THE DANTZIG SELECTOR IN COX'S PROPORTIONAL HAZARDS MODEL

Anestis Antoniadis,

Laboratoire Jean Kuntzmann, Department de Statistique,

Université Joseph Fourier, B.P. 53

38041 Grenoble CEDEX 9, France.

Piotr Fryzlewicz,

Department of Mathematics, University of Bristol

University Walk, Bristol BS8 1TW, UK

and

Frédérique Letué,

Laboratoire Jean Kuntzmann, Department de Statistique,

Université Pierre Mendès France, B.P. 53

38041 Grenoble CEDEX 9, France.

Abstract

The Dantzig Selector is a recent approach to estimation in high-dimensional linear regression models with a large number of explanatory variables and a relatively small number of observations. As in the least absolute shrinkage and selection operator (LASSO), this approach sets certain regression coefficients exactly to zero, thus performing variable selection. However, such a framework, contrary to the LASSO, has never been used in regression models for survival data with censoring. A key motivation of this article is to study the estimation problem for Cox's proportional hazards function regression models using a framework that extends the theory, the computational advantages and the optimal asymptotic rate properties of the Dantzig selector to the class of Cox's proportional hazards under appropriate sparsity scenarios. We perform a detailed simulation study to compare our approach to other methods and illustrate it on a well-known microarray gene expression data set for predicting survival from gene expressions.

Some key words: VARIABLE SELECTION; GENERALIZED LINEAR MODELS; DANTZIG SELECTOR; LASSO; PENALIZED PARTIAL LIKELIHOOD; PROPORTIONAL HAZARDS MODEL;

1 INTRODUCTION

An objective of survival analysis is to identify the risk factors and their risk contributions. Often, many covariates are collected and, to reduce possible modelling bias, a large parametric model is built. An important and challenging task is then variable selection which is a form of model selection in which the class of models under consideration is represented by subsets of covariate components to be included in the analysis. Variable selection methods are well developed in linear regression settings and in recent years many of them have been extended to the context of censored survival data analysis. They include best-subset selection (Jovanovic et al. (1995)), stepwise selection (Delong et al. (1994)), asymptotic procedures based on score tests, Wald tests and other approximate chi-squared testing procedures (Harrell (2001)), bootstrap procedures (Graf et al. (1999)) and Bayesian variable selection (Faraggi and

Simon (1998); Ibrahim et al. (2008)). However, theoretical properties of these methods have not been fully validated (Fan and Li (2002)).

Recently a family of penalized partial likelihood methods, such as the LASSO (Tibshirani (1997)) and the smoothly clipped absolute deviation method (SCAD; Fan and Li (2002)), were proposed for Cox's proportional hazards model. By shrinking some regression coefficients to zero, these methods select important variables and estimate the regression model simultaneously. The LASSO estimator does not possess the oracle properties (Fan and Li (2002)). The SCAD estimator has better theoretical properties than the LASSO, but the nonconvex form of its penalty makes its computation challenging in practice, and the solutions may suffer from numerical instability (see Zou (2008)). An adaptive LASSO method based on a penalized partial likelihood with adaptively weighted L_1 penalties on regression coefficients developed by Zhang and Lu (2007) enjoys the oracle properties of the SCAD estimator but the optimization problem is efficiently solved by standard algorithms.

Recently, Candès and Tao (2007) proposed the Dantzig selector for performing model fitting for linear regression models where the number of variables can be much larger than the sample size but the set of coefficients is sparse, i.e. most of the coefficients are zero. Unlike most other procedures such as the LASSO and the SCAD, which minimize the sum of squared errors subject to a penalty on the regression coefficients, the Dantzig Selector minimizes the L_1 norm of the coefficients subject to a constraint on the error terms. As with the LASSO and the SCAD or the adaptive LASSO, this approach sets certain coefficients exactly to zero, thus performing variable selection. However, unlike the other methods, standard linear programming methods can be used to compute the solution to the Dantzig selector, providing a computationally efficient algorithm and the resulting estimated coefficients enjoy near-optimal ℓ_2 non-asymptotic error bounds. Hence, the Dantzig selector appears to be an appealing estimation procedure for sparse linear regression models and this encourages us to extend the theory and its computational advantages to the class of semi-parametric Cox's proportional hazards models. The proposed method compares favorably with other methods available in the literature, and thus provides a useful addition to the toolbox of estimation and prediction methods for the widely used Cox's model.

The paper is organized as follows. The usual survival data setup for (generalized)

Cox's regression model with time-independent covariates (for cross sectional type data) is introduced in Section 2, recalling the basic ideas of Cox's original proportional model for the hazard rates. In particular, we briefly recall in this section the appropriate framework needed to represent this model in a martingale notation based on theories of counting processes (see e.g. Andersen and Gill (1982)). In Section 3, after outlining the approach behind the Dantzig Selector for linear regression models, we introduce our Dantzig Selector for proportional hazards (PH) models and develop a computationally efficient algorithm for computing the estimator. Section 3 also contains our main assumptions and theoretical results concerning the estimator, the main result relating to its l_2 error, in analogy with Candès and Tao's (Candès and Tao (2007)) results for linear models. In Section 4, we present a simulation study comparing the proposed approach with various competitors, where we also present the application of our method on a well-known microarray gene expression data set, used previously for similar purposes in the literature (Bovelstad et al. (2007)). Proofs of main and intermediate results are in Section 5.

Software (an R script) implementing our Dantzig selector for survival data can be downloaded from

`http://www.maths.bris.ac.uk/~mapzf/dscox/ds_cox.html`.

2 Notation and preliminaries

In order to fix the notation we consider the usual survival data setup. The reader unfamiliar with the concepts described in this Section is referred to the book by Andersen et al. (1993). The survival time X is assumed to be conditionally independent of a censoring time U given the p-dimensional vector of covariates $\mathbf{Z} = (Z^1, Z^2, \ldots, Z^p)^T$ so that the construction of the partial likelihood is justified. We observe n i.i.d. copies $(\tilde{X}_i, D_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$, of the right censored survival time $\tilde{X} = \min(X, U)$ and the censoring indicator $D = \mathbf{I}[X \leq U] = \mathbf{I}[\tilde{X} = X]$. The covariates are assumed to be bounded: there exists a positive constant C such that $\sup_{1 \leq j \leq p} |Z^j| \leq C$. This assumption is fully justified in the fixed design case, and is used in our theoretical calculations regarding the performance of our estimator. However, we emphasise that in practice, our computational algorithm makes no use of either the assumption itself or the (possibly unknown) value of the constant C.

Thus this assumption should not be viewed as restrictive, even in the random design case. We also note that in cases where Z^j represent gene expressions measured on a microarray, they are naturally bounded by virtue of the measurement process.

In the following we will denote by Z the $n \times p$ matrix whose generic term Z_{ij} is the ith observed value of the jth covariate Z^j, and the ith row of Z will be denoted by $\mathbf{z}_i^T$. For simplicity, we will also assume that there are no tied failure times; suitable modifications of the partial likelihood exist for the case of predictable and locally bounded covariate processes and for the case of ties (see Andersen et al. (1993)). Most often in the literature, proportional hazards models are formulated using random variables (as opposed to stochastic processes), and the implied statistical methods are based on maximum (partial) likelihoods. However, we prefer studying such problems in terms of the theory of counting processes, since time and random phenomena occurring in time play an essential role in survival analysis. Moreover, this counting process approach has been facilitated by the work of Andersen and Gill (1982) and permits us to use martingale convergence results in a unified way to demonstrate theoretical properties of our approach.

In the counting process setup, we can represent the observed data as follows. The regression model for survival data, described above, is linked to the multivariate counting process $\mathbf{N} = (N_1, \ldots, N_n)$ of the form, $N_i(t) = \mathbf{I}(\tilde{X}_i \leq t, D_i = 1)$ where the N_i's are independent copies of the single-jump counting process $N(t) = \mathbf{I}(\tilde{X} \leq t, D = 1)$ that registers whether an uncensored failure (or death) has occurred by time t. Let $Y(t) = \mathbf{I}[\tilde{X} \geq t]$ be the corresponding "at risk" indicator. Define the filtration $\mathcal{F}_t = \mathcal{F}_0 \vee \{N(u); u \leq t\}$, where $\mathcal{F}_0 = \sigma(\mathbf{Z})$. Under the true probability measure $\mathbb{P}$ on $\mathcal{F} = \mathcal{F}_t$, the counting processes $N_i(t)$ have intensity processes $\lambda_i(t, \mathbf{z}_i)$ and under the Cox regression model, the conditional intensities $\lambda_i(t, \mathbf{z}_i)$ of N_i given $\mathbf{Z}_i = \mathbf{z}_i$ for t restricted to a fixed time interval $[0, \tau]$ are

$$\lambda_i(t, \mathbf{z}_i) = Y_i(t)\alpha_0(t)\exp(\mathbf{z}_i^T \boldsymbol{\beta}_0) \tag{1}$$

where α_0 is the baseline hazard function and $\boldsymbol{\beta}_0$ is the unknown vector of regression coefficients. For flexibility of fit, the baseline hazard function is left unspecified and our setting is therefore semiparametric. This, in particular, means that

$$M_i(t) = N_i(t) - \int_0^t \lambda_i(u, \mathbf{z}_i)du, \quad t \in [0, \tau],$$

are independent $\mathcal{F}_t$ square-integrable martingales under $\mathbb{P}$ with compensator $V_i(t) = \int_0^t \lambda_i(u, \mathbf{z}_i)du$. In particular, we have

$$\langle M_i, M_i \rangle(t) = \int_0^t \lambda_i(u, \mathbf{z}_i)du = V_i(t).$$

