Statistica Sinica **29** (2019), 1939-1961
doi:https://doi.org/10.5705/ss.202017.0089

PSEUDO-VALUE METHOD FOR ULTRA HIGH-DIMENSIONAL SEMIPARAMETRIC MODELS WITH LIFETIME DATA

Tony Sit[1], Yue Xing[2], Yongze Xu[1] and Minggao Gu[1]

[1]*The Chinese University of Hong Kong and* [2]*Purdue University*

Abstract: We develop a new procedure called the "pseduo-value method" (PVM) for ultra high-dimensional variable selection problems in semiparametric survival models. Currently, the prevailing strategies available for working with ultra high-dimensional lifetime data are the sure independence screening (SIS) strategies. The proposed unified methodology covers a much broader class of survival models, including general transformation models and the accelerated failure time (AFT) model. The proposed method is versatile because the conversion involved easily casts the problem of interest as a regular linear regression. Through this translation, all existing techniques developed for linear regression problems can be leveraged at almost no extra cost. The numerical performance of the PVM shows promising results: in addition to outperforming the (iterative) SIS for the Cox model, the new method accurately selects the effective variables for probit, proportional odds, and AFT models, which have been studied in ultra high-dimensional contexts on a case-by-case basis. We apply our unified method to analyze diffuse large-B-cell lymphoma data, finding genes that may be overlooked, but that could be influential. This finding is potentially of scientific importance on its own.

Key words and phrases: Accelerated failure time model, penalized log-marginal likelihood, semiparametric models, transformation models, variable selection.

1. Introduction

Time-to-event data, which are characterized by the presence of (right-) censored observations, are often collected in clinical studies. Survival analyses attempt to model the dependence of the survival time T of a subject on the covariate variables $\mathbf{Z} = (Z_1, \ldots, Z_p)^\top$, where p indicates the dimensionality of the covariate space. Variable selection has been studied extensively since the mid-1990s, with rapid technological development making the collection of vast amounts of data technically and economically feasible.

For lifetime data, a popular class of semiparametric models is that of transformation models, of which the Cox (1972, 1975) proportional hazards model and the proportional odds model (Bennett (1983)) are special cases. Given conventional high-dimensional data sets, variable selection for the Cox model is usually carried out using the parametric partial likelihood. Here, penalties are often imposed, such as the least absolute shrinkage and selection operator (LASSO, (Tibshirani (1996))), smoothly clipped absolute deviation (SCAD, (Fan and Li (2001))), least angle regression selection (LARS, (Efron et al. (2004))), and elastic net (Zou and Hastie (2005)).

For the Cox model, Li and Luan (2003) proposed a procedure that uses reproducing kernel Hilbert spaces for inferences on censored data. In addition, Gui and Li (2005) and Antoniadis, Fryzlewicz and Letue (2010) employ a threshold gradient descent regularization and the Dantzig selector, respectively, for inference problems using the Cox model. For the proportional odds model, Lu and Zhang (2007) proposed an inference procedure that uses a penalized marginal likelihood of ranks. This was later extended by Li and Gu (2012) to include a more general family of transformation models; see also Li et al. (2014).

An attractive alternative to the Cox model is the accelerated failure time model (AFT), which directly relates the log of the survival time to the covariates; see Ying (1993) and Jin et al. (2003), among others. This model offers a straightforward interpretation and is more appealing than the proportional hazards model in many aspects. Several variable selection methods have been proposed. Huang and Harrington (2005) applied a LASSO-type penalty to a Buckley−James-type estimator. Furthermore, a rank-based variable selection procedure for regular high-dimensional data was studied by Cai, Huang and Tian (2009) and Xu, Leng and Ying (2010) for the AFT model because in this case, unlike the Cox model, the partial likelihood is not available.

Microarray, proteomic, and SNP data from bioimaging technology studies have induced a recent surge of interest in variable selection in ultra high-dimensional settings. Here, problems of interest involve an exponentially growing parameter space with respect to the sample size, that is, $\log(p) = \mathcal{O}(n^{\alpha})$, for $\alpha \in (0, 1/2)$. In view of this new statistical challenge, Fan, Yang and Wu (2010) applied the sure independence screening (SIS; see also (Fan and Lv (2008))) method to Cox's proportional hazard models, and Song et al. (2014) applied SIS to censored rank data for transformation models with an independent censoring assumption. Recently, Khan and Shaw (2016) extended the weighted least squares formulation of Stute (1993, 1996) to a class of elastic net techniques to

handle ultra high-dimensional data. However, there is no unified procedure for other lifetime data models in ultra high-dimension frameworks for a more general class of semiparametric transformation models. Here, the main difficulty remains the handling of censored data. Here, unlike the special case of the Cox model, where the partial likelihood can serve as an effective vehicle for inferences, we may need to handle the likelihood or (rank-based) estimating equations directly to make the variable selection procedure possible.

Ing and Lai (2011) developed the orthogonal greedy algorithm (OGA), which is a stepwise regression modified for (ultra) high-dimensional linear regression models. Coupled with the high-dimensional information criterion (HDIC), the efficient OGA algorithm avoids a potentially restrictive assumption on the maximum eigenvalue of the covariance matrix of the candidate regressors, which may not hold when all regressors are equally correlated. The orthogonal projections carried out in each forward selection step ensure that all remaining variables become perpendicular to the selected variable(s). As a result, the OGA tends to exhibit lower false selection rates compared with its SIS counterpart. Thus, smaller correct models can be selected. To the best of our knowledge, no studies have examined how this powerful tool can be applied to time-to-event data.

