A New Class of Dantzig Selectors for Censored Linear Regression Models

Yi Li, Lee Dicker and Sihai Zhao *

Abstract

The Dantzig variable selector has recently emerged as a powerful tool for fitting regularized regression models. A key advantage is that it does not pertain to a particular likelihood or objective function, as opposed to the existing penalized likelihood methods, and hence has the potential for wide applicability. To our knowledge, limited work has been done for the Dantzig selector when the outcome is subject to censoring. This paper proposes a new class of Dantzig variable selectors for linear regression models for right-censored outcomes. We first establish the finite sample error bound for the estimator and show the proposed selector is nearly optimal in the ℓ_2 sense. To improve model selection performance, we further propose an adaptive Dantzig variable selector and discuss its large sample properties, namely, consistency in model selection and asymptotic normality of the estimator. The practical utility of the proposed adaptive Dantzig selectors is verified via extensive simulations. We apply the proposed methods to a myeloma clinical trial and identify important predictive genes for patients' survival.

KEY WORDS: Adaptive Dantzig variable selector; Censored linear regression; Buckley-James imputation; Model selection consistency; Asymptotic normality.

RUNNING TITLE: Dantzig Selector for Censored Regression

*Yi Li is Associate Professor, Department of Biostatistics, Harvard School of Public Health and Dana-Farber Cancer Institute, Boston, MA 02115, Lee Dicker is a Ph.D candidate, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, and Sihai Zhao is a Ph.D candidate, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115. We thank the editor, the AE and two referees for their insightful comments, which help substantially improve this manuscript. This work was partially supported by U. S. National Cancer Institute grants R01 CA95747.

1 Introduction

Technical advances in biomedicine have produced an abundance of high-throughput data. This has resulted in major statistical challenges and helped bring great attention to the variable selection and estimation problem, where the goal is to discover relevant variables among many potential candidates and obtain high prediction accuracy. For example, variable selection is essential when performing gene expression profiling for cancer patients in order to better understand cancer genomics and design effective therapies (Anderson et al., 2005; Pawitan et al., 2005; Potti et al., 2007).

Penalized likelihood methods, represented by the LASSO, have been extensively studied as a means for simultaneous estimation and variable selection (Tibshirani, 1996). It is known that the LASSO estimator can discover the correct sparse representation of the model (Donoho and Huo, 2002); however, the LASSO estimator is in general biased (Zou, 2006), especially when the true coefficients are relatively large. Several remedies, including the smoothly clipped absolute deviation (SCAD) (Fan and Li 2001) and the adaptive LASSO (ALASSO) (Zou 2006) have been proposed to discover the sparsity of the true models, while producing consistent estimates for nonzero regression coefficients. Though these methods do differ to a great extent, they are all cast in the framework of penalized likelihoods or penalized objective functions.

More recently a new variable selector, namely the Dantzig selector (Candès and Tao, 2007), has emerged to enrich the class of regularization techniques. Though under some general conditions the LASSO and Dantzig may produce the same solution path (James et al., 2008) and are asymptotically equivalent (Bickel et al., 2009), they differ fundamentally in that the Dantzig selector stems directly from an estimating equation, whereas the LASSO requires the specification of a likelihood or an objective function. Moreover, as the Dantzig selection is a linear programming problem, the computational burden is manageable.

To our knowledge, most work on the Dantzig selector has been performed with fully observed outcome variables. In many clinical studies, the outcome variable, e.g. the CD4 counts in an AIDS trial or patients' survival times, may not be fully observed. In a myeloma clinical trial that motivated this research, the goal was to identify predictive genes for patients' survival times,

which were subject to right censoring. Given the infinite-dimensional nuisance parameters in the likelihood function for censored linear regressions, the estimating equation-based Dantzig selector may be a natural choice.

While the vast majority of work in variable selection for censored outcome data has focused on the Cox proportional hazards model (e.g. Tibshirani, 1997; Li and Luan, 2003; Li and Gui, 2004; Gui and Li, 2005a,b), a linear regression model offers a viable alternative, as it directly links the outcome to the covariates. Hence, its regression coefficients have an easier interpretation than those of the Cox model, especially when the response does not pertain to a survival time. Some recent work in censored linear regression can be found in Engler and Li (2009), Cai et al. (2009), Wang et al. (2008), and Ma et al. (2006). To our knowledge, however, results concerning the Dantzig selector suitable for censored linear models are lacking from the literature. Antoniadis et al. (2009) and Martinussen and Sheike (2009) consider using the Dantzig selector to fit survival data and compute the finite sample error bounds of their estimators, but they only deal with the Cox proportional proportional hazards model and the semiparametric additive risk model. Furthermore, large sample properties, e.g. model selection consistency and asymptotic normality, are unavailable.

This paper proposes a new class of Dantzig variable selectors for linear regression models when the response variable is subject to censoring. The proposed method has several attractive features that make it a competing tool for analyzing high-dimensional data with censored outcomes. First, it carries out variable selection and estimation simultaneously, without resorting to maximizing or minimizing a given likelihood function, which is important for some semiparametric models whose likehood functions are often difficult to specify. Second, finite-sample bounds on the error of the estimator can be derived when $p > n$, which are nearly optimal in the ℓ_2 sense. Third, we show that a refined version of the Dantzig selector – the adaptive Dantzig selector – can achieve appealing large sample properties when the tuning parameters follow appropriate rates, providing further support for the theoretical basis of the proposed procedures. Finally, the complex regularization problem has been reduced to a linear programming problem, resulting in computationally efficient algorithms.

The rest of the paper is structured as follows. Section 2 reviews the censored linear regression

model and the Buckley-James estimation approach. Section 3 shows the Buckley-James approach naturally leads to a Dantzig-type selector when the number of covariates exceeds the sample size. We show that the resulting estimator reaches the near-optimal ℓ_2 non-asymptotic error bound. In Section 4, we propose an adaptive Dantzig selector and derive the large sample properties of the estimators. We discuss implementation and the choice of tuning parameters in finite sample settings in Section 5. We conduct numerical simulations in Section 6 and apply the proposal to a myeloma study in Section 7. We conclude with a discussion in Section 8. All technical proofs are relegated to the Appendix.

2 Censored Linear Regression and Buckley-James Estimation

Consider a censored linear regression model,

$$Y_i = \mathbf{X}_i'\boldsymbol{\beta} + \epsilon_i \tag{1}$$

where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})'$ is the covariate vector for the i-th subject and ϵ_i are iid with an unspecified distribution $F(\cdot)$ and survival function $S(\cdot) = 1 - F(\cdot)$. The mean of ϵ_i, denoted by α, is not necessarily 0. Let $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})'$ denote the true $\boldsymbol{\beta}$ and $A = \{j;\ \beta_{0j} \neq 0\}$ be the true model. Suppose that the response Y_i may be right censored by a competing observation C_i and that only $Y_i^* = Y_i \wedge C_i$ and $\delta_i = I(Y_i^* = Y_i)$ are observed for each subject. We assume that Y_i is independent of C_i conditional on $\mathbf{X}_i$. When the response variable pertains to survival time, with both Y_i and C_i measured on the log scale, the model is called the accelerated failure time (AFT) model (Kalbfleisch and Prentice, 2002).

Denote by $e_i(\boldsymbol{\beta}) = Y_i^* - \boldsymbol{\beta}'\mathbf{X}_i$, and consider

$$\tilde{Y}_i(\boldsymbol{\beta}) = E(Y_i | Y_i^*, \delta_i, \mathbf{X}_i, \boldsymbol{\beta}) = Y_i^* + (1 - \delta_i)\frac{\int_{e_i(\boldsymbol{\beta})}^{\infty} S(s, \boldsymbol{\beta})ds}{S\{e_i(\boldsymbol{\beta}), \boldsymbol{\beta}\}}. \tag{2}$$

Clearly,

$$E\left\{\tilde{Y}_i(\boldsymbol{\beta}) | \mathbf{X}_i, \boldsymbol{\beta}\right\} = \alpha + \mathbf{X}_i'\boldsymbol{\beta}.$$

The Buckley-James estimating equation is

$$\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)\left\{\hat{Y}_i(\boldsymbol{\beta}) - \mathbf{X}_i'\boldsymbol{\beta}\right\} = 0,\ j = 1, \ldots, p, \tag{3}$$

where $\bar{X}_j = \frac{1}{n}\sum_{i=1}^n X_{ij}$ for $j = 1, \ldots, p$ and

$$\hat{Y}_i(\boldsymbol{\beta}) = Y_i^* + (1-\delta_i)\frac{\int_{e_i(\boldsymbol{\beta})}^{\infty} \hat{S}(s,\boldsymbol{\beta})ds}{\hat{S}\{e_i(\boldsymbol{\beta}),\boldsymbol{\beta}\}}. \tag{4}$$

is the empirical version of $\tilde{Y}_i(\boldsymbol{\beta})$. Here, $\hat{S}(\cdot,\boldsymbol{\beta})$ is the one-sample Nelson-Aalen estimator based on $(e_i(\boldsymbol{\beta}), \delta_i)$,

$$\hat{S}(t,\boldsymbol{\beta}) = \exp\left\{-\sum_{i=1}^n \int_{-\infty}^t \frac{dN_i(u,\boldsymbol{\beta})}{\bar{Y}(u,\boldsymbol{\beta})}\right\}, \tag{5}$$

where $N_i(u,\boldsymbol{\beta}) = I\{e_i(\boldsymbol{\beta}) \le u, \delta_i = 1\}$ and $\bar{Y}(u,\boldsymbol{\beta}) = \sum_i I\{e_i(\boldsymbol{\beta}) \ge u\}$. Under mild conditions, Lai and Ying (1991) have shown that the Buckley-James estimator $\hat{\boldsymbol{\beta}}_{BJ}$, which solves (3), is $\sqrt{n}$ consistent.

Note that (3) can be written in a more compact form

$$\mathbf{X}'\mathbf{P}_n\{\hat{\mathbf{Y}}(\boldsymbol{\beta}) - \mathbf{X}\boldsymbol{\beta}\} = \mathbf{0}, \tag{6}$$

where $\mathbf{P}_n = \mathbf{I}_n - \mathbf{1}\mathbf{1}'/n$, $\mathbf{I}_n$ is an $n \times n$ identity matrix, $\mathbf{0}$ is a $p \times 1$ vector with all elements being 0, $\mathbf{1}$ is an $n \times 1$ vector with all elements being 1 and $\hat{\mathbf{Y}}(\boldsymbol{\beta}) = (\hat{Y}_1(\boldsymbol{\beta}), \ldots, \hat{Y}_n(\boldsymbol{\beta}))'$. It is clear that the Buckley-James estimator is a direct generalization of the least squares estimator to censored data; it is most efficient when the error terms ϵ_i follow a normal distribution (Lai and Ying, 1991). Unfortunately, when $p > n$, model (1) becomes nonidentifiable and the Buckley-James procedure fails.

3 Dantzig Selector Derived from B-J Estimation and its Non-asymptotic Error Bound

When $p > n$ but the model is sparse, our proposal is to adopt the following constrained optimization problem:

$$\begin{array}{ll} \text{minimize} & \sum_j |\beta_j| \\ \text{subject to} & |\mathbf{X}_{\cdot j}'\mathbf{P}_n\{\hat{\mathbf{Y}}(\boldsymbol{\beta}) - \mathbf{X}\boldsymbol{\beta}\}| \le \gamma,\ j = 1, \ldots, p, \end{array} \tag{7}$$

where $\gamma > 0$ is a constant and $\mathbf{X}_{\cdot j}$ is the j-th column of matrix $\mathbf{X}$. We refer to (7) as the Dantzig selector for censored linear regression (or just, the Dantzig selector), as it is motivated by Candès and Tao's procedure of the same name. We will denote the procedure by DZ and its solution by

$\hat{\boldsymbol{\beta}}$. The goal of this section is to show under proper assumptions on the information matrix for the underlying model, $\hat{\boldsymbol{\beta}}$ exists with high probability and has good finite-sample properties even for $p = O(n^\kappa)$, $\kappa \geq 1$, and in the presence of censoring.

For convenience, we introduce concise notation for referring to subvectors and matrices. For a subset $H \subset \{1, \ldots, p\}$ and a vector $\boldsymbol{\alpha} \in \mathbf{R}^p$, let $\boldsymbol{\alpha}_H = (\alpha_j)_{j \in H}$ be the $|H| \times 1$ vector whose entries are those of $\boldsymbol{\beta}$ indexed by H, where $|H|$ refers to the cardinality of H. For an $n \times p$ matrix $\mathbf{M}$, $\mathbf{M}_H$ is the $n \times |H|$ matrix whose columns are those of $\mathbf{M}$ that are indexed by H. When $\mathbf{M}$ is a $p \times p$ covariance or information matrix, we slightly abuse notation and let $\mathbf{M}_H$ denote the $|H| \times |H|$ submatrix of $\mathbf{M}$ whose rows and columns are both indexed by H. Define the sign vector corresponding to $\boldsymbol{\alpha}$, $\text{sgn}(\boldsymbol{\alpha})$, by $\text{sgn}(\boldsymbol{\alpha})_j = \text{sgn}(\alpha_j)$ (by definition, $\text{sgn}(0) = 0$). Finally, define the ℓ_r-norms $||\cdot||_r$ by $||\boldsymbol{\alpha}||_r = (\sum_{i=1}^p |\alpha_i|^r)^{1/r}$ for $0 < r < \infty$, $||\boldsymbol{\alpha}||_0 = \#\{j : \alpha_j \neq 0\}$ and $||\boldsymbol{\alpha}||_\infty = \max_{1 \leq j \leq p} |\alpha_j|$.

We compute the ℓ_2 error bound of the Dantzig selector using the following steps. Denote the left hand side of (6) by $U(\boldsymbol{\beta})$. We first establish that with a proper tuning parameter γ and with probability going to 1, the true $\boldsymbol{\beta}_0$ is a feasible solution to the Dantzig selector optimization problem (7). Then immediately $\|\hat{\boldsymbol{\beta}}\|_1 \leq \|\boldsymbol{\beta}_0\|_1$. This, coupled with some assumptions on the decomposition of the information matrix of the underlying model, will lead to an error bound for $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2$.

Proposition 1 *Let $\gamma = \sqrt{n(1+a)\log p}$ for some $a > 0$. Then*

$$P(\|U(\boldsymbol{\beta}_0)\|_\infty \leq \gamma) > 1 - 2p \exp\left(-\frac{\gamma^2/n}{L + 2BK\gamma/n}\right),$$

where B, K, L are positive constants defined in Appendix A.0. Moreover, if $p = O(n^\kappa)$ for $\kappa \geq 1$, then

$$P(\|U(\boldsymbol{\beta}_0)\|_\infty \leq \gamma) > 1 - O(n^{-a\kappa}).$$

This proposition is important as it stipulates that with high probability $||U(\boldsymbol{\beta}_0)||_\infty \leq \gamma$ for a proper γ, implying that even when $p > n$ (7) will have a solution on the intersection of $\{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_1 \leq \|\boldsymbol{\beta}_0\|_1\}$ and the closure of $\{\boldsymbol{\beta} : \|U(\boldsymbol{\beta})\|_\infty \leq \gamma\}$. Note that this intersection is nonempty as it contains at least $\boldsymbol{\beta}_0$.

To further characterize the bound of $\hat{\boldsymbol{\beta}}$, first note that $U(\boldsymbol{\beta})$ is approximately equal to $n\tilde{\boldsymbol{\Omega}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)$, where $\tilde{\boldsymbol{\Omega}}$ is defined in (23) and can be viewed as an information matrix (Lai and Ying, 1991). Let $\tilde{\boldsymbol{\Omega}}^{1/2}$ be an $n \times p$ decomposition matrix of the semipositive-definite $\tilde{\boldsymbol{\Omega}}$ such that $\{\tilde{\boldsymbol{\Omega}}^{1/2}\}'\tilde{\boldsymbol{\Omega}}^{1/2} = \tilde{\boldsymbol{\Omega}}$. Now define the constants δ_{V_1} and θ_{V_1,V_2} such that for all disjoint subsets H and $\tilde{H}$ of $\{1, \ldots, p\}$, with respective sizes V_1 and V_2, and all vectors $\mathbf{c}$ and $\tilde{\mathbf{c}}$ with respective lengths V_1 and V_2, δ_{V_1} is the largest quantity such that $\delta_{V_1}\|\mathbf{c}\|_2^2 \leq \|\tilde{\boldsymbol{\Omega}}_H^{1/2}\mathbf{c}\|_2^2$, and θ_{V_1,V_2} is the smallest quantity such that $|(\tilde{\boldsymbol{\Omega}}_H^{1/2}\mathbf{c})'(\tilde{\boldsymbol{\Omega}}_{\tilde{H}}^{1/2}\tilde{\mathbf{c}})| \leq \theta_{V_1,V_2}\|\mathbf{c}\|_2\|\tilde{\mathbf{c}}\|_2$. These are related to the restricted isometry and restricted orthogonality constants of Candès and Tao (2007).

