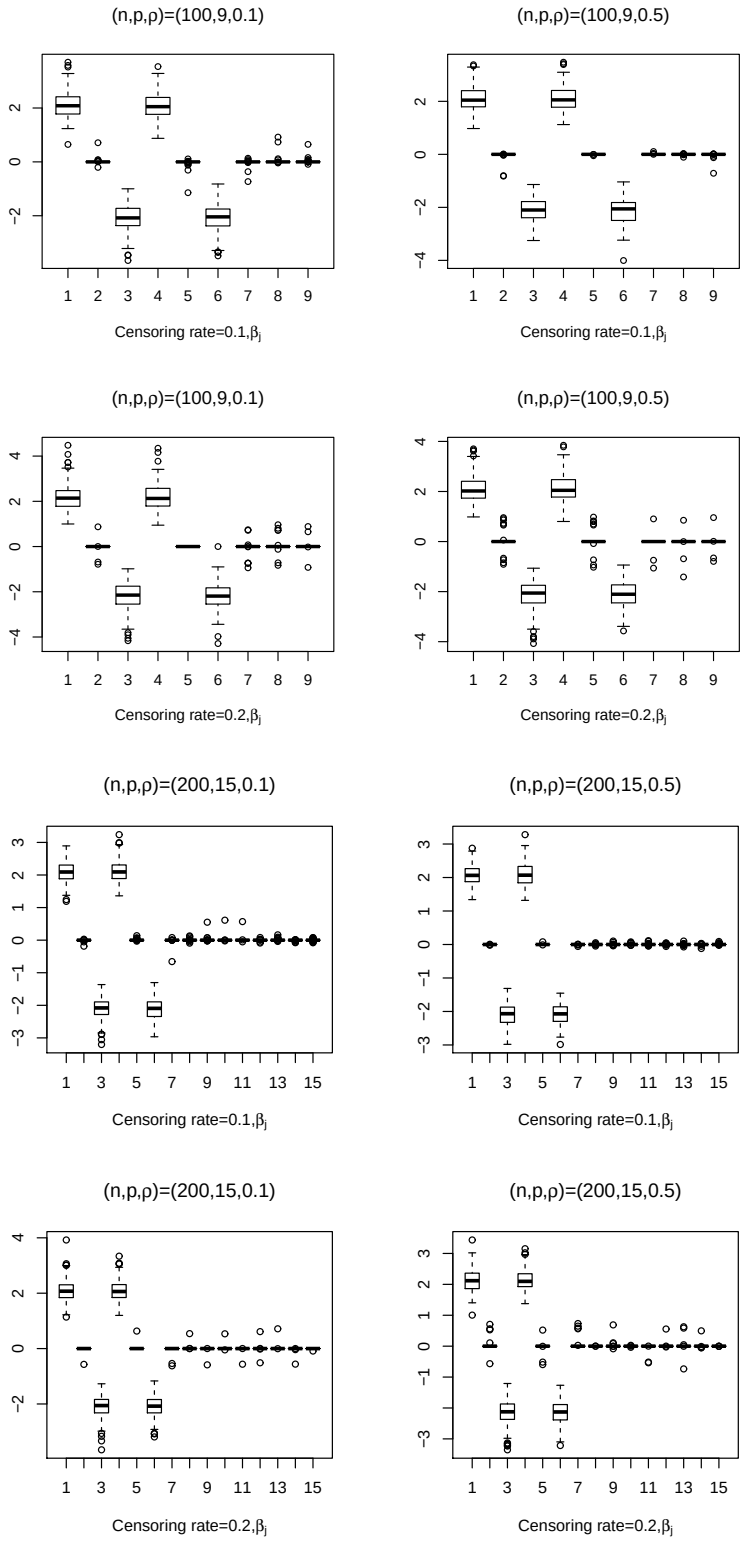

Figure 2: Boxplots of estimates $\hat{\boldsymbol{\beta}}$ based on 200 replicates with sparsity level $d = 4$. Top two rows: sample size $n = 100$ and dimension $p = 9$; Bottom two rows: sample size $n = 200$ and dimension $p = 15$.