Under the above notation, the (rescaled by $-1/n$) Cox's partial loglikelihood function is given by

$$l(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i^T \boldsymbol{\beta} \int_0^\tau dN_i(u) - \int_0^\tau \log\left(\sum_{i=1}^{n} Y_i(u)\exp(\mathbf{z}_i^T\boldsymbol{\beta})\right)\frac{d\bar{N}(u)}{n},$$

where $d\bar{N}(u) = d\sum_{i=1}^n N_i(u)$. Let $S_n(\boldsymbol{\beta}, u) = \sum_{i=1}^n Y_i(u)\exp(\mathbf{z}_i^T\boldsymbol{\beta})$. Then

$$l(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i^T \boldsymbol{\beta} \int_0^\tau dN_i(u) - \int_0^\tau \log\left(S_n(\boldsymbol{\beta}, u)\right)\frac{d\bar{N}(u)}{n}.$$

Define the first and second order partial derivative of $S_n(\boldsymbol{\beta}, u)$ with respect to $\boldsymbol{\beta}$:

$$S_n^1(\boldsymbol{\beta}, u) = \sum_{i=1}^{n} Y_i(u)\exp(\mathbf{z}_i^T\boldsymbol{\beta})\mathbf{z}_i \quad \text{and} \quad S_n^2(\boldsymbol{\beta}, u) = \sum_{i=1}^{n} Y_i(u)\exp(\mathbf{z}_i^T\boldsymbol{\beta})\mathbf{z}_i^{\otimes 2}, \tag{2}$$

where $\mathbf{z}^{\otimes 2} = \mathbf{z}\mathbf{z}^T$. The maximum likelihood estimator of $\boldsymbol{\beta}$ in Cox's model, is found as the solution to the score equation $U(\hat{\boldsymbol{\beta}}) = 0$, where the score process $U(\boldsymbol{\beta})$ is defined by

$$U(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau (\mathbf{z}_i - E(u, \boldsymbol{\beta}))dN_i(u),$$

with $E(u, \boldsymbol{\beta}) = \frac{S_n^1(\boldsymbol{\beta}, u)}{S_n(\boldsymbol{\beta}, u)}$. In particular, for the true parameter $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, we have:

$$(U(\boldsymbol{\beta_0}))_j = \left(\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j}\right)_{\boldsymbol{\beta}_0} = \frac{1}{n}\sum_{i=1}^{n} Z_{ij}\int_0^\tau dM_i(u) - \int_0^\tau \frac{S_n^1(\boldsymbol{\beta_0}, u)}{S_n(\boldsymbol{\beta_0}, u)}\frac{d\bar{M}(u)}{n},$$

where $d\bar{M}(u) = d\sum_{i=1}^n M_i(u)$. Thus the score process evaluated at the true parameter $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ is itself a martingale and this fact, together with standard regularity assumptions, facilitates the study of the asymptotic properties of the MLE estimator of the vector of regression coefficients.

In practice, not all the covariates (components of $\mathbf{Z}$) may contribute to the prediction of survival outcomes: some components of $\boldsymbol{\beta}$ in the true model may be zero. Our Dantzig selection procedure, described in the next section, works under this "sparsity" assumption and produces consistent and easily computable estimates of the relevant coefficients.

3 Dantzig selector for Cox's regression model

Theoretical properties of LASSO and SCAD for Cox's proportional hazard model have been investigated. These penalized partial likelihood methods may be viewed, in an asymptotic sense, as instances of iteratively re-weighted least squares procedures by transferring the objective functions involved in the optimization into asymptotically equivalent least-squares problems. Indeed, as noted by Wang and Leng (2007), when p is fixed and is smaller that n, using the asymptotic theory for the MLE estimator $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ in a standard Cox's regression model, the negative log-likelihood function can be replaced locally by a Taylor series expansion at $\tilde{\boldsymbol{\beta}}$ leading to a least squares penalized criterion which is updated iteratively (LASSO Estimation via Least Squares Approximation (LSA)). As shown by Wang and Leng (2007), their resulting LSA estimators are often asymptotically as efficient as oracle as long as the number of components p remains fixed and the tuning parameters are chosen appropriately. In our case, we do not want to restrict ourselves to the standard $p < n$ setup, but we would also like to examine the case where p may grow with, and exceed, n, i.e. the case of a (fast) growing dimension of the predictor. This is indeed part of our motivation for proposing the Dantzig selector. However, in order to justify the algorithm that numerically implements our procedure, we will make some use of the above remarks about LSA.

3.1 Dantzig selector for linear regression

The Dantzig Selector (Candès and Tao (2007)) was designed for linear regression models

$$\mathbf{Y} = Z\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3}$$

with a large p but a sparse set of coefficients, i.e. where most of the regression coefficients β_j are zero. For the linear regression model given by (3), the Dantzig Selector estimate, $\hat{\boldsymbol{\beta}}$, is defined as the solution to

$$\min_{\boldsymbol{\beta} \in B} \|\boldsymbol{\beta}\|_1 \quad \text{subject to} \quad |\mathbf{Z}^j{}^T(\mathbf{Y} - Z\boldsymbol{\beta})| \leq \lambda,\ j = 1, \ldots, p, \tag{4}$$

where $\|\cdot\|_1$ is the L_1 norm, $\mathbf{Z}^j$ is the jth column of Z, λ is a tuning parameter and B represents the set of possible values for $\boldsymbol{\beta}$, usually taken to be a subset of $\mathbb{R}^p$. The

L_1 norm minimization produces coefficient estimates that are exactly zero in a similar fashion to the LASSO and hence can be used as a variable selection tool. In this setup $\mathbf{Z}^j$ is assumed to be norm one which is rarely the case in practice. However, this difficulty is easily resolved by reparamaterizing (3) such that the $\mathbf{Z}^j$'s do have norm one.

Notice that for Gaussian error terms, (4) can be rewritten as,

$$\min_{\boldsymbol{\beta} \in B} \|\boldsymbol{\beta}\|_1 \quad \text{subject to} \quad |\ell_j'(\boldsymbol{\beta})| \leq \lambda/\sigma^2,\ j = 1, \ldots, p, \tag{5}$$

where ℓ_j' is the partial derivative of the log likelihood function with respect to β_j and $\sigma^2 = \text{Var}(\epsilon_j)$. Hence, an intuitive motivation for the Dantzig Selector, as also observed by James and Radchenko (2009), is that, for $\lambda = 0$, the solution to (5) will return the maximum likelihood estimator. For $\lambda > 0$, the Dantzig Selector searches for the $\boldsymbol{\beta}$ with the smallest L_1-norm that is within a given distance of the maximum likelihood solution, i.e. the sparsest $\boldsymbol{\beta}$ that is still reasonably consistent with the observed data. Notice that even for $p > n$, where the likelihood equation will have infinite possible solutions, this approach can still hope to identify a unique solution, provided $\boldsymbol{\beta}$ is sparse, because it is only attempting to locate the sparsest $\boldsymbol{\beta}$ close to the peak of the likelihood function.

The Dantzig Selector has two main advantages. The first is that (4) can be formulated as a standard linear programming problem. The second main advantage is theoretical. Candès and Tao (2007) proved tight non-asymptotic bounds on the error in the estimator for $\boldsymbol{\beta}$, a result which has recently attracted a lot of attention since it demonstrated that the L_2-error in estimating $\boldsymbol{\beta}$ was within a factor of $\log p$ of that one could achieve if the true model were known. More precisely, suppose that that ϵ_i are i.i.d. $N(0, \sigma^2)$ variables and that $\boldsymbol{\beta}$ has at most S non-zero components. Assume also that a Uniform Uncertainty Principle (UUP) condition holds on the design matrix, i.e. suppose that the Gram matrix $\Psi = \frac{1}{n} Z^T Z$ is such that $\Psi_{ii} = 1$ for all $i = 1, \ldots, p$ and $\max_{i \neq j} |\Psi_{i,j}| \leq \frac{1}{3\alpha S}$ for some $\alpha > 1$ (see Lounici (2008)). Then for any $a \geq 0$ and $\lambda = \sigma\sqrt{2(1+a)(\log p)/n}$, the Dantzig selector estimator satisfies

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \leq (1+a) \cdot C \cdot S \cdot \sigma^2 \cdot (\log p)/n, \tag{6}$$

with probability close to 1. Even if we knew ahead of time which β_j's were non-zero, under the same conditions on the design Gram matrix, it would still be the case that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$ grew at the rate of $S \cdot \sigma^2/n$. Hence the rate is optimal up to a

factor of $\log p$, and we only pay a small price for adaptively choosing the significant variables. As mentioned before, equation (6) holds for Gaussian errors with a linear regression model. Our purpose is to extend the Dantzig estimator, algorithm and the above theoretical bounds to the general class of Cox's proportional hazards regression models introduced in Section 2. To our knowledge this is the first time that bounds of this form have been proposed for such models.