Proposition 2 *Let $\gamma = \sqrt{n(1+a)\log p}$ for some $a > 0$. Define $V = ||\boldsymbol{\beta}_0||_0$ to be the size of the true model. If the constants δ_V and $\theta_{V,2V}$ for $\tilde{\boldsymbol{\Omega}}^{1/2}$ obey $\theta_{V,2V} < \delta_{2V}$, then*

$$P\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 \leq \frac{36V\gamma^2/n^2}{(\delta_{2V} - \theta_{V,2V})^2}\right) > 1 - 2p\exp\left(-\frac{\gamma^2/n}{L + 2BK\gamma/n}\right),$$

where B, K, L are the same constants defined in Proposition 1. Moreover, if $p = O(n^\kappa)$ for $\kappa \geq 1$, then

$$P\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 \leq \frac{36V\gamma^2/n^2}{(\delta_{2V} - \theta_{V,2V})^2}\right) > 1 - O(n^{-a\kappa}),$$

We note that this error bound is of the same order as the bounds derived by Candès and Tao (2007) for the uncensored linear models and by Antoniadis et al. (2009) for the Cox models. Even if we knew the true subset of covariates of size V, it would be the case that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2$ grew at the rate of V/n (Lai and Ying, 1993). Hence, the rate guaranteed in Proposition 2 reaches the optimal non-asymptotic bound (Candès and Tao, 2007), meaning we only pay a small price (up to a factor of $\log p$) for not knowing the true model.

The condition of $\theta_{V,2V} < \delta_{2V}$ is similar to the uniform uncertainty principle required by Candès and Tao (2007), though it is applied to the information matrix rather than the design matrix. In particular, δ_V puts a lower bound on the minimum singular values of the submatrices of $\tilde{\boldsymbol{\Omega}}$ with less than V columns, indicative of how precisely the model can be estimated (Silvey, 1968). Hence, the $\theta_{V,2V} < \delta_V$ condition can be interpreted as requiring our data to allow a minimum level of precision for our model fitting, though verifying this condition is much involved in practice (Antoniadis et al., 2009; Cai and Lv, 2007).

4 Variable selection and the Adaptive Dantzig Selector for Censored Linear Regression

We note that the error bound established in Proposition 2 refers to the mean-squared error of the point estimate, which does not directly translate into optimal variable selection. As we report in our simulation studies, we found that the Dantzig selector may effectively reduce the size of the model. However, given the relatively large false-positive rate (namely, the proportion of true zero coefficients estimated to be nonzero), it appears that the Dantzig selector tends to select models that are too large.

Indeed, following the asymptotic equivalence of the Dantzig selector and LASSO (Bickel et al., 2009), Dicker and Lin (2009) have confirmed that the Dantzig selector, like LASSO (Zhao and Yu, 2006), may not consistently select the true model for fully observed data. Furthermore, Dicker and Lin have shown that the Dantzig selector is not asymptotically normal. To address these issues, we consider the adaptive Dantzig selector – a modified Dantzig selector for censored linear regression which, given the existence of a "reliable" initial estimate, is consistent for model selection and is asymptotically normal in large samples.

Let $\hat{\boldsymbol{\beta}}^{(0)}$ be some initial estimator for $\boldsymbol{\beta}_0$. The adaptive Dantzig selector is the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \sum_j w_j|\beta_j| \\ \text{subject to} & |\mathbf{X}'_{\cdot j}\mathbf{P}_n\{\hat{\mathbf{Y}}(\hat{\boldsymbol{\beta}}^{(0)}) - \mathbf{X}\boldsymbol{\beta}\}| \leq \gamma w_j,\ j = 1, ..., p. \end{array} \tag{8}$$

Here, $\gamma > 0$ is the tuning constant and w_j are data driven weights that should be chosen to vary inversely with the magnitude of β_{0j}. If we take $w_j = |\hat{\beta}_j^{(0)}|^{-\eta}$ for some $\eta > 0$, then (8) requires us to nearly solve the j-th score equation (where the surrogate vector $\hat{\mathbf{Y}}(\hat{\boldsymbol{\beta}}^{(0)})$ is treated as a fully observed outcome vector) when $|\hat{\beta}_j^{(0)}|$ is large and heavily penalizes non-zero estimates of β_{0j} when $|\hat{\beta}_j^{(0)}|$ is small.

Recall that $A = \{j : \beta_{0j} \neq 0\}$. A variable selector $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0$ in a generic model

$$Y_i \sim \sum_{j=1}^{p} X_{ij}\beta_j$$

is considered to have reasonable large sample behavior if (i) it can identify the right subset model with a probability tending to 1, i.e. $P(\{j : \hat{\beta}_j \neq 0\} = A) \to 1$ as the sample size $n \to \infty$,

and (ii) $\sqrt{n}(\hat{\boldsymbol{\beta}}_A - \boldsymbol{\beta}_A) \to N(0, \Sigma^*)$ where Σ^* is some $|A| \times |A|$ covariance matrix. Property (i) is often considered to be the consistency property, while property (ii) involves the efficiency of the estimator. If properties (i) and (ii) hold, and Σ^* is optimal (by some criterion), the variable selection procedure is said to have the oracle property (Fan and Li, 2001). We show that if $\hat{\boldsymbol{\beta}}^{(0)}$ is $\sqrt{n}$-consistent for $\boldsymbol{\beta}_0$, then the adaptive Dantzig selector may have the oracle property.

The oracle properties of the adaptive Dantzig selector for the appropriate tuning parameter γ have been established when the response vector $\mathbf{Y}$ is fully observed. Until now, however, it has been unclear whether these properties hold when the response $\mathbf{Y}$ is subject to censoring. A fundamental theoretical difficulty is that $\hat{Y}_i(\hat{\boldsymbol{\beta}}^{(0)})$ is only a surrogate for the unobserved outcome Y_i, preventing the direct applications of the existing Dantzig selector results obtained for fully observed outcomes.

In the ensuing theoretical development, we first quantify the "distance" between the surrogate and the true outcomes, and show that the average difference between the imputed $\hat{Y}_i(\hat{\boldsymbol{\beta}}^{(0)})$ and true Y_i is bounded by a random variable with order of $n^{-1/2}$. This turns out to be essential for establishing the consistency and oracle properties of the Dantzig selector estimator. Given this random bound, we then show that the existing Dantzig selector results for the non-censored case can be extended to the censored case, leading to the desirable oracle property. Note that for all of the asymptotic results in this section, we assume that p is fixed.

4.1 Quantify the "Distance" Between the Imputed and "True" Responses.

We first use Lemma 1 to bound the difference between the surrogate and the true outcomes.

Proposition 3 *Under the regularity conditions listed in Lemma 1,* $\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\{\hat{Y}_i(\hat{\boldsymbol{\beta}}^{(0)}) - Y_i\} = O_p(n^{-1/2})$ *if* $\hat{\boldsymbol{\beta}}^{(0)} = \boldsymbol{\beta}_0 + O_p(n^{-1/2})$.

Several points are worth noting. First, the result can be succinctly rephrased as $\mathbf{X}'(\hat{\mathbf{Y}} - \mathbf{Y}) = O_p(n^{1/2})$, where $\hat{\mathbf{Y}} = \hat{\mathbf{Y}}(\hat{\boldsymbol{\beta}}^{(0)})$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)'$. Second, the result further implies $(\mathbf{P}_n\mathbf{X})'(\hat{\mathbf{Y}} - \mathbf{Y}) = O_p(n^{1/2})$, where $\mathbf{X}$ is replaced by its centralized version; this will facilitate the proof of consistency of model selection. Finally, as the validity of Proposition 3 requires $\hat{\boldsymbol{\beta}}^{(0)}$ to be $\sqrt{n}$ consistent, taking the $\hat{\boldsymbol{\beta}}^{(0)}$ equal to the Buckley-James estimate, which is $\sqrt{n}$ consistent, will suffice.

4.2 Consistency and Oracle Properties

To ease notation in what follows, we use $\hat{\mathbf{Y}}$ to denote $\hat{\mathbf{Y}}(\hat{\boldsymbol{\beta}}^{(0)})$. Observe that the adaptive Dantzig selector for data with a censored response, (8), can be rewritten compactly as

$$\begin{array}{ll} \text{minimize} & ||\mathbf{W}\boldsymbol{\beta}||_1 \\ \text{subject to} & ||\mathbf{Z}'(\hat{\mathbf{Y}} - \mathbf{Z}\mathbf{W}\boldsymbol{\beta})||_\infty \leq \gamma, \end{array} \tag{9}$$

where $\mathbf{W} = \text{diag}(w_1, ..., w_p)$ and $\mathbf{Z} = \mathbf{P}_n\mathbf{X}\mathbf{W}^{-1}$. The optimization problem (9) is a linear programming problem, which means that there is a corresponding dual linear programming problem. Specifically, the solution to (9), denoted by $\hat{\boldsymbol{\beta}}$, can be characterized in terms of primal and dual feasibility and complementary slackness conditions as shown below.

Proposition 4 *If there is $\hat{\boldsymbol{\mu}} \in \mathbf{R}^p$ such that,*

$$||\mathbf{Z}'(\hat{\mathbf{Y}} - \mathbf{Z}\mathbf{W}\hat{\boldsymbol{\beta}})||_\infty \leq \gamma, \tag{10}$$

$$||\mathbf{Z}'\mathbf{Z}\hat{\boldsymbol{\mu}}||_\infty \leq 1, \tag{11}$$

$$\hat{\boldsymbol{\mu}}'\mathbf{Z}'\mathbf{Z}\mathbf{W}\hat{\boldsymbol{\beta}} = ||\mathbf{W}\hat{\boldsymbol{\beta}}||_1, \tag{12}$$

$$\hat{\boldsymbol{\mu}}'\mathbf{Z}'(\hat{\mathbf{Y}} - \mathbf{Z}\mathbf{W}\hat{\boldsymbol{\beta}}) = \gamma||\hat{\boldsymbol{\mu}}||_1, \tag{13}$$

then the vector $\hat{\boldsymbol{\beta}} \in \mathbf{R}^p$ solves (9).

The parameter $\boldsymbol{\mu}$ in Proposition 4 is the dual variable and may be viewed as a Lagrangian multiplier. Inequalities (10) and (11) correspond to primal and dual feasibility respectively, while (12) and (13) concerns with complementary slackness. By inspecting (10)- (13), we prove that the adaptive Dantzig selector is selection consistent, provided γ and $(w_1, ..., w_p)$ follow an appropriate rate.

Proposition 5 *Suppose that $\boldsymbol{\beta}_0$ is the true parameter value and $A = \{j;\ \beta_{0j} \neq 0\}$. Also assume that $\frac{1}{n}\mathbf{X}'\mathbf{P}_n\mathbf{X}$ converges in probability to some positive definite matrix. Suppose further that*

$$\frac{\gamma}{\sqrt{n}}w_j \xrightarrow{P} \infty \text{ if } j \notin A \text{ and } \gamma w_j = O_P(\sqrt{n}) \text{ if } j \in A.$$

Denote by $\bar{A}$ the complement of A in $\{1, \ldots, p\}$. Then, with probability tending to 1, a solution to the adaptive Dantzig selector, $\hat{\boldsymbol{\beta}}$, and the corresponding $\hat{\boldsymbol{\mu}}$ from Lemma 2 are given by

$$\hat{\boldsymbol{\mu}}_A = (\mathbf{Z}_A'\mathbf{Z}_A)^{-1}\text{sgn}(\boldsymbol{\beta}_0)_A \tag{14}$$

$$\hat{\boldsymbol{\mu}}_{\bar{A}} = 0 \tag{15}$$

and

$$\begin{aligned}\hat{\boldsymbol{\beta}}_A &= \mathbf{W}_A^{-1}\left\{(\mathbf{Z}_A'\mathbf{Z}_A)^{-1}\mathbf{Z}_A'\hat{\mathbf{Y}} - \gamma(\mathbf{Z}_A'\mathbf{Z}_A)^{-1}\mathrm{sgn}(\hat{\boldsymbol{\mu}})_A\right\} \\ &= (\mathbf{X}_A'\mathbf{P}_n\mathbf{X}_A)^{-1}\mathbf{X}_A'\mathbf{P}_n\hat{\mathbf{Y}} - \gamma(\mathbf{X}_A'\mathbf{P}_n\mathbf{X}_A)^{-1}\mathbf{W}_A\mathrm{sgn}(\hat{\boldsymbol{\mu}})_A \qquad (16) \\ \hat{\boldsymbol{\beta}}_{\bar{A}} &= 0. \qquad (17)\end{aligned}$$

Corollary 1 *(consistency of model selection) Suppose that the conditions of Proposition 5 hold and let $\hat{\boldsymbol{\beta}}$ be any sequence of solutions to (9). Then $P(\{j;\ \hat{\beta}_j \neq 0\} = A) \to 1$.*

We make a few remarks about Proposition 5 and Corollary 1. First, to ensure that the conditions in Proposition 5 hold, one selects data-driven weights w_j and an appropriate γ. Examples of weights and γ such that these conditions hold include $w_j = |\hat{\boldsymbol{\beta}}^{(0)}|^{-\eta}$, where $\hat{\boldsymbol{\beta}}^{(0)}$ is $\sqrt{n}$-consistent for $\boldsymbol{\beta}_0$ and $\eta > 0$, and γ such that $n^{-1/2}\gamma = O(1)$ and $n^{(\eta-1)/2}\gamma \to \infty$. Also note that though Proposition 5 makes no uniqueness claims about solutions to (9), it can be shown that in "most" cases (9) has a unique solution (Dicker and Lin, 2009). Furthermore, Corollary 1 states that regardless of whether or not there is a unique solution, the adaptive Dantzig selector is consistent for model selection.

The estimator defined in (16) and (17) solves (9) in probability. This expression may be leveraged to obtain the large sample distribution of $\sqrt{n}$-standardized adaptive Dantzig selector estimates estimates. However, since $\hat{\boldsymbol{\beta}}^{(0)}$ is not consistent for model selection, the solution to (9), $\hat{\boldsymbol{\beta}}$, which we refer to in what follows as the *one-iteration estimator*, may not achieve optimal efficiency. To remedy this, we propose two modified estimators which do possess the oracle property.

To proceed, let $T = \{j;\ \hat{\beta}_j \neq 0\}$ be the index set of non-zero estimated coefficients from the one-iteration estimator $\hat{\boldsymbol{\beta}}$, and $\bar{T}$ be the complement of T in $\{1, \ldots, p\}$. Define the *intermediate estimator* $\hat{\boldsymbol{\beta}}^{(0,T)}$ so that $\hat{\boldsymbol{\beta}}_{\bar{T}}^{(0,T)} = 0$ and $\hat{\boldsymbol{\beta}}_T^{(0,T)}$ is the Buckley-James estimate obtained by solving (6) with $\mathbf{X}$ replaced by $\mathbf{X}_T$. That is, we perform a Buckley-James estimation based on the subset of covariates selected by the one-iteration estimator. Now, let $\hat{\mathbf{Y}}^{(1)} = \hat{\mathbf{Y}}(\hat{\boldsymbol{\beta}}^{(0,T)})$ be the imputed value of $\mathbf{Y}$, defined in (4) and based on $\hat{\boldsymbol{\beta}}^{(0,T)}$. The *two-iteration estimator* $\hat{\boldsymbol{\beta}}^*$ is then defined to be the solution to (9) with $\hat{\mathbf{Y}}$ replaced by $\hat{\mathbf{Y}}^{(1)}$ and $\mathbf{X}$ by $\mathbf{X}_T$. The rationale is that since T is consistent for A, the intermediate estimator $\hat{\boldsymbol{\beta}}^{(0,T)}$ and the two-iteration estimator $\hat{\boldsymbol{\beta}}^*$ will be

model selection consistent and efficient. As summarized in the following proposition and corollary, $\hat{\boldsymbol{\beta}}^{(0,T)}$ and $\hat{\boldsymbol{\beta}}^{*}$ achieve the oracle property.

Proposition 6 *(oracle property) Assume that the conditions of Proposition 5 hold. Let* $T = \{j;\ \hat{\beta}_j \neq 0\}$*, where* $\hat{\boldsymbol{\beta}}$ *is the one-iteration estimator for* $\boldsymbol{\beta}_0$ *and let* $\boldsymbol{\beta}_{0,A}$ *be the non-zero subvector of* $\boldsymbol{\beta}_0$*. Define* $\hat{\boldsymbol{\beta}}^{(0,A)}$ *so that* $\hat{\boldsymbol{\beta}}_{\bar{A}}^{(0,A)} = 0$ *and* $\hat{\boldsymbol{\beta}}_{A}^{(0,A)}$ *is the Buckley-James estimate obtained by solving (6) with* $\mathbf{X}$ *replaced by* $\mathbf{X}_A$*. Then the intermediate estimator* $\hat{\boldsymbol{\beta}}^{(0,T)}$ *satisfies*

$$P\left(\hat{\boldsymbol{\beta}}^{(0,T)} = \hat{\boldsymbol{\beta}}^{(0,A)}\right) \to 1$$

and

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_{A}^{(0,T)} - \boldsymbol{\beta}_{0,A}\right) \to N(0, \boldsymbol{\Sigma}_A)$$

weakly, where $\boldsymbol{\Sigma}_A = \boldsymbol{\Omega}_A^{-1}\boldsymbol{\Lambda}_A\boldsymbol{\Omega}_A^{-1}$ *and* $\boldsymbol{\Omega}_A$ *and* $\boldsymbol{\Lambda}_A$ *are the submatrices of* $\boldsymbol{\Omega}$ *and* $\boldsymbol{\Lambda}$ *[defined in (32) and (33)] corresponding to index set* A*.*

Corollary 2 *(oracle property) Let* $\hat{\boldsymbol{\beta}}^{*}$ *be the two-iteration estimator. Assume that the conditions of Proposition 6 holds and, additionally, that* $\gamma w_j = o_P(\sqrt{n})$ *for* $j \in T$*. Then*

$$P(\{j;\ \hat{\beta}_j^{*} \neq 0\} = A) \to 1$$

and

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{A}^{*} - \boldsymbol{\beta}_{0,A}) \to N(0, \boldsymbol{\Sigma}_A)$$

weakly, where $\boldsymbol{\Sigma}_A$ *is defined in Proposition 6.*

Note that $\boldsymbol{\Sigma}$ in Proposition 6 and Corollary 2 is the asymptotic variance of the Buckley-James estimator given the true subset of covariates; see Lai and Ying (1991).

