The Dantzig Selector for Censored Linear Regression Models

Yi Li[1], Lee Dicker[2], and Sihai Dave Zhao[3]

[1] *University of Michigan,* [2] *Rutgers University,* [3] *Harvard University*

Abstract: The Dantzig variable selector has recently emerged as a powerful tool for fitting regularized regression models. To our knowledge, most work involving the Dantzig selector has been performed with fully-observed response variables. This paper proposes a new class of adaptive Dantzig variable selectors for linear regression models when the response variable is subject to right censoring. This is motivated by a clinical study to identify genes predictive of event-free survival in newly diagnosed multiple myeloma patients. Under some mild conditions, we establish the theoretical properties of our procedures, including consistency in model selection (i.e. the right subset model will be identified with a probability tending to 1) and the optimal efficiency of estimation (i.e. the asymptotic distribution of the estimates is the same as that when the true subset model is known a priori). The practical utility of the proposed adaptive Dantzig selectors is verified via extensive simulations. We apply our new methods to the aforementioned myeloma clinical trial and identify important predictive genes.

Key words and phrases: Buckley-James imputation; Censored linear regression; Dantzig selector; Oracle property.

1. Introduction

Technical advances in biomedicine have produced an abundance of high-throughput data. This has resulted in major statistical challenges and brought 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 gene therapy (Anderson et al., 2005; Pawitan et al., 2005).

Penalized likelihood methods, represented by the LASSO, have been extensively studied as a means of simultaneous estimation and variable selection (Tibshirani, 1996). It is known that the LASSO estimator can discover the right

sparse representation of the model (Zhao and Yu, 2006); 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 (Candés and Tao, 2007), has emerged to enrich the class of regularization techniques. The Dantzig selector can be implemented as a linear programming problem, making the computational burden manageable. Though under some general conditions the LASSO and Dantzig may produce the same solution path (James et al., 2008), they differ conceptually in that the Dantzig stems directly from an estimating equation, whereas the LASSO stems from a likelihood or an objective function.

The Dantzig selector has been most thoroughly studied with fully observed outcome variables. But 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 motivates this research, the goal was to identify genes predictive of a patient's event-free survival.

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; Antoniadis et al., 2010), 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 on regularized linear regression models for censored data can be found in Ma et al. (2006), Johnson et al. (2008), Wang et al. (2008), Cai et al. (2009), Engler and Li (2009), and Johnson (2009).

Most of these methods operate under the penalization framework. Given that a censored linear regression does not pertain to a likelihood function, the Dantzig selector may be a natural choice. Johnson et al. (2008) approached the problem using a penalized estimation equation approach, but Johnson (2009) noted that

their procedure gives only an approximate *root-n consistent estimator. To our knowledge, it remains unclear whether the Dantzig selector can also be used to estimate linear regression models with censored outcome data. Johnson et al. (2011) studied such a procedure but did not provide theoretical support. It is therefore of interest to (i) explore the utility of the Dantzig selector in censored linear regression models, (ii) rigorously evaluate its theoretical properties, and (iii) compare its numerical numerical properties to similar methods developed under the lasso/penalization-based framework.*

This paper proposes a new class of Dantzig variable selectors for linear regression models when the response variable is subject to right censoring. Dicker (2011) proposed the adaptive Dantzig selector for the linear model, and here we develop a similar procedure for use with censored outcomes. First, our proposed method carries out simultaneous variable selection and estimation, and is motivated from the estimating equation perspective, which may be important for some semiparametric models whose likelihood functions are often difficult to specify. Second, the proposed selectors possess the oracle property when the tuning parameters follow some appropriate rates, providing the theoretical justification for the proposed procedures. Thirdly, 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 Dantzig selector for noncensored linear regression models, as well as its connection with the penalized likelihood methods. Section 3 considers its extension to the linear regression models when the response variable is subject to censoring. In Section 4, we discuss the large sample properties and prove the consistency of variable selection and the optimal efficiency of the estimators. We discuss the choice of tuning parameters for the finite sample situations in Section 5. We conduct numerical simulations in Section 6 and apply the proposal to a myeloma study in Section 7. We conclude the paper with a discussion in Section 8. All the technical proofs are relegated to a web supplement.

2. Penalized Likelihood Methods and the Dantzig Selector

We begin by considering a linear regression model with p predictors

$$Y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \epsilon_i, \tag{2.1}$$

where ϵ_i are iid mean zero residuals for $i = 1, \ldots, n$. Denote the truth by $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})$ and define $A = \{j : \beta_{0j} \neq 0\}$. The goal of the model selection in this context is to identify A, often referred to as the "true model."

A variable selector $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}_0$ 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 $\boldsymbol{\beta}_A$ is the subvector of $\boldsymbol{\beta}$ extracted by the subset A of $\{1, \ldots, p\}$ and Σ^* is some $|A| \times |A|$ covariance matrix (here, $|A|$ denotes the cardinality of the set A). 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.

Concise notation is to be used for referring to sub-vectors and sub-matrices. For a subset $T \subset \{1, \ldots, p\}$ and $\boldsymbol{\beta} \in \mathbf{R}^p$, let $\boldsymbol{\beta}_T = (\beta_j)_{j \in T}$ be the $|T| \times 1$ vector whose entries are those of $\boldsymbol{\beta}$ indexed by T. For an $n \times p$ matrix, $\mathbf{X}$, $\mathbf{X}_T$ is the $n \times |T|$ matrix whose columns are those of $\mathbf{X}$ that are indexed by T. Additionally, let $\mathbf{X}_i$ and $\mathbf{X}_{\cdot j}$ denote the i^{th} row and j^{th} column of $\mathbf{X}$, respectively, for $i = 1, ..., n$ and $j = 1, ..., p$. Denote the complement of T in $\{1, ..., p\}$ by $\bar{T}$. Other common notation includes the norms $\|\boldsymbol{\beta}\|_r = (\sum_{i=1}^{p} |\beta_i|^r)^{1/r}$ for $0 < r < \infty$, $\|\boldsymbol{\beta}\|_0 = \#\{j : \beta_j \neq 0\}$ and $\|\boldsymbol{\beta}\|_\infty = \max_{1 \leq j \leq p} |\beta_j|$; and $\mathrm{sgn}(\boldsymbol{\beta})$, the sign vector corresponding to $\boldsymbol{\beta}$, where $\mathrm{sgn}(\boldsymbol{\beta})_j = \mathrm{sgn}(\beta_j)$ (by definition, $\mathrm{sgn}(0) = 0$). For a diagonal matrix $\mathbf{W} = \mathrm{diag}(w_1, ..., w_p)$, we define $\mathbf{W}_{T,T} = \mathrm{diag}(w_j; j \in T)$.