3.2 Survival Dantzig Selector

We have already observed that for Gaussian errors in a linear regression model, the inner product between the jth covariate and the vector of residuals, $\mathbf{Z}^{j^T}(\mathbf{Y} - Z\boldsymbol{\beta})$ is proportional to the jth component $\ell_j'(\boldsymbol{\beta})$ of the score vector. Hence, the Dantzig optimization criteria given by (4) and (5) can be extended to the class of Cox's PH regression models in a natural fashion by computing the solution $\hat{\boldsymbol{\beta}}$ of

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{\beta}\|_1 \text{ subject to } \|U(\boldsymbol{\beta})\|_\infty \leq \gamma, \tag{7}$$

where $\gamma \geq 0$ and $U(\boldsymbol{\beta})$ is the score process. Note that such a solution exists because the negative of the loglikelihood is a convex function of $\boldsymbol{\beta}$. We will call the resulting procedure the Survival Dantzig Selector (SDS for short). The purpose of this subsection is to show that, under appropriate assumptions on the information matrix of the corresponding point process, the resulting SDS estimator maintains all the important properties of the Dantzig selector.

In order to prove our main results we will partially proceed along similar lines to Candès and Tao's (2007) original result on the DS and we will need for that the fact that $\|\hat{\boldsymbol{\beta}}\|_1 \leq \|\boldsymbol{\beta}_0\|_1$. However, while for Gaussian errors in a sparse linear regression model, such an inequality is "automatic" (it follows from obvious concentration properties of centered Gaussian measures), this is not the case in our general point process setup, and, indeed, it is implied by Lemma 3.1 stated below and proved in Section 7. The number of predictors $p = p_n$ is allowed to grow (fast) with the sample size n.

Lemma 3.1 *Assume that the dimension of predictor in Cox's PH model satisfies $p_n = O(n^\xi)$, $n \to \infty$, for some $1 < \xi$. Assume also that the number S of effective predictors, i.e. the number of $\boldsymbol{\beta}_{0j,n} \neq 0$ is independent of n and finite (S-sparsity of $\boldsymbol{\beta}_0$). Let $\gamma = \gamma_{n,p} = \frac{\sqrt{(1+a)\log p_n}}{\sqrt{n}}$ for some $a > 0$. Under the additional assumptions that*

- *the baseline hazard function in eq. (1) is such that* $\int \alpha_0(u)du < +\infty$
- $\sup_{1\leq i\leq n} \sup_{1\leq j\leq p_n} |Z_{ij}| \leq C$,

it follows that

$$\mathbb{P}\{\|U(\boldsymbol{\beta}_0)\|_\infty \geq \gamma_{n,p}\} \leq p_n \exp\left(-\frac{n\gamma_{p,n}^2}{2(2C\gamma_{p,n}+K)}\right) = O\left(n^{-a\xi}\right),$$

with $K > 0$ *a suitable constant. It follows that, as* $n \to \infty$*, with probability tending to 1, the true* $\boldsymbol{\beta}_0$ *is admissible for problem (7) , i.e.* $\|U(\boldsymbol{\beta}_0)\|_\infty < \gamma$ *and in particular* $\|\hat{\boldsymbol{\beta}}\|_1 \leq \|\boldsymbol{\beta}_0\|_1$.

Remark 3.1 *The scaling* $1/\sqrt{n}$ *in* $\gamma_{n,p}$ *in the above lemma comes from the scaling* $1/n$ *we chose in the log-likelihood. This choice is also made by Bickel et al. (2008) and Lounici (2008). Note also that the result of Lemma 3.1 is taken for granted in the extension of the DS to the class of generalized linear models derived by James and Radchenko (2009), but it is not automatically true. Finally, note that we allow for a large predictor dimension relative to the sample size* n *as long as* $\xi > 1$ *and the S-sparsity assumption holds. The other assumptions about the boundedness of the predictor variables and the baseline hazard are standard under Cox's PH model (Andersen et al. (1993)).*

In order to obtain error bounds on the components selected by our Survival Dantzig Selector, we introduce a few definitions that are closely related to those from Candès and Tao (2007).

Given an $n \times p$ matrix A and an index set $T \subset \{1, \ldots, p\}$ we will write A_T for the $n \times |T|$ matrix constructed by extracting the columns of A corresponding to the indices in T. The quantities defined below depend on A but this will be omitted to simplify the notation. If this dependency is needed we will denote them with a superscript A. As in Candès and Tao (2007), for any integer $S \leq p$, δ_S is the largest quantity such that

$$\|A_T\mathbf{c}\|_2^2 \geq \delta_S\|\mathbf{c}\|_2^2$$

for all subsets T with $|T| \leq S$ and all vectors $\mathbf{c}$ of length $|T|$. If A is an orthonormal matrix, then $\|A_T\mathbf{c}\|^2 = \|\mathbf{c}\|^2$ for all $T, \mathbf{c}$ and hence $\delta_S = 1$. If some columns of A are linearly dependent then for a certain T and $\mathbf{c}$, $\|A_T c\|^2 = 0$ and hence $\delta_S = 0$.

If $S + S' \leq p$, we also define $\theta_{S,S'}$ as the smallest quantity such that

$$|(A_T\mathbf{c})^T A_{T'}\mathbf{c}'| \leq \theta_{S,S'}\|\mathbf{c}\|_2\|\mathbf{c}'\|_2$$

for all disjoint subsets T and T' with $|T| \leq S$ and $|T'| \leq S'$ and all corresponding vectors $\mathbf{c}$ and $\mathbf{c}'$. Note that when the columns of A are orthogonal then $\theta_{S,S'} = 0$.

Before stating our main result, we recall that the $p \times p$ observed "information" matrix up to time τ corresponding to Cox's proportional model is given by (see e.g. Andersen and Gill (1982)):

$$J(\boldsymbol{\beta}, \tau) = J_n(\boldsymbol{\beta}, \tau) = \int_0^\tau [\frac{S_n^2}{S_n}(\boldsymbol{\beta}, u) - (\frac{S_n^1}{S_n})^{\otimes 2}(\boldsymbol{\beta}, u)] \frac{d\bar{N}_n(u)}{n},$$

with notation as in (2). For a fixed sparsity parameter S, as n tends to infinity, it tends in probability (see Theorem VII.2.2 in Andersen et al. (1993)) to the $p \times p$ matrix of rank S

$$I(\boldsymbol{\beta}, \tau) = \int_0^\tau [\frac{s^2}{s}(\boldsymbol{\beta}, u) - (\frac{s^1}{s})^{\otimes 2}(\boldsymbol{\beta}, u)] s(\boldsymbol{\beta}, u) \alpha_0(u) du,$$

where $s(\boldsymbol{\beta}, u) = \mathbb{E}(S_n(\boldsymbol{\beta}, u)/n)$, $s^1(\boldsymbol{\beta}, u) = \mathbb{E}(S_n^1(\boldsymbol{\beta}, u)/n)$, $s^2(\boldsymbol{\beta}, u) = \mathbb{E}(S_n^2(\boldsymbol{\beta}, u)/n)$. Finally, when derivatives defining $s(\boldsymbol{\beta}, u)$, $s^1(\boldsymbol{\beta}, u)$ and $s^2(\boldsymbol{\beta}, u)$ are computed only with respect to the components of the true S-dimensional vector $\boldsymbol{\beta}_0$, the true $S \times S$ information matrix, not be confused with the $p \times p$ matrix $I(\boldsymbol{\beta}_0, \tau)$ (of rank S) which is the asymptotic limit of $J(\boldsymbol{\beta}_0, \tau)$, will be denoted by $\mathcal{I}(\boldsymbol{\beta}_0, \tau)$. Applying Theorem 7.2.6 of Horn and Johnson (1985) with $k = 2$, we will denote hereafter $V^{1/2}$ the unique (semi)definite positive square root matrix of a (semi)definite positive matrix V.

Let $\gamma = \gamma_{n,p}$ be a tuning parameter. We now state our main theoretical result in Theorem 3.1 below. The proof is in Section 7.