4.3 A Coherent Two-stage Procedure When $p > n$

Like other adaptive variable selectors, the adaptive Dantzig selector theoretically requires a $\sqrt{n}$-consistent initial estimator $\hat{\boldsymbol{\beta}}^{(0)}$. This is straightforward in the fixed p, large n regime, where the Buckley-James estimator or other rank based estimators can be a natural choice for $\hat{\boldsymbol{\beta}}^{(0)}$.

When $p > n$, a consistent initial estimator may not be well defined, but we propose the following two-stage estimating procedure. The first stage uses the Dantzig selector defined in (7) to screen out unimportant covariates and reduce the number of parameters. At the second stage, we use the covariates selected from the first stage to implement the the one-iteration, intermediate, or two-iteration adaptive Dantzig selector procedures described above. The corresponding selectors will be denoted by DZ-ADZ-1, DZ-ADZ-INT, DZ-ADZ-2, respectively. We have found these two-stage procedures to work quite well. Section 6 contains more extensive simulations to evaluate the performance of these estimators, along with their competitors.

5 Computational Considerations

5.1 Tuning Parameter Selection For Finite Sample Cases

In practice, it is very important to select an appropriate tuning parameter γ in order to obtain good performance. For regularized linear regression without censoring, Tibshirani (1996) and Fan and Li (2001) proposed the following generalized cross-validation (GCV) statistic:

$$GCV^*(\gamma) = \frac{AR(\gamma)}{\{1 - d(\gamma)/n\}^2}$$

where $AR(\gamma)$ is the average residual sum of squares $\frac{1}{n}||\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\gamma)||_2^2$, $\hat{\boldsymbol{\beta}}(\gamma)$ is the estimate of $\boldsymbol{\beta}$ under γ and $d(\gamma)$ is the effective number of parameters, i.e. the number of non-zero components of the LASSO estimates (Zou et al. (2007)). When the data are censored, we adopt an inverse reweighting scheme to account for censoring. Assume the potential censoring C_i are iid and have a common survival function G_i, which is a reasonable assumption for clinical trials where most censoring is due to administrative censoring. As suggested by Johnson et al. (2008), we approximate the unobserved $AR(\gamma)$ by

$$\widehat{AR}(\gamma) = \frac{\sum_{i=1}^n \delta_i \{Y_i^* - \hat{\alpha}^{(0)} - \mathbf{X}_i'\hat{\beta}(\gamma)\}^2 / \hat{G}(Y_i^*)}{\sum_{i=1}^n \delta_i / \hat{G}(Y_i^*)}$$

where $\hat{G}(\cdot)$ is the Kaplan-Meier estimator for $G(\cdot)$, and $\hat{\alpha}^{(0)} = \frac{1}{n}\sum_{i=1}^n \{Y_i(\hat{\boldsymbol{\beta}}^{0)}) - \mathbf{X}_i'\hat{\boldsymbol{\beta}}^{(0)}\}$. Conditional on $(Y_i, C_i, \mathbf{X}_i)$, the expected value of $\delta_i/G(Y_i^*)$ is one, and hence, the expected values of the numerator and the denominator of $\widehat{AR}(\gamma)$ are equal to the expected value of $\sum_{i=1}^n \{Y_i - \hat{\alpha}^{(0)} - \mathbf{X}_i'\hat{\boldsymbol{\beta}}(\gamma)\}^2$ and n, respectively. Elementary probability implies that $\widehat{AR}(\gamma)$

and $AR(\gamma)$ have the same limit, justifying the use of the inverse reweighting scheme. To obtain an estimate of the effective number of parameters for the estimator from (9), we follow Zou et al. (2007). The expressions (16)-(17) suggests that $\hat{d}(\gamma) = \text{trace}\{\mathbf{X}_T(\mathbf{X}_T'\mathbf{P}_n\mathbf{X}_T)^{-1}\mathbf{X}_T'\mathbf{P}_n\} = ||T||_0$, where $T = \{j; \hat{\beta}_j \neq 0\}$, is a consistent estimator for $d(\gamma)$. In the ensuing data analysis and simulation studies, we propose to select the γ that yields the smallest GCV, defined as

$$GCV(\gamma) = \frac{\widehat{AR}(\gamma)}{\{1 - \hat{d}(\gamma)/n\}^2}. \tag{18}$$

Similar GCV schemes have been proposed by Nan et al. (2006), Wang et al. (2008), and Johnson et al. (2008) in various contexts.

5.2 Implementation

The proposed two-stage procedures for censored linear regression can be easily implemented. The first stage iterates between imputing the outcome vector $\hat{\mathbf{Y}}(\boldsymbol{\beta})$ and solving the optimization problem (7) via the linear programming algorithm of James and Radchenko (2009). In our numerical experiments, convergence is often achieved within a few iterations. Using the covariates selected by the first stage, the second stage imputes the $\hat{\mathbf{Y}}(\boldsymbol{\beta})$ and uses the Buckely-James estimates as the weights w_j for the adaptive Dantzig selector. For each γ, it can again be programmed using linear programming (see James and Radchenko, 2009), or using the DASSO algorithm (see, e.g. James et al. 2008) after replacing the original design matrix $\mathbf{X}$ with $\mathbf{X}\mathbf{W}^{-1}$.

6 Simulation Studies

6.1 Simulation Set-up

We examine the finite sample performance of the proposed methods through simulation studies. For $i = 1, \ldots, n$ we generate the true response Y_i (after the exponential transform) from an exponential distribution with hazard rate $\exp(-\boldsymbol{\beta}_0'\mathbf{X}_i)$, i.e., $Y_i = \boldsymbol{\beta}_0'\mathbf{X}_i + e_i$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})'$ is generated from a multivariate normal with mean zero and covariance matrix $\Sigma = (\sigma_{jj'})_{p\times p} = (\rho^{|j-j'|})$, and e_i follows the standard extreme value distribution. This model falls into both the censored linear regression and Cox model families. We consider $p = 2n$, mimicking our data example in Section 7, and set all components of $\boldsymbol{\beta}_0$ to zero except for the first V. To model different levels of sparsity, we consider $V = 3$ or 5. To model weak and moderate

associations between the predictors and the response, we set each of the V non-zero components to $\beta_{0j} = 1, 1.5$, or 3 for $j = 1, \ldots, S$. Finally, we let ρ equal 0, 0.5, or 0.9, corresponding to zero, moderate, and strong collinearity among the predictors. The censoring variable C_i (after exponential transform) is generated from a uniform$[0, \xi]$, where ξ is chosen to achieve about 50% of censoring.

To test the robustness of our method when the working model, ie. the censored linear regression model, is misspecified, we use the same simulation settings except that we generate $n = 50$ observations from a Cox proportional hazards model with a piecewise constant hazard function $\lambda_i(t) = 0.3\{\sum_{k=1}^{6} I(t \leq 0.4 \times k) - \sum_{k=1}^{4} I(t \leq 2.4 + 0.4 \times k)\} \exp(\boldsymbol{\beta}_0' \mathbf{X}_i)$.

6.2 Competing Methods and Measures of Performance

For each scenario, the following proposed estimation procedures are evaluated based on 500 simulated datasets with sample size $n = 50$ or 100: the proposed one-stage Danzig selector DZ (defined in (7)), the proposed two-stage procedures for the censored linear regression model, namely DZ-ADZ-1, DZ-ADZ-INT, DZ-ADZ-2 (defined in Section 4.3), the Dantzig selector for the Cox model (Antoniadis et al., 2009), the adaptive LASSO for censored linear regression, and the adaptive LASSO for the Cox model (Zhang and Lu, 2007). We feel that the selected competing methods cover the spectrum of existing methods, especially the adaptive methods, reasonably well: the Dantzig Cox selector is spiritually similar to our methods, the adaptive LASSO for the censored linear regression model has been previously implemented (Cai et al., 2009), while the adaptive LASSO for the Cox model seems to be a simple and standard method for the regularized survival analysis.

The penalty parameters used in these regularized estimators are selected based on the generalized cross-validation function (18). The adaptive LASSO for the censored linear regression model is performed by following Datta et al. (2007) and replacing the observed survival times Y_i with inverse probability of censoring-weighted (IPW) times $\delta_i \log(Y_i)/\hat{G}(Y_i)$, where $\hat{G}(Y_i)$ is the Kaplan-Meier estimator of the censoring survival function. Because initial estimates are needed, the adaptive LASSO methods were not originally designed for the situations of $p > n$. Nevertheless, we follow the suggestion of Datta et al. (2007) and use ridge regression to estimate the

initial weights when $p > n$.

We evaluate the accuracy and precision of the estimates based on mean-squared errors (MSE) and prediction errors (PE). For the censored linear regression model-based methods, we calculated the prediction error as $\text{PE} = n^{-1}\sum_{i=1}^{n}\{Y_i^{val} - \hat{\alpha} - \hat{\boldsymbol{\beta}}'\mathbf{X}_i^{val}\}^2$, where $Y_i^{val}, \mathbf{X}_i^{val}$ are the log survival time and the covariate vector, respectively, in a validation sample with the same number of subjects. For the Cox model-based methods, we followed Heller and Simonoff (1992) and calculated $\text{PE} = n^{-1}\sum_{i=1}^{n}\{Y_i^{val} - \hat{Y}_i^{med}\}^2$, where $\hat{Y}_i^{med}$ is the log median survival time predicted by the fitted survival curve for the ith subject. To examine how well the proposed procedures perform with respect to variable selection, we recorded the frequencies of truly zero regression coefficients being incorrectly set to non-zero, leading to the false positive (FP) rates. The false negative (FN) rates were defined analogously.

6.3 Results of Simulations

The simulation results are summarized in Tables 1 through 3, which exhibit several notable patterns.

First, the two-stage adaptive procedures (DZ-ADZ-1, DZ-ADZ-INT, and DZ-ADZ-2) improve the one-stage DZ estimator by greatly reducing the false positive rate and moderately reducing the MSE, at the cost of a slight increase in the false negative rate. For example, Figure 1 plots each component $\hat{\beta}_j$ of $\hat{\boldsymbol{\beta}}$ against its coordinate $j = 1, \ldots, p$ over all 500 simulations when the samples size $n = 100$, the first $V = 5$ non-zero elements of $\boldsymbol{\beta}_0$ are 1.5 and $\rho = 0.5$. The average mean-squared error for the estimates was 2.450, or 21.8% of $\|\boldsymbol{\beta}_0\|^2$, while the average false positive rate is 35.8% and false-negative rate is 0.6%. On the other hand, the mean-squared errors for the estimates obtained by DZ-ADZ-1, DZ-ADZ-INT and DZ-ADZ-2 reduce to 1.232, 1.624 and 1.335, with much reduced false positive rates of 4.5%, 4.5% and 4.3%, respectively, and slightly increased false-negative rates of 0.8%, 0.8% and 0.9%, respectively; see Table 2 for detail. This suggests that the two-stage adaptive procedure may be a more effective variable selector than the one-stage DZ.

Second, among all the examined methods, our proposed two-stage methods perform the best in terms of prediction error. The advantage becomes more obvious with stronger signals, e.g. larger

values of non-zero components of $\boldsymbol{\beta}$. With regard to variable selection, when the collinearity among covariates is low (e.g. $\rho = 0$), nearly all of the methods give very low false negative rates, while our proposed two-stage methods give the lowest false positive rates. When the collinearity is high, a mixed result is present. Our proposed two-stage methods still give the lowest false positive rates, at the cost of higher false negative rates, indicating that our methods select much smaller models.

Finally, when the working model is misspecified as a censored linear regression model while the data are truly generated from a Cox PH model with a piecewise constant hazard, our two-stage methods still behave reasonably. They tend to select small models, while achieving low prediction error in most cases examined.

It is worth noting the computational efficiency and stability of the proposed estimators, compared with that of the competing methods. For example, both the adaptive LASSO and the Dantzig selector for the Cox model require the Cholesky decomposition of the Cox partial likelihood information matrix, and this decomposition is slow and unstable when the number of parameters is large, resulting failures of the methods on a nonnegligible fraction of our simulated datasets. In contrast, our proposed two-stage selectors avoid such a decomposition in computation, and are fast and stable based on our numerical experiences.

7 Example of Myeloma Patients' Survival Prediction

Multiple myeloma is a progressive hematologic disease, characterized by excessive numbers of abnormal plasma cells in the bone marrow and overproduction of intact monoclonal immunoglobulin. Myeloma patients are typically characterized by wide clinical and pathophysiologic heterogeneities, with survival ranging from a few months to more than 10 years. Gene expression profiling of multiple myeloma patients has offered an effective way of understanding myeloma's genetic basis and designing gene therapy. Identifying risk groups with a high predictive power could contribute to personalized medicine.

For this purpose, we 'train' the models on a total of 188 subjects with late stage multiple myeloma recruited in a clinical trial run by Millennium Pharmaceuticals (Mulligan et al., 2005). We refer to this dataset as our training dataset. Here, the main endpoint was overall survival

and the median followup was 15 months. During the study, a total of 119 deaths were observed. Subject RNA expression levels were measured using Affymetrix microarrays prior to receiving the treatment.

Of interest is the detection of predictive genes among those genes with highly variable expressions, namely, those with sample standard deviation-mean ratio (SDM) larger than 0.5, a clinically meaningful cutoff for the gene expression variability (Novaka etal., 2002; Cheung et al., 2003). A total of $p = 400$ genes meet such a criterion, which is roughly twice as the sample size $n = 188$. We use the adaptive Dantzig selector and the adaptive LASSO for the censored linear regression and Cox models to select predictive genes among these p candidate genes. The Cox Dantzig selector returns a final model of 97 genes, the censored linear regression adaptive LASSO selects 152, and the Cox adaptive LASSO selects 73. In contrast, the three adaptive Dantzig selectors for the censored linear regression model (one-iteration, intermediate, and two-iteration) select only 6 genes, which are presented in Table 4.

To validate the results, we use the obtained models to predict the risks of death in an independent validation dataset, consisting of 351 multiple myeloma patients recruited later but with similar clinical characteristics, e.g. stage of cancer and treatment (Shaughnessy et al., 2007), and with microarrays processed on the same platform. A subject is classified to be of high or low risk based on whether the model-based predicted risk exceeds the median value in the training dataset. Figure 2 depicts the comparisons of the Kaplan-Meier curves for the high and low risk groups defined by the competing methods. It is noted that the risk score based on the much fewer genes selected by the adaptive Dantzig selectors performs markedly better than those based on the genes selected by the other methods. For example, the p-values for comparing high and low risk groups are 0.0110 for DZ-ADZ-1, 0.0120 for DZ-ADZ-INT, 0.0036 for DZ-ADZ-2. In contrast, the Cox Dantzig selector, the censored linear regression adaptive LASSO, and the Cox adaptive LASSO selected far more genes, but the associated gene scores only moderately distinguish the high and low risk groups, yielding p-values of 0.2230, 0.1063 and 0.0382, respectively.

8 Discussion

Several issues merit further investigations. First, more research is needed on the evaluation of the variation of the estimator for small or moderate sample size. Proposition 5 gives a nearly closed-form solution of the Dantzig selector estimators (16) and (17), and this may be useful in estimating the covariance of $\hat{\boldsymbol{\beta}}$, along the line of Tibshirani (1997). Conditional on the model selected by $T = \{j : \hat{\beta}_j \neq 0\}$, a reasonable covariance estimator might be

$$\widehat{cov}(\hat{\boldsymbol{\beta}}_T) = (\mathbf{Z}_T'\mathbf{Z}_T)^{-1}\mathbf{Z}_T' var(\hat{\mathbf{Y}})\mathbf{Z}_T(\mathbf{Z}_T'\mathbf{Z}_T)^{-1}.$$

For the components of $\hat{\boldsymbol{\beta}}$ with zero values, the estimated standard errors would be set to zero, coinciding with Tibshirani (1997), Fan and Li (2001) and Zou (2006). However, the form of $var(\hat{\mathbf{Y}})$ is elusive, making it difficult to use in practice. Furthermore, assigning zero variance estimates to covariates with zero coefficients is not satisfactory. An obvious remedy is through the bootstrap as proposed in Huang et al. (2006); however, this lacks theoretical justifications. Future work towards obtaining reliable standard error estimates is warranted. Indeed, for most variable selection and estimation procedures, including penalized likelihood procedures, the question of standard errors remains open.

Second, our results have shown that methods based on Cox and censored linear regression models differ in their predictive ability for different datasets. Some work has been done on the issue of when each model should be used (see for example Heller and Simonoff, 1992), but the high-dimensional data setting has not yet been addressed.

Finally, one potential advantage of the Dantzig selector over penalized likelihood methods such as LASSO is that it can be naturally extended to the settings where no explicit likelihoods or loss functions are available. We envision that our work can be extended to apply the Dantzig selector to more general estimating equations.