2.1. Penalized Likelihood Methods

The LASSO is a benchmark penalized likelihood procedure. LASSO works by minimizing an L_2 loss function $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ subject to an L_1 constraint: $\|\boldsymbol{\beta}\|_1 = \sum_j |\beta_j| \leq s$, where $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is the response vector, $\mathbf{X}$ is the $n \times p$ design matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the $p \times 1$ vector of coefficients and s is a nonnegative tuning parameter. Equivalently, the LASSO estimate can be

obtained by minimizing

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.2}$$

where λ is a nonnegative tuning parameter. It is known that the LASSO performs variable selection, but in general does not possess the oracle property. A remedy is to utilize an adaptive LASSO that minimizes

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j|, \tag{2.3}$$

where w_j is a data-driven weight. For example, we can take $w_j = |\hat{\beta}_j^{(0)}|^{-\gamma}$ for $\gamma > 0$ and where $\hat{\boldsymbol{\beta}}^{(0)}$ is some $\sqrt{n}$-consistent estimate of β_0, such as the ordinary least square estimate when $n \geq p$. Note that w_j tends to be large when the true $\beta_{0j} = 0$; in this case, the nonzero estimates of $\beta_{j,0}$ are heavily penalized in (2.3). Conversely, if $\beta_{j,0} \neq 0$, then w_j tends to be small which ensures that the nonzero estimates of $\beta_{j,0}$ are moderately penalized in (2.3). Obtaining these initial estimates is much more difficult when $p > n$, and for that we defer our discussion to Section 3.2.

2.2. Adaptive Dantzig Selector

Derived directly from the score equations, the Dantzig selector also belongs to the class of regularization methods in regression. Specifically, the Dantzig selector estimator is defined to be the solution to

$$\begin{aligned} &\text{minimize} && \|\boldsymbol{\beta}\|_1 \\ &\text{subject to} && \|\mathbf{X}'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\|_\infty \leq \lambda, \end{aligned}$$

which strikes a balance between nearly solving the score equation, $\mathbf{X}'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = 0$ and minimizing the L_1 norm of $\boldsymbol{\beta}$. The Dantzig selector and LASSO are closely related. Connections between the Dantzig selector and the LASSO have been discussed in James et al. (2008), where it is shown that under some general conditions the Dantzig selector and the LASSO produce the same solution path. Note that in general the Dantzig selector does not have the oracle property.

As a remedy, a modified Dantzig selector, analogous to the adaptive LASSO, is proposed by Dicker (2011):

$$\begin{aligned} &\text{minimize} && \textstyle\sum_j w_j |\beta_j| \\ &\text{subject to} && |\mathbf{X}'_{\cdot j}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})| \leq \lambda w_j, j = 1, \ldots, p. \end{aligned}$$

As the Dantzig selector and LASSO are related, so are the adaptive Dantzig selector and adaptive LASSO. Indeed, the adaptive Dantzig selector and adaptive LASSO are equivalent to instances of the Dantzig selector and LASSO, respectively, where $\mathbf{X}$ is replaced with $\mathbf{X}\mathbf{W}^{-1}$, $\boldsymbol{\beta}$ is replaced with $\mathbf{W}\boldsymbol{\beta}$, and $\mathbf{W} = \text{diag}(w_1, ..., w_p)$. The key to the adaptive Dantzig selector is to strike a balance between minimizing the weighted L_1 norm, which promotes sparsity, and approximately solving the weighted normal equations. Weights w_j in the adaptive Dantzig selector should be chosen according to the same principles which determine weights in the adaptive LASSO. When the response vector $\mathbf{Y}$ is fully observed, Dicker (2011) established the oracle property of the adaptive Dantzig selector for an appropriately chosen tuning parameter λ. It is unclear, however, whether this property hold when the response $\mathbf{Y}$ is subject to censoring.

3. Adaptive Dantzig Selector for Censored Linear Regression

Consider a slightly modified version of (2.1),

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

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 denoted by $F(\cdot)$, with survival function $S(\cdot) = 1 - F(\cdot)$. The mean of ϵ_i, denoted by α, is not necessarily 0. As above, $\boldsymbol{\beta}_0$ denotes the true $\boldsymbol{\beta}$ and $A = \{j;\ \beta_{0j} \neq 0\}$ is the true model. Suppose that 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, both Y_i and C_i are commonly measured on the log scale, and the model is called the accelerated failure time model (Kalbfleisch and Prentice 2002).

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

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

Clearly,

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

The Buckley-James estimating equation is

$$\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)\left\{\hat{Y}_i(\boldsymbol{\beta}) - \mathbf{X}_i'\boldsymbol{\beta}\right\} = 0,\ j = 1, \ldots, p, \tag{3.1}$$

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

$$\hat{Y}_i(\boldsymbol{\beta}) = Y_i^* + (1-\delta_i)\frac{\int_{e_i(\boldsymbol{\beta})}^{\infty} \hat{S}(s,\boldsymbol{\beta})ds}{\hat{S}\{e_i(\boldsymbol{\beta}),\boldsymbol{\beta}\}} \tag{3.2}$$

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 $(e_i(\boldsymbol{\beta}),\delta_i)$,

$$\hat{S}(t,\boldsymbol{\beta}) = \exp\left\{-\sum_{i=1}^n \int_{-\infty}^{t} \frac{dN_i(u,\boldsymbol{\beta})}{\bar{Y}(u,\boldsymbol{\beta})}\right\}, \tag{3.3}$$

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, which solves (3.1), is $\sqrt{n}$-consistent. To facilitate the ensuing development, note that (3.1) can be written in a more compact form

$$\mathbf{X}'\mathbf{P}_n\{\hat{\mathbf{Y}}(\boldsymbol{\beta}) - \mathbf{X}\boldsymbol{\beta}\} = 0, \tag{3.4}$$

where $\mathbf{P}_n = \mathbf{I}_n - \mathbf{1}\mathbf{1}'/n$, $\mathbf{I}_n$ is an $n \times n$ identity matrix, $\mathbf{1}$ is an $n \times 1$ vector with all elements being 1 and $\hat{\mathbf{Y}}(\boldsymbol{\beta}) = (\hat{Y}_1(\boldsymbol{\beta}), \ldots, \hat{Y}_n(\boldsymbol{\beta}))'$.