The key objective of this study is to develop a general approach that utilizes "pseduo-values" as a bridge between inference problems for survival data and those appearing in conventional linear regression models in a ultra high-dimensional setting. Specifically, pseudo values can be regarded as a set of educated guesses for the response variables, some of which are not fully observable owing to censoring. We make two contributions to the literature. First, the proposed method offers a solution to open challenges for modeling ultra high-dimensional lifetime data. Second, the concept of pseudo values facilitates the use of existing variable selection tools used in linear models for more general model settings. The code for the simulations and numerical studies, composed in `MATLAB`, are available upon request.

The remainder of the paper is organized as follows: Section 2 describes the proposed pseudo value method (PVM) using two popular classes of models for lifetime data: (i) general transformation models, and (ii) the AFT model. Sections 3 and 4 present simulations and analyses based on Stanford heart transplant data and diffuse large-B-cell lymphoma data, respectively. Concluding remarks are presented in Section 5.

2. Methodology and Algorithm

To facilitate the discussion, we first introduce some standard notation. In this paper, we let $(T, C, \mathbf{Z})$ denote a triplet of the survival time, the censoring time, and their associated covariates, respectively. Under the conditional independence censoring mechanism, that is, $C \perp T \mid \mathbf{Z}$, we can only observe $\tilde{T} = \min(T, C)$, with $\Delta = I(T \leq C)$ as the censoring indicator. The observed data set includes independent and identically distributed (i.i.d.) samples of the triplet $(\tilde{T}, \Delta, \mathbf{Z})$, denoted by $\{(\tilde{T}_i, \Delta_i, \mathbf{Z}_i)\}_{i=1,\ldots,n}$, where $\dim(\mathbf{Z}_i) = p$. For $i = 1, \ldots, n$, we also let $Y_i = \mathbf{Z}_i^\top \boldsymbol{\beta}$.

Traditionally, estimations of $\boldsymbol{\beta}$ in regression problems are performed using (convex) optimization on an appropriate likelihood/loss function, say $L^*(\boldsymbol{\beta} \mid \mathcal{Z}) = L(\mathcal{Z}^\top \boldsymbol{\beta} \mid \mathcal{Z})$, where $\mathcal{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ is of dimension $p \times n$. However, in (ultra-) high dimensional settings, it is difficult to estimate the optimizer $\hat{\boldsymbol{\beta}}$ directly with respect to L^*. Thus, the key contribution of our methodology lies in the observation that the effect of $\boldsymbol{\beta}$ is manifested through $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top = \mathcal{Z}^\top \boldsymbol{\beta}$. As a result, we first obtain a set of values $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)^\top$ that maximizes L. Then, we use these optimizers to estimate $\hat{\boldsymbol{\beta}}$. These optimizers $\hat{\mathbf{Y}}$ are the "pseduo-values". Note that the proposed method significantly reduces the dimension of the problem of interest because $\dim(\hat{\mathbf{Y}}) = n \ll p$. Using these pseudo values, we can apply the OGA with a high-dimensional information criterion (Ing and Lai (2011)) in the second stage of the inference without needing further modifications, owing to the model-specific settings.

A generic algorithm for the proposed method is composed of three stages:

Stage 1: Obtain an initial set of pseudo values that maximizes the objective function L. Recall that our problem of interest is to infer $\boldsymbol{\beta}$ in $H(T) = \mathbf{Z}^\top \boldsymbol{\beta} + \epsilon$ based on assumptions on the monotone transformation function $H(\cdot)$ and the residual ϵ for different models. Therefore, at this stage of the optimization, it is natural to impose a restriction that the pseudo values should lie on the linear span of $\mathcal{Z}$, denoted as $\text{span}(\mathcal{Z})$. In other words, we reparametrize the parameters of interest to resolve challenges resulting from the high dimensionality of $\boldsymbol{\beta}$.

Stage 2: We introduce a penalty at this stage to reduce the dimension of the problem. Traditionally, as in the LASSO, hard thresholding, and SCAD, a penalty $p_\lambda(\cdot)$ is imposed on the regression coefficient $\boldsymbol{\beta}$. Our procedure first treats $\mathbf{Y}$ as the parameter of interest. Thus, we suggest the following conversion. Because $\mathbf{Y} = \mathcal{Z}^\top \boldsymbol{\beta}$, we can write $\boldsymbol{\beta} = T_{\mathcal{Z}} \mathbf{Y}$, where $T_{\mathbf{Z}} = (\mathcal{Z}\mathcal{Z}^\top)^+ \mathcal{Z}$

and A^+ denotes the Moore−Penrose inverse of a matrix A. Generically, the optimization carried out in this stage can be expressed as

$$\hat{\mathbf{Y}} = \arg\max_{\mathbf{Y}} \left[\log\{L(\mathbf{Y})\} - P_\lambda(\mathbf{Y})\right], \tag{2.1}$$