Theorem 3.1 *Suppose that the true vector of coefficients $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is a nonzero S-sparse coefficient vector with S independent of n, such that the coefficients δ and θ for the matrix $I^{1/2}(\boldsymbol{\beta}_0, \tau)$ obey $\theta_{S,2S} < \delta_{2S}$. Assume that the assumptions used in Lemma 3.1 hold and let $\hat{\boldsymbol{\beta}}$ be the estimate from the SDS using tuning parameter $\gamma = \gamma_{n,p}$ with $\gamma_{n,p}$ as in Lemma 3.1. Then, as long as the information matrix $\mathcal{I}(\boldsymbol{\beta}_0, \tau)$ is positive definite at $\boldsymbol{\beta}_0$, we have:*

$$\mathbb{P}\left(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 > 64S(\frac{\gamma}{\delta_{2S} - \theta_{S,2S}})^2 \right) \leq O(n^{-a\xi}).$$

The assumptions of Theorem 3.1 are similar to the assumption $\delta + \theta < 1$ made for the Dantzig selector in standard linear models by Candès and Tao (2007) and the assumption $\Delta_K > 0$ made for sparse generalized models by James and Radchenko (2009). The positive-definiteness of $\mathcal{I}(\boldsymbol{\beta}_0, \tau)$ is classical in survival analysis (condition VII.2.1 of Andersen et al. (1993)).

The above theorem depends on the rate at which p is allowed to increase with the number of observations n. Under the usual regularity assumptions for our point process (similar to those of Andersen and Gill (1982), Theorem 4.1) our choice of the threshold γ leads to an optimal (a rate that is similar to the one obtained for the classical Dantzig selector in linear models by Candès and Tao (2007)), up to a $\log p$ factor, squared error bound for the SDS estimator $\hat{\boldsymbol{\beta}}$, provided that S remains small. Under such conditions the SDS will give accurate results even for values of p that are larger than n.

3.3 An algorithm for computing the SDS

In this section, we propose an iterative weighted Dantzig selector algorithm for computing the SDS solution for a given value of γ.

Note that the constraints in (7) are non-linear, so linear programming software cannot directly be used to compute the SDS solution. As noted in the Introduction, in a standard GLM setting, an iterative weighted least squares algorithm is usually used to solve the system of score equations. More precisely, given a current estimate for $\hat{\boldsymbol{\beta}}$, an adjusted dependent variable is computed, and a new estimate for $\boldsymbol{\beta}$ is then computed using weighted least squares. This procedure is iterated until $\hat{\boldsymbol{\beta}}$ converges. For more details the reader is referred to McCullagh and Nelder (1989). An analogous iterative approach works well in computing the SDS solution. We can describe it as follows.

For any fixed γ:

1. At the $k+1$st iteration, compute the gradient vector $U(\hat{\boldsymbol{\beta}}^{(k)})$ and the Hessian matrix $J(\hat{\boldsymbol{\beta}}^{(k)}, \tau)$, where (k) denotes the corresponding estimate from the kth iteration. Consider the unique square root of the matrix $J(\hat{\boldsymbol{\beta}}^{(k)}, \tau)$, i.e. $J(\hat{\boldsymbol{\beta}}^{(k)}, \tau) = A_{(k)}^2$, and set the pseudo response vector $\mathbf{Y} = (A_{(k)})^{-}\{J(\hat{\boldsymbol{\beta}}^{(k)}, \tau)\hat{\boldsymbol{\beta}}^{(k} - U(\hat{\boldsymbol{\beta}}^{(k)})\}$, where V^{-} denotes the Moore-Penrose generalized inverse of V. This amounts to approximating Cox's partial likelihood at the current estimate by the quadratic form
$$\frac{1}{2}(\mathbf{Y} - A_{(k)}\boldsymbol{\beta})^T(\mathbf{Y} - A_{(k)}\boldsymbol{\beta}). \tag{8}$$
2. Re-parameterize $A_{(k)}$ say to $A^{\star}_{(k)}$ such that its columns have norm one and modify accordingly $\mathbf{Y}$ to $\mathbf{Y}^{\star}$ to produce the SDS estimate of $\boldsymbol{\beta}$ at the original scale.

3. Use Candes and Tao's (2007) Dantzig selector to compute $\hat{\boldsymbol{\beta}}^{(k+1)}$ with $\mathbf{Y}^\star$ as the response and $A^\star_{(k)}$ as the design matrix.

4. Repeat steps 1 through 3 until convergence.

Note that step 3 only requires a linear programming algorithm to compute.

This algorithm gives exact zeros for some coefficients and it converges quickly based on our empirical experience. However, as with the standard GLM iterative algorithm, there is no theoretical proof that the algorithm is guaranteed to converge to the global minimizer of (7). Especially in the case $n < p$, instead of using a Moore-Penrose inverse for the possibly semi-positive definite matrix $A_{(k)}$ in the previous algorithm, we could have used, as it is done in ridge regression, the square root of the positive definite matrix $J(\hat{\boldsymbol{\beta}}^{(k)}, \tau) + \mu I_p$ for a small $\mu > 0$.

To estimate the tuning parameter γ, we use generalized cross-validation (Craven and Wahba (1979)). Let $\nu = \gamma^{-1}$ and $V(\hat{\boldsymbol{\beta}})$ be the diagonal matrix with diagonal entries $1/\hat{\beta}_i^2$ when $\hat{\beta}_i^2 > 0$ and 1 when $\hat{\beta}_i = 0$. At convergence, the minimizer of (8) in step 1 can be approximated by the ridge solution $(J(\hat{\boldsymbol{\beta}}, \tau) + \nu V(\hat{\boldsymbol{\beta}}))^{-1} A^T \mathbf{Y}$. Therefore, the number of effective parameters in the SDS estimator can be approximated by $p(\nu) = \text{tr}\left((J(\hat{\boldsymbol{\beta}}, \tau) + \nu V(\hat{\boldsymbol{\beta}}))^{-1} J(\hat{\boldsymbol{\beta}}, \tau)\right)$ and the generalized cross-validation function is $GCV(\nu) = -\ell(\hat{\boldsymbol{\beta}})/[n(1 - p(\nu)/n)^2]$. If $\hat{\nu}$ minimizes $GCV(\nu)$ then γ is chosen to be $1/\hat{\nu}$. We used the above algorithm both in the simulation study and in the real data analysis, reported below.

4 Simulation study

In this section, we present the results of a simulation study conducted to evaluate the performance of the SDS in comparison with three other approaches which include both state-of-the-art and classical methods. In the simulations, we focused on finding the best prediction rule for the time to an adverse event using all the available covariates measurements. To keep the scope of the study manageable, we only included a limited number of methods in our comparison. We feel that the current selection covers the spectrum of existing methods reasonably well: one of them is similar to the Lasso but better, the other one is known to be an excellent predictor while the third one is simple and standard. We briefly describe below the methods to which the comparisons with

SDS are made, namely Partial Cox regression with one or two retained components (PLS Cox), Cox regression with the subset of 20 "best" genes (Cox20) and the threshold gradient descent procedure (TGD) for the Cox model by Gui and Li (2005).

Partial Cox Regression. Nguyen and Rocke (2002) proposed the use of the partial least squares (PLS) algorithm for the prediction of survival with gene expression. This method, however, does not handle the censoring aspect of the survival data properly. We adopted the approach of Park et al. (2002) in which the full likelihood for Cox's model is reformulated as the likelihood of a Poisson model, i.e. a generalized linear model (GLM). This reformulation enables application of the iteratively reweighted partial least squares (IRPLS) procedure for GLM (Marx (1996)). We used the implementation of the PLS algorithm of Park et al. (2002) in R provided by Boulesteix and Strimmer (2007) where the PLS components depend solely on the gene expressions. The interpretation of components is generally not straightforward, especially if the number of genes that contribute to the component becomes large. Aside from this difficulty, PLS components may be excellent survival time predictors.

Cox with univariate gene selection. Possibly the most straightforward and intuitive approach to handling high-dimensional data consists of carrying out univariate gene selection and using the obtained (small) subset of genes as covariates in the standard Cox model. Such an approach was adopted by Jenssen et al. (2002) and van Wieringen et al. (2008). We order genes based on the p-value obtained using Wald's test in univariate Cox regression and, similarly to van Wieringen et al. (2008), we select a pre-fixed number of genes (20 in the present study) rather than genes whose p-values fall below a threshold. This ensures having a set of genes of a convenient size for any training set. A partial justification for selecting 20 covariates comes from the work of van Wieringen et al. (2008), which indicates that using more covariates may lead to more variable results. Furthermore, the univariate Cox regression model is estimated based on the training data only, which is a universally recommended approach.

TGD Cox. The threshold gradient descent procedure for the Cox regression analysis in the high-dimensional and low-sample size setting approximates the Lasso or LARS estimates, while selecting more relevant genes, which is also the

reason why we did not include Lasso directly in our simulation study. The method is described in Gui and Li (2005). The approach has two parameters but they rarely need to be tuned, and can instead be chosen by minimizing a cross-validated partial likelihood. The complete method, including the dimensional reduction and the ability to capture correlated genes, is discussed in details in the above cited paper and implemented as an R script available at http://www.cceb.upenn.edu/~hli/prog.html.