Solving (3.4) does not directly render an automatic variable selection procedure. But because the adaptive Dantzig selector is derived directly from score equations instead of a loss function, it naturally presents an appealing solution to our problem. Applying it to (3.4) gives

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

Unfortunately, this is no longer a simple linear programming problem, because (3.4) is a discrete estimating equation. One strategy uses an iterative algorithm, starting with initial value of $\boldsymbol{\beta}$, as in Wang et al. (2008), but such methods have numerical and theoretical difficulties (Johnson 2009).

3.1. Low-dimensional Setting

Instead of implementing the adaptive Dantzig selector with the true estimating equation (3.4), we propose to use a $\sqrt{n}$-consistent initial estimator $\boldsymbol{\beta}_0$, denoted $\hat{\boldsymbol{\beta}}^{(0)}$ to construct an imputed version of the true response $\mathbf{Y}$, denoted $\hat{\mathbf{Y}}(\hat{\boldsymbol{\beta}}^{(0)})$; we then employ a version of the adaptive Dantzig selector replacing (3.4) with $\mathbf{X}'\mathbf{P}_n\{\hat{\mathbf{Y}}(\hat{\boldsymbol{\beta}}^{(0)}) - \mathbf{X}\boldsymbol{\beta}\} = 0$. In the low-dimensional setting, where

$p < n$, we can obtain $\boldsymbol{\beta}_0$ using unpenalized Buckley-James estimation, or rank-based procedures with Gehan or logrank weights (Jin et al., 2003). A similar one-step imputation strategy was used by Johnson (2009), though there the imputed $\hat{\mathbf{Y}}(\hat{\boldsymbol{\beta}}^{(0)})$ were used to construct a loss function which was then used with a LASSO penalty. In contrast, here we proceed directly from the estimating equation.

Our new version of the adaptive Dantzig selector is thus given by

$$\begin{array}{ll} \text{minimize} & \sum_j w_j|\beta_j| \\ \text{subject to} & |\mathbf{X}'_{\cdot j}\mathbf{P}_n\{\hat{\mathbf{Y}}(\hat{\boldsymbol{\beta}}^{(0)}) - \mathbf{X}\boldsymbol{\beta}\}| \leq \lambda w_j,\ j = 1, ..., p. \end{array} \tag{3.5}$$

Again, w_j are data driven weights and should be chosen to vary inversely with the magnitude of β_{0j}. For instance, as with the adaptive Dantzig selector, we can take $w_j = |\hat{\beta}_j^{(0)}|^{-\gamma}$ for some $\gamma > 0$. Then when $|\hat{\beta}_j^{(0)}|$ is large, (3.5) 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, and heavily penalizes non-zero estimates of β_{0j} when $|\hat{\beta}_j^{(0)}|$ is small. Note that when the response is fully observed, that is $\delta_i \equiv 1$ for all i, the imputation-based Dantzig selector, (3.5) reduces to the adaptive Dantzig selector for linear regression models and the result is an effective variable selection procedure. However, the censoring present difficulties that are worth investigating.

3.2. High-dimensional Setting

In the high-throughput datasets that now characterize modern medicine, however, it is rare that the number of covariates is smaller than the sample size. But in this high-dimensional setting with $p > n$, it is difficult to obtain initial $\sqrt{n}$-consistent estimates $\hat{\boldsymbol{\beta}}^{(0)}$. Our strategy here is to first reduce the number of the covariates to be smaller than n using a sure screening procedure. We then calculate the $\hat{\boldsymbol{\beta}}^{(0)}$ using the retained covariates and proceed as in Section 3.1.

Specifically, we employ the screening procedure of Zhu et al. (2011), which can provide sure screening for any single-index model, which includes the AFT model. To choose the number of covariates to retain after screening, we follow the combined soft- and hard-thresholding rule of Zhu et al. (2011), which chooses up to $n/\log(n)$ covariates using a procedure involving randomly generated auxiliary variables.

Recently, Johnson et al. (2011) also studied Buckley-James estimation using

the Dantzig selector with an initial $\sqrt{n}$-consistent estimator. To deal with the high-dimensional problem, they proposed a strategy they termed "B initialization", in which they select a subset of important covariates to use in calculating the initial estimate. While this is similar to our proposal, our work has several advantages. First, we propose an adaptive Dantzig selector, which has practical and theoretical advantages over the nonadaptive version. Second, by using the procedure of Zhu et al. (2011) to select the subset of important covariates, we can take advantage of the sure screening property of their method, which states that under certain conditions on the design matrix, the probability that the selected covariates contains the truly important covariates approaches 1.

4. Theoretical Results

A fundamental difficulty of extending the Dantzig selector to the censored regression setting is that $\hat{Y}_i(\hat{\boldsymbol{\beta}}^{(0)})$ is only a surrogate for the unobserved outcome Y_i. This prevents 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 the true Y_i is bounded by a random variable of order $n^{-1/2}$. This turns out to be essential for establishing the oracle property 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 oracle property.

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

Before stating the main result of the section, we state a lemma, which implies that even though $\hat{S}$, defined in (3.3), is a discontinuous function, a first order asymptotic linearization exists. This is useful for bounding the difference between the surrogate and the true outcomes.