where $P_\lambda(\mathbf{Y}) = \sum_{i=1}^{p} p_\lambda\{|T_{\mathcal{Z}}\mathbf{Y}_i|\}$ is an appropriate penalty. This step works properly because (2.1) is concave in $\mathbf{Y}$; see Boyd and Vandenberghe (2004). The results obtained in Stage 1 can be used as an initial set of values for this round of optimization. From another perspective, the introduction of this penalty adjusts for the fact that minimizing $\|\mathbf{Y} - \mathbf{Y}_0\|^2$ with respect to $\mathbf{Y}$ is not equivalent to minimizing $\|L(\mathbf{Y}) - L(\mathbf{Y}_0)\|^2$ with respect to $\mathbf{Y}$. In general, popular penalty approaches, including the LASSO, SCAD, and adaptive LASSO, can be applied to this method. Note that although the penalty applied here borrows the idea of penalizing the parameter of interest β, the main goal is to obtain pseudo values upon which the variable selection methods developed for linear models can be applied. In our simulation exercise in Section 3, we demonstrate that these three approaches all produce similar and desirable results.

Stage 3: Perform the variable selection procedure on $\boldsymbol{\beta}$ by solving a high-dimensional linear regression problem with $\hat{\mathbf{Y}}$ as the response variable; that is, we estimate $Y = \mathbf{Z}^\top\boldsymbol{\beta} + \epsilon$. Here, any effective variable selection procedure can be applied. After obtaining the active set, the set of nonzero $\boldsymbol{\beta}$, estimate the final regression coefficient using the classical procedures for low-dimensional settings.

In summary, the PVM procedure maximizes the corresponding objective function using $\hat{\mathbf{Y}}$ with an appropriate regularization penalty and linear constraints. The last stage is to apply the OGA to the transformed $\hat{\mathbf{Y}}$ on the original covariates in the linear regression to complete the final variable selection procedure. As shown in Sections 3 and A2, we divide the algorithm into three stages to demonstrate the computation time needed for each step. In the following subsections, we provide two generic examples (Sections 2.1 and 2.2) to show how to use this method for variable selection problems that appear in general transformation models and the AFT model.

2.1. General transformation models

General transformation models assume that the true underlying failure time T is related to the covariates in the following form:

$$S_{\mathbf{Z}}(t) = \Phi\{S_0(t), \mathbf{Z}, \boldsymbol{\beta}\}, \tag{2.2}$$

where $S_0(\cdot)$ is an unknown continuous baseline survival function, $S_{\mathbf{Z}}(\cdot)$ is the survival function of T, given $\mathbf{Z}$, and $\Phi(u, v, w)$ is assumed to be known, with $\Phi(0, v, w) = 0$ and $\Phi(1, v, w) = 1$ for any v and w. By letting $\Phi(u, v, w) = g^{-1}\{g(u) - v^\top w\}$, where $g^{-1}(\cdot) = 1 - F(\cdot) = \Pr\{\epsilon \geq \cdot\}$, we can rewrite (2.2) as $H(T) = \mathbf{Z}^\top \boldsymbol{\beta} + \epsilon$, which is the traditional transformation model; see Gu, Sun and Zuo (2005). This generalized class of transformation models covers a class of frailty models, heteroscedastic hazards regression models (Hsieh (2001)), and general heteroscedastic rank regression models/probit rank regression models (Chen and Little (2001)).

Suppose we have n i.i.d. observations $\{(\tilde{T}_i, \mathbf{Z}_i, \Delta_i)\}_{i=1,\ldots,n}$. We define $k_n = \sum_{i=1}^n \Delta_i$ as the total number of observed failure times. We also denote $\mathcal{R}_n^*$ as the partial ranking of the k_n observed failure times and the specific observations between each pair of uncensored failure times, and $\mathcal{R}_n$ as the complete ranking of the underlying failure times $T_n = \{T_1, \ldots, T_n\}^\top$. Given the observed $\mathcal{R}_n^*$, we define $\mathcal{S}_n$ as a set composed of all possible complete rankings $\mathcal{R}_n$ and

$$\begin{aligned}\mathcal{C}_n = \{&\boldsymbol{t}_n = (t_1, t_2, \ldots, t_n) : t_{i_1} < t_{i_2} < \cdots < t_{i_{k_n}},\, t_j \geq t_{i_r},\\ &\text{for } j \in \mathcal{L}_{i_r} \text{ and } 0 \leq r \leq k_n\}\end{aligned}$$

as a time set that is consistent with the order restriction $\mathcal{R}_n^*$. Here, i_r denotes the rth ordered observed failure and $\mathcal{L}_{i_r}$ is the set of censored observations in $[T_{i_r}, T_{i_{r+1}})$, with $T_{i_0} = 0$ and $T_{i_{k_n+1}} = \infty$. It follows that the marginal likelihood can be rewritten as

$$\begin{aligned}L(\mathbf{Y}) &= \Pr\{\mathcal{R}_n \in \mathcal{S}_n \mid \mathcal{R}_n^*\} = \Pr\{\boldsymbol{t}_n \in \mathcal{C}_n \mid \mathcal{R}_n^*\}\\ &= (-1)^n \int_{\boldsymbol{t}_n \in \mathcal{C}_n} \prod_{i=1}^n \phi\{S_0(t_i), Y_i\} \prod_{i=1}^n dS_0(t_i)\\ &= \int_\xi \prod_{i=1}^n \phi(1 - u_i, Y_i) \prod_{i=1}^n du_i,\end{aligned} \tag{2.3}$$

where $\phi_u(u, v) = \partial\Phi(u, v)/\partial u$. Here, ξ is the corresponding collection of Uni(0, 1) vectors consistent with the order restriction specified in $\mathcal{C}_n$.