The methods are compared in a simulation study. As in van Wieringen et al. (2008) two artificial data sets are used. In the first data set the survival times are generated independently of the gene expression data. Its results give an indication of the performance of the tested algorithms when there is no predictive power in the expression data. The other simulated data set was introduced by Bair et al. (2006), also for evaluation purposes.

Design of artificial data sets

Each artificial data set used in the simulation study consists of $p = 500$ variables and $n = 100$ samples. The survival times and covariate values are distributed as follows.
Data set 1: The columns of the design matrix are samples from a multivariate normal distribution with a given non-diagonal covariation matrix. The survival and censoring times (with censoring probability 1/3) are exponentially distributed. They are independent from each other as well as from the covariates data. Hence, there is no prediction power in the covariates.
Data set 2: Following Bair et al. (2006) the covariate data are distributed as:

$$\log(Z_{ij}) = \begin{cases} 3 + \epsilon_{ij} & \text{if} \quad i \leq n/2, j \leq 30 \\ 4 + \epsilon_{ij} & \text{if} \quad i > n/2, j \leq 30 \\ 3.4 + \epsilon_{ij} & \text{if} \quad j > 30 \end{cases}$$

where the ϵ_{ij} are drawn from a standard normal distribution. The survival and censoring times (with censoring probability 1/3) are generated from an accelerated failure model in which only the values of covariates 1 to 30 (with additional noise) contribute. In other words, only the first 30 covariates determine the survival.

As noted in van Wieringen et al. (2008), it is not straightforward to evaluate or compare prediction methods in the presence of censoring. The standard mean-squared-error or misclassification rate criteria used in regression or classification cannot be applied to censored survival times. In the simulations, we used three measures to evaluate the prediction of the compared methods: the p-value (likelihood ratio test) of Bair et al. (2006), which is in fact the probability of drawing the observed data under the null-hypothesis that the covariates have no effect on survival (the lower the p-value, the more probable that the null hypothesis is not true); a goodness-of-fit measure for the proportional hazard model based on the variance of the martingale residuals proposed by Barlow and Prentice (1988) (the smaller the better); and the integrated Brier-Score introduced by Graf et al. (1999). The values of the Brier-Score are between 0 and 1 and good predictions result in small Brier-Scores. A detailed description of these measures is given in van Wieringen et al. (2008). The first two measures are based on the Cox model, while the Brier score uses the predicted survival curves, which can be derived via other approaches. For applying the evaluation measures to our prediction methods, we simply extract the predicted median survival time from the predicted survival curves and use it as a predictor in a univariate Cox model. This approach, though possibly suboptimal, allows to compare all the prediction methods with these three evaluation measures.

Simulation results

The data sets described above were generated 50 times, and randomly split into training and test sets with a 7:3 ratio. The survival prediction methods were applied to the training sets, and the test set was then used for calculation of the evaluation measures (p-value, variance of martingale residuals and Brier score as implemented in the R package `ipred`). The hyperparameters needed for the TGD and the DS methods were determined by cross-validation on the training sample.

The results are plotted and summarized in the figures and tables that follow. Figures 4.1, 4.2 and 4.3 show evaluation measure boxplots for the results of each of the five methods. The boxplots are grouped by method: two boxplots for the two artificial data sets per method. The coding of the methods underneath the boxplots is explained in Tables 1, 2 and 3 which also contain the summary statistics of the results for the three evaluation measures. The median and IQR are given to match the characteristic

features of the boxplots.

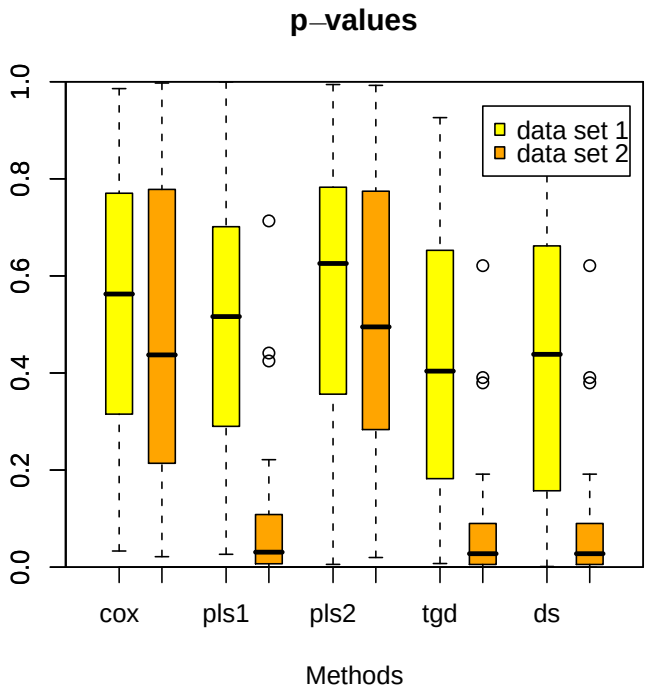

Figure 4.1: Box plots of the p-values for each method over the 50 simulations of each data set. The lower the p-value, the more probable is that the covariates have predictive power.

With respect to the variance of the martingale residuals, no method clearly stands out. They all perform more or less alike. Hence, the variance of the martingale residuals is not very discriminative as an evaluation measure for survival prediction methods.

The smaller the Brier score, the better the survival prediction. Focusing on the second data set where the expression data contains predictive information on survival, we observe that PLS1, PLS2 and DS have a similar good performance. Exceptions are the Cox with 20 genes method and the TGD Cox regression, which do not perform so well, even falling behind the simple Cox regression with univariate feature selection. A closer look at this method revealed that for data set 2 sometimes no features were selected, leading to poor evaluation measures. We believe this is partially due to the choice of the tuning parameters in the cross-validation, forcing the method to choose between either the maximum (no features included) or a value that leads to a poor

Method	Coded as	Data set	Median	IQR
Cox regression with 20 best genes	COX	ds1	0.563	0.445
Cox regression with 20 best genes	COX	ds2	0.437	0.555
PLS Cox (1 comp)	PLS1	ds1	0.516	0.399
PLS Cox (1 comp)	PLS1	ds2	0.031	0.099
PLS Cox (2 comp)	PLS2	ds1	0.626	0.412
PLS Cox (2 comp)	PLS2	ds2	0.495	0.483
TGD Cox regression	TGD	ds1	0.404	0.121
TGD Cox regression	TGD	ds2	0.028	0.084
Dantzig Selector	DS	ds1	0.438	0.492
Dantzig Selector	DS	ds2	0.027	0.084

Table 1: Results for the simulated data sets: p-values.

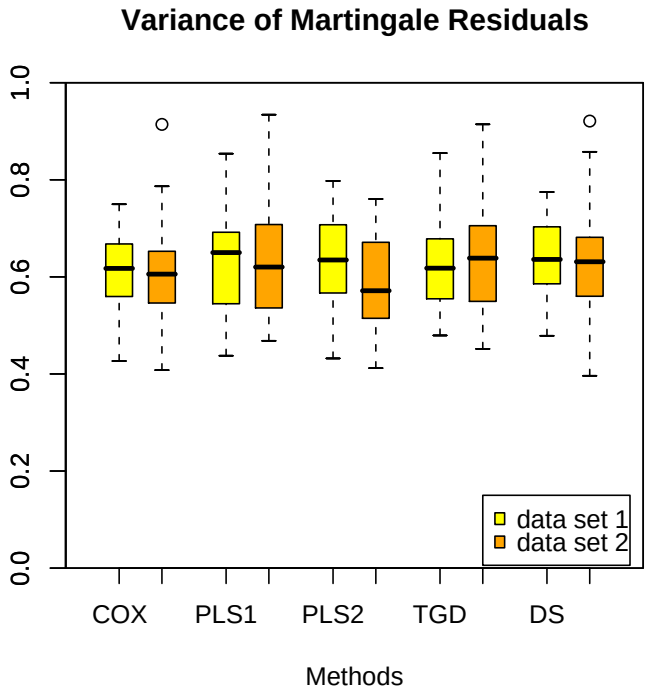

Figure 4.2: Box plots of the variance of martingale residuals for each method over the 50 simulations of each data set.

Method	Coded as	Data set	Median	IQR
Cox regression with 20 best genes	COX	ds1	0.617	0.106
Cox regression with 20 best genes	COX	ds2	0.606	0.104
PLS Cox (1 comp)	PLS1	ds1	0.650	0.144
PLS Cox (1 comp)	PLS1	ds2	0.620	0.165
PLS Cox (2 comp)	PLS2	ds1	0.635	0.138
PLS Cox (2 comp)	PLS2	ds2	0.571	0.156
TGD Cox regression	TGD	ds1	0.618	0.121
TGD Cox regression	TGD	ds2	0.639	0.150
Dantzig Selector	DS	ds1	0.636	0.114
Dantzig Selector	DS	ds2	0.631	0.120

Table 2: Results for the simulated data sets: variance of martingale residuals.

