Low-dimensional Confounder Adjustment and High-dimensional Penalized Estimation for Survival Analysis

Xiaochao Xia[1], Binyan Jiang[2], Jialiang Li[2], Wenyang Zhang[3]

[1] *College of Mathematics and Statistics, Chongqing University, China*

[2] *Department of Statistics and Applied Probability, National University of Singapore, Singapore*

[3] *Department of Mathematics, University of York, United Kingdom*

Abstract: High-throughput profiling is now common in biomedical research. In this paper we consider the layout of an etiology study composed of a failure time response, and gene expression measurements. In current practice, a widely adopted approach is to select genes according to a preliminary marginal screening and a follow-up penalized regression for model building. Confounders, including for example clinical risk factors and environmental exposures, usually exist and need to be properly accounted for. We propose covariate-adjusted screening and variable selection procedures under the accelerated failure time model. While penalizing the high-dimensional coefficients to achieve parsimonious model forms, our procedure also properly adjust the low-dimensional confounder effects to achieve more accurate estimation of regression coefficients. We establish the asymptotic properties of our proposed methods and carry out simulation studies to assess the finite sample performance. Our methods are illustrated with a real gene expression data analysis where proper adjustment of confounders produces more meaningful results.

Key words and phrases: Accelerated failure time model; Confounder adjustment; Gene expression; Independent screening; Variable selection.

Low-dimensional Confounder Adjustment and High-dimensional Penalized Estimation for Survival Analysis

Abstract: High-throughput profiling is now common in biomedical research. In this paper we consider the layout of an etiology study composed of a failure time response, and gene expression measurements. In current practice, a widely adopted approach is to select genes according to a preliminary marginal screening and a follow-up penalized regression for model building. Confounders, including for example clinical risk factors and environmental exposures, usually exist and need to be properly accounted for. We propose covariate-adjusted screening and variable selection procedures under the accelerated failure time model. While penalizing the high-dimensional coefficients to achieve parsimonious model forms, our procedure also properly adjust the low-dimensional confounder effects to achieve more accurate estimation of regression coefficients. We establish the asymptotic properties of our proposed methods and carry out simulation studies to assess the finite sample performance. Our methods are illustrated with a real gene expression data analysis where proper adjustment of confounders produces more meaningful results.

Key words and phrases: Accelerated failure time model; Confounder adjustment; Gene expression; Independent screening; Variable selection.

1 Introduction

High-throughput profiling is now routinely conducted in biomedical studies. Available measurements include mRNA gene expression, SNP, rare variant, exome sequencing, and many others. In what follows, we focus on studies using mRNA gene expression to predict failure time outcome; the proposed methods, however, are directly applicable to measurements of other types. In the study of an ultra-high dimensional gene expression, the data analysis usually consists of two stages: (i) rank the importance of genes based on their marginal associations with the outcome variable and screen out unimportant genes from the ordered list; (ii) build a parsimonious regression model with sufficient complexity using variable selection methods, mostly based on penalties. Such a two-step procedure has been widely adopted and enjoyed theoretical and practical supports (eg. Cheng et al. [2014], Fan and Lv [2008], Fan et al. [2009], Li et al. [2012], among others).

In genetic epidemiology studies, we usually collect data for high-dimensional as well as

low-dimensional covariates. The low-dimensional covariates may include demographical variables, risk factors, environmental exposures or other variables (Cheng et al. [2009]) that may be regarded as confounders (VanderWeele and Shpitser [2013]). The common practice in an observation subject is to include as many relevant confounding variables as possible so that we can obtain more accurate estimate for the coefficients of the main variables of interest (Gordis [2008]). Even if some confounders are not significant in the final model, including them may produce better results for the estimation of the regression coefficients of variables of interest. Furthermore, we may compare the results across different studies easily since most studies on the same response would adjust a similar set of confounders. However, in the two-step procedure described previously, very often low-dimensional covariates are ignored and how to fully incorporate low-dimensional covariates has not been systematically addressed in the literature.

Though variable selection methods for continuous and binary outcomes are abundant, the related development in survival analysis has been relatively sparse and most existing programs focus on Cox proportional hazards model only (Bradic et al. [2011], Fan and Li [2002], Lian et al. [2014]). There are other useful models widely accepted in medical research when proportional hazards assumption does not hold for the failure process. Specifically, we consider the accelerated failure time (AFT) model which is introduced in most standard textbooks. In an earlier development when the two-step procedure has not become the prevailing practice, Huang et al. [2006] considered using AFT model with the Lasso penalty which is now known to be selection inconsistent. Cai et al. [2009] considered a similar setting with Lasso penalty but proposed a more general rank-based estimator for AFT model. Johnson et al. [2008] studied the AFT model with the smoothly clipped absolute deviation (SCAD) penalty function which we will consider in this paper. Most recently Huang and Ma [2010] extended the regularized estimation of the AFT model under the bridge penalty. Hu and Chai [2013] considered a high dimensional penalization with MCP penalty in the AFT model. Apart from the recent work of Hu and Chai [2013], in their methodology, however, none of these previous authors considered the ultra-high dimension setting where screening deserves pursuing before the penalized estimation. Only *ad hoc* treatments were provided in a few case studies. In our opinion the screening step (i) is almost inevitable for gene expression studies and a formal methodology construction is necessary for the AFT model.

The main contribution of this paper can be summarized as follows. A formal methodology to properly adjust the low-dimensional covariates in the familiar two-step procedure

for survival analysis is proposed. Having studied the real data analysis, we discover that the set of markers selected with and without covariate adjustment could be remarkably different, indicating that the contribution of some important markers selected without covariate adjustment could be confounded with the covariate effects and these markers should be dropped in the presence of available low-dimensional covariates. As far as we have reviewed, confounder-adjusted variable screening and variable selection have never been thoroughly discussed for continuous and binary outcomes yet. Our proposal thus may be equally applicable for those settings as well. In particular, the effects of confounders are modelled as additive nonparametric terms in the AFT model. Treating nuisance parameters as nonparametric functions is quite desirable in many applications, yielding more reliable estimation for covariate effects. For the high-dimensional covariates, we still consider a linear parametric form which may provide a lucid interpretation of the gene effects in practice. The selection of important genes is realized using the SCAD penalty (Fan and Li [2001, 2002]).

The remainder of this paper is arranged as follows. A detailed procedure for our methodology is proposed in Section 2. The relevant theoretical justification is provided in Section 3. Simulation studies are carried out in Section 4. In Section 5, an analysis of a Lung cancer dataset using our method is presented. All the proofs are relegated to the Appendix.

2 Two-step analysis procedure

Denote by $\mathbf{X} = (X_1, \cdots, X_p)^T$ the expression of p genes where $\log(p) = O(n^c)$ for $c > 0$. We adopt the common sparsity assumption and believe that only a small subset of these p genes are indeed related to a particular disease outcome. In practice, we may collect data on low dimensional covariates such as demographic information as well. Denote by $\mathbf{U} = (U_1, \ldots, U_d)^T$ the d confounders where $d << n$. Directly applying penalized estimation with p genes is infeasible in most statistical programs. It is necessary to first implement a screening procedure and cut the number p from the non-polynomial order to a much smaller (polynomial) order. Furthermore, we intend to conduct screening and variable selection in the presence of confounders. Detailed methodology follows.

2.1 Variable screening

Consider the partially linear additive accelerated failure time model

$$T_i = \mathbf{X}_i^T \boldsymbol{\beta} + \sum_{j=1}^{d} g_j(U_{ij}) + \varepsilon_i, \qquad i = 1, \cdots, n; \tag{1}$$

where T_i is the logarithm of the failure time and $\mathbf{X}_i$ is a p-dimensional covariate vector for the ith subject in a random sample of size n, $g_j(\cdot)$ is an unknown and nonlinear function depending merely on a univariate U_j for $j = 1, \ldots, d$, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the regression parameter and ε_i is random error. Assuming that T_i is subject to right censoring, we can only observe $\{(Y_i, \delta_i, \mathbf{X}_i, \mathbf{U}_i) : i = 1, \cdots, n\}$ with $Y_i = \min(T_i, C_i)$, where C_i is the logarithm of random censoring time and $\delta_i = I(T_i \leq C_i)$ is the censoring indicator.