Reference

Anderson, E., Miller, P., Ilsley, D., Marshall, W., Khvorova, A., Stein, C. and and Benimetskaya, L. (2006) Gene profiling study of G3139- and Bcl-2-targeting siRNAs identifies a unique G3139 molecular signature. *Cancer Gene Therapy*, 13, 406-414.

Antoniadis, A., Fryzlewicz, P., and Letue, F. (2009). The Dantzig selector in Cox's proportional hazards model. To appear in *Scand. J. Statist.*

Buckley,J., and James,I. (1979). Linear regression with censored data. *Biometrika,* 66, 429-436.

Cai, T., Huang, J. and Tian, L. (2009) Regularized Estimation for the Accelerated Failure Time Model. *Biometrics*, 65, 394-404.

Cai, T.T. and Lv, J. (2007). Discussion: the Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2365–2369.

Candès, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.,* **35,** 2313-2351.

Cheung V., Conlin, L., Weber, T., Arcaro, M., Jen, K., Morley, M. and Spielman, R. (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature Genetics*, 33, 422-425.

Datta, S., Le-Rademacher, J., and Datta, S. (2007). Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. *Biometrics* **63**, 259–271.

Dicker, L. and Lin, X. (2009) Variable selection using the Danzig Selector: Asymptotic theory and extensions. *submitted.*

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression (with discussion). Ann. Statist., 32, 407-451.

Engler, D. and Li, Y. (2009) Survival Analysis With High Dimensional Covariates: An Application In Microarray Studies. *Statistical Applications in Genetics and Molecular Biology*, 8, Iss. 1, Article 14.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.,* 96, 1348-1360.

Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space (with discussions). *Journal of Royal Statistical Society, Series B.,* 70, 849-91.

Gui, J. and Li, H. (2005a) Threshold gradient descent method for censored data regression, with applications in pharmacogenomics. *Pacific Symposium on Biocomputing,* 10, 272-283.

Gui, J. and Li, H. (2005b) Penalized Cox regression analysis in the highdimensional and low-sample size settings, with applications to microarray gene expression data. Bioinformatics, 21, 3001-3008.

Heller, G. and Simonoff, J.S. (1992) Prediction in censored survival data: a comparison of the proportional hazards and linear regression models. *Biometrics* **48**, 101–115.

Huang, J., Ma, S., and Xie, H. (2006) Regularized Estimation in the Accelerated Failure Time Model with High-Dimensional Covariates. *Biometrics*, 62, 813-820.

James, G., Radchenko, P. and Lv, J. (2008) DASSO: connections between the Dantzig selector and LASSO. JRSSB, 71, 127-142.

James, G. and Radchenko, P. (2009) A generalized Dantzig selector with shrinkage tuning. *Biometrika* **96**, 323–337.

Jin, Z., Lin, D., Wei, L.J. and Ying, Z. (2003) Rank-based inference for the accelerated failure time model *Biometrika,* 90, 341-353

Johnson, B., Lin, D. and Zeng, D. (2008) Penalized Estimating Functions and Variable Selection in Semiparametric Regression Models. *Journal of the American Statistical Association,* 103, 672-680.

Kalbfleisch, J. and Prentice, R. (2002) *The Statistical Analysis of Failure Time Data*, John Willey and Sons, New York.

Kurtz, T and Protter, P. (1991) Weak limit theorems for stochastic integrals and stochastic differential equations. *Annals of Probability*, 19, 1035-1070.

Lai, T. and Ying, Z. (1991) Large Sample Theory of a Modified Buckley-James Estimator for Regression Analysis with Censored Data. *Annals of Statistics,* 19, 1370-1402.

Li, H., and Luan, Y. (2003). Kernel Cox regression models for linking gene expression profiles to censored survival data. Pacific Symposium of Bio- computing, 8, 65-76.

Li, H. and Gui, J. (2004). Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, 20, 208-215.

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

Martinussen, T. and Scheike, T.H. (2009). Covariate Selection for the Semiparametric Additive Risk Model. *Scand. J. Statist.* **36**, 602-619.

Meinshausen, N. (2007) Relaxed LASSO. Computnl Statist. Data Anal., 52, 374-393.

Meinshausen, N. and Buhlmann, P. (2006) High dimensional graphs and variable selection with the LASSO. Ann. Statist., 34, 1436-1462.

Mulligan, G. et al. (2005) Gene expression profiling and correlation with outcome in clinical trials of the proteasome inhibitor bortezomib. *Blood* **109**, 3177-3188.

Novaka, J., Sladeka, R. and Hudson, T. (2002) Characterization of Variability in Large-Scale Gene Expression Data: Implications for Study Design. *Genomics* **79,** 104-113.

Pawitan, Y. et al. (2005) Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research,* 7, 953-964.

Potti, A., et al. (2007) Genomic signatures to guide the use of chemotherapeutics. *Nature Medicine*, 12, 1294-1300.

Shaughnessy, J.D. et al. (2007) A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. *Blood* **109**, 2276-2284.

Silvey, S.D. (1969). Multicollinearity and Imprecise Estimation. *J. R. Statist. Soc. B.* **31**, 539–552.

Tibshirani, R. (1996) Regression shrinkage and selection via the LASSO. *J. R. Statist. Soc. B,* 58, 267-288.

Tibshirani, R. (1997) The LASSO method for variable selection in the Cox model. *Statistics in*

Medicine, 16, 385-395

Wang, S., Nan, B., Zhu, J., and Beer, D.G. (2008). Doubly penalized Buckley-James method for survival data with high-dimensional covariates. *Biometrics,* **64,** 132-140.

Weissfeld, L.A. (1989). A multicollinearity diagnostic for models fit to censored data. *Comm. Statist. -Theory Meth.* **18**, 2073–2085.

Ying, Z. (1993) A Large Sample Study of Rank Estimation for Censored Regression Data *Annals of Statistics,* 21, 76-99.

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

Zhao, P. and Yu, B. (2006). On model selection consistency of LASSO. *The Journal of Machine Learning Research* **7**, 2541–2563.

Zhao, S.D. and Li, Y. (2009). Survival analysis with ultra-high dimensional covariates: identifying predictive genes for cancer disease. *Submitted.*

Zou, H. (2006) The adaptive LASSO and its oracle properties. J. Am. Statist. Ass., 101, 1418-1429.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. J. R. Statist. Soc. B, 67, 301-320.

Zou, H., Hastie, T. and Tibshirani, R. (2007) On the"degrees of freedom"of the LASSO. *Annals of Statistics,* 35, 2173-2192.

Appendix

A.0: Regularity Conditions

The following assumptions made for Propositions 1 and 2.

1. Assumptions 1–4 of Ying (1993, p. 80).

2. Assumptions (3.1)–(3.5), (5.1), and (3.19) of Lai and Ying (1991). In particular, Assumption (3.1) states that each X_{ij} is bounded by a nonrandom $B > 0$.

3. Define $\tilde{\epsilon}_i = \tilde{Y}_i(\boldsymbol{\beta}_0) - \alpha - \mathbf{X}_i'\boldsymbol{\beta}_0$, the imputation of (centered) ϵ_i defined in (1), where $\tilde{Y}_i(\boldsymbol{\beta}_0)$ is as defined as in (2). Assume that there exists a constant $K > 0$ such that $E(|\tilde{\epsilon}_i|^m) < m!K^{m-2}L/2$ for every $m \geq 2$, where $L = var(\tilde{\epsilon}_i)$.

4. The size V of the true model, i.e. the number of non-zero coefficients, is finite and is independent of n.

Assumption 3 is a standard Bernstein type condition, which simply requires that the high moments of $\tilde{\epsilon}_i$ do not increase too quickly and is generally satisfied when Y_i has a 'thin' tail. The other conditions are standard for censored linear regression models, while the S-sparsity condition was assumed by Candes and Tao (2007) and Fan and Lv (2008).

A.1: Lemma 1 and the Proof

We state a lemma, which will be repeatedly used in our later proofs. It implies that even though $\hat{S}$ [defined in (5)] is a discontinuous function, a first order asymptotic linearization exists.

Lemma 1 *Assume the conditions 1-4 of Ying (1993, p.80). Also suppose that the derivative of the hazard function $\lambda(s)$ with respect to s is continuous for $-\infty < s < \infty$. Then,*

$$\begin{aligned}\hat{S}(s_1, \boldsymbol{\beta}_1) - S(s_0) &= S(s_0)\{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)^T \mathcal{A}(s_0, \boldsymbol{\beta}_0) - \lambda(s_0)(s_1 - s_0) \\ &\quad + n^{-1/2} Z(s_0)\} + o\{\max(n^{-1/2}, |s_1 - s_0| + ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0||)\},\end{aligned}$$

with probability 1 uniformly for any $(s_1, \boldsymbol{\beta}_1) \in \mathcal{B} = \{(s, \boldsymbol{\beta}) : |s - s_0| + ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| < Cn^{-1/2}\}$, where $C > 0$ is any arbitrary constant, $\mathcal{A}$ is a $p \times 1$ nonrandom function, $\lambda(s)$ is the hazard function for $S(s)$ and the stochastic process $Z(s)$ is a version of $\mathcal{W}(v(s))$. Here, $\mathcal{W}(\cdot)$ is the Wiener process and $v(\cdot)$ is defined in (20).

Proof:

Decompose $\hat{S}(s_1, \boldsymbol{\beta}_1) - S(s_0) = \hat{S}(s_1, \boldsymbol{\beta}_1) - \hat{S}(s_0, \boldsymbol{\beta}_0) + \hat{S}(s_0, \boldsymbol{\beta}_0) - S(s_0)$. First study $\hat{S}(s_1, \boldsymbol{\beta}_1) - \hat{S}(s_0, \boldsymbol{\beta}_0)$. Using the arguments of Lai and Ying (1988), it follows that with probability 1,

$$\sup_{(s_1, \boldsymbol{\beta}_1) \in \mathcal{B}} |\hat{S}(s_1, \boldsymbol{\beta}_1) - \hat{S}(s_0, \boldsymbol{\beta}_0) - \xi(s_1, \boldsymbol{\beta}_1) + \xi(s_0, \boldsymbol{\beta}_0)| = o(n^{-1/2}),$$

where

$$\xi(s,\boldsymbol{\beta}) = \exp\left\{-\sum_{i=1}^{n}\int_{-\infty}^{s_1}\frac{dE_xN_i(s,\boldsymbol{\beta})}{E_x\bar{Y}(s,\boldsymbol{\beta})}\right\} = \exp\left\{-\frac{\sum_i G_i(s+\boldsymbol{\beta}'\mathbf{X}_i)f(s+(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\mathbf{X}_i)}{\sum_i G_i(s+\boldsymbol{\beta}'\mathbf{X}_i)S(s+(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\mathbf{X}_i)}ds\right\}$$

where E_x denote the expectation conditional on X, G_i is the survival function of C_i conditional on $\mathbf{X}_i$ and f is the density function of S. Note that $\xi(s,\boldsymbol{\beta}_0) = S(s)$.

Now denote by $\mathbf{d} = \boldsymbol{\beta} - \boldsymbol{\beta}_0$ and by $\lambda(\cdot)$ the hazard function for S. Note that

$$\begin{aligned}
\xi(s,\boldsymbol{\beta}) &= \exp\left\{\int_{-\infty}^{s} -\frac{\sum_i G_i(s+\boldsymbol{\beta}_0'\mathbf{X}_i+\mathbf{d}'\mathbf{X}_i)S(s+\mathbf{d}'\mathbf{X}_i)(\lambda(s+\mathbf{d}'\mathbf{X}_i)-\lambda(s))}{\sum_i G_i(s+\boldsymbol{\beta}_0'\mathbf{X}_i+\mathbf{d}'\mathbf{X}_i)S(s+\mathbf{d}'\mathbf{X}_i)}ds - \int_{-\infty}^{s}\lambda(s)ds\right\} \\
&= S(s)\exp\left\{\int_{-\infty}^{s} -\frac{\sum_i G_i(s+\boldsymbol{\beta}_0'\mathbf{X}_i+\mathbf{d}'\mathbf{X}_i)S(s+\mathbf{d}'\mathbf{X}_i)\lambda^{(1)}(s+d_*'\mathbf{X}_i)\mathbf{d}'\mathbf{X}_i}{\sum_i G_i(s+\boldsymbol{\beta}_0'\mathbf{X}_i+\mathbf{d}'\mathbf{X}_i)S(s+\mathbf{d}'\mathbf{X}_i)}ds\right\} \\
&= S(s)\left\{1-\mathbf{d}'\int_{-\infty}^{s}\frac{\sum_i G_i(s+\boldsymbol{\beta}_0'\mathbf{X}_i+\mathbf{d}'\mathbf{X}_i)S(s+\mathbf{d}'\mathbf{X}_i)\lambda^{(1)}(s+d_*'\mathbf{X}_i)\mathbf{X}_i}{\sum_i G_i(s+\boldsymbol{\beta}_0'X_i+\mathbf{d}'\mathbf{X}_i)S(s+\mathbf{d}'\mathbf{X}_i)}ds\right\} + o(||\mathbf{d}||),
\end{aligned}$$

where $\lambda^{(1)}(\cdot)$ denotes the first derivative of $\lambda(\cdot)$. Hence,

$$\begin{aligned}
& \xi(s_1,\boldsymbol{\beta}_1) - \xi(s_0,\boldsymbol{\beta}_0) \\
= \;& \xi(s_1,\boldsymbol{\beta}_1) - \xi(s_1,\boldsymbol{\beta}_0) + \xi(s_1,\boldsymbol{\beta}_0) - \xi(s_0,\boldsymbol{\beta}_0) \\
= \;& -S(s_1)\mathbf{d}'\int_{-\infty}^{s}\frac{\sum_i G_i(s+\boldsymbol{\beta}_0'X_i+\mathbf{d}'\mathbf{X}_i)S(s+\mathbf{d}'\mathbf{X}_i)\lambda^{(1)}(s+\mathbf{d}_*'\mathbf{X}_i)\mathbf{X}_i}{\sum_i G_i(s+\boldsymbol{\beta}_0'\mathbf{X}_i+\mathbf{d}'\mathbf{X}_i)S(s+\mathbf{d}'\mathbf{X}_i)}ds \\
& +S(s_1) - S(s_0,\boldsymbol{\beta}_0) + o(||\mathbf{d}||) \\
= \;& -S(s_0)\mathbf{d}'\int_{-\infty}^{s}\frac{\sum_i G_i(s+\boldsymbol{\beta}_0'\mathbf{X}_i+\mathbf{d}'\mathbf{X}_i)S(s+\mathbf{d}'\mathbf{X}_i)\lambda^{(1)}(s+\mathbf{d}_*'\mathbf{X}_i)\mathbf{X}_i}{\sum_i G_i(s+\boldsymbol{\beta}_0'\mathbf{X}_i+\mathbf{d}'\mathbf{X}_i)S(s+\mathbf{d}'\mathbf{X}_i)}ds \\
& -f(s_0,\boldsymbol{\beta}_0)(s_1-s_0) + o(||\mathbf{d}|| + |s_1-s_0|),
\end{aligned}$$

where $||\mathbf{d}_*|| \leq ||\mathbf{d}||$. Denote by

$$\Gamma^{(r)}(s,\boldsymbol{\beta}_0) = \text{plim}\frac{1}{n}\sum_i G_i(s+\boldsymbol{\beta}_0'\mathbf{X}_i)\mathbf{X}_i^{\otimes r} \tag{19}$$

for $r = 0,1,2$, where for a vector $\mathbf{a}$ $\mathbf{a}^{\otimes 0} = 1, \mathbf{a}^{\otimes 1} = \mathbf{a}$ and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}'$ and plim denote the probabilistic limit. The argument of Ying (1993, p.87) leads to

$$\int_{-\infty}^{s}\frac{\sum_i G_i(s+\boldsymbol{\beta}_0'\mathbf{X}_i+\mathbf{d}'\mathbf{X}_i)S(s+\mathbf{d}'\mathbf{X}_i)\lambda^{(1)}(s+\mathbf{d}_*'\mathbf{X}_i)\mathbf{X}_i}{\sum_i G_i(s+\boldsymbol{\beta}_0'\mathbf{X}_i+\mathbf{d}'\mathbf{X}_i)S(s+\mathbf{d}'\mathbf{X}_i)}ds = \int_{-\infty}^{s}\frac{\Gamma^{(1)}(s,\boldsymbol{\beta}_0)}{\Gamma^{(0)}(s,\boldsymbol{\beta}_0)}d\lambda(s) + o(||\mathbf{d}||)$$

Hence, $\hat{S}(s_1,\boldsymbol{\beta}_1) - \hat{S}(s_0,\boldsymbol{\beta}_0) = (\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)'\left\{-\int_{-\infty}^{s}\frac{A_1(s,\boldsymbol{\beta}_0)}{A_2(s,\boldsymbol{\beta}_0)}d\lambda(s)\times S(s_0)\right\} - f(s_0,\boldsymbol{\beta}_0)(s_1-s_0) + o(n^{-1/2}, ||\boldsymbol{\beta}_1-\boldsymbol{\beta}_0|| + |s_1-s_0|)$.