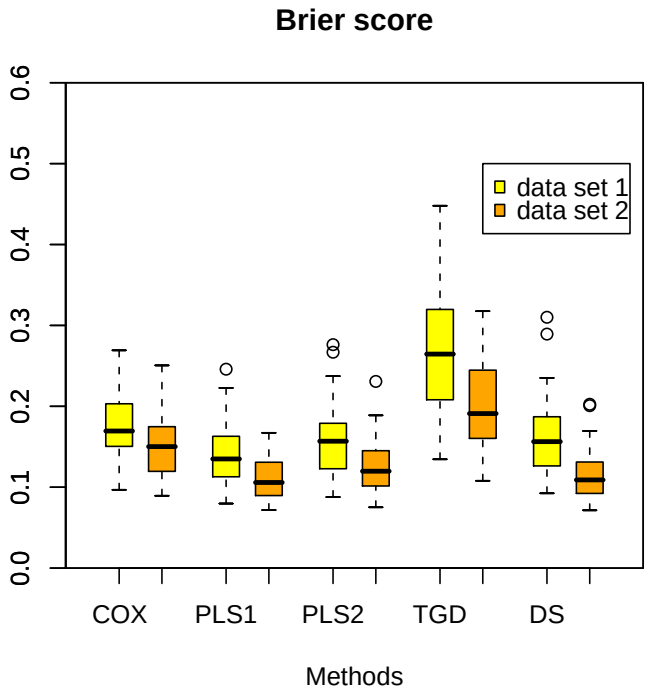

Figure 4.3: Box plots of the Brier prediction score for each method over the 50 simulations of each data set.

Method	Coded as	Data set	Median	IQR
Cox regression with 20 best genes	COX	ds1	0.169	0.052
Cox regression with 20 best genes	COX	ds2	0.150	0.054
PLS Cox (1 comp)	PLS1	ds1	0.135	0.049
PLS Cox (1 comp)	PLS1	ds2	0.106	0.040
PLS Cox (2 comp)	PLS2	ds1	0.157	0.055
PLS Cox (2 comp)	PLS2	ds2	0.120	0.043
TGD Cox regression	TGD	ds1	0.264	0.107
TGD Cox regression	TGD	ds2	0.191	0.082
Dantzig Selector	DS	ds1	0.156	0.060
Dantzig Selector	DS	ds2	0.109	0.037

Table 3: Results for the simulated data sets: Brier scores (the lower the better).

prediction.

In the simulations, we focused on finding the best prediction rule for the time to an adverse event using all the available covariates measurements. However, if we bear in mind that in many studies, the main focus is on finding a small subset of the covariates that are the most important ones for predicting survival, we find the survival Dantzig selector very interesting, as it also is a variable selection method. Note that the SDS selector picked on average as few as 15 genes (median over the 50 splits) for the second data set and as few as 3 genes for the first data set.

5 Analysis of a real-life data set

In this section, we compare the performance of the prediction methods on a real-life data set from survival gene expression data. As in Bovelstad et al. (2007), we use a well known real-life data set, namely the **Dutch breast cancer data** used by Van't Veer et al. (2002) to build a model to predict the time to metastasis of breast cancer in patients based on microarray data from 78 patients. The expression levels of $p = 4919$ genes were available for this study (consisting of 78 patients). In order to evaluate the methods we divided the data set randomly into two parts; a training set of about 2/3 of the patients used for estimation and a test set of about 1/3 of the patients used

for evaluation or testing of the prediction capability of the estimated model. The split was done 50 times and in such a way that the proportion of censored observations in the original data set was respected. The results are plotted and summarized in the following figures and tables.

As shown in the simulations, the variance of the martingale residuals was not highly discriminative as an evaluation measure for survival prediction. Bearing this in mind, for this real-data case we only used the p-values and the Brier scores as evaluation measures of predictive performance. Figures 5.4 and 5.5 show boxplots for the evaluation measures of the results for each of the five methods. Table 4 and Table 5 contain the summary statistics of the results for the two evaluation measures over the 50 iterations. The median and IQR are given to match the characteristic features of the boxplots.

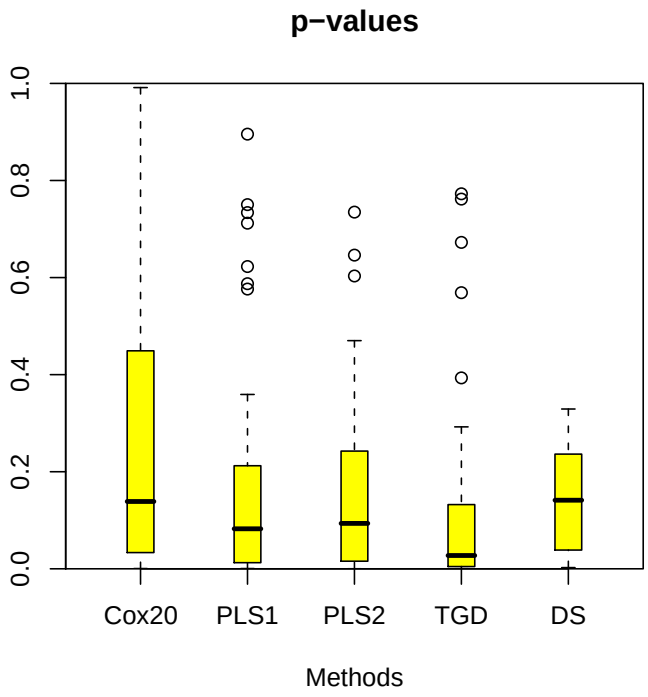

Figure 5.4: Box plots of the p-values for each method over the 50 simulations for the Breast Cancer data set.

With respect to the variance of the martingale residuals, as for the simulation, no method clearly stands out. Both the boxplots in Figure 5.5 and Table 5 indicate that

Method	Median	IQR
COX	0.139	0.406
PLS1	0.082	0.181
PLS2	0.094	0.217
TGD	0.027	0.120
DS	0.141	0.194

Table 4: Results for the Breast Cancer data: p-values over the 50 splits.

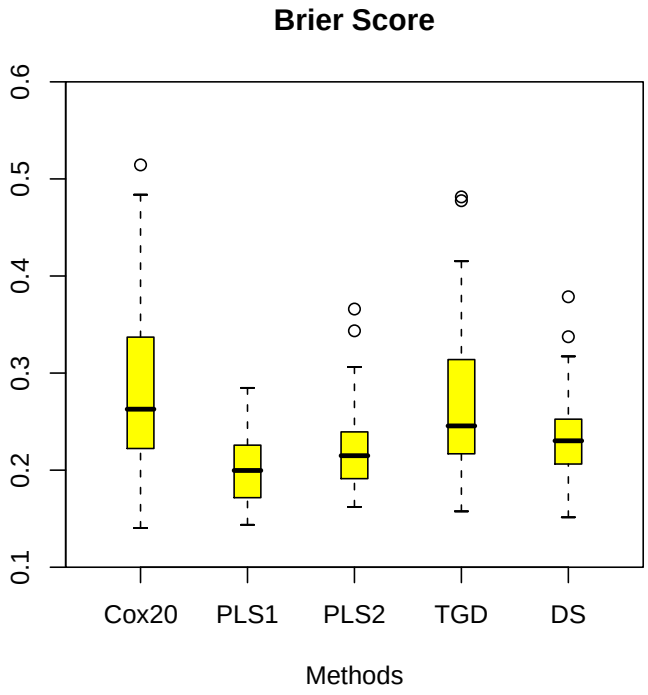

Figure 5.5: Box plots of the Brier prediction score for each method over the 50 simulations for the Breast Cancer data set.

the PLS based methods and the Dantzig selector have the smallest Brier score, with the Dantzig selector also having the smallest IQR. Remembering that the PLS components are built out of a combination of genes, the Dantzig selector is therefore preferable in terms of interpretability for the breast cancer data set.

Method	Median	IQR
COX	0.263	0.113
PLS1	0.199	0.052
PLS2	0.215	0.047
TGD	0.246	0.093
DS	0.230	0.045

Table 5: Results for the Breast Cancer data: Brier scores over the 50 splits.

6 Conclusions

We compared our Dantzig selector method for survival data to several previously published methods for predicting survival and applied it on some simulated data and also on a survival study based on microarray data. Our method performed well on simulations and for real data compared to these existing methods. Another important advantage of the Dantzig selector is that it selects a subset of the genes to use as predictors. The PLS based method that had a comparable predicting power, by contrast, require the use of all (or a large number) of the genes.

We close with a few further remarks. We acknowledge that previous work (Lounici (2008); James and Radchenko (2009)) established links between the Danztig selector and LASSO for linear models, also as variable selectors. We note that establishing a possible similar connection between the two procedures in Cox's model appears challenging and is out of scope of the present work. It is also unclear to us whether or how it is possible to rapidly compute entire solution paths for the Survival Dantzig Selector; we note that generalised path algorithms for penalised optimisation problems for loss functions different from least-squares are not obvious to construct or known to exist (Rosset and Zhu (2007)).