We rank the p ultra-high dimensional markers for their marginal importance in the presence of confounders and remove unimportant markers from further consideration. Specifically, we consider fitting marginal regression in the following manner. Suppose that $Y_{(1)} \leq \cdots \leq Y_{(n)}$ are the ordered statistic of Y_i's and $\delta_{(1)}, \cdots, \delta_{(n)}$ are the associated censoring indicators of the ordered Y_i's, and $\mathbf{X}_{(1)}, \cdots, \mathbf{X}_{(n)}$ and $U_{(1)j}, \cdots, U_{(n)j}$ for $j = 1, \ldots, d$ are defined similarly. Let $\widehat{F}_n$ be the Kaplan-Meier (KM) estimator of the distribution function F of the failure time T. Then $\widehat{F}_n$ can be written as $\widehat{F}_n(y) = \sum_{i=1}^{n} w_{ni} I(Y_{(i)} \leq y)$, where the weights $\{w_{ni}; i = 1, \ldots, n\}$ are given by

$$w_{n1} = \frac{\delta_{(1)}}{n}, \qquad \text{and} \qquad w_{ni} = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} \Big(\frac{n-j}{n-j+1} \Big)^{\delta_{(j)}}, \qquad i = 2, \ldots, n.$$

Suppose that $\mathbf{B}(\cdot) = (B_1(\cdot), \cdots, B_{M_n}(\cdot))^T$ is an M_n-dimensional vector of basis functions. Then, for every smoothing function $g_j(u)$, we can approximate it by

$$g_j(u) \approx \mathbf{B}(u)^T \boldsymbol{\gamma}_j, \qquad \text{for } j = 1, \ldots, d \tag{2}$$

where $\boldsymbol{\gamma}_j$ is a vector of length M_n. Thus, we can obtain a benchmark initial estimate of the nonparametric functions by $g_j^*(\cdot) = \mathbf{B}(\cdot)^T \boldsymbol{\gamma}_j^*$, where

$$\boldsymbol{\gamma}^* = \arg \min_{\gamma = (\gamma_1^T, \cdots, \gamma_d^T)^T} \Big\{ \sum_{i=1}^{n} w_{ni} \big[Y_{(i)} - \sum_{j=1}^{d} \mathbf{B}(U_{(i)j})^T \boldsymbol{\gamma}_j \big]^2 \Big\}. \tag{3}$$

Denote the partial residuals from minimizing (3) by $Y_{(i)}^* = Y_{(i)} - \sum_{j=1}^{d} \mathbf{B}(U_{(i)j})^T \boldsymbol{\gamma}_j^*$ for $i =$

$1, \ldots, n$. At the next step, we may solve

$$\widehat{\beta}_j^* = \arg\min_{\boldsymbol{\beta}_j} \Big\{ \sum_{i=1}^{n} w_{ni} \big(Y_{(i)}^* - X_{(i)j}\beta_j\big)^2 \Big\}, \tag{4}$$

and select a set of variables

$$\widehat{\mathcal{M}} = \big\{ 1 \le j \le p : |\widehat{\beta}_j^*| \ge v_n \big\}, \tag{5}$$

where v_n is a predefined threshold value. In practice, we often rank the features by $|\beta_j^*|$ and keep the top $\lfloor n/log(n) \rfloor$ features, where $\lfloor a \rfloor$ denotes the integer part of a.

We note that in the above marginal screening we have controlled the variation in the response due to confounders. The top genes in the ranked list thus reflect the strong correlation with the response in the presence of various confounding factors. Intuitively, these genes are more likely to have non-zero coefficients in the true model (1) where the outcome-generating mechanism clearly acknowledges the contributions from the low-dimensional covariates.

2.2 Variable selection

In the previous section, we have carried out a very important step to reduce the cardinality of the set of candidate genes, usually below the total sample size n. With a slight abuse of notation, in the following model, we continue using p to denote the dimension of gene expressions kept in the reduced set and using model (1) to denote the true model. In fact, methodology research for variable selection under the AFT model normally kicks off at this point.

To fit model (1), one common approach is the Stute's estimator which has a resemblance to the weighted least squares. Using the same notations defined in the preceding section, we can obtain the estimators for $\boldsymbol{\beta}$ and $g_j(\cdot)$ by minimizing the following objective function

$$\frac{1}{2} \sum_{i=1}^{n} w_{ni} \Big[Y_{(i)} - \mathbf{X}_{(i)}^T \boldsymbol{\beta} - \sum_{j=1}^{d} g_j(U_{(i)j}) \Big]^2. \tag{6}$$

Still applying the approximation from the spline functions (2) , we may re-write (6) as

$$\frac{1}{2} \sum_{i=1}^{n} w_{ni} \Big[Y_{(i)} - \mathbf{X}_{(i)}^T \boldsymbol{\beta} - \sum_{j=1}^{d} \mathbf{B}(U_{(i)j})^T \boldsymbol{\gamma}_j \Big]^2. \tag{7}$$

Denote by $\mathbf{Y} = (Y_{(1)}, \cdots, Y_{(n)})^T$, $\mathbf{X} = (\mathbf{X}_{(1)}, \cdots, \mathbf{X}_{(n)})^T$, $\mathbf{B}_i = \mathbf{B}(\mathbf{U}_{(i)}) = (\mathbf{B}(U_{(i)1})^T, \cdots, \mathbf{B}(U_{(i)d})^T)^T$, $\mathbf{B} = (\mathbf{B}_1, \cdots, \mathbf{B}_n)^T$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^T, \cdots, \boldsymbol{\gamma}_d^T)^T$. Then the objective function (7) becomes

$$\frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\boldsymbol{\gamma})^T\mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\boldsymbol{\gamma})$$

where $\mathbf{W} = \text{diag}\{w_{n1}, \cdots, w_{nn}\}$.

Next, for the purpose of variable selection, we incorporate a penalty into (7) and define

$$Q(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \frac{1}{2n}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\boldsymbol{\gamma})^T\mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{B}\boldsymbol{\gamma}) + \sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|), \tag{8}$$

where $p_\lambda(\cdot)$ is a penalty function and λ is the regularization parameter. An appealing choice for $p_\lambda(\cdot)$ is the SCAD penalty (Fan and Li [2001, 2002]), which is defined by

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta|, & |\beta| \le \lambda \\ -\frac{\beta^2 - 2a\lambda|\beta| + \lambda^2}{2(a-1)}, & \lambda < |\beta| \le a\lambda \\ \frac{(a+1)\lambda^2}{2}, & |\beta| > a\lambda, \end{cases}$$

where $a > 2$ is a constant. This penalty function is nonconcave and contains coherent theoretical properties, including unbiasedness, continuity and sparsity (Fan and Li [2001]).

Furthermore, we notice that it is hard to solve the problem (8) since the SCAD penalty function $p_{\lambda_n}(\cdot)$ is irregular at the origin and has no continuous second-order derivative. We develop an iterative algorithm to find the solution of problem (8). Specifically, we utilize the local quadratic approximation in Fan and Li [2001] for the SCAD penalty function $p_\lambda(\cdot)$ as the following

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^{(0)}|) + \frac{p'_\lambda(|\beta_j^{(0)}|)}{2|\beta_j^{(0)}|}(\beta_j^2 - \beta_j^{(0)2}) \qquad \text{for } \beta_j \approx \beta_j^{(0)}. \tag{9}$$

We denote $\widetilde{\mathbf{X}} = (\mathbf{X}, \mathbf{B})$, $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$, and $\mathrm{D}_\lambda(\boldsymbol{\beta}^{(0)}) = \text{diag}\left\{\frac{p'_\lambda(|\beta_1^{(0)}|)}{|\beta_1^{(0)}|}, \cdots, \frac{p'_\lambda(|\beta_p^{(0)}|)}{|\beta_p^{(0)}|}\right\}$. Then, minimizing (8) reduces to minimize the following quadratic objective function

$$Q(\boldsymbol{\theta}|\boldsymbol{\beta}^{(0)}) = n^{-1}(\mathbf{Y} - \widetilde{\mathbf{X}}\boldsymbol{\theta})^T\mathbf{W}(\mathbf{Y} - \widetilde{\mathbf{X}}\boldsymbol{\theta}) + \boldsymbol{\beta}^T\mathrm{D}_{\lambda_n}(\boldsymbol{\beta}^{(0)})\boldsymbol{\beta}. \tag{10}$$

We now summarize our computing algorithm as follows:

- Step 1. Find an initial solution $\boldsymbol{\beta}^{(0)}$ using an un-penalized ridge regression estimation, i.e.,

$$\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}^{(0)T}, \boldsymbol{\gamma}^{(0)T})^T = (\widetilde{\mathbf{X}}^T\mathbf{W}\widetilde{\mathbf{X}} + r_{\text{ridge}}\mathbf{I})^{-1}\widetilde{\mathbf{X}}^T\mathbf{W}\mathbf{Y},$$

where r_{ridge} is a ridge tuning parameter and $\mathbf{I}$ is a $(p+dM_n)\times(p+dM_n)$ identity matrix.

- Step 2. At current iteration $k \geq 1$, update $\boldsymbol{\beta}$ by minimizing (10) and obtain the solution $\boldsymbol{\beta}^{(k)}$ for $\boldsymbol{\beta}$ as

$$\boldsymbol{\theta}^{(k)} = (\boldsymbol{\beta}^{(k)T}, \boldsymbol{\gamma}^{(k)T})^T = \arg\min_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\beta^{(k-1)}) = [\widetilde{\mathbf{X}}^T\mathbf{W}\widetilde{\mathbf{X}} + n\Lambda_{\lambda_n}(\boldsymbol{\beta}^{(k-1)})]^{-1}\widetilde{\mathbf{X}}^T\mathbf{W}\mathbf{Y},$$

where $\Lambda_\lambda(\boldsymbol{\beta}) = \text{diag}\Big\{\frac{p'_\lambda(|\beta_1|)}{|\beta_1|}, \cdots, \frac{p'_\lambda(|\beta_p|)}{|\beta_p|}, 0, \cdots, 0\Big\}$.