Finally, note that

$$\begin{aligned}\hat{S}(s_0,\boldsymbol{\beta}_0)-S(s_0) &= S(s_0)\int_{-\infty}^{s_0}\frac{\sum_i dM_i(u,\boldsymbol{\beta}_0)}{\bar{Y}(u,\boldsymbol{\beta}_0)}+o_p(n^{-1/2})\\ &= n^{-1/2}S(s_0)Z(s_0)+o_p(n^{-1/2}),\end{aligned}$$

where the last equality comes from the Martingale CLT, $M_i(u,\boldsymbol{\beta}_0)=N_i(u,\boldsymbol{\beta}_0)-\int_{-\infty}^{u}Y_i(u,\boldsymbol{\beta}_0)\lambda(u)du$ and $Z(s)$ is a version of $\mathcal{W}(v(s))$, where $\mathcal{W}(\cdot)$ is the Wiener process and

$$v(t)=\int_{-\infty}^{t}\lambda(s)ds/\pi(s,\boldsymbol{\beta}_0). \tag{20}$$

Here, $\pi(s,\boldsymbol{\beta}_0)=plim\frac{1}{n}\bar{Y}(s,\boldsymbol{\beta}_0)=S(s)\Gamma^{(0)}(s,\boldsymbol{\beta}_0)$.

Hence, the result follows by denoting $\mathcal{A}(s_0,\boldsymbol{\beta}_0)=-\int_{-\infty}^{s}\frac{\Gamma^{(1)}(s,\boldsymbol{\beta}_0)}{\Gamma^{(0)}(s,\boldsymbol{\beta}_0)}d\lambda(s)$. □

A.2: Proof of Proposition 1

Define

$$\tilde{Y}_i^0=E(Y_i\mid Y_i^*,\delta_i,\mathbf{X}_i,\boldsymbol{\beta}_0)=Y_i^*+(1-\delta_i)\frac{\int_{e_i(\boldsymbol{\beta}_0)}^{\infty}S(s)ds}{S\{e_i(\boldsymbol{\beta}_0)\}},$$

and let $U_j(\boldsymbol{\beta}_0)$ be the jth component of $U(\boldsymbol{\beta}_0)$. Then

$$U_j(\boldsymbol{\beta}_0)=\sum_{i=1}^{n}(X_{ij}-\bar{X}_j)\left\{\left[\hat{Y}_i(\boldsymbol{\beta}_0)-\tilde{Y}_i^0\right]+\left[\tilde{Y}_i^0-\mathbf{X}'\boldsymbol{\beta}_0\right]\right\}.$$

Taylor expansion gives that $\hat{Y}_i(\boldsymbol{\beta}_0)-\tilde{Y}_i^0$ is asymptotically equal to

$$\begin{aligned}&\frac{1-\delta_i}{S\{e_i(\boldsymbol{\beta}_0)\}}\left\{\int_{e_i(\boldsymbol{\beta}_0)}^{\infty}\left[\hat{S}(s,\boldsymbol{\beta}_0)-S(s)\right]ds\right\}-\\ &\frac{(1-\delta_i)\int_{e_i(\boldsymbol{\beta}_0)}^{\infty}S(s)ds}{S^2\{e_i(\boldsymbol{\beta}_0)\}}\left\{\hat{S}\{e_i(\boldsymbol{\beta}_0),\boldsymbol{\beta}_0\}-S\{e_i(\boldsymbol{\beta}_0)\}\right\}+O_p(1).\end{aligned}$$

By Lemma 1, the first term is equal to

$$n^{-1/2}\frac{1-\delta_i}{S\{e_i(\boldsymbol{\beta}_0)\}}\left\{\int_{e_i(\boldsymbol{\beta}_0)}^{\infty}S(s)Z(s)ds+o_p(1)\right\},$$

where $Z(s)$ is a version of $\mathcal{W}(v(s))$ with $v(s)$ defined in (20). Also by Lemma 1, the second term is equal to

$$n^{-1/2}\frac{(1-\delta_i)\int_{e_i(\boldsymbol{\beta}_0)}^{\infty}S(s)ds}{S^2\{e_i(\boldsymbol{\beta}_0)\}}\left[S\{e_i(\boldsymbol{\beta}_0)\}Z\{e_i(\boldsymbol{\beta}_0)\}+o_p(1)\right].$$

Let the sum of these two approximations be denoted by $n^{-1/2}\zeta_i$. Then by Chebyshev's inequality, for every $M > 0$,

$$P\left(\left[\sum_{i=1}^{n}(X_{ij}-\bar{X}_j)(\hat{Y}_i(\boldsymbol{\beta}_0)-\tilde{Y}_i^0)\right]^2 > M\right) < M^{-1}\text{var}\left(\sum_{i=1}^{n}(X_{ij}-\bar{X}_j)(\hat{Y}_i(\boldsymbol{\beta}_0)-\tilde{Y}_i^0)\right).$$

But

$$\begin{aligned}\text{var}\left(\sum_{i=1}^{n}(X_{ij}-\bar{X}_j)(\hat{Y}_i(\boldsymbol{\beta}_0)-\tilde{Y}_i^0)\right) &= \text{E}\left(\text{var}\left[\sum_{i=1}^{n}(X_{ij}-\bar{X}_j)(\hat{Y}_i(\boldsymbol{\beta}_0)-\tilde{Y}_i^0)\middle|\mathbf{X}_i\right]\right)\\ &+ \text{var}\left(\text{E}\left[\sum_{i=1}^{n}(X_{ij}-\bar{X}_j)(\hat{Y}_i(\boldsymbol{\beta}_0)-\tilde{Y}_i^0)\middle|\mathbf{X}_i\right]\right)\\ &= \text{E}\left(n^{-1}(X_{ij}-\bar{X}_j)^2\sum_{i=1}^{n}\text{var}(\zeta_i)\middle|\mathbf{X}_i\right)\\ &\leq 4B^2\text{var}(\zeta_i) < \infty,\end{aligned}$$

by Assumption 2 of our regularity conditions, which include (3.1) of Lai and Ying (1991) stating that each X_{ij} is bounded by a nonrandom $B > 0$. Therefore, $P([\sum_{i=1}^{n}(X_{ij}-\bar{X}_j)(\hat{Y}_i(\boldsymbol{\beta}_0)-\tilde{Y}_i^0)]^2 > M)$ is independent of j, and we can claim that it approaches 0 as $M \to \infty$. We can conclude that

$$U_j(\boldsymbol{\beta}_0) = O_p(1) + \sum_{i=1}^{n}(X_{ij}-\bar{X}_j)\left[\tilde{Y}_i^0 - \mathbf{X}'\boldsymbol{\beta}_0\right].$$

Now let $w_i^{(j)}(\boldsymbol{\beta}_0) = (X_{ij}-\bar{X}_j)\left[\tilde{Y}_i^0 - \mathbf{X}_i'\boldsymbol{\beta}_0\right]$. Conditional on $\mathbf{X}_i$, $E(w_i^{(j)}(\boldsymbol{\beta}_0)|\mathbf{X}_i) = (X_{ij}-\bar{X}_j)\alpha$. In addition, Assumptions 2 and 3 of our regularity conditions imply $E(|(X_{ij}-\bar{X}_j)\tilde{\epsilon}_i|^m) < m!(2BK)^{m-2}L/2$ for every $m > 2$. Thus Berstein's inequality gives

$$\begin{aligned}P\left(|\sum_{i=1}^{n}w_i^{(j)}(\boldsymbol{\beta}_0)| > \gamma\middle|\mathbf{X}\right) &= P\left(|\sum_{i=1}^{n}\left[w_i^{(j)}(\boldsymbol{\beta}_0)-(X_{ij}-\bar{X}_j)\alpha\right]| > \gamma\middle|\mathbf{X}\right)\\ &= P\left(|\sum_{i=1}^{n}\left[(X_{ij}-\bar{X}_j)\tilde{\epsilon}_i\right]| > \gamma\middle|\mathbf{X}\right)\\ &\leq 2\exp\left(-\frac{\gamma^2/n}{L+2BK\gamma/n}\right).\end{aligned}$$

Since the probability bound is independent of $\mathbf{X}$, we can marginalize over $\mathbf{X}$ to conclude that

$$P\left(\left|\sum_{i=1}^{n}w_i^{(j)}(\boldsymbol{\beta}_0)\right| > \gamma\right) \leq 2\exp\left(-\frac{\gamma^2/n}{L+2BK\gamma/n}\right).$$

Now, since $|U_j(\boldsymbol{\beta}_0)| \leq |\sum_{i=1}^{n}w_i^{(j)}(\boldsymbol{\beta}_0)| + O_p(1)$, and since we have shown that the $O_p(1)$ doesn't depend on j, then $\sup_j|n^{-1}U_j(\boldsymbol{\beta}_0)| \leq \sup_j|n^{-1}\sum_{i=1}^{n}w_i^{(j)}(\boldsymbol{\beta}_0)| + o_p(1)$. In addition, for

any $\epsilon > 0$, $P(R_1 + R_2 \geq c) \leq P(R_1 \geq c - \epsilon) + P(R_2 \geq \epsilon)$, where R_1 and R_2 are two random variables and c is a constant. Hence

$$P\left(\sup_j \left|\sum_{i=1}^n w_i^{(j)}(\boldsymbol{\beta}_0)\right| + o_p(1) > \gamma\right) \leq P\left(\sup_j \left|\sum_{i=1}^n w_i^{(j)}(\boldsymbol{\beta}_0)\right| > \gamma - \epsilon\right) + P(o_p(1) \geq \epsilon).$$

The second term on the right-hand side can be made arbitrarily close to zero, so

$$P(\|U(\boldsymbol{\beta}_0)\|_\infty > \gamma) \leq P\left(\sup_j \left|\sum_{i=1}^n w_i^{(j)}(\boldsymbol{\beta}_0)\right| > \gamma\right) \leq 2p \exp\left(-\frac{\gamma^2/n}{L + 2BK\gamma/n}\right).$$

The choice of γ concludes the proof. Moreover, when $p = O(n^\kappa)$, $\kappa \geq 1$, the result follows immediately. □

A.3: Proof of Proposition 2

Let $U_0(\boldsymbol{\beta})$ be the smoothed version of $U(\boldsymbol{\beta})$ such that $U_0(\boldsymbol{\beta}_0) = 0$. The explicit form of $U_0(\boldsymbol{\beta})$ can be found in (3.8) in Lai and Ying (1991). Then for $0 < \iota < 1/32$, Lai and Ying (1991) show that

$$\sup_\beta \|U(\boldsymbol{\beta}) - U_0(\boldsymbol{\beta})\|_\infty = o_p(n^{5/8})\text{a.s.} \tag{21}$$

and

$$\lim_{n\to\infty} n^{-3/4} \left\{\inf_{\|\beta-\beta_0\|_2 \geq n^{-\iota}} \|U_o(\boldsymbol{\beta})\|_\infty\right\} = \infty. \tag{22}$$

Define

$$\tilde{\boldsymbol{\Omega}} = \int_{-\infty}^{\infty} \left[\Gamma_n^{(2)}(t, \boldsymbol{\beta}_0) - \frac{\{\Gamma_n^{(1)}(t, \boldsymbol{\beta}_0)\}^{\otimes 2}}{\Gamma_n^{(0)}(t, \boldsymbol{\beta}_0)}\right] \frac{\int_t^\infty (1 - F(s))ds}{1 - F(t)} \left\{\frac{d \log f(t)}{dt} + \frac{f(t)}{1 - F(t)}\right\} dF(t), \tag{23}$$

where $\Gamma_n^{(r)}(s, \boldsymbol{\beta}_0) = \frac{1}{n}\sum_{i=1}^n G_i(s + \boldsymbol{\beta}_0'\mathbf{X}_i)\mathbf{X}_i^{\otimes r}$. Then the arguement in Lai and Ying (1991) also leads to

$$U_o(\boldsymbol{\beta}) = -\tilde{\boldsymbol{\Omega}} n(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + o(1)\text{a.s. uniformly in } \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq n^{-\iota}. \tag{24}$$

To relate $\boldsymbol{\beta}$ to the score equation we invoke (24). We now show that the set of feasible $\boldsymbol{\beta}$ must be within $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq n^{-\iota}$ almost surely for n sufficiently large. If not, the inequality $\|U(\boldsymbol{\beta})\|_\infty \leq \gamma$ implies that $\|n^{-3/4}U_0(\boldsymbol{\beta})\|_\infty - o(n^{-1/8}) \leq 2n^{-3/4}\gamma$ a.s. by (21), while (22) implies that on the set $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 > n^{-\iota}$ the left-hand side tends to ∞. With γ, the right-hand side

tends to zero, contradicting to $\|U(\boldsymbol{\beta})\|_\infty \leq \gamma$. Thus for n sufficiently large, $P(\{\boldsymbol{\beta} : \|U(\boldsymbol{\beta})\|_\infty \leq \gamma\} \cap \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \geq n^{-\iota}\}) = 0$, i.e. $\hat{\boldsymbol{\beta}}$ must be in $\{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\|_2 \leq n^{-\iota}\}$ almost surely.

Restricting our attention to this set, (24) and (21) give that for a feasible $\boldsymbol{\beta}$, with probability 1,

$$\begin{aligned}
\|\tilde{\boldsymbol{\Omega}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\|_\infty &\leq o(n^{-1}) + \|n^{-1}(U_0(\boldsymbol{\beta}) - U_0(\boldsymbol{\beta}_0))\|_\infty \\
&= o(n^{-1}) + n^{-1}\|(U_0(\boldsymbol{\beta}) - U(\boldsymbol{\beta})) + (U(\boldsymbol{\beta}) - U(\boldsymbol{\beta}_0)) + (U(\boldsymbol{\beta}_0) - U_0(\boldsymbol{\beta}_0))\|_\infty \\
&\leq o(n^{-1}) + o(n^{-3/8}) + \|n^{-1}U(\boldsymbol{\beta})\|_\infty + \|n^{-1}U(\boldsymbol{\beta}_0)\|_\infty + o(n^{-3/8}) \\
&\leq o(n^{-3/8}) + 2\gamma/n.
\end{aligned}$$

Let $\mathbf{h} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$. As γ/n dominates $o(n^{-3/8})$ for n large enough and because $\hat{\boldsymbol{\beta}}$ is feasible by definition and $\boldsymbol{\beta}_0$ is feasible with high probability by Proposition 1, we immediately have that $\|\{\tilde{\boldsymbol{\Omega}}^{1/2}\}'\tilde{\boldsymbol{\Omega}}^{1/2}\mathbf{h}\|_\infty \leq 3\gamma/n$.

Now suppose T_0 is a set of cardinality V with $\delta_V + \theta_{V,2V} < 1$, and T_1 consists of the V largest components of $\mathbf{h}$ outside of T_0. Let $T_{01} = T_0 \cup T_1$. Then Cauchy-Schwarz inequality gives $\|\{\tilde{\boldsymbol{\Omega}}_{T_{01}}^{1/2}\}'\tilde{\boldsymbol{\Omega}}^{1/2}\mathbf{h}\|_2 \leq (2V)^{1/2}3\gamma/n$. By a trivial modification of Lemma 3.1 of Candes and Tao (2007),

$$\|\mathbf{h}_{T_{01}}\|_2 \leq \frac{(2V)^{1/2}3\gamma/n}{\delta_{2V}} + \frac{\theta_{V,2V}}{\delta_{2V}V^{1/2}}\|\mathbf{h}_{\bar{T}_0}\|_1,$$

and

$$\|\mathbf{h}\|_2^2 \leq \|\mathbf{h}_{T_{01}}\|_2^2 + V^{-1}\|\mathbf{h}_{\bar{T}_0}\|_1^2$$

where $\bar{T}_0$ is the completement of T_0 in $\{1, \ldots, p\}$. Using the fact that $\|\hat{\boldsymbol{\beta}}\|_1 \leq \|\boldsymbol{\beta}_0\|_1$, one can show that $\|\mathbf{h}_{\bar{T}_0}\|_1 \leq V^{1/2}\|\mathbf{h}_{T_0}\|_2$. Subtracting $\|\mathbf{h}_{T_{01}}\|_2\theta_{V,2V}/\delta_{2V}V^{1/2}$ from both sides of the previous inequalities gives

$$\begin{aligned}
\|\mathbf{h}_{T_{01}}\|_2 &\leq \frac{\delta_{2V}}{\delta_{2V} - \theta_{V,2V}}\left\{\frac{(2V)^{1/2}3\gamma/n}{\delta_{2V}} + \frac{\theta_{V,2V}}{\delta_{2V}}\left(\|\mathbf{h}_{T_0}\|_2 - \|\mathbf{h}_{T_{01}}\|_2\right)\right\} \\
&\leq \frac{(2V)^{1/2}3\gamma/n}{\delta_{2V} - \theta_{V,2V}}.
\end{aligned}$$

Since

$$\|\mathbf{h}\|_2^2 \leq \|\mathbf{h}_{T_{01}}\|_2^2 + V^{-1}\|\mathbf{h}_{\bar{T}_0}\|_1^2 \leq \|\mathbf{h}_{T_{01}}\|_2^2 + V^{-1}V\|\mathbf{h}_{T_0}\|_2^2 \leq 2\|\mathbf{h}_{T_{01}}\|_2^2,$$

this gives, along with the probability bound of Proposition 1,

$$P\left(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_2^2 \leq \frac{36V\gamma^2/n^2}{(\delta_{2V}-\theta_{V,2V})^2}\right) > 1-2p\exp\left(-\frac{\gamma^2/n}{L+2BK\gamma/n}\right).$$

Moreover, when $p=O(n^\kappa)$, $\kappa \geq 1$, the result follows immediately. □

A.4: Proof of Proposition 3

Denote by $\boldsymbol{\beta}_0$ the truth and

$$\tilde{Y}_i^0 = E(Y_i|Y_i^*,\delta_i,\mathbf{X}_i,\boldsymbol{\beta}_0) = Y_i^* + (1-\delta_i)\frac{\int_{e_i(\boldsymbol{\beta}_0)}^\infty S(s)ds}{S\{e_i(\boldsymbol{\beta}_0)\})},$$

where S is the (true) survival function corresponding to the distribution function F, ie $S(.) = 1-F(.)$. Then

$$\sum_{i=1}^n \mathbf{X}_i\left\{\hat{Y}_i(\hat{\boldsymbol{\beta}}^{(0)})-Y_i\right\} = \sum_{i=1}^n \mathbf{X}_i\left\{\hat{Y}_i(\hat{\boldsymbol{\beta}}^{(0)})-\tilde{Y}_i^0\right\} + \sum_{i=1}^n \mathbf{X}_i(\tilde{Y}_i^0-Y_i) \tag{25}$$

The second term on the right hand side of (25) is $O_p(n^{1/2})$ by the CLT, we only need to consider the first term on the right hand side of (25), which is equal to

$$\sum_i \mathbf{X}_i(1-\delta_i)\left\{\frac{\int_{e_i(\hat{\boldsymbol{\beta}}^{(0)})}^\infty \hat{S}\{s,\hat{\boldsymbol{\beta}}^{(0)}\}ds}{\hat{S}\{e_i(\hat{\boldsymbol{\beta}}^{(0)}),\hat{\boldsymbol{\beta}}^{(0)}\}} - \frac{\int_{e_i(\boldsymbol{\beta}_0)}^\infty S(s)ds}{S\{e_i(\boldsymbol{\beta}_0)\})}\right\}, \tag{26}$$

where $\hat{S}(t,\boldsymbol{\beta})$ is the Nelson-Aalen estimator based on data $(Y_i^*-\mathbf{X}_i'\boldsymbol{\beta},\delta_i), i=1,\ldots,n$.