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,*

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

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* $W(v(s))$. *Here,* $W(\cdot)$ *is the Wiener process and* $v(\cdot)$ *is defined in (S1.2) in the web supplement.*

Proposition 1 *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 1 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. Selection-consistent Adaptive Dantzig Selector

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, (3.5), can be rewritten compactly as

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

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

Lemma 2 *If there is* $\hat{\boldsymbol{\mu}} \in \mathbf{R}^p$ *such that,*

$$\begin{aligned} \|\mathbf{Z}'(\hat{\mathbf{Y}} - \mathbf{ZW}\hat{\boldsymbol{\beta}})\|_\infty &\leq \lambda, & (4.1)\\ \|\mathbf{Z}'\mathbf{Z}\hat{\boldsymbol{\mu}}\|_\infty &\leq 1, & (4.2)\\ \hat{\boldsymbol{\mu}}'\mathbf{Z}'\mathbf{ZW}\hat{\boldsymbol{\beta}} &= \|\mathbf{W}\hat{\boldsymbol{\beta}}\|_1, & (4.3)\\ \hat{\boldsymbol{\mu}}'\mathbf{Z}'(\hat{\mathbf{Y}} - \mathbf{ZW}\hat{\boldsymbol{\beta}}) &= \lambda\|\hat{\boldsymbol{\mu}}\|_1, & (4.4) \end{aligned}$$

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

The parameter $\boldsymbol{\mu}$ in Lemma 2 is the dual variable and may be viewed a a Lagrangian multiplier. Inequalities (4.1) and (4.2) correspond to primal and dual feasibility respectively, while (4.3) and (4.4) concerns with complementary slackness. Inspecting (4.1)–(4.4), the following proposition proves that (SADS) is selection consistent, provided λ and $(w_1, ..., w_p)$ follow an appropriate rate.

Proposition 2 *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*

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

Then, with probability tending to 1, a solution to (SADS), $\hat{\boldsymbol{\beta}}$*, and the corresponding* $\hat{\boldsymbol{\mu}}$ *from Lemma 2 are given by*

$$\begin{aligned} \hat{\boldsymbol{\mu}}_A &= (\mathbf{Z}'_A\mathbf{Z}_A)^{-1}\mathrm{sgn}(\boldsymbol{\beta}_0)_A & (4.5)\\ \hat{\boldsymbol{\mu}}_{\bar{A}} &= 0 & (4.6) \end{aligned}$$

and

$$\begin{aligned} \hat{\boldsymbol{\beta}}_A &= \mathbf{W}^{-1}_{A,A}\left\{(\mathbf{Z}'_A\mathbf{Z}_A)^{-1}\mathbf{Z}'_A\hat{\mathbf{Y}} - \lambda(\mathbf{Z}'_A\mathbf{Z}_A)^{-1}\mathrm{sgn}(\hat{\boldsymbol{\mu}})_A\right\}\\ &= (\mathbf{X}'_A\mathbf{P}_n\mathbf{X}_A)^{-1}\mathbf{X}'_A\mathbf{P}_n\hat{\mathbf{Y}} - \lambda(\mathbf{X}'_A\mathbf{P}_n\mathbf{X}_A)^{-1}\mathbf{W}_{A,A}\mathrm{sgn}(\hat{\boldsymbol{\mu}})_A & (4.7)\\ \hat{\boldsymbol{\beta}}_{\bar{A}} &= 0. & (4.8) \end{aligned}$$

Corollary 1 (consistency of model selection) *Suppose that the conditions of Proposition 2 hold and let* $\hat{\boldsymbol{\beta}}$ *be any sequence of solutions to (SADS). Then* $P(\{j;\ \hat{\beta}_j \neq 0\} = A) \to 1$.

We make a few remarks about Proposition 2 and Corollary 1. First, to ensure that the conditions in Proposition 2 hold, one selects data-driven weights w_j and an appropriate λ. Examples of weights and λ such that these conditions hold include $w_j = |\hat{\boldsymbol{\beta}}^{(0)}|^{-\gamma}$, where $\hat{\boldsymbol{\beta}}^{(0)}$ is $\sqrt{n}$-consistent for $\boldsymbol{\beta}_0$ and $\gamma > 0$, and λ such that $n^{-1/2}\lambda = O(1)$ and $n^{(\gamma-1)/2}\lambda \to \infty$. Also note that though Proposition 2 makes no uniqueness claims about solutions to (SADS), it can be shown that in "most" cases (SADS) has a unique solution (Dicker 2011). Furthermore, Corollary 1 states that regardless of whether or not there is a unique solution, (SADS) is consistent for model selection.

4.3. Oracle Adaptive Dantzig Selector

The estimator defined in (4.7) and (4.8) solves (SADS) in probability. This expression may be leveraged to obtain the large-sample distribution of $\sqrt{n}$-standardized (SADS) estimates. However, though the solution to (SADS) is selection consistent, it may not achieve optimal efficiency. In other words, the variances of the nonzero components of $\hat{\boldsymbol{\beta}}$ could be larger than the corresponding variances obtained from the oracle estimator. To remedy this, we propose the oracle adaptive Dantzig selector (OADS), which does 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 SADS estimator $\hat{\boldsymbol{\beta}}$. Define the OADS estimator $\hat{\boldsymbol{\beta}}^{(0,T)}$ so that $\hat{\boldsymbol{\beta}}^{(0,T)}_{\bar{T}} = 0$ and $\hat{\boldsymbol{\beta}}^{(0,T)}_{T}$ is the Buckley-James estimate obtained by solving (3.4) 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 SADS estimator. This is similar to the Gauss-Dantzig selector of Candés and Tao (2007), in which ordinary linear regression is performed on the covariates selected by the Dantzig selector. As summarized in the following proposition, $\hat{\boldsymbol{\beta}}^{(0,T)}$ achieves the oracle property.