- Step 3. Repeat step 2 until convergence and denote the final solution by $\widehat{\boldsymbol{\beta}}$.

For the choice of the regulating parameter λ_n, we adopt the Bayesian information criterion (BIC) selector in Wang et al. [2009] which is defined by

$$\text{BIC}(\lambda) = \log(\text{RSS}_\lambda) + \text{df}_\lambda n^{-1}\log(n)C_n, \tag{11}$$

where $C_n = \log\log p$, RSS_λ stands for the residual sum of squares, and df_λ is the effective number of parameters. The optimal λ_n is the one that minimizes (11). Our numerical studies suggest BIC produces very stable estimation results and may be slightly more appropriate than other alternative approaches such as the cross-valdiation or the generalized cross-validation.

3 Asymptotic results

In this section, we establish the asymptotic theory for our estimator. Let

$$\boldsymbol{\beta}_0 = (\beta_{10}, \cdots, \beta_{p0})^T = (\boldsymbol{\beta}_{10}^T, \boldsymbol{\beta}_{20}^T)^T \tag{12}$$

be the true regression coefficient vector for the high-dimensional markers. Without loss of generality we partition the p-vector so that the first s elements $\boldsymbol{\beta}_{10}$ are non-zero and the remaining $p-s$ elements $\boldsymbol{\beta}_{20} = \mathbf{0}$. In correspondence we may partition each $\mathbf{X}_i = ((\mathbf{X}_i^{(1)})^T, (\mathbf{X}_i^{(2)})^T)^T$ for the high dimensional covariates.

For the nonparametric components, we focus our asymptotic analysis for $\mathcal{G}_k$ which is a space of spline functions. Extension to general basis expansions can be obtained with slight modification. Define $\rho_n = \max_{1\leq k\leq d}\inf_{g\in\mathcal{G}_k}\|g_{k0} - g\|_{L_2}$ where g_{k0} is the kth true function, $1 \leq k \leq d$. Thus ρ_n characterizes the approximation error due to spline approximation. Let $r_n = (M_n/n)^{1/2}$. The proofs of the following theorems are contained in the Appendix.

Theorem 1. *Under assumptions (C1)-(C5) in the Appendix,* $\lim_{n\to\infty}\rho_n = 0$,

$$\lim_{n\to\infty} n^{-1}M_n \log(M_n) = 0,$$

$\lambda_n \to 0$, *and* $\lambda_n/\max(r_n,\rho_n) \to \infty$, *we have the following:*

a. $\hat{\beta}_k = 0$, $s+1 \le k \le p$, *with probability approaching 1.*

b. $\|\hat{\beta}_k - \beta_{k0}\| = O_p(\max(r_n,\rho_n))$, $1 \le k \le s$.

c. $\|\hat{g}_k - g_{k0}\|_{L_2} = O_p(\max(r_n,\rho_n))$, $1 \le k \le d$.

Part a of Theorem 1 indicates the selection consistency of our procedure since we can identify the zero coefficients with probability tending to 1. Parts b and c provide the rate of convergence in estimating the nonzero coefficients and nonparametric functions, respectively.

We then establish the asymptotic distribution results. Denote by H the distribution of the observable Y's, and let $\tau_H = \inf\{y : H(y) = 1\}$ be the least upper bound for the support of H. Also denote by A the set of atoms of H. Introduce the following sub-distribution functions:

$$\begin{aligned}
\tilde{H}_1(\mathbf{x}^{(1)},\mathbf{u},y) &= P(\mathbf{X}^{(1)} \le \mathbf{x}^{(1)}, \mathbf{U} \le \mathbf{u}, Y \le y, \delta = 1)\\
\tilde{H}_0(y) &= P(Y \le y, \delta = 1).
\end{aligned}$$

Put $\tilde{\mathbf{x}}^{(1)} = ((\mathbf{x}^{(1)})^T, \mathbf{B}(u_1)^T, \cdots, \mathbf{B}(u_d)^T)^T$, $\boldsymbol{\theta}^* = (\boldsymbol{\beta}_1^T, \boldsymbol{\gamma}^T)^T$ and

$$\begin{aligned}
\xi_0(y) &= \exp\left\{\int_0^{y-} \frac{\tilde{H}_0(dz)}{1-H(z)}\right\}\\
\xi_{1j}^*(y;\boldsymbol{\theta}^*) &= \frac{1}{1-H(y)}\int_{w>y}(w - (\tilde{\mathbf{x}}^{(1)})^T\boldsymbol{\theta}^*)\tilde{x}_j\xi_0(w)\tilde{H}_1(d\tilde{\mathbf{x}}^{(1)}, dw)\\
\xi_{2j}^*(y;\boldsymbol{\theta}^*) &= \int\int \frac{I(v<y, v<w)(w - (\tilde{\mathbf{x}}^{(1)})^T\boldsymbol{\theta}^*)\tilde{x}_j\xi_0(w)}{(1-H(v))^2}\tilde{H}_0(dv)\tilde{H}_1(d\tilde{\mathbf{x}}^{(1)}, dw)\\
\xi_{l,\beta}^*(y;\boldsymbol{\theta}^*) &= (\xi_{l1}^*(y;\boldsymbol{\theta}^*), \cdots, \xi_{l,s}^*(y;\boldsymbol{\theta}^*)), \qquad l = 1,2,\\
\xi_{l,\gamma}^*(y;\boldsymbol{\theta}^*) &= (\xi_{l,s+1}^*(y;\boldsymbol{\theta}^*), \cdots, \xi_{l,s+M_nd}^*(y;\boldsymbol{\theta}^*)), \qquad l = 1,2.
\end{aligned}$$

Next, define $\tilde{\mathbf{Y}} = E(\mathbf{Y}|\mathbf{X},\mathbf{U})$, $\tilde{\boldsymbol{\theta}} = [\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{X}}]^{-1}\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{Y}} = (\tilde{\boldsymbol{\beta}}^T, \tilde{\boldsymbol{\gamma}}^T)^T$, and $\tilde{g}_k(u) = \mathbf{B}(u)^T\tilde{\boldsymbol{\gamma}}_k$. Denote $\mathbf{B}^*(\mathbf{u}) = \mathrm{diag}(\mathbf{B}(u_1)^T, \cdots, \mathbf{B}(u_d)^T)$, $\hat{\mathbf{g}}(\mathbf{u}) = (\hat{g}_1(u_1), \cdots, \hat{g}_d(u_d))^T = \mathbf{B}^*(\mathbf{u})\hat{\boldsymbol{\gamma}}$ and $\tilde{\mathbf{g}}(\mathbf{u}) = (\tilde{g}_1(u_1), \cdots, \tilde{g}_d(u_1))^T = \mathbf{B}^*(\mathbf{u})\tilde{\boldsymbol{\gamma}}$.

Theorem 2. *Suppose assumptions (C1)-(C6) in the Appendix hold,* $\lim_{n\to\infty}\rho_n = 0$,

$$\lim_{n\to\infty} n^{-1}M_n\log(M_n) = 0,$$

$\lambda_n \to 0$, *and* $\lambda_n/\max(r_n,\rho_n)\to\infty$. *Then as* $n\to\infty$, *we have*

a. $\sqrt{n}(\hat{\boldsymbol{\beta}}_1 - \tilde{\boldsymbol{\beta}}_1) \to N(0,\Psi)$ *in distribution, where* $\Psi = (\mathbf{H}^{(1)})^{-1}\Sigma^*_\beta(\mathbf{H}^{(1)})^{-1}$, $\mathbf{H}^{(1)} = E(\mathbf{X}_i^{(1)}(\mathbf{X}_i^{(1)})^T)$ *and* $\Sigma^*_\beta = Var[\delta_i\xi_0^*(Y_i)(Y_i - E(Y_i|\mathbf{X}_i^{(1)},\mathbf{U}_i))\mathbf{X}_i^{(1)} + (1-\delta_i)\xi^*_{1,\beta}(Y_i;(\tilde{\boldsymbol{\beta}}_1^T,\tilde{\boldsymbol{\gamma}}^T)^T) - \xi^*_{2,\beta}(Y_i;(\tilde{\boldsymbol{\beta}}_1^T,\tilde{\boldsymbol{\gamma}}^T)^T)]$.

b. $\sqrt{n}(\hat{\mathbf{g}}(\mathbf{u}) - \tilde{\mathbf{g}}(\mathbf{u})) \to N(0,\Gamma(\mathbf{u}))$ *in distribution, where* $\Gamma(\mathbf{u}) = \mathbf{B}^*(\mathbf{u})\mathbf{C}^{-1}\Sigma^*_\gamma\mathbf{C}^{-1}\mathbf{B}^*(\mathbf{u})^T$, $\mathbf{C} = E(\mathbf{B}_i\mathbf{B}_i^T)$ *and* $\Sigma^*_\gamma = Var[\delta_i\xi_0^*(Y_i)(Y_i - E(Y_i|\mathbf{X}_i^{(1)},\mathbf{U}_i))\mathbf{B}_i + (1-\delta_i)\xi^*_{1,\gamma}(Y_i;(\tilde{\boldsymbol{\beta}}_1^T,\tilde{\boldsymbol{\gamma}}^T)^T) - \xi^*_{2,\gamma}(Y_i;(\tilde{\boldsymbol{\beta}}_1^T,\tilde{\boldsymbol{\gamma}}^T)^T)]$.