As suggested in Gu and Kong (1998) and Gu, Wu and Huang (2014), the Monte Carlo method can be used to maximize (2.3). We assume that $\Phi(u, v)$ is twice differentiable with respect to u and v, and define $\phi_v(u, v) = \partial\phi(u, v)/\partial v$ and $S_i(\mathbf{Y}) = \partial \log L(\mathbf{Y})/\partial Y_i$ as the ith element of the score function $S(\mathbf{Y}) = \partial \log L(\mathbf{Y})/\partial \mathbf{Y}$. It follows that, for $i = 1, \ldots, n$ and $\mathbf{u} = (u_1, \ldots, u_n)^\top$,

$$S_i(Y) = \int_{\varepsilon} H(Y_i; u_i) p(\mathbf{u}, \mathbf{Y}) d\mathbf{u}, \tag{2.4}$$

where

$$H(Y_i, u_i) = \frac{\phi_v(1-u_i, Y_i)}{\phi_u(1-u_i, Y_i)},$$

with

$$p(\mathbf{u}, \mathbf{Y}) = \{L(\mathbf{Y})\}^{-1} \prod_{i=1}^{n} \phi_u(1-u_i, Y_i) I(\mathbf{u} \in \epsilon), \tag{2.5}$$

denotes the conditional density of $\mathbf{u}$.

To implement the PVM, we first solve the following optimization problem without performing variable selection. At this stage, we can obtain preliminary estimates for our pseudo values that we will use in the next stage of our procedure. The maximum likelihood estimates of $\mathbf{Y}$, given (2.3), can be obtained by maximizing

$$\log\{L(\mathbf{Y})\} \quad \text{subject to} \quad (\mathbf{I} - \mathbf{H}_{\mathcal{Z}})\mathbf{Y} = 0,$$

where $\mathbf{H}_{\mathcal{Z}} = \mathcal{Z}(\mathcal{Z}\mathcal{Z}^{\top})^{+}\mathcal{Z}^{\top}$. The constraint imposed here restricts $\mathbf{Y}$ to lie on the linear span of $\mathcal{Z}$. Equivalently, we can solve

$$S(\mathbf{Y}) = 0 \quad \text{subject to} \quad (\mathbf{I} - \mathbf{H}_{\mathcal{Z}})\mathbf{Y} = 0. \tag{2.6}$$

In many cases, $L(\mathbf{Y})$ is a log-concave function. For example, in the Cox proportional hazards model, we have

$$\phi_u(u, v) = \frac{\exp\{\log(u)e^{-v}\}e^{-v}}{u}.$$

Because $\phi(u, v)$ is a log-concave function for v, using Theorem 6 in Prekopa (1973), we have that $L(\mathbf{Y})$ is also log-concave. This problem can be solved using Newton's method with equality constraints; see Boyd and Vandenberghe (2004).

A difficulty with this optimization is that the integral is usually of a very high dimension and the normalising constant in $p(\mathbf{u}, \mathbf{Y})$ defined in (2.5) has no closed analytic expression. Thus, the computation of (2.6) cannot be solved trivially using standard numerical methods. Here, we adopt the Markov chain Monte Carlo-Stochastic Approximation algorithm (MCMC-SA), following Gu and Kong (1998) and Gu, Sun and Zuo (2005), for our initial pseudo value estimation. The corresponding algorithm is described as follows:

Step 1: Choose positive integers λ, m, and κ_1, an initial value $\mathbf{Y}^{(0)}$, an initial matrix $\mathbf{\Gamma}^{(0)}$, an initial data $\boldsymbol{U}_m^{(0)}$, and a sequence $\nu_k \downarrow 0$. Repeat (i) and (ii) κ_1 times:

(i) For a fixed k, set $\boldsymbol{U}_0^{(k)} = \boldsymbol{U}_m^{(k-1)}$. For $i = 1, \ldots, m$, generate $\boldsymbol{U}_i^{(k)}$ from the transition probability $\Pi_{\mathbf{Y}^{(k-1)}}\{\boldsymbol{U}_{i-1}^{(k)}\}$. The construction of the Markov transition probability is similar to the procedure discussed in Li and Gu (2012).

(ii) Update the estimate $\hat{Y}$ iteratively using

$$\mathbf{Y}^{(k)} = \mathbf{Y}^{(k-1)} + \nu_k \Delta \mathbf{Y}^{(k)},$$

where

$$\begin{bmatrix} \Delta \mathbf{Y}^{(k)} \\ \boldsymbol{\omega}^{(k)+} \end{bmatrix} = \begin{bmatrix} -\boldsymbol{\Gamma}^{(k)} & (\mathbf{I} - \mathbf{H}_{\mathcal{Z}})^\top \\ \mathbf{I} - \mathbf{H}_{\mathcal{Z}} & \mathbf{0} \end{bmatrix}^{-1} \times \begin{bmatrix} -\bar{H}\{\hat{\mathbf{Y}}^{(k-1)}, \boldsymbol{U}^{(k)}\} \\ -(\mathbf{I} - \mathbf{H}_{\mathcal{Z}})\hat{\mathbf{Y}}^{(k-1)} \end{bmatrix}$$