Proposition 3 (oracle property) *Assume that the conditions of Proposition 2 hold. Let $T = \{j;\ \hat{\beta}_j \neq 0\}$, where $\hat{\boldsymbol{\beta}}$ is the SADS 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 $\hat{\boldsymbol{\beta}}^{(0,A)}_{\bar{A}} = 0$ and $\hat{\boldsymbol{\beta}}^{(0,A)}_{A}$ is the Buckley-James estimate obtained by solving (3.4) with $\mathbf{X}$ replaced by $\mathbf{X}_A$. Then the OADS estimator $\hat{\boldsymbol{\beta}}^{(0,T)}$ satisfies*

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

and

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

weakly, where $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Omega}^{-1}$. *Here,*

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

and

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

where $\Gamma^{(r)}(t, \boldsymbol{\beta}_0)$ *for* $r = 0, 1$ *are defined as in (S1.1) in the web supplement.*

Note that $\boldsymbol{\Sigma}$ in Proposition 3 is the asymptotic variance of the Buckley-James estimator given the true subset of covariates; see Lai and Ying (1991). Finally, note that these theoretical results all depend on the existence of $\sqrt{n}$-consistent $\hat{\boldsymbol{\beta}}^{(0)}$. In high dimensions we estimate $\hat{\boldsymbol{\beta}}^{(0)}$ from the covariates retained after screening using the method of Zhu et al. (2011). Then by the sure screening property, the probability that $\hat{\boldsymbol{\beta}}^{(0)}$ is estimated from the truly important covariates, and is thus $\sqrt{n}$-consistent, will approach 1.

In practice we propose using the covariance matrix estimated from the second-stage Buckley-James fit to estimate the covariance of the nonzero components of $\hat{\boldsymbol{\beta}}^{(0,T)}$, while we set the zero components to have zero variance. This ad-hoc estimator ignores the variability coming from the imputation of $\hat{Y}(\hat{\boldsymbol{\beta}}^{(0)})$ as well as the variability from the first-stage SADS model selection, and so will in general underestimate the true variance of the OADS estimator. However, as the probability of selecting the true model increases, this variance estimator will clearly approach $\boldsymbol{\Sigma}$, the variance of the oracle estimator.

5. Tuning Parameter Selection

In practice, it is very important to select an appropriate tuning parameter λ in order to obtain good performance. For the uncensored linear regression (2.1), Tibshirani (1996) and Fan and Li (2001) proposed the following generalized cross-validation (GCV) statistic:

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

where $AR(\lambda)$ is the average residual sum of squares $\frac{1}{n}\|\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}(\lambda)\|_2^2$, $\hat{\boldsymbol{\beta}}(\lambda)$ is the estimate of $\boldsymbol{\beta}$ under λ and $d(\lambda)$ 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 $AR(\lambda)$ by

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

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 $\widehat{AR}(\lambda)$ are equal to the expected value of $\sum_{i=1}^n\{Y_i-\hat{\alpha}^{(0)}-\mathbf{X}_i'\hat{\boldsymbol{\beta}}(\lambda)\}^2$ and n, respectively. Elementary probability implies that $\widehat{AR}(\lambda)$ and $AR(\lambda)$ have the same limit, justifying the utility of the inverse reweighting scheme. To obtain an estimate of the effective number of parameters for the SADS estimator, we follow Zou et al. (2007). The expression (4.7)-(4.8) suggests that $\hat{d}(\lambda)=\text{trace}\{\mathbf{X}_T(\mathbf{X}_T'\mathbf{P}_n\mathbf{X}_T)^{-1}\mathbf{X}_T'\mathbf{P}_n\}=\|T\|_0$, where $T=\{j;\hat{\beta}_j\neq 0\}$, is a consistent estimator for $d(\lambda)$. In the ensuing data analysis and simulation studies, we propose to select λ that yield the smallest GCV defined as

$$GCV(\lambda)=\frac{\widehat{AR}(\lambda)}{\{1-\hat{d}(\lambda)/n\}^2}. \tag{5.1}$$

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

6. Simulations and Comparisons

6.1. Simulation Set-up

We examined the finite sample performance of the proposed methods in low- and high-dimensional settings. Mimicking the simulation setup of Tibshirani (1997) and Cai et al. (2009), for $i=1,\ldots,n$ we generated the true response Y_i (after the exponential transform) from the exponential distribution with rate $\lambda_i=\exp(-\boldsymbol{\beta}_0'\mathbf{X}_i)$, i.e., $Y_i=\boldsymbol{\beta}_0'\mathbf{X}_i+e_i$. In the low-dimensional setting, we let

$p = 9$, and to model weak and moderate associations between the predictors and the response we considered: $\boldsymbol{\beta}_0 = (0.35, 0.35, 0, 0, 0, 0.35, 0, 0, 0)'$ and $\boldsymbol{\beta}_0 = (0.7, 0.7, 0, 0, 0, 0.7, 0, 0, 0)'$. In the high-dimensional setting, we let $p = 10000$ and considered $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_0^{low}, \boldsymbol{\beta}_0^{low}, \mathbf{0})$, where $\boldsymbol{\beta}_0^{low}$ are the 9×1-dimensional true parameter vectors from the low-dimensional setting.

We generated covariates $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})'$ from a multivariate normal with mean zero and a compound symmetry covariance matrix $\Sigma = (\sigma_{jj'})_{p\times p} = (\rho)$ and e_i follows the standard extreme value distribution. In low dimensions, we varied ρ to be 0, 0.5, and 0.9, corresponding to zero, moderate, and strong collinearity among the predictors. *In high dimensions, good performance becomes difficult to achieve with a p as large as ours, so we simulated slightly easier settings with ρ equal to either 0, 0.3, or 0.5.*

The censoring variable C_i (after exponential transform) was generated from a uniform$[0, \xi]$, where ξ was chosen to achieve about 40% censoring. The initial estimate $\hat{\boldsymbol{\beta}}^{(0)}$ is obtained via the Buckley-James procedure. We simulated sample sizes of $n = 50$ or $n = 200$ and generated 200 independent datasets under each simulation setting.

6.2. Comparisons of Competing Methods

For each scenario, the following proposed estimation procedures were evaluated: the selection-consistent adaptive Dantzig selector $\hat{\boldsymbol{\beta}}$ (SADS) and the oracle adaptive Dantzig selector $\hat{\boldsymbol{\beta}}^{(0,T)}$. For the OADS estimator, we first tuned the SADS estimator and then fit a Buckley-James estimate to the selected covariates. We used the unpenalized Buckley-James procedure to obtain the initial $\sqrt{n}$-consistent estimates $\hat{\boldsymbol{\beta}}^{(0)}$.