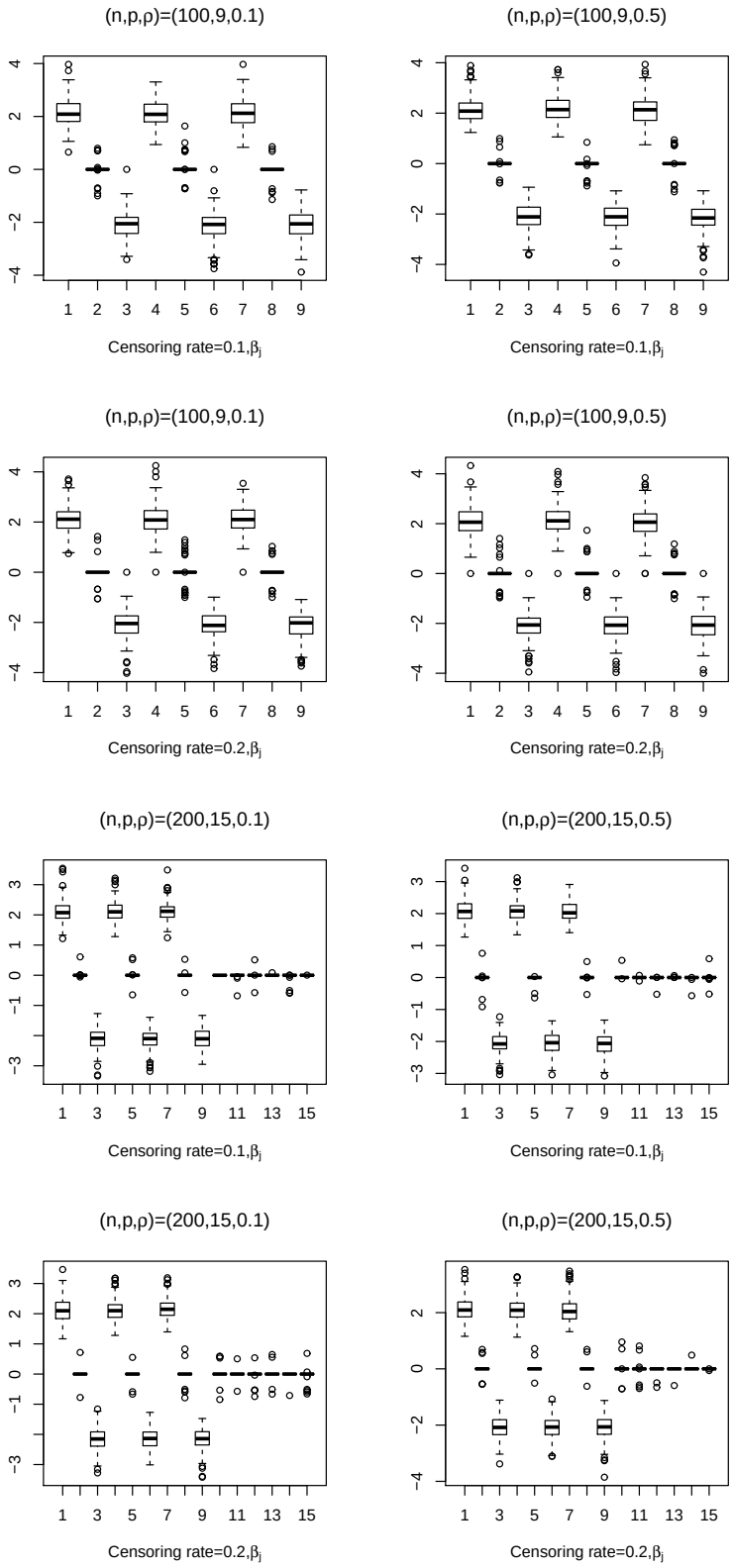

Figure 3: Boxplots of estimates $\hat{\boldsymbol{\beta}}$ based on 200 replicates with sparsity level $d = 6$. Top two rows: sample size $n = 100$ and dimension $p = 9$; Bottom two rows: sample size $n = 200$ and dimension $p = 15$.