Equation (26) is asymptotically equal to

$$\sum_i \frac{\mathbf{X}_i(1-\delta_i)}{S\{e_i(\boldsymbol{\beta}_0)\}}\left\{\int_{e_i(\hat{\boldsymbol{\beta}}^{(0)})}^\infty \hat{S}(s,\hat{\boldsymbol{\beta}}^{(0)})ds - \int_{e_i(\boldsymbol{\beta}_0)}^\infty S(s)ds\right\} \tag{27}$$

$$-\sum_i \frac{\mathbf{X}_i(1-\delta_i)\int_{e_i(\boldsymbol{\beta}_0)}^\infty S(s)ds}{S^2\{e_i(\boldsymbol{\beta}_0)\}}\left[\hat{S}\{e_i(\hat{\boldsymbol{\beta}}^{(0)}),\hat{\boldsymbol{\beta}}^{(0)}\}-S\{e_i(\boldsymbol{\beta}_0)\}\right] + O_p(1) \tag{28}$$

Next consider (27). Note that

$$\begin{aligned}
&\int_{e_i(\hat{\boldsymbol{\beta}}^{(0)})}^\infty \hat{S}\{s,\hat{\boldsymbol{\beta}}^{(0)}\}ds - \int_{e_i(\boldsymbol{\beta}_0)}^\infty S(s)ds \\
=\ &\int_{e_i(\boldsymbol{\beta}_0)}^{e_i(\hat{\boldsymbol{\beta}}^{(0)})} \hat{S}\{s,\hat{\boldsymbol{\beta}}^{(0)}\}ds + \int_{e_i(\boldsymbol{\beta}_0)}^\infty \hat{S}\{s,\hat{\boldsymbol{\beta}}^{(0)}\}-S(s)ds \\
=\ &\int_{e_i(\boldsymbol{\beta}_0)}^{e_i(\hat{\boldsymbol{\beta}}^{(0)})} S\{e_i(\boldsymbol{\beta}_0)\}ds + \int_{e_i(\boldsymbol{\beta}_0)}^{e_i(\hat{\boldsymbol{\beta}}^{(0)})} \hat{S}\{s,\hat{\boldsymbol{\beta}}^{(0)}\}-S\{e_i(\boldsymbol{\beta}_0)\}ds \\
&+\int_{e_i(\boldsymbol{\beta}_0)}^\infty \hat{S}\{s,\hat{\boldsymbol{\beta}}^{(0)}\}-S(s)ds.
\end{aligned}$$

It is obvious the first term in the above equation is equal to

$$\int_{e_i(\boldsymbol{\beta}_0)}^{e_i(\hat{\boldsymbol{\beta}}^{(0)})} S\{e_i(\boldsymbol{\beta}_0)\}ds = S\{e_i(\boldsymbol{\beta}_0)\}\{e_i(\hat{\boldsymbol{\beta}}^{(0)}) - e_i(\boldsymbol{\beta}_0)\} = -S\{e_i(\boldsymbol{\beta}_0)\}\mathbf{X}_i'(\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0).$$

For the second term, applying Lemma 1 and noting that $\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0 = O_p(n^{-1/2})$ yields

$$\int_{e_i(\boldsymbol{\beta}_0)}^{e_i(\hat{\boldsymbol{\beta}}^{(0)})} \hat{S}\{s, \hat{\boldsymbol{\beta}}^{(0)}\} - S\{e_i(\boldsymbol{\beta}_0)\}ds = o_p(n^{-1/2}).$$

Finally, for the third term as

$$\sqrt{n}[\hat{S}\{s, \hat{\boldsymbol{\beta}}^{(0)}\} - S(s)] \to S(s)Z(s)$$

weakly, where $Z(s)$ is a version of $\mathcal{W}(v(s))$, by the weak convergence of stochastic integrals (e.g Theorem 2.2 of Kurtz and Protter (1991)) and the Skorohod representation theorem, we have that

$$\int_{e_i(\boldsymbol{\beta}_0)}^{\infty} [\hat{S}\{s, \hat{\boldsymbol{\beta}}^{(0)}\} - S(s)]ds = n^{-1/2}\int_{e_i(\boldsymbol{\beta}_0)}^{\infty} S(s)Z(s)ds + o_p(n^{-1/2}).$$

When applying Theorem 2.2 of Kurtz and Protter (1991), we need to verify that the variance of the integrand of the last integral (or the "change of the time" in the Gaussian process), which is $var\{S(t)Z(t)\} = S^2(t)v(t)$ is bounded at ∞. That is $\limsup_{t\to\infty} S^2(t)v(t) < \infty$. Indeed,

$$S^2(t)v(t) < \int_{-\infty}^{t} S^2(s)\frac{\lambda_0(s)}{\pi(s, \boldsymbol{\beta}_0)}ds < \int_{-\infty}^{\infty} \frac{dF(s, \boldsymbol{\beta}_0)}{\Gamma^{(0)}(s, \boldsymbol{\beta}_0)} < \infty$$

by the regularity condition. Hence, (27) is equal (in distribution) to

$$\begin{aligned} & \sum_{i=1}^{n} \mathbf{X}_i(1-\delta_i)\Big([\tilde{A}_i(\boldsymbol{\beta}_0) - S\{e_i(\boldsymbol{\beta}_0)\}\mathbf{X}_i]'(\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0) \\ & + n^{-1/2}\int_{e_i(\boldsymbol{\beta}_0)}^{\infty} \frac{S(t)}{S\{e_i(\boldsymbol{\beta}_0)\}}\mathcal{W}\{v(t)\}dt\Big) + o_p(n^{1/2}) \\ = \; & O_p(n^{1/2}), \end{aligned}$$

where $\tilde{A}_i(\boldsymbol{\beta}_0) = \int_{e_i(\boldsymbol{\beta}_0)}^{\infty} \frac{S(s)}{S\{e_i(\boldsymbol{\beta}_0)\}}\mathcal{A}(\boldsymbol{\beta}_0, s)ds$ and the last equality stems from that $\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0 = O_p(n^{-1/2})$.

Finally consider (28). Using Lemma 1, it is follows that

$$\begin{aligned} & \hat{S}\{e_i(\hat{\boldsymbol{\beta}}^{(0)}), \hat{\boldsymbol{\beta}}^{(0)}\} - S\{e_i(\boldsymbol{\beta}_0)\} \\ \stackrel{d}{=} \; & S\{e_i(\boldsymbol{\beta}_0)\}([\mathcal{A}\{e_i(\boldsymbol{\beta}_0), \boldsymbol{\beta}_0\} - \lambda\{e_i(\boldsymbol{\beta}_0)\}\mathbf{X}_i]'(\hat{\boldsymbol{\beta}}^{(0)} - \boldsymbol{\beta}_0) + n^{-1/2}\mathcal{W}[v\{e_i(\boldsymbol{\beta}_0)\}]) + o_p(n^{-1/2}), \end{aligned}$$

where $\stackrel{d}{=}$ is for equal in distribution.

Hence, (28) is equal, in distribution, to

$$\begin{aligned}
&\sum_i \mathbf{X}_i(1-\delta_i)\frac{\int_{e_i(\boldsymbol{\beta}_0)}^{\infty} S(s)ds}{S\{e_i(\boldsymbol{\beta}_0)\}}\Big([\mathcal{A}\{e_i(\boldsymbol{\beta}_0),\boldsymbol{\beta}_0\}+\lambda\{e_i(\boldsymbol{\beta}_0)\}\mathbf{X}_i]'(\hat{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}_0)\\
&+n^{-1/2}\mathcal{W}[v\{e_i(\boldsymbol{\beta}_0)\}]\Big)+o_p(n^{1/2})\\
=\;&\sum_i \mathbf{X}_i(\tilde{Y}_i(\boldsymbol{\beta}_0)-Y_i^*)([\mathcal{A}\{e_i(\boldsymbol{\beta}_0),\boldsymbol{\beta}_0\}+\lambda\{e_i(\boldsymbol{\beta}_0)\}\mathbf{X}_i]'(\hat{\boldsymbol{\beta}}^{(0)}-\boldsymbol{\beta}_0)+O_p(n^{-1/2}))+o_p(n^{1/2})\\
=\;&O_p(n^{1/2}).
\end{aligned}$$

Combining (27) and (28) yields the result. □

A.5: Proof of Lemma 4

Define the Lagrangian

$$L(\boldsymbol{\beta},\boldsymbol{\mu})=||\mathbf{W}\boldsymbol{\beta}||_1+\boldsymbol{\mu}'\mathbf{Z}'(\hat{\mathbf{Y}}-\mathbf{ZW}\boldsymbol{\beta})-\gamma||\boldsymbol{\mu}||_1.$$

Then (12) and (13) imply

$$||\mathbf{W}\hat{\boldsymbol{\beta}}||_1=L(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\mu}})=\hat{\boldsymbol{\mu}}'\mathbf{Z}'\hat{\mathbf{Y}}-\gamma||\hat{\boldsymbol{\mu}}||_1.$$

Since

$$\begin{aligned}
\inf_{\boldsymbol{\beta}} L(\boldsymbol{\beta},\boldsymbol{\mu}) &= \boldsymbol{\mu}'\mathbf{Z}'\hat{\mathbf{Y}}-\gamma||\boldsymbol{\mu}||_1+\inf_{\boldsymbol{\beta}}\,(\mathrm{sgn}(\boldsymbol{\beta})-\boldsymbol{\mu}\mathbf{Z}'\mathbf{Z})'\mathbf{W}\boldsymbol{\beta}\\
&= \begin{cases} \boldsymbol{\mu}'\mathbf{Z}'\hat{Y}-\gamma||\boldsymbol{\mu}||_1 & \text{if } ||\mathbf{Z}'\mathbf{Z}\boldsymbol{\mu}||_\infty\le 1\\ -\infty & \text{otherwise,}\end{cases}
\end{aligned}$$

and because (11) holds, we have

$$||\mathbf{W}\hat{\boldsymbol{\beta}}||_1=\inf_{\boldsymbol{\beta}} L(\boldsymbol{\beta},\hat{\boldsymbol{\mu}})\le\sup_{\boldsymbol{\mu}}\inf_{\boldsymbol{\beta}} L(\boldsymbol{\beta},\boldsymbol{\mu})\le\sup_{\boldsymbol{\mu}} L(\tilde{\boldsymbol{\beta}},\boldsymbol{\mu})$$

for any $\tilde{\boldsymbol{\beta}}$. This, the inequality (10), and the fact that

$$\begin{aligned}
\sup_{\boldsymbol{\mu}} L(\boldsymbol{\beta},\boldsymbol{\mu}) &= ||\mathbf{W}\boldsymbol{\beta}||_1+\sup_{\boldsymbol{\mu}}\;\boldsymbol{\mu}'[\mathbf{Z}'(\hat{\mathbf{Y}}-\mathbf{ZW}\boldsymbol{\beta})-\gamma\mathrm{sgn}(\boldsymbol{\mu})]\\
&= \begin{cases} ||\mathbf{W}\boldsymbol{\beta}||_1 & \text{if } ||\mathbf{Z}'(\hat{\mathbf{Y}}-\mathbf{ZW}\boldsymbol{\beta})||_\infty\le\gamma\\ \infty & \text{otherwise}\end{cases}
\end{aligned}$$

imply that $||\mathbf{W}\hat{\boldsymbol{\beta}}||_1\le||\mathbf{W}\boldsymbol{\beta}||_1$ whenever $|\mathbf{Z}'(\hat{\mathbf{Y}}-\mathbf{ZW}\boldsymbol{\beta})|\le\gamma$. This means that $\hat{\boldsymbol{\beta}}$ solves (9). □

A.6: Proof of Proposition 5

With $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\beta}}$ as in (14)-(17), we check that (10)-(13) hold with probability tending to 1. First note that

$$\mathbf{Z}_A'\mathbf{Z}\hat{\boldsymbol{\mu}} = \mathbf{Z}_A'\mathbf{Z}_A\hat{\boldsymbol{\mu}}_A = \mathrm{sgn}(\boldsymbol{\beta}_0)_A$$

and

$$\begin{aligned}
\mathbf{Z}_{\bar{A}}'\mathbf{Z}\hat{\boldsymbol{\mu}} &= \mathbf{Z}_{\bar{A}}\mathbf{Z}_A(\mathbf{Z}_A'\mathbf{Z}_A)^{-1}\mathrm{sgn}(\boldsymbol{\beta}_0)_A \\
&= (\mathbf{W}_{\bar{A},\bar{A}}^{-1})'\mathbf{X}_{\bar{A}}'\mathbf{P}_n\mathbf{X}_A(\mathbf{X}_A'\mathbf{P}_n\mathbf{X}_A)^{-1}\mathbf{W}_A\mathrm{sgn}(\boldsymbol{\beta}_0)_A \\
&= o_P(1).
\end{aligned}$$

This implies that (11) holds with probability tending to 1. To see that (10) holds with probability approaching 1, first observe that

$$\mathbf{Z}_A'(\hat{\mathbf{Y}} - \mathbf{Z}\mathbf{W}\hat{\boldsymbol{\beta}}) = \gamma\mathrm{sgn}(\hat{\boldsymbol{\mu}})_A. \tag{29}$$

Furthermore,

$$\begin{aligned}
\mathbf{Z}_{\bar{A}}'(\hat{\mathbf{Y}} - \mathbf{Z}\mathbf{W}\hat{\boldsymbol{\beta}}) &= \mathbf{Z}_{\bar{A}}'\left[\mathbf{I} - \mathbf{Z}_A(\mathbf{Z}_A'\mathbf{Z}_A)^{-1}\mathbf{Z}_A'\right]\hat{\mathbf{Y}} + \gamma\mathbf{Z}_{\bar{A}}'\mathbf{Z}_A(\mathbf{Z}_A'\mathbf{Z}_A)^{-1}\mathrm{sgn}(\hat{\boldsymbol{\mu}})_A \\
&= \mathbf{W}_{\bar{A},\bar{A}}^{-1}\mathbf{X}_{\bar{A}}'\mathbf{P}_n\left[\mathbf{I} - \mathbf{P}_n\mathbf{X}_A(\mathbf{X}_A'\mathbf{P}_n\mathbf{X}_A)^{-1}\mathbf{X}_A'\mathbf{P}_n\right]\hat{\mathbf{Y}} \\
&\quad + \gamma\mathbf{W}_{\bar{A},\bar{A}}^{-1}\mathbf{X}_{\bar{A}}\mathbf{P}_n\mathbf{X}_A(\mathbf{X}_A\mathbf{P}_n\mathbf{X}_A)^{-1}\mathbf{W}_A\mathrm{sgn}(\hat{\boldsymbol{\mu}})_A.
\end{aligned} \tag{30}$$