Note that Ψ is exactly the same asymptotic covariance matrix of the nonpenalized weighted least squares estimate using only markers with nonzero coefficients. Hence Theorem 2 implies the oracle property of our estimator, i.e., the SCAD estimator can perform as well as the estimator obtained when the correct submodel was known. The result of Theorem 2 may also be used to construct confidence intervals and perform hypothesis tests for regression coefficients.

4 Simulation

We simulate sample dataset from Model (1) where $\mathbf{X}_i = (X_{i1},\cdots,X_{ip})^T, i = 1,\ldots,n$ are i.i.d. from multivariate normal distributions with mean $\mathbf{0_p}$ and the covariance between the components X_{ij} and X_{ik} is set to be $\rho_{jk} = \rho^{|j-k|}$. The coefficient $\boldsymbol{\beta}$ is specified to be $(0.5,1,1.5,2,2.5)^T$ for the first five elements and zero for the remaining components. The random error ϵ is generated from the following distributions: (I) the standard extreme value distribution: consequently $\exp(T_i)$ follows the Weibull distribution; (II) the standard normal distribution: consequently $\exp(T_i)$ follows the log-normal distribution.

We note that proportional hazards (PH) assumption is satisfied under (I) and fitting a Cox PH model is also appropriate for the data. In contrast, Case (II) violates the PH assumption. The censoring time C is uniformly distributed such that the censoring rate is about 25% for each simulation. In the following simulations, we set $\rho = 0.6$, $p = 1000$ and we consider $n = 200$ and 400. We investigate two scenarios for nonparametric components: (i) $d = 1$ with $g(u) = (1-u)^{-1}$; (ii) $d = 3$ with $g_1(u) = cos(2\pi u), g_2(u) = (1-u)^{-1}$ and

$g_3(u) = -\exp(-4u)$. The index variables $U_{ij}, 1 \le j \le d$ are generated independently from the uniform distribution $U(0,1)$.

We then compare five different methods for screening and variable selection:

(a). perform the approach proposed in this paper, i.e. covariate-adjusted screening and covariate-adjusted SCAD variable selection in an AFT model with nonparametric adjustment of confounders;

(b). still perform the covariate-adjusted screening and covariate-adjusted SCAD variable selection in an AFT model, adjusting confounders with linear regression components;

(c). perform marginal screening and SCAD variable selection in an AFT model, ignoring the confounders;

(d). perform the covariate-adjusted screening and covariate-adjusted SCAD variable selection in a Cox PH model, adjusting confounders with linear regression components;

(e). perform marginal screening and SCAD variable selection in a Cox PH model, ignoring the confounders.

We notice that (c) and (e) are existing approaches to variable screening and variable selection, available in packages. The other three methods are all new, where the theoretical justification for (a) and (b) are provided in this paper and that for (d) may need further work.

Each case is repeated $N = 1000$ times and the following quantities over 1000 replications are reported:

- For variable selection performance we report: UF, the proportion of underfitted models; OF, the proportion of overfitted models; CZ, the percentage of correctly estimated zero coefficients; IZ, the percentage of incorrectly estimated nonzero coefficients; TP, the true positive fraction; FP, the false positive fraction.

- For model-fitting performance we report: REE, the relative estimation error defined by

$$N^{-1}\sum_{k=1}^{N}\{(\hat{\boldsymbol{\beta}}_{(k)}-\boldsymbol{\beta})^T\hat{E}(\mathbf{X}_{(k)}\mathbf{X}_{(k)}^T)(\hat{\boldsymbol{\beta}}_{(k)}-\boldsymbol{\beta})\}/\{\boldsymbol{\beta}^T\hat{E}(\mathbf{X}_{(k)}\mathbf{X}_{(k)}^T)\boldsymbol{\beta}\},$$

  where $\hat{E}\mathbf{X}_{(k)}\mathbf{X}_{(k)}^T = 1/n\sum_{i=1}^{n}\mathbf{X}_{i,(k)}\mathbf{X}_{i,(k)}^T$, where $\mathbf{X}_{i,(k)}$ is the ith observation in the kth replication, $\hat{\boldsymbol{\beta}}_{(k)}$ stands for the parametric estimation in the kth replication.; MSE, the

mean squared error, defined by

$$\hat{E}||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2$$

where $\|\cdot\|$ is the European norm; MME, the median of model errors defined by

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \hat{E}\mathbf{X}\mathbf{X}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

and MAD is its median absolute deviation.

Simulation results for the $d = 1$ case are given in Table 1 and those for the $d = 3$ case are given in Table 2. We make the following observations.

1. Without adjustment of confounders or with only a linear adjustment, the selection accuracy can be affected. The FP values of (b), (c) and (e) are all unacceptably high and the performance of these methods is even worse in Table 2 where the confounder effects are stronger.

2. With proper adjustment of confounders, (a) and (d) both enjoy selection consistency for the extreme value distribution. When proportional hazards assumption fails, (d) may perform less satisfactorily.

3. Sample size is critical. All methods improve with increasing sample sizes.

4. In view of the REE, MSE, MME and MAE values in both tables, we notice that (a) has the smallest estimation errors in all cases.

In conclusion, method (a) outperforms the other methods in both variable selection and model fitting.

5 Lung cancer data analysis

Lung cancer represents the leading cause of cancer death for both men and women in the United States and many other Western countries. The 5-year survival is only 15% and has little improvement over the past decades. This is mainly because approximately two-thirds of lung cancer cases are diagnosed at advanced stages where surgical resection is no longer an option. Accurate early detection is thus crucial for lung cancer treatment. Prognostic gene expression signatures for survival in early-stage lung cancer have been proposed for clinical

application. Such technologies have identified potential biomarkers and gene signatures for classifying patients with significantly different survival outcomes (Chen et al. [2007], Lu et al. [2006]).

Individual lung cancer profiling studies may have relatively small sample sizes and lead to unreliable results. To increase sample size, Shedden et al. [2008] conducted a large retrospective, multi-site, blinded study, using a total of 442 lung adenocarcinomas, the specific type of lung cancer that is increasing in incidence. Gene expression data were generated by four different laboratories under a common protocol. The same data set has been used as a validation sample for a separate analysis (Xie et al. [2011]).

In addition to genetic mutations and defects factors, multiple clinical and environmental risk factors may contribute to lung cancer progression. In this analysis, the low-dimensional covariates include age, gender, cancer stage, adjuvant chemotherapy treatment and smoking history. Subjects with missing measurements in overall survival or confounders are removed from analysis. A total of 437 subjects are included in downstream analysis. The median follow-up time is 46 months. The overall censoring rate is 46.22%. The Kaplan-Meier estimate of the survival distribution is plotted in Figure 1. For each subject, the expressions of 22283 genes are available.

We first use a marginal screening and SCAD variable selection in an AFT model, ignoring the confounders. The estimation results are summarized in Table 3. We then use our proposed covariate-adjusted screening and covariate-adjusted SCAD variable selection in an AFT model. The estimation results are summarized in Table 4. Under the two approaches, the selected genes are completely different. After adjusting for covaraites, none of the genes in Table 3 are selected in Table 4. The main message is that the variation captured by the genes in Table 3 may be completely due to the sample heterogeneity, resulted from the variation of low-dimensional factors. After controlling the effects of various confounders, a new list of genes are identified. Considering the sample is obtained from multiple observational studies with non-homogeneous populations, we believe the results in Table 4 should be emphasized more in practice in order to achieve meaningful biomarker discovery in the presence of heterogeneously distributed covariates.

6 Concluding Remarks

Stute estimator can be shown to be equivalent to the inverse-probability weighted (IPW) estimator and the proposal in this paper can be easily extended for IPW estimation for AFT model. Besides Stute estimator, there exist other approaches such as rank-based estimator and Buckley-James estimators, among others. The technical conditions and theoretical justification may need to be established in a different fashion from this article. More efforts are needed to completely investigate all these related cases.

We assume constant coefficients for high-dimensional covariates for their lucid interpretation. But for some applications, it has been noticed that using functional coefficients may be more flexible. Variable screening and variable selection for functional coefficients may follow a similar construction as our proposed methods. The statistical properties need to be formally studied in future research.

References

J. Bradic, J. Fan, and J. Jiang. Regularization for Cox's proportional hazards model with NP-dimensionality. *Annals of Statistics*, 39:3092–3120, 2011.

T. Cai, J. Huang, and L. Tian. Regularized estimation for the accelerated failure time model. *Biometrics*, 65:394–404, 2009.