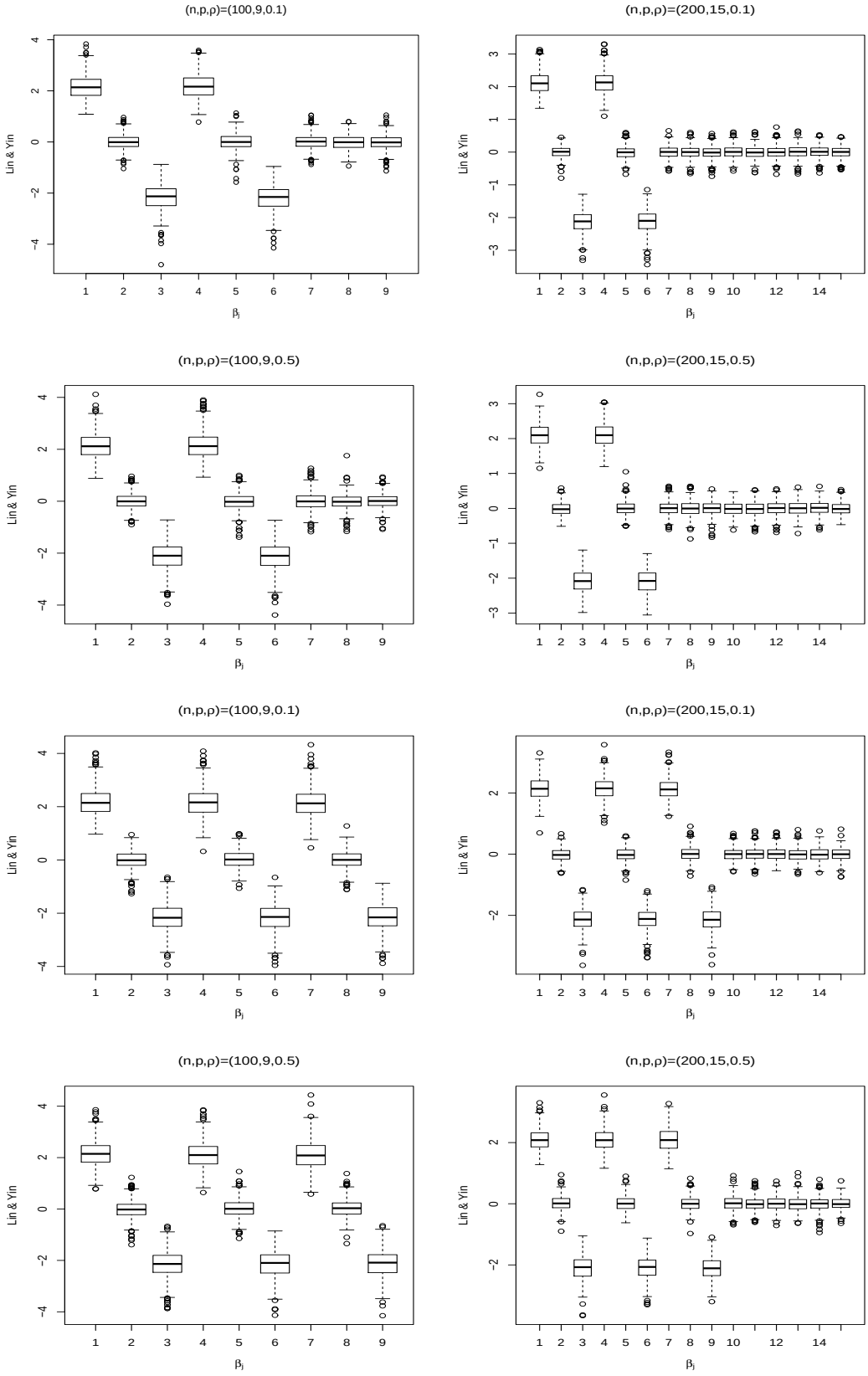

Figure 4: Boxplot of estimates $\hat{\beta}^{LY}$ based on 500 simulations. Top two rows: sparsity dimension $d = 4$; Bottom two rows: sparsity dimension $d = 6$. Left panel: sample size $n = 100$; Right panel: sample size $n = 200$.