$$\boldsymbol{\Gamma}^{(k)} = \bar{\mathbf{I}}_0\{\mathbf{Y}^{(k-1)}, \boldsymbol{U}^{(k)}\}$$

$$\boldsymbol{\omega}^{(k)} = \boldsymbol{\omega}^{(k-1)} + \boldsymbol{\nu}_k \left\{ \boldsymbol{\omega}^{(k)+} - \boldsymbol{\omega}^{(k-1)} \right\},$$

with

$$\bar{H}\{\mathbf{Y}, \boldsymbol{U}^{(k)}\} = m^{-1} \sum_{i=1}^{n} H\left\{\mathbf{Y}, \boldsymbol{U}_i^{(k)}\right\}$$

$$\bar{\mathbf{I}}_0\{Y, \boldsymbol{U}^{(k)}\} = -m^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \mathbf{Y}} H\left\{\mathbf{Y}, \boldsymbol{U}_i^{(k)}\right\}.$$

At the end of this stage, we obtain $\hat{\mathbf{Y}}_0$ as the average of the last 10% of the sequence $\{\hat{\mathbf{Y}}^{(k)}\}_{k=1,\ldots,\kappa_1}$.

Step 2: Our main task is to solve the optimization problem (2.1), subject to the space constraint on $\mathbf{Y}$:

$$S(\mathbf{Y}) - \frac{\partial P_\lambda(\mathbf{Y})}{\partial \mathbf{Y}} = 0$$
$$\text{subject to} \quad (\mathbf{I} - \mathbf{H}_{\mathcal{Z}})\mathbf{Y} = 0. \tag{2.7}$$

Similar to Step 1, we propose an MCMC-SA algorithm to solve (2.7). In contrast to (2.6), we have to address the problem that the penalty function $p_\lambda(\boldsymbol{\beta})$ is irregular at the origin and may not be twice differentiable at some points. This problem can be solved by the approach of Fan and Li (2001) and Fan, Yang and Wu (2010) and applying a local quadratic approximation to the objective function. Because the iterative update procedure is similar to Step 1, we defer a description of the algorithm to Appendix A1.

Step 3: After the first two steps of our proposed algorithm, we obtain $\hat{\mathbf{Y}}$, which contains the pseudo values for the true $\mathbf{Y} = \mathcal{Z}^\top \boldsymbol{\beta}$. In other words,

the transformation of a high-dimensional semiparametric problem to a high-dimensional linear regression problem is complete. The remaining problem is to estimate the effective regression parameter. This can be done using existing variable selection methods designed for a high-dimensional linear regression model: $\hat{\mathbf{Y}} = \mathcal{Z}^\top \boldsymbol{\beta} + \epsilon$. Here, our experience suggests that the OGA offers fast and accurate results. After choosing the effective regression coefficients, say $\hat{\boldsymbol{\beta}}^*$, we then need to use the corresponding selected variable $\mathbf{Z}^*$ as a new input to the algorithm of Gu, Sun and Zuo (2005).

2.2. The AFT model

The PVM can be applied to ultra high-dimensional problems under the AFT model framework. The AFT model has the form:

$$\log(T) = \mathbf{Z}^\top \boldsymbol{\beta} + \epsilon, \tag{2.8}$$

where ϵ is an error term with an unspecified distribution.

For low-dimensional cases, Jin et al. (2003) proposed that the parameter of interest can be estimated by minimizing $L_G^*(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^n \sum_{j=1}^n \Delta_i \{e_i^*(\boldsymbol{\beta}) - e_j^*(\boldsymbol{\beta})\}^-$, where $e_i^*(\boldsymbol{\beta}) = \log(\tilde{T}_i) - \mathbf{Z}_i^\top \boldsymbol{\beta}$ and $a^- = |a| I(a < 0)$. To apply the PVM to the AFT model in a high-dimensional setting, we rewrite $e_i(Y_i)$ as $e_i^*(\boldsymbol{\beta})$, that is, $e_i(Y_i) = \log(\tilde{T}_i) - Y_i$ for $i = 1, \ldots, n$. Similarly to the procedure introduced in Section 2.1, we can equivalently minimize

$$L(\mathbf{Y}) = \sum_{i=1}^n \sum_{j=1}^n \Delta_i |e_i(Y_i) - e_j(Y_j)| + \left| M - \sum_{k=1}^n \sum_{l=1}^n \Delta_k (\mathbf{I}_l - \mathbf{I}_k)^\top \mathbf{Y} \right|,$$

subject to the constraint $(\mathbf{I} - \mathbf{H}_{\mathcal{Z}})\mathbf{Y} = 0$, where I_l denotes the lth column vector of $\mathbf{I}$, for $l = 1, \ldots, n$, and M is a large constant. Because the gradient of $L(\mathbf{Y})$ in this case is not differentiable, to implement the PVM using the procedure introduced in Section 2.1, we use an identity matrix to approximate its gradient; that is, we replace $\boldsymbol{\Gamma}$ by $\mathbf{I}$ in the aforementioned MCMC-SA algorithm.