H. Y. Chen, S. L. Yu, and et al. A five-gene signature and clinical outcome in non-small-cell lung cancer. *New England Journal of Medicine*, 356:11–20, 2007.

M. Y. Cheng, W. Zhang, and L. H. Chen. Statistical estimation in generalized multiparameter likelihood models. *Journal of the American Statistical Association*, 104:1179–1191, 2009.

M. Y. Cheng, T. Honda, J. Li, and H. Peng. Nonparametric independence screening and structure identification for ultra-high dimensional longitudinal/clustered data. *Annals of Statistics*, 42:1819–1849, 2014.

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

J. Fan and R. Li. Variable selection for cox's proportional hazards model and frailty model. *Annals of Statistis*, 30:74–99, 2002.

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

J. Fan, R. Samworth, and Y. Wu. Ultrahigh dimensional feature selection: Beyond the linear model. *Journal of Machine Learning Research*, 10:2013–2038, 2009.

L. Gordis. *Epidemiology*. Saunders; 4th edition, 2008.

J. Hu and H. Chai. Mathematical Statistics. *Journal of Multivariate Analysis*, 122:96–114, 2013.

J. Huang and S. Ma. Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Analysis*, 16:176–195, 2010.

J. Huang, S. Ma, and H. Xie. Regularized estimation in the accelerated failure time model with high dimensional covariate. *Biometrics*, 62:813–820, 2006.

J. Z. Huang, C. O. Wu, and L. Zhou. Polynomial spline estimation and inference for varying-coefficient models with longitudinal data. *Statistica Sinica*, 14:763–788, 2004.

B. A. Johnson, D. Y. Lin, and D. Zeng. Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association*, 103: 672–680, 2008.

G. R. Li, H. Peng, J. Zhang, and L. X. Zhu. Robust rank correlation based screening. *Annals of Statistics*, 40:1846–1877, 2012.

H. Lian, J. Li, and X. Tang. SCAD-penalized regression in additive partially linear proportional hazards models with an ultra-high-dimensional linear part. *Jounral of Multivariate Analysis*, 125:50–64, 2014.

Y. Lu, W. Lemon, and et al. A gene expression signature predicts survival of subjects with state i non-small cell lung cancer. *PLoS Medicine*, 3:2229–2243, 2006.

V. Petrov. *Sums of Independent Random Variables*. New York: Springer-Verlag, 1975.

K. Shedden, J. M. G. Taylor, and et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature Medicine*, 14:822–827, 2008.

W. Stute. Consistent estimation under random censorship when covariates are present. *Journal of Multivariate Analysis*, 45:89–103, 1993.

W. Stute. Distributional convergence under random censorship when covariables are present. *Scandinavian Journal of Statistics*, 23:461–471, 1996.

T. J. VanderWeele and I. Shpitser. On the definition of a confounder. *Annals of Statistics*, 41:196–220, 2013.

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*, 71:671–683, 2009.

Y. Xie, G. Xiao, and et al. Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of non-small-cell lung cancer patients. *Clinical Cancer Research*, 17:5705–5714, 2011.

Appendix: Conditions and Proofs

We write $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$. Define $\tilde{\mathbf{Y}} = E(\mathbf{Y}|\mathbf{X}, \mathbf{U})$, $\tilde{\boldsymbol{\theta}} = [\widetilde{\mathbf{X}}^T\mathbf{W}\widetilde{\mathbf{X}}]^{-1}\widetilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{Y}} = (\tilde{\boldsymbol{\beta}}^T, \tilde{\boldsymbol{\gamma}}^T)^T$, and $\tilde{g}_k(u) = \mathbf{B}(u)^T\tilde{\boldsymbol{\gamma}}_k$.

We need the following assumptions.

(C1) $E(\epsilon_i|\mathbf{X}_i, \mathbf{U}_i) = 0$ and $E(T_i^2)$ is finite.

(C2) T_i and C_i are independent and $P(Y_i \leq C_i|\mathbf{X}_i, \mathbf{U}_i, Y_i) = P(Y_i \leq C_i|Y_i)$.

(C3) The eigenvalues of the matrix $E(\mathbf{X}_i\mathbf{X}_i^T)$ is bounded away from zero and finite.

(C4) $\tau_T < \tau_C$ or $\tau_T = \tau_C = \infty$.

(C5) The true parameter $\boldsymbol{\beta}_0$ lives in a compact space Θ. Each true function g_{k0} is from a second order Sobolev space.

(C6) $E\left\{\epsilon^2\delta\mathbf{X}^{(1)}(\mathbf{X}^{(1)})^T\right\} < \infty$ and $E\left\{|\epsilon\mathbf{X}^{(1)}|\sqrt{R(Y)}\right\} < \infty$ where $R(y) = \int_0^{y-}\{(1-H(w))(1-G(w))\}^{-1}G(dw)$ and G is the distribution function of the censoring time C.

The estimator $\hat{\boldsymbol{\theta}}$ is of a form of weighted least squares estimator. However, the Kaplan-Meier weights $\{w_{ni} : i = 1, \cdots, n\}$ do not satisfy the assumption which are usually required in weighted least squares estimation. We need Lemma 1 to ensure the validity of convergence argument used in the proof of Theorems 1 and 2. The proof of the lemma may follow Stute [1993].

Lemma 3. *For an integrable function ϕ, define a functional $\mathcal{S}_n\phi = \sum_{i=1}^n w_{ni}\phi(Y_i, \mathbf{X}_i, \mathbf{U}_i)$. Under (C1) and (C2), with probability one and in the mean we have*

$$\lim_{n\to\infty} \mathcal{S}_n\phi = \int_{Y<\tau_H} \phi(Y, \mathbf{X}, \mathbf{U})dP + I(\tau_H \in A)\int_{Y=\tau_H} \phi(\tau_H, \mathbf{X}, \mathbf{U})dP. \tag{A.1}$$

We need the following lemma which summarizes necessary properties of the polynomial spline functions. The proof of the lemma may follow lemmas A.3 in Huang et al. [2004].

Lemma 4. *Assume $\lim_{n\to\infty} n^{-1}M_n \log(M_n) = 0$. Except on an event whose probability tends to zero, all the eigenvalues of $M_n/n\sum_{i=1}^n \mathbf{B}_i\mathbf{B}_i^T$ are bounded away from zero and infinity.*

The next lemma establish the consistency of the estimator.

Lemma 5. *Assume the same conditions as Theorem 1. Then $\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\| = O_p(r_n + (\lambda_n\rho_n)^{1/2})$.*

Proof of Lemma 3. We note

$$\begin{aligned} Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) &= \frac{1}{n}\left[(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\theta}})^T\mathbf{W}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\theta}}) - (\mathbf{Y} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}})^T\mathbf{W}(\mathbf{Y} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}})\right] \\ &\quad + \sum_{k=1}^p \{p_{\lambda_n}(|\hat{\beta}_k|) - p_{\lambda_n}(|\tilde{\beta}_k|)\} \\ &= \frac{1}{n}\left[-2(\mathbf{Y} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}})^T\mathbf{W}\tilde{\mathbf{X}}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) + (\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})^T\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{X}}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})\right] \\ &\quad + \sum_{k=1}^p \{p_{\lambda_n}(|\hat{\beta}_k|) - p_{\lambda_n}(|\tilde{\beta}_k|)\} \\ &= -2\boldsymbol{\epsilon}^T\mathbf{W}\tilde{\mathbf{X}}\mathbf{v}M_n^{1/2}\delta_n/n + \delta_n^2/nM_n\mathbf{v}^T\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{X}}\mathbf{v} \\ &\quad + \sum_{k=1}^p \{p_{\lambda_n}(|\hat{\beta}_k|) - p_{\lambda_n}(|\tilde{\beta}_k|)\} \le 0, \end{aligned} \tag{A.2}$$

where in the third equality we write $\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} = \delta_n M_n^{1/2}\mathbf{v}$, with δ_n a scalar and $\mathbf{v}$ a vector satisfying $\|\mathbf{v}\| = 1$, and use the fact that $\tilde{\mathbf{X}}^T\mathbf{W}(\mathbf{Y} - \boldsymbol{\epsilon} - \tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}}) = 0$. The inequality follows from the definition of $\hat{\boldsymbol{\theta}}$. We first show that $\delta_n = O_p(r_n + \lambda_n)$. To this end, we can show easily

$$\frac{M_n^{1/2}}{n}\boldsymbol{\epsilon}^T\mathbf{W}\tilde{\mathbf{X}}\mathbf{v} = \frac{M_n^{1/2}}{n}\sum_{i=1}^n \epsilon_i w_{ni}\tilde{\mathbf{X}}_{(i)}\mathbf{v} = O_p(r_n). \tag{A.3}$$

By assumption (C3) and Lemma 2, there exists a positive c_1 such that

$$\frac{M_n}{n}\mathbf{v}^T\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{X}}\mathbf{v} = \frac{M_n}{n}\sum_{i=1}^n w_{in}\mathbf{v}^T\left(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)} + \mathbf{B}_{(i)}^T\mathbf{B}_{(i)}\right)\mathbf{v} \ge c_1, \tag{A.4}$$

with probability approaching 1. Using inequality $|p_\lambda(a) - p_\lambda(b)| \le \lambda|a-b|$, we obtain

$$\sum_{k=1}^{p}\{p_{\lambda_n}(|\hat{\beta}_k|) - p_{\lambda_n}(|\tilde{\beta}_k|) \;\; \geq \;\; \sum_{k=1}^{p} -\lambda_n|\hat{\beta}_k - \tilde{\beta}_k| \asymp -\lambda_n\delta_n. \tag{A.5}$$

Therefore, $-O_p(r_n)\delta_n + c_1\delta_n^2 - \lambda_n\delta_n \le 0$ with probability approaching 1, which implies that $\delta_n = O_p(r_n + \lambda_n)$.