Acknowledgements

Piotr Fryzlewicz would like to thank Anestis Antoniadis for his hospitality while visiting the Department of Statistics, LJK to carry out this work. Financial support form the IAP research network Nr. P6/03 of the Belgian government (Belgian Federal Science Policy) is gratefully acknowledged. The authors thank Anne-Laure Boulesteix

for kindly providing the R code for the PLS Cox regression described in Boulesteix and Strimmer (2007) and the simulation designs and also Gareth James for kindly providing his Dantzig selector R code implementation for GLM.

7 Appendix: Proofs

This section is devoted to the proofs of our main theoretical results stated in the paper. **Proof of Lemma 3.1**. We have to control $\mathbb{P}(\|U(\boldsymbol{\beta}_0)\|_\infty < \gamma)$ as $n, p \to \infty$. That is, we are studying the event

$$\sup_j \left| \frac{1}{n} \sum_{i=1}^n \int_0^\tau dM_i(u) \left[\sum_{k=1}^n \left\{ Z_{ij} - Z_{kj} \right\} w_k(\boldsymbol{\beta}_0, u) \right] \right| \geq \gamma,$$

where

$$w_k(\boldsymbol{\beta}, u) = \frac{\exp(\mathbf{z}_k^T \boldsymbol{\beta}) Y_k(u)}{\sum_l \exp(\mathbf{z}_l^T \boldsymbol{\beta}) Y_l(u)}.$$

Note that the $w_k(\boldsymbol{\beta}, u), u \in [0, \tau)$ are nonegative and sum to one. Let

$$g_{n,i,j}(u) = \sum_{k=1}^n (Z_{ij} - Z_{kj}) w_k(\boldsymbol{\beta}, u).$$

Note that $g_{n,i,j}(u)$ inherits from $Y_k(u)$ all measurability properties, so it is a predictable process. Thus, for each i, j, $\int_0^\tau g_{n,i,j}(u) dM_i(u)$ is a martingale, which implies that $M_{n,j} = \frac{1}{n} \sum_{i=1}^n \int_0^\tau g_{n,i,j}(u) dM_i(u)$ is a martingale. We use now Lemma 2.1 from van de Geer (1995), which comes from Shorack and Wellner (1986). For that purpose, we need to compute the quantities $\Delta M_{n,j}(u)$ (magnitude of a jump in $M_{n,j}$ if it occurs at time u) and $V_{n,j}(u)$ (the variation process of $M_{n,j}(u)$).

Since the jumps of the processes M_i do not occur at the same time and are all of magnitude one, we have

$$|\Delta M_{n,j}(u)| \leq \sup_{1 \leq i \leq n} \frac{\|g_{n,i,j}\|_\infty}{n} \leq \sup_{i,j,k} \frac{|Z_{i,j} - Z_{k,j}|}{n} \sum_{k=1}^n w_k(u) \leq 2 \sup_j \frac{\|\mathbf{z}^j\|_2}{n} = \frac{2C}{n}.$$

For the variation process, we use the fact that

$$\langle \int_0^\tau H_u dM_u, \int_0^\tau H'_u dM'_u \rangle = \int_0^\tau H_u H'_u d\langle M, M' \rangle_u,$$

where H, H' are square integrable predictable processes, and M and M' are square integrable martingales. Since the M_i are independent, we have

$$\begin{aligned} V_{n,j}(\tau) &= \frac{1}{n^2}\sum_{i=1}^{n}\int_0^\tau g_{n,i,j}^2(u)d\langle M_i, M_i\rangle_u \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\int_0^\tau g_{n,i,j}^2(u)\exp(\mathbf{z}_i^T\boldsymbol{\beta}_0)Y_i(u)\alpha_0(u)du \\ &\leq \frac{4}{n^2}\|\mathbf{z}^j\|_2^2 \sup_{u\in[0,\tau]}\{S_n(\boldsymbol{\beta}_0,u)\}\|\alpha_0\|_1. \end{aligned}$$

We have

$$\sup_{u\in[0,\tau]}\{S_n(\boldsymbol{\beta}_0,u)\} \leq \sum_{i=1}^{n}\exp(\mathbf{z}_i^T\boldsymbol{\beta}_0) \leq n\exp(S\|\beta_0\|_\infty \sup_j\|\mathbf{z}^j\|_2) = O(n),$$

so that $V_{n,j}(\tau) \leq \frac{K}{n}$ for a suitable constant K. We will now use the exponential inequality from Shorack and Wellner (1986).

$$\begin{aligned} &P\left(\sup_j\left|\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau dM_i(u)\left[\sum_{k=1}^{n}\left\{Z_{ij}-Z_{kj}\right\}w_k(\boldsymbol{\beta}_0,u)\right]\right| \geq \gamma\right) \leq \\ &\sum_j P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau dM_i(u)\left[\sum_{k=1}^{n}\left\{Z_{ij}-Z_{kj}\right\}w_k(\boldsymbol{\beta}_0,u)\right]\right| \geq \gamma\right) = \\ &\sum_j P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau dM_i(u)\left[\sum_{k=1}^{n}\left\{Z_{ij}-Z_{kj}\right\}w_k(\boldsymbol{\beta}_0,u)\right]\right| \geq \gamma \cap V_{n,j}(\tau) \leq \frac{K}{n}\right) \leq \\ &p\exp\left(-\frac{n\gamma^2}{2(2C\gamma+K)}\right). \end{aligned}$$

Our choice of γ allows us to conclude.

Proof of Theorem 3.1. To prove the result, we will also need the following Lemma which we state with no proof since it is a straightforward generalization of Lemma 3.1 in Candès and Tao (2007).

Lemma 7.1 *Let A be an $n \times p$ matrix and suppose $T_0 \subset \{1,\ldots,p\}$ is a set of cardinality S. For a vector $h \in \mathbb{R}^p$, let T_1 be the S' largest positions of h outside of T_0 and put $T_{01} = T_0 \cup T_1$. Then*

$$\begin{aligned} \|h_{T_{01}}\|_2 &\leq \frac{1}{\delta_{S+S'}}\|A_{T_{01}}^T Ah\|_2 + \frac{\theta_{S',S+S'}}{\delta_{S+S'}(S')^{1/2}}\|h_{T_0^c}\|_1 \\ \|h\|_2^2 &\leq \|h_{T_{01}}\|_2^2 + (S')^{-1}\|h_{T_0^c}\|_1^2. \end{aligned}$$

To prove the Theorem we need to establish that $\|U(\beta_0)\|_\infty \leq \gamma$ implies that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2^2 \leq 64S(\frac{\gamma}{\delta_{2S} - \theta_{S,2S}})^2$. Assume that $\|U(\beta_0)\|_\infty \leq \gamma$ where

$$\|U(\boldsymbol{\beta}_0)\|_\infty = \sup_j \left| \frac{1}{n} \sum_{i=1}^n \int_0^\tau dM_i(u) \left[\sum_{k=1}^n \left\{ Z_{ij} - Z_{kj} \right\} w_k(u) \right] \right|.$$

Recall here that for any consistent estimator $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$, we may write:

$$\begin{aligned} J(\tilde{\boldsymbol{\beta}}, \tau) - I(\boldsymbol{\beta}_0, \tau) &= \int_0^\tau (V_n(\tilde{\boldsymbol{\beta}}, u) - v(\tilde{\boldsymbol{\beta}}, u)) \frac{d\bar{N}(u)}{n} && (9) \\ &+ \int_0^\tau (v(\tilde{\boldsymbol{\beta}}, u) - v(\boldsymbol{\beta}_0, u)) \frac{d\bar{N}(u)}{n} && (10) \\ &+ \int_0^\tau v(\boldsymbol{\beta}_0, u) \frac{d\bar{M}(u)}{n} && (11) \\ &+ \int_0^\tau v(\boldsymbol{\beta}_0, u) (\frac{S_n(\boldsymbol{\beta}_0, u)}{n} - s(\boldsymbol{\beta}_0, u)) \alpha_0(u) du, && (12) \end{aligned}$$

where $V_n(\boldsymbol{\beta}, u) = \frac{S_n^2}{S_n}(\boldsymbol{\beta}, u) - (\frac{S_n^1}{S_n})^{\otimes 2}(\boldsymbol{\beta}, u)$ and $v(\boldsymbol{\beta}, u) = \frac{s^2}{s}(\boldsymbol{\beta}, u) - (\frac{s^1}{s})^{\otimes 2}(\boldsymbol{\beta}, u)$. Since $\boldsymbol{\beta}_0$ is a nonzero S-sparse vector with S independent of n and since the true information matrix $\mathcal{I}(\boldsymbol{\beta}_0, \tau)$ is positive definite at $\boldsymbol{\beta}_0$, for any $\boldsymbol{\beta}^*$ in an Euclidian ball $B_r = B(\boldsymbol{\beta}_0, r)$ centered at $\boldsymbol{\beta}_0$ and of radius at most $r = 8\sqrt{S}\frac{\gamma}{\delta_{2S} - \theta_{S,2S}}$, the regularity conditions of Theorem 3.4 in Huang (1996) hold and it follows that

$$\sup_{\tilde{\boldsymbol{\beta}} \in B_r} \|J(\boldsymbol{\beta}^\star, \tau) - I(\boldsymbol{\beta}_0, \tau)\|_\infty = O_P(n^{-1/2}) \tag{13}$$

as n tends to ∞.