2.3. Asymptotic properties

We adopt the conditions required in Ing and Lai (2011), which ensure the convergence of the OGA for linear regression models of the form of $Y = \mathbf{Z}\boldsymbol{\beta} + \epsilon$. Specifically, we assume that $p = p_n \to \infty$ and impose the following six conditions:

(C1) $\log p_n = o(f(n) \wedge n)$, for some function f of n.

(C2) Model-specific assumptions that guarantee $P(\|\tilde{\beta} - \beta\|_1 > \lambda) = o(1)$, with $\lambda = O(\log p_n / n)$.

(C3) $|z_j| \le C_{\max} < \infty$ for $j = 1, \ldots, p_n$. This also implies $\limsup_{n\to\infty} \max_{1\le j\le p_n} \mathbb{E}(\exp(s_1 z_j^2)) < \infty$, for some $s_1 > 0$.

(C4) Weak sparse assumption: $\sup_{n\ge 1} \sum_{j=1}^{p_n} |\beta_j \sigma_j| < \infty$, where σ_j is the standard deviation for the jth attribute.

(C5) Sparse assumption: there exists $0 \le \gamma < 1$ such that $n^\gamma = o((n/\log p_n)^{1/2})$ and

$$\liminf_{n\to\infty} n^\gamma \min_{1\le j\le p_n; \beta_j \ne 0} \beta_j^2 \sigma_j^2 > 0.$$

(C6) Define J as a set of selected attributes, $\Gamma(J) = \mathbb{E}(z(J)z(J)^T)$, $g_i(J) = \mathbb{E}(z_i z(J))$. Then,

$$\min_{1\le \#(J)\le K_n} \lambda_{\min}(\Gamma(J)) > \delta, \max_{1\le \#(J)\le K_n, i\notin J} \|\Gamma^{-1}(J) g_i(J)\|_1 < M.$$

Remark 1. Because different models, penalties, and chosen attributes relate to the possible bound of p_n, we express this simply as $f(n)$ in (C1).

Remark 2. For the Cox model, we can take a large constant with respect to λ to satisfy $P(\|\tilde{\beta} - \beta\|_1 > \lambda) = o(1)$, according to Huang et al. (2013), for the LASSO penalty. For general transformation models, Klaassen, Kueck and Spindler (2017), note that the LASSO can also provide the desired result in (C2), given $\log p_n = o(n^{1/4})$. In this case, $f(n) = n^{1/4}$ in (C1). For the AFT model, using the method described in Xu, Leng and Ying (2010), we have that $\tilde{\beta}_i$ is $n^{-1/2}$-consistent for $i = 1, \ldots, s$ and $\tilde{\beta}_i = \beta_i = 0$ for $i = s+1, \ldots, p_n$. If p_n increases with n, we have

$$P(\|\tilde{\beta} - \beta\|_1 > \lambda) \le P(\sqrt{s}\|\tilde{\beta} - \beta\|_2 \le \lambda) \le O\left(\frac{s^2/n}{\lambda^2}\right) = O\left(\frac{s^2}{\log p_n}\right) = o(1).$$

The following two theorems provide a justification for the proposed procedure. The corresponding proofs are provided in the Supplementary Material.

Theorem 1. *Under* (C1) *to* (C6), *suppose* $K_n/n^\gamma \to \infty$, *such that* $K_n = O((n/\log p_n)^{1/2}) \wedge p_n$. *Then,* $\lim_{n\to\infty} P(N_n \subseteq \hat{J}_{K_n}) = 1$, *where* $N_n = \{1 \le j \le p_n : \beta_j \ne 0\}$ *denotes the set of relevant input variables.*

The HDIC introduced in Ing and Lai (2011) is defined as $HDIC(J) = \sum_{t=1}^n (y_t - \hat{y}_{t:J})^2 \{1 + n^{-1}(\#(J) w_n \log p_n)\}$, where w_n satisfies $w_n \to \infty$, $w_n \log p_n = o(n^{1-2\gamma})$. By minimizing the HDIC using the OGA, we have the following result:

Theorem 2. *Under* (C1) *to* (C6), *define* K_n *as in Theorem* 1. *Then,*

$$\lim_{n\to\infty} P(\hat{k} \ge \tilde{k}) = 1,$$

where

$$\hat{k} = \arg \min_{1 \le k \le K_n} HDIC(\hat{J}_k)$$

and

$$\tilde{k} = \min\{k : 1 \le k \le K_n, N_n \subseteq \hat{J}_k\}.$$

Furthermore,

$$P(N_n \subseteq \hat{N}_n) = 1.$$

Remark 3. The above theorems are modified results based on Theorems 3, 4, and 5 of Ing and Lai (2011) under the PVM framework. These results guarantee that all relevant variables can be selected using our proposed method. Although our theoretical results cannot completely eliminate cases with over-selection, our numerical experience suggests that, with the regularization in the early step of the PVM, the use of the OGA can select substantially fewer variables, almost all of which are relevant; see also Section 3.

3. Simulations

To demonstrate the finite-sample performance of the proposed method, we conducted an extensive simulation study based on the two classes of models discussed in Section 2. We consider three special cases for the transformation models, namely, Cox's proportional hazards model, the probit model, and the proportional odds (PO) model. In addition, we also include the results for the AFT model under various settings.