Now we notice for $1 \le k \le p$, $|\hat{\beta}_k - \tilde{\beta}_k| = o_p(1)$. Next we can see

$$\begin{aligned} |\|\tilde{\gamma}_k\| - \|g_k\|_{L_2}| &\le |\|\tilde{g}_k\|_{L_2} - \|g_k\|_{L_2}| && \text{(A.6)} \\ &\le \|\tilde{g}_k - g_k\|_{L_2} = O_p(\rho_n) = o_p(1). && \text{(A.7)} \end{aligned}$$

It then follows that $\hat{\beta}_k \to \beta_{k0}$, $\tilde{\beta}_k \to \beta_{k0}$, $\|\hat{g}_k\|_{L_2} \to \|g_{k0}\|_{L_2}$ and $\|\tilde{g}_k\|_{L_2} \to \|g_{k0}\|_{L_2}$ in probability. Because $|\beta_{k0}| > 0$ for $1 \le k \le s$ and $\lambda \to 0$, we have that, with probability approaching 1, $|\hat{\beta}_k| > a\lambda_n$ and $|\tilde{\beta}_k| > a\lambda_n$ for $1 \le k \le s$. On the other hand, $\beta_{k0} = 0$ for $s+1 \le k \le p$, so the previous results imply $\tilde{\beta}_k = O_p(\rho_n)$. Since $\lambda_n/\rho_n \to \infty$, we have $|\tilde{\beta}_k| < \lambda_n$, $s+1 \le k \le p$. Consequently by the definition of $p_\lambda(\cdot)$, we have $P(p_{\lambda_n}(|\tilde{\beta}_k|) = p_{\lambda_n}(|\hat{\beta}_k|)) \to 1$ when $1 \le k \le s$; and $P(p_{\lambda_n}(|\tilde{\beta}_k|) = \lambda_n|\tilde{\beta}_k|) \to 1$ when $s+1 \le k \le p$. Therefore

$$\sum_{k=1}^{p}\{p_{\lambda_n}(|\hat{\beta}_k|) - p_{\lambda_n}(|\tilde{\beta}_k|)\} = \lambda_n \sum_{k=s+1}^{p} |\tilde{\beta}_k| \ge -O_p(\lambda_n\rho_n). \tag{A.8}$$

Combining with previous results, we have

$$Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}) \ge -O_p(r_n)\delta_n + c_1\delta_n^2 - O(\lambda_n\rho_n), \tag{A.9}$$

which implies that $\delta_n = O_p(r_n + (\lambda_n\rho_n)^{1/2})$. □

Proof of Theorem 1. We prove part a by contradiction. Suppose that for a sufficiently large n there exists a constant $\eta > 0$ such that with probability at least η there exists a $k^* > s$ such that $\hat{\beta}_{k^*} \neq 0$. Let $\hat{\boldsymbol{\theta}}^*$ be a vector constructed by replacing $\hat{\beta}_{k^*}$ with 0 in $\hat{\boldsymbol{\theta}}$. Then

$$Q(\hat{\boldsymbol{\theta}}) - Q(\hat{\boldsymbol{\theta}}^*) = \frac{1}{n}\left[(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\theta}})^T\mathbf{W}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\theta}}) - (\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\theta}}^*)^T\mathbf{W}(\mathbf{Y} - \tilde{\mathbf{X}}\hat{\boldsymbol{\theta}}^*)\right] + p_{\lambda_n}(|\hat{\beta}_{k^*}|) \tag{A.10}$$

By Lemma 1 and the fact that $\beta_{k^*0} = 0$, $\hat{\beta}_{k^*} = O_p(r_n + (\lambda_n\rho_n)^{1/2})$. Because $\lambda_n/\max(r_n, \rho_n) \to \infty$, we have $|\hat{\beta}_{k^*}| < \lambda_n$ and thus $p_{\lambda_n}(|\hat{\beta}_{k^*}|) = \lambda_n|\hat{\beta}_{k^*}|$ with probability approaching 1. For the

first term in (A.10), simple algebra leads to

$$\begin{aligned} & (\mathbf{Y}-\tilde{\mathbf{X}}\hat{\boldsymbol{\theta}})^T\mathbf{W}(\mathbf{Y}-\tilde{\mathbf{X}}\hat{\boldsymbol{\theta}})-(\mathbf{Y}-\tilde{\mathbf{X}}\hat{\boldsymbol{\theta}}^*)^T\mathbf{W}(\mathbf{Y}-\tilde{\mathbf{X}}\hat{\boldsymbol{\theta}}^*) \\ \geq\ & -(\mathbf{Y}-\tilde{\mathbf{X}}\hat{\boldsymbol{\theta}})^T\mathbf{W}\tilde{\mathbf{X}}(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}^*) \\ =\ & -2(\mathbf{Y}-\tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}})^T\mathbf{W}\tilde{\mathbf{X}}(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}^*) \\ & -2(\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}^*)\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{X}}(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}^*). \end{aligned} \tag{A.11}$$

By the Cauchy-Schwartz inequality,

$$\begin{aligned} & (\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}^*)\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{X}}(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}^*) \\ \leq\ & \{(\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}^*)\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{X}}(\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}^*)\}^{1/2} \\ & \times\{(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}^*)\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{X}}(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}^*)\}^{1/2} \\ \leq\ & c_2\frac{n}{M_n}\|\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}^*\|\|\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}^*\|. \end{aligned}$$

From the triangle inequality and Lemma 1, it follows that

$$\begin{aligned} \|\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}^*\| &\leq \|\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}\|+|\tilde{\beta}_{k^*}| \\ &= O_p(M_n^{1/2}\{r_n+(\lambda_n\rho_n)^{1/2}+\rho_n\}), \end{aligned}$$

thus

$$\frac{1}{n}(\tilde{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}^*)\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{X}}(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}^*)=O_p(M_n^{-1/2}(r_n+(\lambda_n\rho_n)^{1/2}+\rho_n))|\hat{\beta}_{k^*}|. \tag{A.12}$$

We can also show that

$$|(\mathbf{Y}-\tilde{\mathbf{X}}\tilde{\boldsymbol{\theta}})^T\mathbf{W}\tilde{\mathbf{X}}(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}^*)|=|\boldsymbol{\epsilon}^T\mathbf{W}\tilde{\mathbf{X}}(\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}^*)|=O_p(\frac{nr_n}{M_n^{1/2}})|\hat{\beta}_{k^*}|. \tag{A.13}$$

Combining (A.10) to (A.13), we arrive at

$$\begin{aligned} Q(\hat{\boldsymbol{\theta}})-Q(\hat{\boldsymbol{\theta}}^*) \ \geq\ & \frac{\lambda_n}{M_n^{1/2}}|\hat{\beta}_{k^*}|-O_p(\frac{r_n}{M_n^{1/2}})|\hat{\beta}_{k^*}| \\ & -O_p(\frac{r_n+(\lambda_n\rho_n)^{1/2}+\rho_n}{M_n^{1/2}})|\hat{\beta}_{k^*}|. \end{aligned} \tag{A.14}$$

We note that the first term on the right hand side of (A.14) dominates the other two terms since $\lambda_n/\max(\rho_n,r_n)\to\infty$. This contradicts the fact that $Q(\hat{\boldsymbol{\theta}})-Q(\hat{\boldsymbol{\theta}}^*)\leq 0$. Hence the proof of part a is completed.