Define $h = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ and let T_0 be the support of $\boldsymbol{\beta}_0$. According to Lemma 3.1, we have $\|\hat{\boldsymbol{\beta}}\|_1 \leq \|\boldsymbol{\beta}_0\|_1$ and this inequality implies that $\|h_{T_0^c}\|_1 \leq \|h_{T_0}\|_1$, which yields, by Cauchy inequality,

$$\|h_{T_0^c}\|_1 \leq \|h_{T_0}\|_1 \leq S^{1/2} \|h_{T_0}\|_2. \tag{14}$$

By assumption, we have $\|U(\boldsymbol{\beta}_0)\|_\infty \leq \gamma$ and by construction of the estimator, $\|U(\hat{\boldsymbol{\beta}})\|_\infty \leq \gamma$. Adding up the two inequalities (triangle inequality)

$$\|U(\boldsymbol{\beta}) - U(\hat{\boldsymbol{\beta}})\|_\infty \leq 2\gamma$$

By Andersen and Gill (1982), formula (2.6), we have, Taylor-expanding the LHS of the above,

$$\left\| J(\boldsymbol{\beta}^*, \tau)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \right\|_\infty \leq 2\gamma, \tag{15}$$

where $\boldsymbol{\beta}^*$ lies within the segment between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$.

Now, using our remark (13) on the behavior of the matrix $I(\boldsymbol{\beta}_0, \tau)$ at the neighborhood of $\boldsymbol{\beta}_0$ we have

$$\begin{aligned}
\left\| I(\boldsymbol{\beta_0}, \tau)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \right\|_\infty &\leq \left\| (J(\boldsymbol{\beta}^*, \tau) - I(\boldsymbol{\beta_0}, \tau))(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \right\|_\infty + \left\| J(\boldsymbol{\beta}^*, \tau)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0}) \right\|_\infty \\
&\leq Dn^{-1/2} \left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0} \right\|_1 + 2\gamma, \\
&\leq 4\gamma,
\end{aligned}$$

for n large enough, since $\left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta_0} \right\|_1 \leq \left\| \hat{\boldsymbol{\beta}} \right\|_1 + \|\boldsymbol{\beta_0}\|_1 \leq 2\|\boldsymbol{\beta_0}\|_1$. Hence, if $A = I(\beta_0, \tau)^{1/2}$ denotes the squared root of the (semi)definite positive matrix $I(\beta_0, \tau)$, we have

$$\|AAh\|_\infty \leq 4\gamma.$$

This, again by Cauchy inequality, implies $\|A^T_{T_{01}} Ah\|_2 \leq 4(S + S')^{1/2}\gamma$. Take $S' = S$. By the first inequality of Lemma 7.1 and inequality (14), we have

$$\begin{aligned}
\|h_{T_{01}}\|_2 &\leq \frac{4}{\delta_{2S}}(2S)^{1/2}\gamma + \frac{\theta_{S,2S}}{\delta_{2S} S^{1/2}} S^{1/2} \|h_{T_0}\|_2 \\
&\leq \frac{4}{\delta_{2S}}(2S)^{1/2}\gamma + \frac{\theta_{S,2S}}{\delta_{2S}} \|h_{T_{01}}\|_2.
\end{aligned}$$

Rearranging for $\|h_{T_{01}}\|_2$, we get

$$\begin{aligned}
\|h_{T_{01}}\|_2 \left(1 - \frac{\theta_{S,2S}}{\delta_{2S}}\right) &\leq \frac{4}{\delta_{2S}}(2S)^{1/2}\gamma \\
\|h_{T_{01}}\|_2 &\leq \frac{4}{\delta_{2S} - \theta_{S,2S}}(2S)^{1/2}\gamma.
\end{aligned}$$

By the second inequality of Lemma 7.1 and inequality (14), we have

$$\|h\|_2^2 \leq \|h_{T_{01}}\|_2^2 + S^{-1}S\|h_{T_0}\|_2^2 \leq 2\|h_{T_{01}}\|_2^2 \leq 64S\left(\frac{\gamma}{\delta_{2S} - \theta_{S,2S}}\right)^2,$$

which completes the proof of the Theorem.

References

ANDERSEN, P. K., BORGAN, O., GILL, R. D. and KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. Springer, New York.

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

BAIR, E., HASTIE, T., PAUL, D. and TIBSHIRANI, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, **101** 119–137.

BARLOW, W. E. and PRENTICE, R. L. (1988). Residuals for relative risk regression. *Biometrika*, **75** 65–74.

BICKEL, P., RITOV, Y. and TSYBAKOV, A. (2008). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, **To appear**.

BOULESTEIX, A. and STRIMMER, K. (2007). Partial least squares: A versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, **8** 24–32.

BOVELSTAD, H., NYGARD, S., STORVOLD, H., ALDRIN, M., BORGAN, O., FRIGESSI, A. and LINGJAERDE, O. C. (2007). Predicting survival from microarray data - a comparative study. *Bioinformatics*, **23** 2080–2087.

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

CRAVEN, P. and WAHBA, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.*, **31**.

DELONG, D., GUIRGUIS, G. and SO, Y. (1994). Efficient computation of subset selection probabilities with application to Cox regression. *Biometrika*, **81** 607–611.

FAN, J. and LI, R. (2002). Variable selection for Cox's proportional hazards model and frailty model. *Annals of Statistics*, **30** 74–99.

FARAGGI, D. and SIMON, R. (1998). Bayesian variable selection method for censored survival data. *Biometrics*, **54** 1475–1485.

GRAF, E., SCHMOOR, C., SAUERBREI, W. and SCHUMACHER, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, **18** 2529–2545.

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

HARRELL, F. E. (2001). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer, New York.

HORN, R. A. and JOHNSON, C. R. (1985). *Matrix Analysis*. Cambridge University Press.

HUANG, J. (1996). Efficient estimation for the Cox model with interval censoring. *Annals of Statistics*, **24** 540–568.

IBRAHIM, J., CHEN, M.-H. and KIM, S. (2008). Bayesian variable selection for the Cox regression model with missing covariates. *Lifetime Data Analysis*, **14** 496–520.

JAMES, G. and RADCHENKO, P. (2009). A generalized Dantzig selector with shrinkage tuning. *Biometrika*, **To appear**.

JENSSEN, T., KUO, W., STOKKE, T. and HOVIG, E. (2002). Associations between gene expressions in beast cancer and patient survival. *Human Genetics*, **111** 411–420.

JOVANOVIC, B. D., HOSMER, D. and BUONACCORSI, J. P. (1995). Equivalence of several methods for efficient best subsets selection in generalized linear models. *Computational Statistics and Data Analysis*, **20** 59–64.

LOUNICI, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, **2** 90–102.

MARX, B. D. (1996). Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics*, **38** 374–381.

MCCULLAGH, P. and NELDER, J. (1989). *Generalized linear models*. 2nd ed. Chapman and Hall.

NGUYEN, D. V. and ROCKE, D. M. (2002). Partial least squares proportional hazard regression for application to DNA microarray survival data. *Bioinformatics*, **18** 1625–1632.

PARK, P., TIAN, L. and KOHANE, I. (2002). Linking expression data with patient survival times using partial least squares. *Bioinformatics*, **18** 120–127.

ROSSET, S. and ZHU, J. (2007). Piecewise linear regularized solution paths. *Annals of Statistics*, **35** 1012–1030.

SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.

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

VAN DE GEER, S. (1995). Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Annals of Statistics*, **23**.

VAN WIERINGEN, D., KUN, D., HAMPEL, R. and BOULESTEIX, A.-L. (2008). Survival prediction using gene expression data: a review and comparison. *Computational Statistics and Data Analysis*, **To appear**.

VAN'T VEER, L. J., DAI, H., VAN DE VIJVER, M. J., HE, Y., HART, A., MAO, M., PETERSE, H., VAN DER KOOY, K., MARTON, M., WITTEVEEN, A., SCHREIBER, G., KERKHOVEN, R., ROBERTS, C., LINSLEY, P., BERNARDS, R. and FRIEND, S. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415** 530–536.

WANG, H. and LENG, C. (2007). Unified lasso estimation via least square approximation. *Journal of American Statistical Association*, **102** 1039–1048.

ZHANG, H. H. and LU, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, **94** 691–703.

ZOU, H. (2008). A note on path-based variable selection in the penalized proportional hazards model. *Biometrika*, **95** 241–247.