In the following examples, after obtaining the pseudo values $\hat{\mathbf{Y}}$ in stage two, we apply the OGA method with HDIC as the selection criterion to perform the variable selection on the regression parameter $\boldsymbol{\beta}$. Note that we only report our selection results here, because the final estimation performance is determined by the classical low-dimensional regression procedure. The simulation results were obtained using a standard desktop computer equipped with an i7-2600 3.40GHz CPU and 8.00Gb RAM. Note that although the PVM utilizes the MCMC-SA optimization procedure, the mean computation time needed for the model selection based on a data set with a sample size of 400 and 5,000 covariates is only around 3,485 seconds (58 minutes). The computational burden for our proposed method is much less demanding than it appears, especially when parallel computing is employed. We relegate the details about the computation time to section A2 of the Appendix. The simulation results are shown in Tables 1–5 in the main text.

For the transformation models, similar to Case 5 of the simulation discussed

Table 1. Performance between the PVM on (a) Cox proportional hazards model, (b) probit model, (c) PO model, and (d) the AFT model under ultra high-dimensional settings, *i.e.* $n \ll p$. Frequency, in 100 simulations, of including all relevant variables (Correct), selecting exactly the relevant variables (E), selecting all relevant variables and i irrelevant variables ($E+i$), and selecting some relevant variables with i relevant variables omitted ($E-i$). The column "Correct" specifies the number of cases where all relevant variables are selected.

n	p	E	$E+1$	$E+2$	Correct	$E-1$	$E-2$	n	p	E	$E+1$	$E+2$	Correct	$E-1$	$E-2$
(a) Cox Proportional Hazards Model								(b) Probit Model							
150	1,000	96	1	1	98	2	0	150	1,000	97	3	0	100	0	0
200	1,000	99	1	0	100	0	0	200	1,000	100	0	0	100	0	0
400	1,000	100	0	0	100	0	0	400	1,000	100	0	0	100	0	0
200	5,000	99	0	0	99	1	0	200	5,000	99	0	0	99	1	0
200	10,000	100	0	0	100	0	0	200	10,000	100	0	0	100	0	0
400	5,000	100	0	0	100	0	0	400	5,000	100	0	0	100	0	0
400	10,000	100	0	0	100	0	0	400	10,000	100	0	0	100	0	0
(c) Proportional odds Model								(d) The Accelerated failure time (AFT) model							
300	1,000	96	4	0	100	0	0	200	1,000	100	0	0	100	0	0
400	1,000	100	0	0	100	0	0	400	1,000	100	0	0	100	0	0
300	5,000	99	0	0	99	1	0	200	5,000	98	0	0	98	2	0
300	10,000	94	0	0	94	4	2	200	10,000	100	0	0	100	0	0
400	5,000	100	0	0	100	0	0	400	5,000	100	0	0	100	0	0
400	10,000	100	0	0	100	0	0	400	10,000	100	0	0	100	0	0

in Fan, Yang and Wu (2010), we generated variables $\mathcal{Z}$ as a $p \times n$ matrix from a multivariate Gaussian distribution $N(0, V_1)$, where V_1 is a $p \times p$ matrix with diagonal elements equal to one, and all other elements equal to 0.5. We set $p = 1{,}000$ and $n = 150, 200$, and 400 for each model. The survival time T is generated from model (2.2), with $H^{-1}(\cdot)$ in $\Phi(u, v, w) = H^{-1}\{H(u) + v^\top w\}$, using the standard extreme value survival function, standard normal survival function, and standard logistic survival function for Cox's proportional hazards model, probit model, and PO models, respectively. The true regression parameters are set to the values considered in Fan, Yang and Wu (2010), that is, $\boldsymbol{\beta}_0 = (-1.5140, 1.2799, -1.5307, 1.5164, -1.3020, 1.15833, 0, \ldots, 0)^\top$, which is a p-column vector with only the first six elements nonzero.

In Step 1 of the MCMC-SA procedure, as in Gu, Sun and Zuo (2005), m was chosen to be 50. Different penalties were applied for the three cases to demonstrate that the proposed method works well for various penalties. In particular, for the Cox proportional hazards model, we adopted the LASSO penalty, and chose $\gamma_k = k^{-1/6}$; for the probit model, we added the SCAD penalty, and set $\gamma_k = k^{-1}$; for the PO model, we imposed the adaptive LASSO penalty, with $\gamma_k = k^{-1/6}$. Our numerical results show that the choice of penalty function

does not affect the selection result significantly. The corresponding tuning parameter was determined using the GCV specified in Li and Gu (2012). The censoring times were generated from an exponential distribution with mean 10, which yielded an average censoring rate of 25%. As presented in Table 1, the PVM performed well for the transformation models.

For the AFT model, we mimicked the example investigated in Xu, Leng and Ying (2010), except that we considered an ultra high-dimensional setting ($p = 1{,}000$, 5,000, and 10,000). In particular, we set $\boldsymbol{\beta}_0 = (3, 1.5, 0, 0, 2, 0, \ldots, 0)$, where $\boldsymbol{\beta}_0$ is a p-column vector; $\mathbf{Z}$ is generated from a multivariate Gaussian distribution $N(0, V_2)$, with V_2 as a $p \times p$ matrix with element (i, j) equal to $0.5^{|i-j|}$.