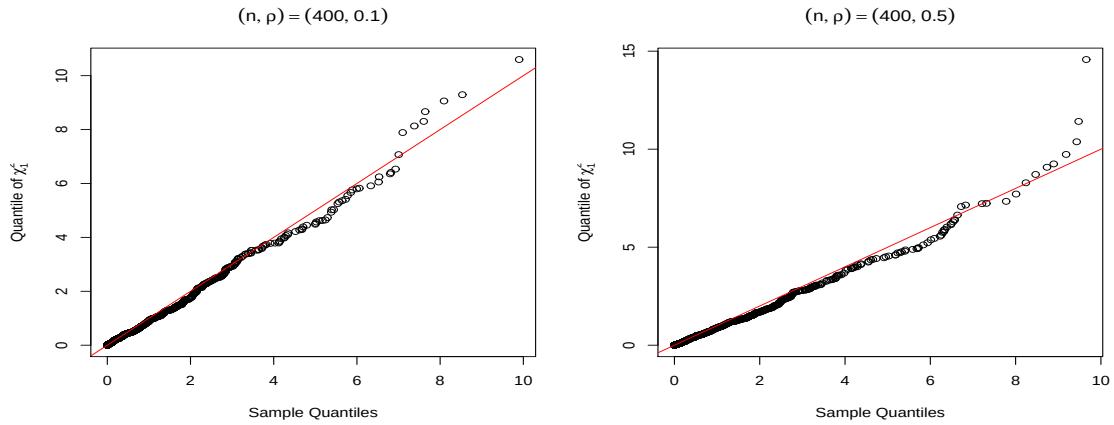

Figure 5: QQplots of the penalized empirical likelihood ratio statistics against χ^2_1 distribution under null hypothesis H_0 when sample size $n = 400$, $p = 21$, $d = 4$, $k = 2$ and $\rho = 0.1$ (Left) and $\rho = 0.5$ (Right).

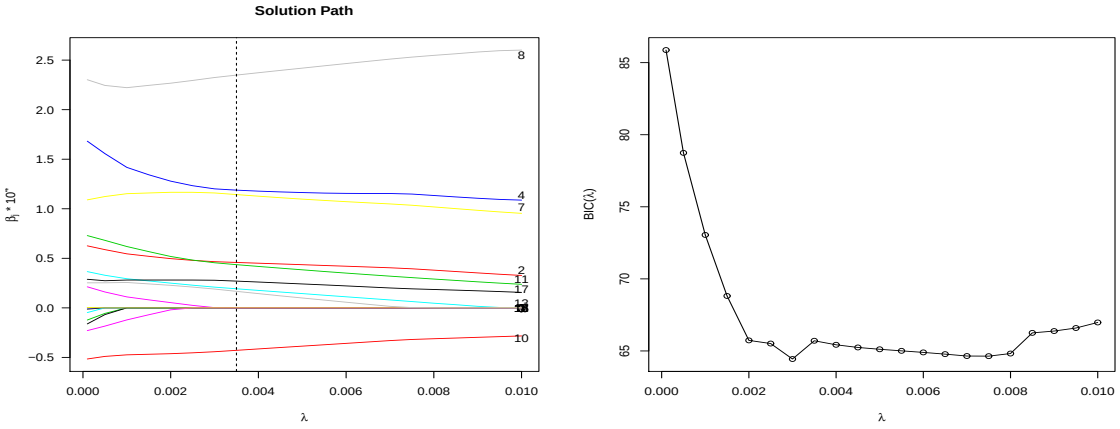

Figure 6: Analysis of the PBC data. Left panel: the solution paths. Right panel: the BIC scores. Dotted lines: optimal tuning parameter chosen by (16).