Proposition 3 implies that

$$\begin{aligned}
\mathbf{X}_{\bar{A}}'\mathbf{P}_n\left[\mathbf{I} - \mathbf{P}_n\mathbf{X}_A(\mathbf{X}_A'\mathbf{P}_n\mathbf{X}_A)^{-1}\mathbf{X}_A'\mathbf{P}_n\right]\hat{\mathbf{Y}} &= \mathbf{X}_{\bar{A}}'\mathbf{P}_n\left[\mathbf{I} - \mathbf{P}_n\mathbf{X}_A(\mathbf{X}_A'\mathbf{P}_n\mathbf{X}_A)^{-1}\mathbf{X}_A'\mathbf{P}_n\right]\mathbf{Y} \\
&\quad + O_P(\sqrt{n}) \\
&= O_P(\sqrt{n}),
\end{aligned}$$

where the second equality above holds because

$$\mathbf{X}_{\bar{A}}'\mathbf{P}_n\left[\mathbf{I} - \mathbf{P}_n\mathbf{X}_A(\mathbf{X}_A'\mathbf{P}_n\mathbf{X}_A)^{-1}\mathbf{X}_A'\mathbf{P}_n\right]\mathbf{Y} = \mathbf{X}_{\bar{A}}'\mathbf{P}_n\left[\mathbf{I} - \mathbf{P}_n\mathbf{X}_A(\mathbf{X}_A'\mathbf{P}_n\mathbf{X}_A)^{-1}\mathbf{X}_A'\mathbf{P}_n\right]\epsilon$$

and

$$\left|\mathbf{X}_j'\mathbf{P}_n\left[\mathbf{I} - \mathbf{P}_n\mathbf{X}_A(\mathbf{X}_A'\mathbf{P}_n\mathbf{X}_A)^{-1}\mathbf{X}_A'\mathbf{P}_n\right]\mathbf{P}_n\mathbf{X}_j\right| \leq \mathbf{X}_j'\mathbf{X}_j,$$

which implies that

$$\frac{1}{\sqrt{n}}\mathbf{X}_{\bar{A}}'\mathbf{P}_n\left[\mathbf{I}-\mathbf{P}_n\mathbf{X}_A(\mathbf{X}_A'\mathbf{P}_n\mathbf{X}_A)^{-1}\mathbf{X}_A'\mathbf{P}_n\right]\mathbf{Y}$$

has mean 0 and bounded variance. Since $\mathbf{W}_{\bar{A},\bar{A}}^{-1}=o_P(\gamma/\sqrt{n})$, it follows that

$$\mathbf{W}_{\bar{A},\bar{A}}^{-1}\mathbf{X}_{\bar{A}}'\mathbf{P}_n\left[\mathbf{I}-\mathbf{P}_n\mathbf{X}_A(\mathbf{X}_A'\mathbf{P}_n\mathbf{X}_A)^{-1}\mathbf{X}_A'\mathbf{P}_n\right]\hat{\mathbf{Y}}=o_P(\gamma).$$

Combining this with (30) and the fact that

$$\gamma\mathbf{W}_{\bar{A},\bar{A}}^{-1}\mathbf{X}_{\bar{A}}\mathbf{P}_n\mathbf{X}_A(\mathbf{X}_A\mathbf{P}_n\mathbf{X}_A)^{-1}\mathbf{W}_A\mathrm{sgn}(\hat{\boldsymbol{\mu}})_A=o_P(\gamma)$$

gives

$$\mathbf{Z}_{\bar{A}}'(\hat{\mathbf{Y}}-\mathbf{ZW}\hat{\boldsymbol{\beta}})=o_P(\gamma).$$

This fact, plus (29), implies that (10) holds with probability tending to 1. Since

$$\hat{\boldsymbol{\mu}}'\mathbf{Z}'\mathbf{ZW}\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\mu}}_A'\mathbf{Z}_A'\mathbf{Z}_A\mathbf{W}_A\hat{\boldsymbol{\beta}}_A=\mathrm{sgn}(\boldsymbol{\beta})_A'\mathbf{W}_A\hat{\boldsymbol{\beta}}_A$$

and $\mathrm{sgn}(\hat{\boldsymbol{\beta}})_A\xrightarrow{P}\mathrm{sgn}(\boldsymbol{\beta}_0)_A$, the probability that (12) holds converges to 1. Lastly,

$$\hat{\boldsymbol{\mu}}'\mathbf{Z}'(\mathbf{P}_n\hat{\mathbf{Y}}-\mathbf{ZW}\hat{\boldsymbol{\beta}})=\hat{\boldsymbol{\mu}}_A'\mathbf{Z}_A'(\mathbf{P}_n\hat{\mathbf{Y}}-\mathbf{ZW}\hat{\boldsymbol{\beta}})=\gamma\hat{\boldsymbol{\mu}}_A'\mathrm{sgn}(\hat{\boldsymbol{\mu}})_A=\gamma||\hat{\boldsymbol{\mu}}||_1,$$

which implies that (13) holds. We conclude that (10)-(13) hold with probability tending to 1 and the proposition is proved. □

A.7: Proof of Corollary 1

Let $\hat{\boldsymbol{\beta}}$ be any sequence of solutions to (9), let $T=\{j;\ \hat{\beta}_j^1\neq 0\}$ and let $E=\{j;\ |\mathbf{Z}_j'(\hat{\mathbf{Y}}-\mathbf{ZW}\hat{\boldsymbol{\beta}})|=\gamma\}$. Proposition 5 implies that by slightly perturbing $\hat{\boldsymbol{\beta}}$ if necessary, we can assume that $E\subseteq A\subseteq T$. The conditions (13)-(10) in Lemma 2 imply that there exists $\mathbf{t}\in\{\pm 1\}^{|T|}$ such that

$$||\mathbf{W}^{-1}\mathbf{X}'\mathbf{P}_n\mathbf{X}_T(\mathbf{X}_T'\mathbf{P}_n\mathbf{X}_T)^{-1}\mathbf{W}_T\mathbf{t}||_\infty\leq 1. \tag{31}$$

Since $w_j/w_k\xrightarrow{P}\infty$, whenever $j\in\bar{A}$ and $k\in A$, it follows that $T=A$, with probability tending to 1. Thus, $P(T=A)\to 1$ and $\hat{\boldsymbol{\beta}}$ is consistent for model selection.

A.8: Proof of Proposition 6

Define

$$\mathbf{\Omega} = \int_{-\infty}^{\infty} \left[\Gamma^{(2)}(t, \boldsymbol{\beta}_0) - \frac{\{\Gamma^{(1)}(t, \boldsymbol{\beta}_0)\}^{\otimes 2}}{\Gamma^{(0)}(t, \boldsymbol{\beta}_0)} \right] \frac{\int_t^{\infty}(1 - F(s))ds}{1 - F(t)} \left\{ \frac{d \log f(t)}{dt} + \frac{f(t)}{1 - F(t)} \right\} dF(t), \quad (32)$$

$$\mathbf{\Lambda} = \int_{-\infty}^{-\infty} \left[\Gamma^{(2)}(t, \boldsymbol{\beta}_0) - \frac{\{\Gamma^{(1)}(t, \boldsymbol{\beta}_0)\}^{\otimes 2}}{\Gamma^{(0)}(t, \boldsymbol{\beta}_0)} \right] \left\{ \frac{\int_t^{\infty}(1 - F(s))ds}{1 - F(t)} \right\}^2 dF(t), \quad (33)$$

where $f(\cdot)$ is the density function for $F(\cdot)$, the CDF of ϵ_i, and $\Gamma^{(r)}(t, \boldsymbol{\beta}_0)$ for $r = 0, 1$ are defined as in (19).

Since $\{T = A\} \subset \left\{ \hat{\boldsymbol{\beta}}^{(0,T)} = \hat{\boldsymbol{\beta}}^{(0,A)} \right\}$, coupled with $P(T = A) \to 1$ implied by Proposition 5, it follows immediately that

$$P\left(\hat{\boldsymbol{\beta}}^{(0,T)} = \hat{\boldsymbol{\beta}}^{(0,A)} \right) \to 1.$$

Therefore, $\sqrt{n}(\hat{\boldsymbol{\beta}}_A^{(0,T)} - \boldsymbol{\beta}_{0,A}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_A^{(0,A)} - \boldsymbol{\beta}_{0,A}) + o_P(1)$. Further, as Theorem 4 of Lai and Ying (1991) implies that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_A^{(0,A)} - \boldsymbol{\beta}_{0,A}) \xrightarrow{d} N(0, \mathbf{\Sigma}_A),$$

the original claim is thus proved. □

A.9: Proof of Corollary 2

Since $\hat{\boldsymbol{\beta}}^{(0,T)}$ is $\sqrt{n}$-consistent for $\boldsymbol{\beta}$ by Proposition 6, Lemma 2 implies that, with probability tending to 1, $\hat{\boldsymbol{\beta}}_{\bar{A}}^* = 0$ and

$$\hat{\boldsymbol{\beta}}_A^* = (\mathbf{X}_A' \mathbf{P}_n \mathbf{X}_A)^{-1} \mathbf{X}_A' \mathbf{P}_n \hat{\mathbf{Y}}^{(1)} - \gamma (\mathbf{X}_A' \mathbf{P}_n \mathbf{X}_A)^{-1} \mathbf{W}_A \mathbf{r} = \hat{\boldsymbol{\beta}}_A^{(0,A)} - \gamma (\mathbf{X}_A' \mathbf{P}_n \mathbf{X}_A)^{-1} \mathbf{W}_A \mathbf{r},$$

where $\mathbf{r} \in \mathbf{R}^{|A|}$ and $||\mathbf{r}||_\infty \leq 1$. By assumption, $(\mathbf{X}_A' \mathbf{P}_n \mathbf{X}_A)^{-1} = O_p(1/n)$. Since $\mathbf{r}$ is bounded and $\gamma \mathbf{W}_A = o_p(\sqrt{n})$,

$$\gamma (\mathbf{X}_A' \mathbf{P}_n \mathbf{X}_A)^{-1} \mathbf{W}_A \mathbf{r} = o_p(1/\sqrt{n}).$$

It follows that $P(\{j;\ \hat{\beta}_j^* \neq 0\} = A) \to 1$ and

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_A^* - \hat{\boldsymbol{\beta}}_A^{(0,A)}) \xrightarrow{P} 0.$$

□

Table 1. Simulation results for $p = 2n$ and $V = 3$ with correctly specified working model (based on 500 simulations)

		$\beta_{0j} = 1$				$\beta_{0j} = 1.5$				$\beta_{0j} = 3$			
n	Method	MSE	FN	FP	PE	MSE	FN	FP	PE	MSE	FN	FP	PE
								$\rho = 0$					
50	DZ	1.199	0.491	0.347	2.960	2.062	0.047	0.345	3.997	2.633	0.014	0.337	4.717
	DZ-ADZ-1	1.361	0.793	0.062	3.161	1.585	0.131	0.047	3.335	1.558	0.043	0.045	3.457
	DZ-ADZ-INT	1.851	0.793	0.062	3.706	1.878	0.131	0.047	3.645	1.863	0.043	0.045	3.795
	DZ-ADZ-2	1.497	0.802	0.059	3.313	1.643	0.135	0.045	3.389	1.597	0.044	0.044	3.513
	Dantzig Cox	1.656	0.039	0.227	3.620	1.579	0.004	0.204	5.057	5.890	0.000	0.162	14.987
	ALASSO CLR	2.242	0.111	0.229	5.288	3.608	0.020	0.241	7.953	12.116	0.002	0.265	21.823
	ALASSO Cox	1.570	0.056	0.132	3.515	1.673	0.007	0.112	5.089	10.865	0.002	0.104	16.579
100	DZ	1.215	0.329	0.387	2.941	1.833	0.002	0.380	3.731	2.233	0.000	0.379	4.261
	DZ-ADZ-1	1.219	0.642	0.059	2.923	1.046	0.002	0.053	2.749	1.003	0.000	0.051	2.693
	DZ-ADZ-INT	1.753	0.642	0.059	3.485	1.592	0.002	0.053	3.299	1.545	0.000	0.051	3.241
	DZ-ADZ-2	1.393	0.650	0.056	3.104	1.199	0.003	0.051	2.902	1.143	0.000	0.048	2.837
	Dantzig Cox	1.108	0.000	0.229	3.071	0.871	0.000	0.210	4.472	2.349	0.000	0.170	13.465
	ALASSO CLR	1.857	0.007	0.239	5.097	3.238	0.000	0.254	7.857	11.472	0.000	0.281	23.932
	ALASSO Cox	1.103	0.000	0.132	3.069	1.089	0.000	0.115	4.593	8.866	0.000	0.109	14.909
								$\rho = 0.5$					
50	DZ	1.133	0.392	0.330	2.877	1.985	0.035	0.327	4.232	2.424	0.011	0.331	5.174
	DZ-ADZ-1	1.239	0.675	0.054	2.943	1.554	0.083	0.046	3.231	1.426	0.022	0.044	3.185
	DZ-ADZ-INT	1.701	0.675	0.054	3.420	1.825	0.083	0.046	3.535	1.686	0.022	0.044	3.462
	DZ-ADZ-2	1.351	0.682	0.051	3.050	1.652	0.093	0.044	3.313	1.485	0.023	0.042	3.241
	Dantzig Cox	1.462	0.015	0.189	4.292	1.291	0.000	0.173	7.615	3.365	0.000	0.135	26.603
	ALASSO CLR	2.120	0.078	0.214	6.841	3.964	0.032	0.241	11.929	14.764	0.006	0.297	38.532
	ALASSO Cox	1.175	0.020	0.097	4.258	1.763	0.007	0.087	7.802	13.703	0.007	0.086	29.529
100	DZ	1.278	0.249	0.378	3.044	1.849	0.003	0.373	4.088	2.226	0.001	0.366	4.820
	DZ-ADZ-1	1.122	0.506	0.053	2.837	1.009	0.004	0.047	2.663	0.996	0.001	0.046	2.719
	DZ-ADZ-INT	1.595	0.506	0.053	3.333	1.456	0.004	0.047	3.116	1.452	0.001	0.046	3.173
	DZ-ADZ-2	1.248	0.525	0.050	2.964	1.131	0.006	0.046	2.783	1.123	0.001	0.044	2.829
	Dantzig Cox	1.062	0.000	0.196	3.890	0.933	0.000	0.183	6.772	1.350	0.000	0.149	24.764
	ALASSO CLR	2.032	0.007	0.241	7.041	3.633	0.000	0.256	12.063	13.781	0.000	0.303	40.561
	ALASSO Cox	0.849	0.000	0.102	3.933	1.406	0.000	0.095	7.088	13.249	0.003	0.098	27.822
								$\rho = 0.9$					
50	DZ	3.192	0.517	0.248	3.197	4.174	0.192	0.256	4.056	4.914	0.135	0.249	4.635
	DZ-ADZ-1	2.065	0.709	0.042	2.820	3.991	0.365	0.042	3.091	4.481	0.260	0.042	3.135
	DZ-ADZ-INT	2.778	0.709	0.042	3.157	4.166	0.365	0.042	3.279	4.673	0.260	0.042	3.363
	DZ-ADZ-2	2.141	0.718	0.038	2.852	4.076	0.379	0.040	3.100	4.586	0.273	0.039	3.173
	Dantzig Cox	2.246	0.110	0.100	5.078	2.680	0.042	0.090	10.171	7.038	0.013	0.031	40.462
	ALASSO CLR	5.046	0.339	0.192	8.253	9.093	0.267	0.215	15.266	31.205	0.204	0.232	50.904
	ALASSO Cox	1.556	0.171	0.040	5.070	3.033	0.161	0.038	10.417	17.637	0.205	0.041	42.760
100	DZ	4.106	0.397	0.322	3.337	4.912	0.097	0.319	4.295	5.634	0.057	0.314	5.271
	DZ-ADZ-1	0.942	0.656	0.020	2.192	2.137	0.213	0.023	2.359	2.070	0.104	0.022	2.389
	DZ-ADZ-INT	1.267	0.656	0.020	2.390	2.161	0.213	0.023	2.526	2.109	0.104	0.022	2.544
	DZ-ADZ-2	0.974	0.673	0.018	2.194	2.173	0.222	0.022	2.382	2.114	0.111	0.021	2.409
	Dantzig Cox	1.260	0.022	0.100	4.915	1.431	0.001	0.094	9.405	3.937	0.000	0.019	38.290
	ALASSO CLR	4.788	0.198	0.197	7.992	8.997	0.139	0.229	15.459	27.901	0.094	0.231	56.400
	ALASSO Cox	1.040	0.059	0.041	4.880	2.516	0.070	0.040	9.680	17.538	0.101	0.043	40.903

DZ: one-stage Dantzig selector defined in (7);
DZ-ADZ-1: one-iteration adaptive Dantzig selector at the second stage (defined in section 4.3);
DZ-ADZ-INT: intermediate adaptive Dantzig selector at the second stage (defined in section 4.3);
DZ-ADZ-2: two-iteration adaptive Dantzig selector at the second stage (defined in section 4.3);
Dantzig Cox: the Dantzig selector proposed in Antoniadis et al. (2009);
ALASSO CLR: Adaptive LASSO for censored linear regression (Datta et al., 2007);
ALASSO Cox: Adaptive LASSO for Cox PH models (Zhang and Lu, 2007)

Table 2. Simulation results for $p = 2n$ and $V = 5$ with correctly specified working model (based on 500 simulations)