To compare these methods to a penalization-based approaches, we also evaluated the adaptive penalized Buckley-James estimator (APBJ) of Johnson (2009), which uses the $\hat{\boldsymbol{\beta}}^{(0)}$ to impute outcomes $\hat{\mathbf{Y}}(\hat{\boldsymbol{\beta}}^{(0)})$ and then applies the adaptive LASSO penalty to the least-square loss function constructed using the $\hat{\mathbf{Y}}(\hat{\boldsymbol{\beta}}^{(0)})$. Here we followed Johnson (2009) and used the Gehan estimator (Gehan 1965) to obtain the initial $\hat{\boldsymbol{\beta}}^{(0)}$. In the high-dimensional setting, we used the same screening procedure of Zhu et al. (2011) on the APBJ estimator as well as our proposed adaptive Dantzig selectors.

We evaluated the accuracy and precision of the parameter estimates based

Table 6.1: Comparisons of Methods with Different Signal Strengths in Low Dimensions

	$\beta_{0j} = 0.7$					$\beta_{0j} = 0.35$				
Method	MSE	FP	FN	% Correct	C-stat	MSE	FP	FN	% Correct	C-stat
					$n = 50, \rho = 0$					
SADS	0.46	1.36	0.24	0.24	0.72	0.42	1.58	1.36	0.04	0.59
OADS	0.54	1.36	0.24	0.24	0.72	0.55	1.58	1.36	0.04	0.59
APBJ	0.44	1.42	0.23	0.24	0.72	0.39	1.60	1.23	0.04	0.60
					$n = 200, \rho = 0$					
SADS	0.11	0.56	0.01	0.74	0.75	0.10	0.80	0.34	0.32	0.64
OADS	0.08	0.56	0.01	0.74	0.75	0.11	0.80	0.34	0.32	0.64
APBJ	0.11	0.87	0	0.64	0.75	0.10	0.82	0.34	0.30	0.64
					$n = 50, \rho = 0.5$					
SADS	1.03	1.60	0.55	0.10	0.78	0.61	1.27	1.62	0	0.66
OADS	1.14	1.60	0.55	0.10	0.78	0.85	1.27	1.62	0	0.66
APBJ	0.94	1.77	0.48	0.10	0.78	0.54	1.29	1.51	0.02	0.67
					$n = 200, \rho = 0.5$					
SADS	0.18	0.88	0.06	0.57	0.81	0.18	0.93	0.74	0.12	0.69
OADS	0.17	0.88	0.06	0.57	0.81	0.21	0.93	0.74	0.12	0.69
APBJ	0.18	1.10	0.02	0.48	0.81	0.16	0.96	0.55	0.20	0.69
					$n = 50, \rho = 0.9$					
SADS	4.31	1.90	1.41	0	0.82	2.25	1.52	1.96	0	0.71
OADS	5.04	1.90	1.41	0	0.81	3.32	1.52	1.96	0	0.70
APBJ	3.68	2.12	1.26	0	0.82	1.85	1.52	1.88	0.01	0.71
					$n = 200, \rho = 0.9$					
SADS	1.20	1.55	0.78	0.04	0.83	0.59	0.65	1.98	0	0.72
OADS	1.35	1.55	0.78	0.04	0.83	0.86	0.65	1.98	0	0.72
APBJ	1.05	1.47	0.71	0.08	0.83	0.53	0.74	1.77	0.01	0.72

on the mean squared errors MSE $= E(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2)$. To examine how well the proposed procedures perform with respect to variable selection, we recorded the average number of of zero regression coefficients being incorrectly set to non-zero and non-zero regression coefficients being incorrectly set to zero, leading to the average number of false positives (FP) and false negatives (FN). We also recorded for each method the probability, across the 200 simulations, of selecting exactly the correct model. Finally, we compared the predictive abilities of the fitted models by estimating their C-statistics (Uno et al., 2011) on independent test datasets.

The results for the low-dimensional setting are summarized in Table 6.1. The Dantzig selector-based estimators appeared to have better model selection

performance. When $n = 50$, all three methods performed alike in terms of model selection, and not surprisingly had a difficult time selecting the correct model. However, when $n = 200$ and $\beta_{0j} = 0.7$, the SADS and OADS estimators were able to select the correct model up to 74% of the time. The APBJ method of Johnson (2009) performed worse than the other estimators except when $\rho = 0.9$.

However, the APBJ method of Johnson (2009) appears to have the best estimation accuracy in general, as its average mean squared errors were usually lower than those of OADS and SADS. The OADS estimator could indeed outperform the SADS estimator, but apparently only when the probability of selecting the true model was sufficiently high. Indeed, in such situations the OADS even outperformed the APBJ estimator. However, when this was not the case, such as in the simulation settings with $n = 50$, it was actually detrimental to fit a Buckley-James estimator to the covariates selected by SADS.

Finally, the methods did not exhibit appreciable differences in predictive abilities. The selection performance and mean squared errors of all three methods improved with increasing sample size and degraded with increasing correlation between the covariates. On the other hand, the predictive abilities were not affected much by the sample size, and not surprisingly also improved with increasing correlation.

The results for the high-dimensional setting are summarized in Table 6.2. All of the methods performed poorly in variable selection. However, this is not surprising given the difficulty of the simulation settings. On the other hand, when $n = 200$ and $\beta_{0j} = 0.7$, they were able to achieve fairly good estimation accuracy, as the MSE's of the three methods are lower than $\|\boldsymbol{\beta}_0\|_2^2 = 2.94$ for $\rho = 0$ and $\rho = 0.3$. Again, the APBJ estimator gave the most accurate parameter estimates. Finally, we see that the different methods give fitted models with very similar predictive abilities. When $n = 50$ and $\rho = 0$, the C-statistics were very low because of the noise involved in the screening step, but with larger n and higher ρ the C-statistics were around 80% even though very few of the truly important covariates were selected.