To prove parts b and c, we define the oracle version of $\tilde{\boldsymbol{\theta}}$,

$$\tilde{\boldsymbol{\theta}}_\Omega = \arg \min_{\boldsymbol{\theta}=(\boldsymbol{\beta}_1^T, \mathbf{0}^T, \boldsymbol{\gamma}^T)} \frac{1}{n}(\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\theta})^T \mathbf{W} (\mathbf{Y} - \tilde{\mathbf{X}}\boldsymbol{\theta}) \tag{A.15}$$

which is obtained as if the information of the nonzero components were given. By the construction and Lemma 2, we have $\|\tilde{\boldsymbol{\theta}}_\Omega - \boldsymbol{\theta}_0\| = O_p(\rho_n)$ and $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = o_p(1)$. Thus, with probability approaching 1, $\tilde{\beta}_{k,\Omega} \to \beta_{k0}$, $\hat{\beta}_k \to \beta_{k0}$ $(1 \le k \le s)$, $\|\tilde{g}_{k,\Omega}\| \to \|g_{k0}\|$, and $\|\hat{g}_k\| \to \|g_{k0}\|$ $(1 \le k \le d)$. On the other hand, for $s+1 \le k \le p$, by the definition $\tilde{\beta}_{k,\Omega} = 0$, and by part a, with probability approaching 1, $\hat{\beta}_k = 0$. Consequently, we have

$$\sum_{k=1}^{p} p_{\lambda_n}(|\tilde{\beta}_{k,\Omega}|) = \sum_{k=1}^{p} p_{\lambda_n}(|\hat{\beta}_k|) \tag{A.16}$$

with probability approaching 1. Now write $\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}_\Omega = \delta_n M_n^{1/2} \mathbf{v}$, with $\|\mathbf{v}\| = 1$. By (A.2) and (A.16),

$$\begin{aligned} 0 &\ge Q(\hat{\boldsymbol{\theta}}) - Q(\tilde{\boldsymbol{\theta}}_\Omega) \\ &= -2\boldsymbol{\epsilon}\mathbf{W}\tilde{\mathbf{X}}\mathbf{v}M_n^{1/2}\delta_n/n + \delta^2/nM_n\mathbf{v}^T\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{X}}\mathbf{v} \\ &\ge -O_p(r_n)\delta_n + c_1\delta_n^2. \end{aligned}$$

Thus $\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}_\Omega\| \asymp \delta_n = O_p(r_n)$, which, together with $\|\tilde{\boldsymbol{\theta}}_\Omega - \boldsymbol{\beta}_0\| = O_p(\rho_n)$, implies that $\|\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}\| = O_p(\rho_n + r_n)$. Hence the claims in parts b and c follow. □

To prove Theorem 2, we need the following lemma which gives the asymptotic behavior of the AFT estimator under Kaplan-Meier weights. Denote the right hand side of (A.1) to be $\mathcal{S}\phi$. Introduce the following sub-distribution functions:

$$\begin{aligned} \tilde{H}_1(\mathbf{x}, \mathbf{u}, y) &= P(\mathbf{X} \le \mathbf{x}, \mathbf{U} \le \mathbf{u}, Y \le y, \delta = 1) \\ \tilde{H}_0(y) &= P(Y \le y, \delta = 1). \end{aligned}$$

Put

$$\begin{aligned} \xi_0(y) &= \exp\left\{ \int_0^{y-} \frac{\tilde{H}_0(dz)}{1 - H(z)} \right\} \\ \xi_1^\phi(y) &= \frac{1}{1 - H(y)} \int_{w>y} \phi(\mathbf{x}, \mathbf{u}, w) \xi_0(w) \tilde{H}_1(d\mathbf{x}, d\mathbf{u}, dw) \\ \xi_2^\phi(y) &= \int\int \frac{I(v < y, v < w)\phi(\mathbf{x}, \mathbf{u}, w)\xi_0(w)}{(1 - H(v))^2} \tilde{H}_0(dv)\tilde{H}_1(d\mathbf{x}, d\mathbf{u}, dw). \end{aligned}$$

Let $\{\phi_1, \cdots, \phi_J\}$ be a set of measurable functions. Write

$$\underline{\mathcal{S}}_n = (\mathcal{S}_n\phi_1, \cdots, \mathcal{S}_n\phi_J)^T$$

and

$$\underline{\mathcal{S}} = (\mathcal{S}\phi_1, \cdots, \mathcal{S}\phi_J)^T.$$

Lemma 6. *Assume that (C1) and (C2) hold. In addition, assume the following two integrability conditions hold for all ϕ_j, $1 \le j \le J$,*

$$\int \phi_j(\mathbf{X}, \mathbf{U}, W)\xi_0(W)\delta^2 d\mathbf{P}_{\mathbf{X},\mathbf{U},Y} < \infty \tag{A.17}$$

$$\int \phi_j(\mathbf{X}, \mathbf{U}, W)\sqrt{R(W)}d\mathbf{P}_{\mathbf{X},\mathbf{U},Y} < \infty. \tag{A.18}$$

Then in distribution

$$\sqrt{n}(\underline{\mathcal{S}}_n - \underline{\mathcal{S}}) \to N(0, \Sigma), \tag{A.19}$$

where $\Sigma = (\sigma_{jj'})$, $\sigma_{jj'} = cov(\psi_j, \psi_{j'})$ and $\psi_j = \phi_j(\mathbf{X}, \mathbf{U}, Y)\xi_0(Y)\delta + \xi_1^{\phi_j}(Y)(1-\delta) - \xi_2^{\phi_j}(Y)$.

The proof of this lemma may follow Stute [1996]. We may now proceed to the proof of Theorem 2.

Proof of Theorem 2. According to the proof of Lemma 3, with probability approaching 1, $|\tilde{\beta}_k| > a\lambda_n$, $|\hat{\beta}_k| > a\lambda_n$ and thus $p_{\lambda_n}(|\tilde{\beta}_k|) = p_{\lambda_n}(|\hat{\beta}_k|)$ for $1 \le k \le s$. By Theorem 1, with probability approaching 1, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T, \hat{\boldsymbol{\gamma}}^T)^T$ is a local minimizer of $Q(\boldsymbol{\theta})$. We may note that $Q(\boldsymbol{\theta})$ is quadratic in $(\boldsymbol{\beta}_1^T, \boldsymbol{\gamma}^T)^T$ when $|\beta_k| > a\lambda_n$ for $1 \le k \le s$. Therefore $\partial Q(\boldsymbol{\theta})/\partial\boldsymbol{\theta}|_{\boldsymbol{\beta}_1=\hat{\boldsymbol{\beta}}_1, \boldsymbol{\beta}_2=\mathbf{0}, \boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}} = \mathbf{0}$, which implies that

$$(\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\gamma}}^T)^T = \left(\sum_{i=1}^n w_{ni} \begin{bmatrix} (\mathbf{X}_i^{(1)})(\mathbf{X}_i^{(1)})^T & (\mathbf{X}_i^{(1)})\mathbf{B}_i^T \\ \mathbf{B}_i(\mathbf{X}_i^{(1)})^T & \mathbf{B}_i\mathbf{B}_i^T] \end{bmatrix}\right)^{-1} \left(\sum_{i=1}^n w_{ni} \begin{bmatrix} (\mathbf{X}_i^{(1)}) \\ \mathbf{B}_i \end{bmatrix} Y_i\right).$$

Next we put

$$(\tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\gamma}}^T)^T = \left(\sum_{i=1}^n w_{ni} \begin{bmatrix} (\mathbf{X}_i^{(1)})(\mathbf{X}_i^{(1)})^T & (\mathbf{X}_i^{(1)})\mathbf{B}_i^T \\ \mathbf{B}_i(\mathbf{X}_i^{(1)})^T & \mathbf{B}_i\mathbf{B}_i^T \end{bmatrix}\right)^{-1} \left(\sum_{i=1}^n w_{ni} \begin{bmatrix} (\mathbf{X}_i^{(1)}) \\ \mathbf{B}_i \end{bmatrix} E\{Y_i|\mathbf{X}_i, \mathbf{U}_i\}\right).$$

Invoking Lemmas 1 and 4 while applying a version of Lindeberge central limit theorem (cf. Petrov [1975]), we obtain that for any vector $\mathbf{c}_n$ with dimension $s + dM_n$ and components not

all 0,

$$\{\mathbf{c}_n^T \Upsilon \mathbf{c}_n\}^{-1/2} \mathbf{c}_n^T \left(\begin{bmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\gamma}} \end{bmatrix} - \begin{bmatrix} \tilde{\boldsymbol{\beta}}_1 \\ \tilde{\boldsymbol{\gamma}} \end{bmatrix} \right) \to_d N(0,1), \tag{A.20}$$

where $\Upsilon = \mathbf{H}^{-1}\Sigma^*\mathbf{H}^{-1}$, $\mathbf{H} = E \begin{bmatrix} (\mathbf{X}_i^{(1)})(\mathbf{X}_i^{(1)})^T & (\mathbf{X}_i^{(1)})\mathbf{B}_i^T \\ \mathbf{B}_i(\mathbf{X}_i^{(1)})^T & \mathbf{B}_i\mathbf{B}_i^T \end{bmatrix}$ and $\Sigma^* = \mathrm{Var}[\delta_i \xi_0^*(Y_i)(Y_i - E(Y_i|\mathbf{X}_i^{(1)}, \mathbf{U}_i))((\mathbf{X}_i^{(1)})^T, \mathbf{B}_i^T)^T + (1-\delta_i)\xi_1^*(Y_i) - \xi_2^*(Y_i)]$. Part a of Theorem 2 follows from (A.20) immediately. Further, if we choose $\mathbf{c}_n = (\mathbf{0}^T, \mathbf{B}(\mathbf{u})^T \mathbf{a}_n)$ such that not all elements of $\mathbf{a}_n$ are 0, we obtain

$$\{\mathbf{a}_n^T \Gamma(\mathbf{u}) \mathbf{a}_n\}^{-1/2} \mathbf{a}_n^T \left\{ \begin{bmatrix} \hat{g}_1(u_1) \\ \cdots \\ \hat{g}_d(u_d) \end{bmatrix} - \begin{bmatrix} \tilde{g}_1(u_1) \\ \cdots \\ \tilde{g}_d(u_d) \end{bmatrix} \right\} \to_d N(0,1)$$

which leads to part b.