		$\beta_{0j} = 1$				$\beta_{0j} = 1.5$				$\beta_{0j} = 3$			
n	Method	MSE	FN	FP	PE	MSE	FN	FP	PE	MSE	FN	FP	PE
								$\rho = 0$					
50	DZ	1.278	0.491	0.344	3.030	3.217	0.090	0.341	5.238	4.314	0.041	0.338	6.620
	DZ-ADZ-1	1.486	0.798	0.060	3.203	2.825	0.253	0.043	4.568	2.964	0.131	0.045	4.719
	DZ-ADZ-INT	1.911	0.798	0.060	3.669	3.000	0.253	0.043	4.753	3.189	0.131	0.045	4.952
	DZ-ADZ-2	1.591	0.806	0.057	3.321	2.865	0.256	0.042	4.611	2.986	0.132	0.044	4.733
	Dantzig Cox	2.432	0.065	0.213	4.860	4.034	0.023	0.193	8.081	19.825	0.004	0.151	26.028
	ALASSO CLR	3.405	0.125	0.242	7.012	6.423	0.064	0.261	11.670	22.966	0.020	0.295	39.186
	ALASSO Cox	2.388	0.106	0.122	4.817	4.710	0.068	0.113	8.465	28.240	0.055	0.106	30.878
100	DZ	1.341	0.319	0.384	3.109	2.586	0.005	0.377	4.634	3.378	0.001	0.373	5.682
	DZ-ADZ-1	1.387	0.644	0.059	3.105	1.252	0.008	0.052	2.958	1.324	0.000	0.052	3.058
	DZ-ADZ-INT	1.924	0.644	0.059	3.683	1.778	0.008	0.052	3.505	1.882	0.000	0.052	3.609
	DZ-ADZ-2	1.561	0.651	0.057	3.287	1.367	0.008	0.049	3.075	1.425	0.000	0.049	3.156
	Dantzig Cox	1.271	0.000	0.215	4.032	1.621	0.000	0.193	6.589	11.070	0.000	0.148	22.571
	ALASSO CLR	2.828	0.019	0.242	7.082	5.332	0.003	0.264	11.457	20.044	0.000	0.298	36.733
	ALASSO Cox	1.306	0.002	0.120	4.051	2.369	0.001	0.110	6.979	22.879	0.000	0.109	26.190
								$\rho = 0.5$					
50	DZ	1.272	0.390	0.324	3.114	2.804	0.062	0.321	5.774	3.697	0.028	0.318	7.621
	DZ-ADZ-1	1.434	0.676	0.049	3.201	2.124	0.125	0.037	3.761	1.995	0.049	0.035	3.697
	DZ-ADZ-INT	1.811	0.676	0.049	3.588	2.230	0.125	0.037	3.936	2.137	0.049	0.035	3.865
	DZ-ADZ-2	1.517	0.690	0.047	3.280	2.202	0.128	0.036	3.826	2.022	0.050	0.034	3.697
	Dantzig Cox	1.587	0.025	0.168	7.093	2.321	0.007	0.146	14.359	14.647	0.000	0.104	57.015
	ALASSO CLR	3.450	0.113	0.234	11.317	7.095	0.071	0.262	21.071	27.457	0.041	0.307	80.614
	ALASSO Cox	1.844	0.062	0.086	7.232	4.816	0.062	0.083	15.583	30.634	0.087	0.089	64.889
100	DZ	1.397	0.234	0.379	3.222	2.450	0.006	0.358	5.396	3.163	0.001	0.342	6.852
	DZ-ADZ-1	1.338	0.492	0.052	2.992	1.232	0.008	0.045	2.927	1.203	0.001	0.044	2.947
	DZ-ADZ-INT	1.780	0.492	0.052	3.452	1.624	0.008	0.045	3.329	1.611	0.001	0.044	3.335
	DZ-ADZ-2	1.467	0.508	0.050	3.106	1.335	0.009	0.043	3.010	1.283	0.001	0.042	2.998
	Dantzig Cox	0.990	0.000	0.179	6.379	1.020	0.000	0.160	12.861	8.237	0.000	0.117	51.953
	ALASSO CLR	3.225	0.022	0.260	11.255	6.638	0.008	0.293	21.302	24.818	0.004	0.304	75.416
	ALASSO Cox	1.314	0.004	0.094	6.637	3.947	0.004	0.090	14.113	29.309	0.013	0.096	59.175
								$\rho = 0.9$					
50	DZ	3.094	0.504	0.248	3.143	5.733	0.245	0.243	5.856	8.133	0.214	0.229	8.080
	DZ-ADZ-1	2.633	0.720	0.040	2.818	5.806	0.393	0.036	3.339	7.250	0.320	0.035	3.588
	DZ-ADZ-INT	3.206	0.720	0.040	3.072	5.918	0.393	0.036	3.518	7.176	0.320	0.035	3.734
	DZ-ADZ-2	2.719	0.729	0.036	2.841	5.937	0.408	0.033	3.358	7.295	0.328	0.033	3.589
	Dantzig Cox	3.414	0.139	0.082	11.468	4.364	0.114	0.051	25.716	21.864	0.086	0.013	117.025
	ALASSO CLR	8.954	0.422	0.208	17.213	18.518	0.388	0.226	34.293	112.656	0.226	0.458	163.348
	ALASSO Cox	2.896	0.303	0.034	11.765	7.044	0.338	0.036	27.141	36.073	0.413	0.043	123.750
100	DZ	4.134	0.402	0.320	3.417	6.719	0.147	0.297	6.546	8.010	0.099	0.257	8.220
	DZ-ADZ-1	1.522	0.664	0.021	2.300	3.544	0.261	0.020	2.532	3.873	0.162	0.020	2.526
	DZ-ADZ-INT	1.791	0.664	0.021	2.465	3.460	0.261	0.020	2.659	3.799	0.162	0.020	2.652
	DZ-ADZ-2	1.614	0.678	0.019	2.318	3.590	0.268	0.020	2.555	3.915	0.165	0.019	2.541
	Dantzig Cox	1.944	0.033	0.088	10.775	2.279	0.014	0.053	24.271	15.829	0.005	0.002	110.641
	ALASSO CLR	8.790	0.300	0.221	17.461	17.472	0.279	0.227	36.905	123.400	0.100	0.544	169.164
	ALASSO Cox	2.437	0.200	0.038	11.165	6.661	0.222	0.041	25.843	35.505	0.273	0.044	116.626

DZ: one-stage Dantzig selector defined in (7);
DZ-ADZ-1: one-iteration adaptive Dantzig selector at the second stage (defined in section 4.3);
DZ-ADZ-INT: intermediate adaptive Dantzig selector at the second stage (defined in section 4.3);
DZ-ADZ-2: two-iteration adaptive Dantzig selector at the second stage (defined in section 4.3);
Dantzig Cox: the Dantzig selector proposed in Antoniadis et al. (2009);
ALASSO CLR: Adaptive LASSO for censored linear regression (Datta et al., 2007)
ALASSO Cox: Adaptive LASSO for Cox PH models (Zhang and Lu, 2007)

Table 3. Simulation results for the misspecified working model with $n = 50$ and $p = 2n$ (based on 500 simulations)

Method	$\beta_{0j} = 1$			$\beta_{0j} = 1.5$			$\beta_{0j} = 3$		
	FN	FP	PE	FN	FP	PE	FN	FP	PE
					$V = 3$				
DZ	0.404	0.239	1.502	0.213	0.268	0.943	0.026	0.301	4.265
DZ-ADZ-1	0.682	0.04	1.648	0.495	0.045	0.951	0.129	0.05	3.24
DZ-ADZ-INT	0.682	0.04	1.648	0.495	0.045	0.951	0.129	0.05	3.24
DZ-ADZ-2	0.682	0.04	1.648	0.495	0.045	0.951	0.129	0.05	3.24
Dantzig Cox	0.201	0.24	1.389	0.09	0.23	1.009	0.003	0.2	7.348
ALASSO CLR	0.638	0.151	1.232	0.475	0.168	8.726	0.16	0.201	6.04
ALASSO Cox	0.251	0.161	1.494	0.102	0.146	1.232	0.004	0.104	7.473
					$V = 5$				
DZ	0.345	0.278	1.097	0.19	0.298	2.172	0.054	0.293	8.89
DZ-ADZ-1	0.66	0.046	1.48	0.494	0.048	2.133	0.192	0.045	4.866
DZ-ADZ-INT	0.66	0.046	1.48	0.494	0.048	2.133	0.192	0.045	4.866
DZ-ADZ-2	0.66	0.046	1.48	0.494	0.048	2.133	0.192	0.045	4.866
Dantzig Cox	0.219	0.227	1.731	0.098	0.21	1.646	0.01	0.171	10.159
ALASSO CLR	0.599	0.169	1.478	0.42	0.188	1.689	0.188	0.224	11.076
ALASSO Cox	0.251	0.146	1.713	0.116	0.121	1.728	0.015	0.085	10.795

DZ: one-stage Dantzig selector defined in (7);
DZ-ADZ-1: one-iteration adaptive Dantzig selector at the second stage (defined in section 4.3);
DZ-ADZ-INT: intermediate adaptive Dantzig selector at the second stage (defined in section 4.3);
DZ-ADZ-2: two-iteration adaptive Dantzig selector at the second stage (defined in section 4.3);
Dantzig Cox: the Dantzig selector proposed in Antoniadis et al. (2009);
ALASSO CLR: Adaptive LASSO for censored linear regression (Datta et al., 2007)
ALASSO Cox: Adaptive LASSO for Cox PH models (Zhang and Lu, 2007)

Table 4. Final models for the multiple myeloma training dataset selected by the adaptive Dantzig selectors

Probeset	DZ-ADZ-1	DZ-ADZ-INT	DZ-ADZ-2
202075_s_at	-0.03	-0.07	-0.01
206871_at	0.10	0.14	0.10
211674_x_at	-0.06	-0.10	-0.08
213674_x_at	0.05	0.11	0.09
214777_at	0.02	0.05	0.01
225626_at	0.05	0.09	0.07

DZ-ADZ-1: one-iteration adaptive Dantzig selector at the second stage (defined in section 4.3);
DZ-ADZ-INT: intermediate adaptive Dantzig selector at the second stage (defined in section 4.3);
DZ-ADZ-2: two-iteration adaptive Dantzig selector at the second stage (defined in section 4.3)

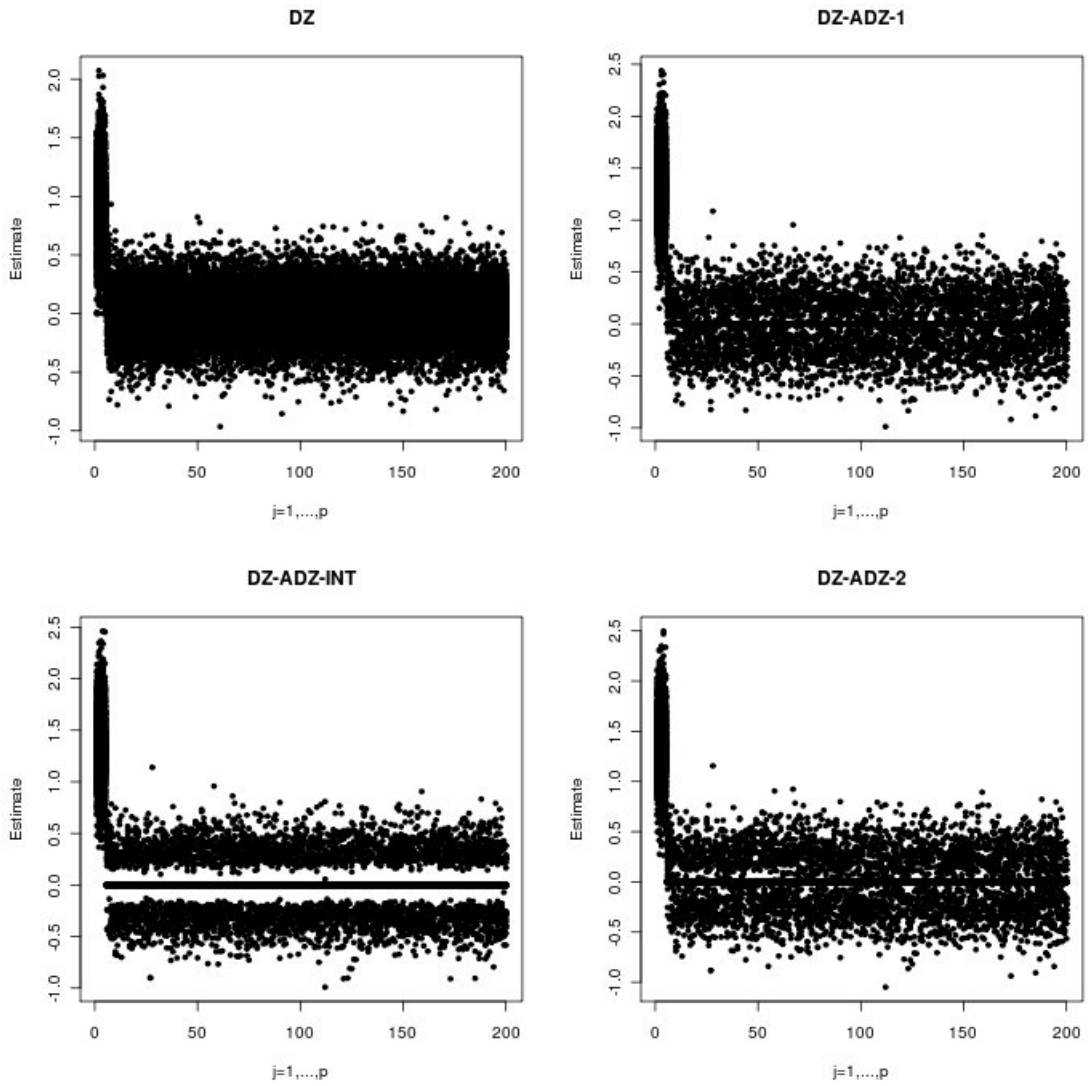

Figure 1: Results of 500 simulations of the one-stage Dantzig selector for the censored linear regression model with $n = 100$, $p = 200$, $\beta_{0j} = 1.5$, and $\rho = 0.5$

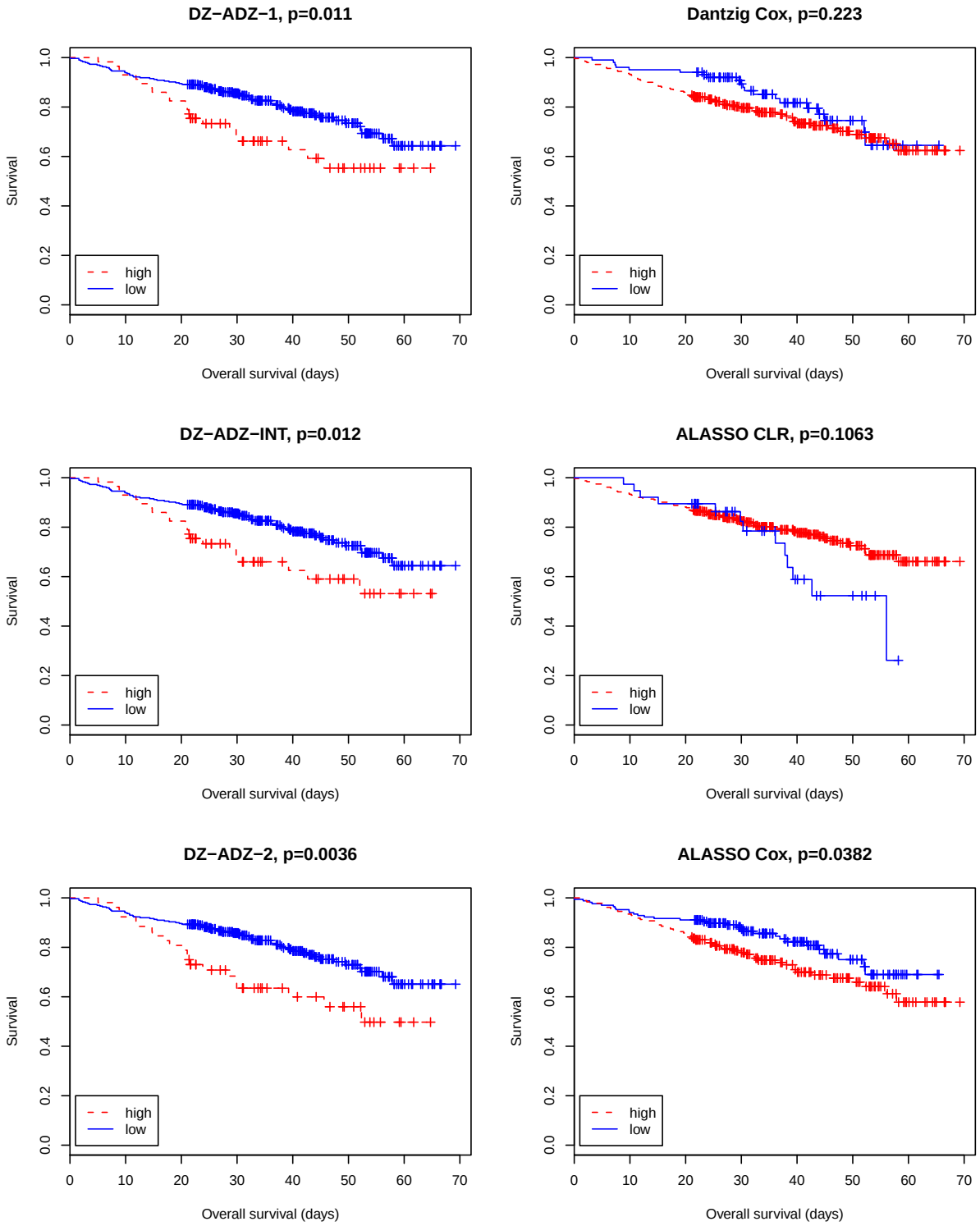

Figure 2: Survival comparison between the high/low risk groups using various selectors on an independent validation dataset (the high or low risk is defined based on whether the model-based predicted risk exceeds the median value in the training dataset)