We also studied the performance of our proposed covariance estimator for the OADS estimators. As we mentioned in Section 4.3, as the SADS stage does a better job of selecting the true model, our variance estimator will approach

Table 6.2: Comparisons of Methods with Different Signal Strengths in High Dimensions

	$\beta_{0j} = 0.7$					$\beta_{0j} = 0.35$				
Method	MSE	FP	FN	% Correct	C-stat	MSE	FP	FN	% Correct	C-stat
					$n = 50, \rho = 0$					
SADS	4.81	7.66	5.50	0	0.55	1.87	7.78	5.91	0	0.54
OADS	4.97	7.66	5.50	0	0.55	2.00	7.78	5.91	0	0.54
APBJ	4.66	8.13	5.51	0	0.55	1.74	8.01	5.91	0	0.54
					$n = 200, \rho = 0$					
SADS	1.37	20.01	0.79	0	0.72	1.26	21.34	3.58	0	0.55
OADS	1.64	20.01	0.79	0	0.70	1.43	21.34	3.58	0	0.55
APBJ	1.34	19.31	0.79	0	0.72	1.18	22.02	3.51	0	0.55
					$n = 50, \rho = 0.3$					
SADS	5.92	7.18	5.79	0	0.73	2.08	5.96	5.92	0	0.67
OADS	6.01	7.18	5.79	0	0.73	2.21	5.96	5.92	0	0.67
APBJ	5.66	7.30	5.80	0	0.73	1.88	6.29	5.92	0	0.67
					$n = 200, \rho = 0.3$					
SADS	2.40	15.62	2.86	0	0.81	1.15	12.64	4.66	0	0.71
OADS	2.57	15.62	2.86	0	0.81	1.20	12.64	4.66	0	0.71
APBJ	2.30	13.45	2.87	0	0.81	1.03	10.57	4.69	0	0.71
					$n = 50, \rho = 0.5$					
SADS	6.90	7.23	5.87	0	0.80	2.52	5.98	5.94	0	0.72
OADS	6.97	7.23	5.87	0	0.80	2.71	5.98	5.94	0	0.72
APBJ	6.50	7.50	5.88	0	0.80	2.29	6.05	5.94	0	0.72
					$n = 200, \rho = 0.5$					
SADS	3.52	17.32	4.06	0	0.85	1.41	12.33	5.28	0	0.76
OADS	3.70	17.32	4.06	0	0.84	1.48	12.33	5.28	0	0.76
APBJ	3.34	15.70	4.07	0	0.84	1.25	9.54	5.35	0	0.76

Table 6.3: Covariance estimators

	Oracle cov.			Ave. OADS cov.		
	β_1	β_2	β_6	β_1	β_2	β_6
$n = 200, \beta_{0j} = 0.7, \rho = 0$, 74% correct						
β_1	0.014	0.004	0.004	0.016	0.002	0.002
β_2	0.004	0.013	0.004	0.002	0.016	0.002
β_6	0.004	0.004	0.016	0.002	0.002	0.016
$n = 50, \beta_{0j} = 0.7, \rho = 0.5$, 10% correct						
β_1	0.127	-0.036	-0.031	0.100	-0.010	-0.011
β_2	-0.036	0.098	-0.020	-0.010	0.108	-0.014
β_6	-0.031	-0.020	0.094	-0.011	-0.014	0.098
$n = 50, \beta_{0j} = 0.35, \rho = 0.9$, 0% correct						
β_1	0.457	-0.206	-0.224	0.144	-0.018	-0.017
β_2	-0.206	0.406	-0.171	-0.018	0.132	-0.011
β_6	-0.224	-0.171	0.416	-0.017	-0.011	0.158

the true variance of the oracle estimator. In Table 6.3 we compare a few average estimated covariance matrices to their corresponding empirical oracle covariance matrices in low dimensions. When $n = 200$, $\beta_{0j} = 0.7$, and $\rho = 0$, the OADS estimator selected the correct model 74% of the time, so the two covariance matrices were very similar. We also included a more difficult case, where $n = 50$, $\beta_{0j} = 0.7$, and $\rho = 0.5$. Even when the model was selected only 10% of the time, our covariance estimator still performed fairly well. In the worst case setting of $n = 50$, $\beta_{0j} = 0.35$, and $\rho = 0$, however, when the OADS estimator never selected the correct model, our estimator was very different from the truth. Thus while our ad-hoc proposal is reasonable for easy or moderately difficult settings, a more appropriate variance estimator would be an interesting subject for further research.

7. Example of Myeloma Patients' Survival Prediction

Multiple myeloma is a progressive hematologic (blood) disease, characterized by excessive numbers of abnormal plasma cells in the bone marrow and overproduction of intact monoclonal immunoglobulin. Myeloma patients are typically characterized with wide clinical and pathophysiologic heterogeneities, with survival ranging from a few months to more than 10 years. Gene expression profiling

Table 7.4: Validation C-statistics on TT3

Method	Model size	C-statistic
SADS	4	0.6067
OADS	4	0.6160
APBJ	13	0.6255

of multiple myeloma patients has offered an effective way of understanding the cancer genomics and designing gene therapy. Identifying risk groups with a high predictive power could contribute to selecting patients for personalized medicine.

To address this issue, we studied event-free survival from newly diagnosed multiple myeloma patients enrolled in trials UARK 98-026 and UARK 2003-33 (Zhan et al. 2006, Shaughnessy et al. 2007). The trials compared the results of two treatment regimes, total therapy II (TT2) and total therapy III (TT3). There were 340 patients in TT2, with 191 events and an average follow-up of 47.1 months, and 214 patients in TT3, with 55 events and an average follow-up of 35.6 months. Gene expression values for 54675 probesets were measured for each subject using Affymetrix U133Plus2.0 microarrays. We retrieved the data from the MicroArray Quality Control Consortium II (Shi et al., 2010) GEO entry (GSE24080).