□

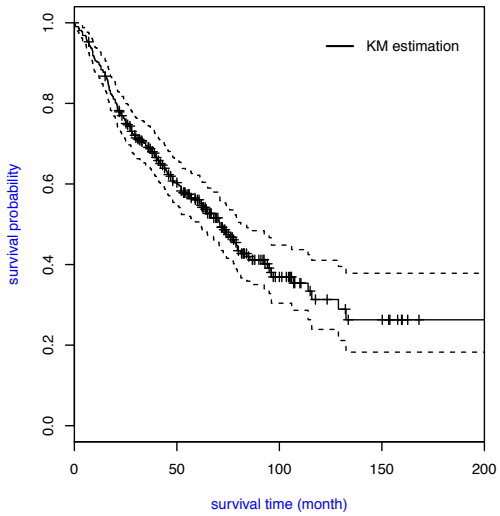

Figure 1: Kaplan Meier estimate of the survival probability.

Method	UF	OF	CZ	IZ	TP	FP	REE	MSE	MME	MAD
$n = 200$ Extreme value distribution										
(a)	0.08	0.72	0.998	0.016	4.92	1.92	0.007	0.289	0.190	0.096
(b)	0.31	0.59	0.997	0.062	4.69	2.75	0.022	0.794	0.610	0.207
(c)	0.99	0.01	0.991	0.424	2.88	8.65	0.386	11.085	12.428	2.163
(d)	0.84	0.16	0.995	0.262	3.69	5.21	0.398	6.106	11.101	3.729
(e)	0.99	0.01	0.990	0.32	3.4	10.22	0.501	8.033	14.964	2.045
$n = 400$ Extreme value distribution										
(a)	0.01	0.68	0.998	0.002	4.99	1.59	0.004	0.134	0.103	0.056
(b)	0.11	0.72	0.998	0.022	4.89	2.29	0.014	0.405	0.387	0.156
(c)	0.98	0.02	0.992	0.306	3.47	7.84	0.265	6.770	8.361	1.649
(d)	0.86	0.14	0.997	0.242	3.79	3.42	0.435	6.23	12.992	2.352
(e)	0.97	0.03	0.995	0.404	2.98	5.36	0.629	8.921	20.080	2.106
$n = 200$ Normal distribution										
(a)	0.02	0.75	0.998	0.004	4.98	1.82	0.005	0.202	0.146	0.067
(b)	0.24	0.61	0.997	0.048	4.76	2.53	0.019	0.710	0.464	0.233
(c)	0.94	0.06	0.994	0.254	3.73	6.12	0.176	5.414	5.171	1.547
(d)	0.83	0.17	0.995	0.260	3.70	4.76	0.416	6.292	12.338	3.710
(e)	0.90	0.1	0.989	0.318	3.41	10.93	0.515	8.303	16.539	2.354
$n = 400$ Normal distribution										
(a)	0.00	0.56	0.999	0.000	5.00	1.10	0.002	0.077	0.051	0.024
(b)	0.04	0.71	0.998	0.008	4.96	1.97	0.010	0.325	0.194	0.093
(c)	0.79	0.20	0.994	0.174	4.13	5.75	0.116	3.120	3.312	0.932
(d)	0.82	0.18	0.996	0.220	3.90	3.67	0.429	6.130	12.957	2.317
(e)	0.97	0.03	0.994	0.420	2.90	5.83	0.640	9.098	20.919	2.095

Note: UF is the proportion of underfitted models; OF is the proportion of overfitted models; CZ is the percentage of correctly estimated zero coefficients; IZ is the percentage of incorrectly estimated nonzero coefficients; TP is the true positive fraction; FP is the false positive fraction; REE is the relative estimation error; MSE is the mean squared error; MME is the median of model errors and MAD is its median absolute deviation.

Table 1: Variable selection and model-fitting performance when $d = 1$.

Method	UF	OF	CZ	IZ	TP	FP	REE	MSE	MME	MAD
$n = 200$ Extreme value distribution										
(a)	0.17	0.65	0.998	0.034	4.83	1.98	0.006	0.304	0.181	0.077
(b)	1.00	0.00	0.988	0.534	2.33	11.41	0.872	31.440	27.118	4.847
(c)	0.99	0.01	0.974	0.586	2.07	25.82	6.365	174.991	199.466	20.744
(d)	0.99	0.00	0.991	0.420	2.90	9.05	0.697	10.457	21.402	3.880
(e)	1.00	0.00	0.996	0.622	1.89	3.67	0.882	12.449	28.695	2.577
$n = 400$ Extreme value distribution										
(a)	0.02	0.52	0.999	0.004	4.98	1.06	0.002	0.094	0.049	0.028
(b)	1.00	0.00	0.983	0.404	2.98	16.45	0.646	24.272	20.904	3.064
(c)	0.95	0.05	0.955	0.412	2.94	44.99	5.383	162.695	172.606	15.438
(d)	1.00	0.00	0.995	0.446	2.77	4.72	0.809	11.271	26.570	3.085
(e)	1.00	0.00	0.999	0.546	2.27	0.34	0.870	11.981	28.388	1.629
$n = 200$ Normal distribution										
(a)	0.11	0.60	0.999	0.022	4.89	1.41	0.004	0.210	0.106	0.057
(b)	0.99	0.01	0.989	0.522	2.39	10.82	0.774	29.227	25.184	4.490
(c)	0.99	0.01	0.974	0.608	1.96	26.28	6.561	176.632	208.530	26.358
(d)	0.97	0.03	0.989	0.388	3.06	10.49	0.677	10.150	20.642	3.224
(2)	1.00	0.00	0.997	0.602	1.99	2.68	0.892	12.511	28.637	1.633
$n = 400$ Normal distribution										
(a)	0.00	0.50	0.999	0.00	5.00	0.96	0.002	0.080	0.040	0.019
(b)	1.00	0.00	0.987	0.344	3.28	14.32	0.520	19.344	16.303	2.837
(c)	0.88	0.12	0.954	0.388	3.06	46.05	5.535	164.404	177.647	12.326
(d)	0.98	0.02	0.993	0.384	3.08	7.38	0.771	10.787	24.937	3.429
(e)	1.00	0.00	0.999	0.526	2.37	0.40	0.872	12.017	27.985	1.632

Note: UF is the proportion of underfitted models; OF is the proportion of overfitted models; CZ is the percentage of correctly estimated zero coefficients; IZ is the percentage of incorrectly estimated nonzero coefficients; TP is the true positive fraction; FP is the false positive fraction; REE is the relative estimation error; MSE is the mean squared error; MME is the median of model errors and MAD is its median absolute deviation.

Table 2: Variable selection and model-fitting performance when $d = 3$.

Gene	Coefficients	Standard Error	Lower	Upper
6141	-2.396	1.258	-4.861	0.070
7265	-3.549	1.542	-6.572	-0.526
7271	-3.390	1.656	-6.635	-0.145
10127	-1.420	1.640	-4.634	1.794
10366	3.459	1.335	0.842	6.076
14683	2.095	1.622	-1.083	5.274
16107	0.666	1.749	-2.761	4.093
16817	-0.870	1.613	-4.032	2.292
16870	-0.932	1.134	-3.156	1.291
17318	2.078	0.518	1.062	3.093
19866	3.536	1.714	0.176	6.896
19867	-4.653	1.397	-7.392	-1.915
20401	4.381	0.872	2.673	6.090
21658	2.085	1.351	-0.563	4.734

Table 3: Estimation results without adjustment of confounders. Standard errors and lower and upper bounds of the 95% confidence intervals are obtained from the bootstrap.

Gene	Coefficients	Standard Error	Lower	Upper
4447	-0.314	0.174	-0.655	0.028
5303	0.176	0.101	-0.021	0.373
6150	0.425	0.195	0.043	0.806
6177	0.189	0.199	-0.201	0.578
6226	-0.271	0.181	-0.626	0.085
9602	0.430	0.087	0.260	0.599
10189	0.181	0.125	-0.065	0.427
14295	0.211	0.169	-0.121	0.544
14716	0.155	0.071	0.016	0.294
15102	-0.187	0.151	-0.482	0.108
16454	-0.255	0.115	-0.481	-0.030
16929	0.201	0.150	-0.092	0.494
17465	0.158	0.074	0.013	0.303
19128	-0.263	0.054	-0.369	-0.157

Table 4: Estimation results adjusted for age, gender, cancer stage, adjuvant chemotherapy treatment and smoking history. Standard errors and lower and upper bounds of the 95% confidence intervals are obtained from the bootstrap.