We used our proposed adaptive Dantzig selector methods to develop risk scores by fitting AFT models to the TT2 patients. We then estimated the C-statistics (Uno et al., 2011) of those models on the TT3 patients, and compared the results to the APBJ estimator. Table 7.4 contains the results, and we see that the SADS, OADS, and APBJ estimators have similar predictive performances, with validation C-statistics of around 61%. Our adaptive Dantzig selectors achieved this using 4 probesets, while the APBJ estimator used 13. Table 7.5 reports the final models and the parameter estimates.

8. Discussion

We have studied variable selection for the AFT model by applying the adaptive Dantzig selector to the Buckley-James estimating equation, and we have provided two estimators. To our knowledge, this is the first theoretical study applying the Dantzig selector to censored linear regression. We showed that our SADS estimator is selection consistent while our OADS has the oracle property.

Table 7.5: Parameter estimates by various selectors

Probeset	Gene name	SADS	OADS	APBJ
205072_s_at	XRCC4	0.057	0.285	0.14
208966_x_at	IFI16	-0.836	-0.678	-0.485
225450_at	AMOTL1	-0.042	-0.18	-0.082
233750_s_at	C1orf25	-0.207	-0.386	-0.191
204700_x_at	C1orf107			-0.055
1565951_s_at	CHML			-0.135
1568907_at	Unknown			0.08
201897_s_at	CKS1B			0.009
222437_s_at	VPS24			0.297
222443_s_at	RBM8A			0.255
209052_s_at	WHSC1			-0.04
228817_at	ALG9			0.236
225834_at	FAM72A /// FAM72B /// GCUD2			-0.046

In simulations we showed that they perform similarly compared to the APBJ method of Johnson (2009), though the Dantzig-based methods may have a slight advantage in variable selection while the penalization-based method has slightly better estimation accuracy. Similar results have been found when comparing the Dantzig selector to the lasso in the complete data case, and we have shown the a similar relationship holds in the censored data case.

When the data are high- or ultrahigh-dimensional, we proposed using a screening procedure for single-index models before applying the SADS or OADS estimators, a strategy justified by the sure screening property of Zhu et al. (2011). Using this approach to analyze the data from the multiple myeloma clinical trial, we showed that our methods could achieve comparable validation C-statistics as the APBJ method, but using far fewer probesets.

Several issues merit further investigations. First, our asymptotic setup in this paper is that the number of predictors is fixed while the sample size approaches infinity. In this work we have appealed to the sure screening procedure of Zhu et al. (2011), but an asymptotic theory with a diverging p seems to be more applicable to problems involving a huge number of predictors, such as microarray analysis and document/image classification.

Second, more research is needed on the evaluation of the variation of the

estimator for small or moderate sample size. We proposed an ad-hoc variance estimator that gives reasonable performance when the signal-to-noise ratio is not too weak. Another possibility is to use a perturbation resampling technique, as in Minnier et al. (2011), though this lacks theoretical justification when applied to Dantzig selector-type regularization.

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, and may be more computationally and theoretically appealing than the penalized estimating equation method of Johnson et al. (2008). We envision that our work can be extended to handle Dantzig selector in the framework of more general estimating equations.

Acknowledgment

This work was partially supported by U.S. National Cancer Institute grant R01 CA95747.

References

Anderson, E., Miller, P., Ilsley, D., Marshall, W., Khvorova, A., Stein, C. 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. (2010). The Dantzig selector in Cox's proportional hazards model. *Scandinavian Journal of Statistics* **37**, 531–552.

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

Candés, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics* **35**, 2313–2351.

Dicker, L. (2011). Regularized regression methods for variable selection and estimation. Ph.D. thesis, Harvard University, USA.

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. *Journal of the American Statistical Association* **96**, 1348–1360.

Gehan, E.A. (1965). A generalized Wilcoxon test for comparing arbitrarily single-censored samples. *Biometrika* **90**, 341–353.

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.

James, G., Radchenko, P. and Lv, J. (2008). DASSO: connections between the Dantzig selector and lasso. *Journal of the Royal Statistical Society, Ser. B* **71**, 127–142.

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.

Johnson, B.A. (2009). On lasso for censored data. *Electronic Journal of Statistics* **3**, 485–506.

Johnson, B.A., Long, Q., and Chung, M. (2011). On path restoration for censored outcomes. *Biometrics* **67**, 1379–1388.

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

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

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

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

Minnier, J., Tian, L., and Cai, T. (2011). A perturbation method for inference on regularized regression estimates. *Journal of the American Statistical Association* **106**, 1371–1382.

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.

Shaughnessy, J.D., Zhan, F., Buring, B., Huang, Y., Colla, S., Hanamura, I., Stewart, J.P., Kordsmeier, B., Randolph, C., Williams, D.R., Xiao, Y., Xu, H., Epstein, J., Anaissie, E., Krishna, S.G., Cottler-Fox, M., Hollmig, K., Mohiuddin, A., Pineda-Roman, M., Tricot, G., van Rhee, F., Sawyer, J., Alsayed, Y., Walker, R., Zangari, M., Crowley, J., and Barlogie, B. (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.

Shi, L., Campbell, G., Jones, W. D., et al. (2010). The MAQC-II Project: A comprehensive study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology* **28**, 827–838.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Ser. B* **58**, 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* **16**, 385–395.

Uno, H., Cai, T., Pencina, M.J., DAgostino, R.B., and Wei, L.J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* **30**, 1105–1117.

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.

Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *Annals of Statistics* **21**, 76–99.

Zhan, F., Huang, Y., Colla, S., Stewart, J., Hanamura, I., Gupta, S., Epstein, J., Yaccoby, S., Sawyer, J., Burington, B., Anaissie, E., Hollmig, K., Pineda-Roman, M., Tricot, G., van Rhee, F., Walker, R., Zangari, M., Crowley, J., Barlogie, B., and Shaughnessy, J.D. (2006). The molecular classification of multiple myeloma. *Blood* **108**, 2020–2028.

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

Zhu, L.P., Li, L., Li, R., and Zhu, L.X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **106**, 1464–1475.

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

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

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109

E-mail: yili@umich.edu

Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ 08854

E-mail: ldicker@stat.rutgers.edu

Department of Biostatistics, University of Pennsylvania, Philadelphia, PA 19104

E-mail: sihai@mail.med.upenn.edu