Table 1 summarizes the performance of the PVM under ultra high-dimensional settings. For cases with relatively small sample sizes, say $n = 200$, the PVM always discovers all relevant variables and, occasionally, one extra variable. This sample size was chosen to represent a similar situation to that considered in our data analysis. It can also be seen from Table 1 that, in most cases, the PVM selects all relevant variables and, occasionally, a few irrelevant variables. Under the setting of $n = 400$ with $p = 5{,}000$ or 10,000, the PVM always selects exactly all of the relevant variables.

We present a comparison between the PVM and some existing methods for the general transformation models (Li et al. (2014)) and the AFT model (Xu, Leng and Ying (2010)) under a high-dimensional setting with $p < n$. The corresponding results are summarized in Table 2. For the transformation models, the PVM outperforms the method of Li et al. (2014), especially for small sample cases, where $n = 150$. With the exception of the PO model, the PVM never omits any relevant variables. In contrast, the approach of Li et al. (2014) tends to omit one to two relevant variables. In addition, for cases where all relevant variables are selected, the PVM produces a smaller active set, as shown in columns $4-7$ of Table 2. A similar, yet more distinctive pattern is also found in a comparison between the PVM and the method of Xu, Leng and Ying (2010).

Under the ultra high-dimensional settings, we also compare the PVM with SIS (Fan, Yang and Wu (2010)) and the method of Song et al. (2014) for Cox's proportional hazards model and the PO model, respectively. For Cox's model, we adopted the same setting shown in Table 1, except that $n < p$. For the PO model, we change the variance matrix to be a $p \times p$ matrix with element (i, j) equal to $0.5^{|i-j|}$. The regression parameters $\boldsymbol{\beta}_0$ follow those in Song et al. (2014), which are $(-1, -0.9, 0, 0, 0, 0, 0, 0, 0.8, 1, 0, \ldots, 0, 0)$, with $H(t) = \log\{0.5(e^{2t} - 1)\}$. The

Table 2. Comparison of performance between the PVM and relevant existing methods on (a) Cox proportional hazards model, (b) probit model, (c) PO model and (d) the AFT model under traditional high-dimensional settings, *i.e.*, $n > p,\ p \approx n$. Frequency, in 100 simulations, of including all relevant variables (Correct), selecting exactly the relevant variables (E), selecting all relevant variables and i irrelevant variables ($E+i$), and selecting some relevant variables with i relevant ones omitted ($E-i$). The notation 3^+ denotes the results with at least three irrelevant variables selected. The column "Correct" specifies the number of cases where all relevant variables are selected.

Method	n	p	E	E+1	E+2	E+3$^+$	Correct	E−1	E−2	E−1+1	E−2+3	E−3+2
(a) Cox Proportional Hazards Model												
PVM	150	100	100	0	0	0	100	0	0	0	0	0
PVM	200	150	100	0	0	0	100	0	0	0	0	0
PVM	400	300	100	0	0	0	100	0	0	0	0	0
Li et al. (2014)	150	100	60	27	2	2	91	4	1	2	1	1
Li et al. (2014)	200	150	94	4	0	0	98	2	0	0	0	0
Li et al. (2014)	400	300	99	1	0	0	100	0	0	0	0	0
(b) Probit Model												
PVM	150	100	98	2	0	0	100	0	0	0	0	0
PVM	200	150	100	0	0	0	100	0	0	0	0	0
PVM	400	300	100	0	0	0	100	0	0	0	0	0
Li et al. (2014)	150	100	88	10	1	0	99	1	0	0	0	0
Li et al. (2014)	200	150	99	1	0	0	99	0	0	0	0	0
Li et al. (2014)	400	300	100	0	0	0	100	0	0	0	0	0
(c) Proportional odds Model												
PVM	150	100	95	2	0	0	97	3	0	0	0	0
PVM	200	150	96	3	0	0	99	1	0	0	0	0
PVM	400	300	100	0	0	0	100	0	0	0	0	0
Li et al. (2014)	150	100	89	5	0	0	94	2	4	0	0	0
Li et al. (2014)	200	150	95	4	0	0	99	1	0	0	0	0
Li et al. (2014)	400	300	100	0	0	0	100	0	0	0	0	0
(d) The Accelerated failure time (AFT) model												
PVM	150	100	99	1	0	0	100	0	0	0	0	0
PVM	200	150	100	0	0	0	100	0	0	0	0	0
PVM	400	300	100	0	0	0	100	0	0	0	0	0
Xu, Leng and Ying (2010)	150	100	64	23	9	3	99	1	0	0	0	0
Xu, Leng and Ying (2010)	200	150	67	23	9	0	99	1	0	0	0	0
Xu, Leng and Ying (2010)	400	300	82	16	2	0	100	0	0	0	0	0

censoring times were generated from uniform distributions Uniform$(0, 5.8)$ and Uniform$(0, 1.9)$ to achieve censoring rates of 15% and 40%, respectively. Table 3 presents the results. Note that the two contenders assign a rank to each regression covariate. The corresponding performance is usually measured by reporting the minimum number of variables selected such that all relevant variables are included. Therefore, no relevant variables are ever missing. This explains why the E^- column in Table 3 is always zero for these two methods. We also summarize the mean numbers of extra irrelevant variables, E^+. The figures in Table 3 reveal that the PVM offers a high probability of selecting the important variables, and

Table 3. Comparison of performance between the PVM and (a) SIS for Cox proportional hazards model and (b) Song et al.'s (2014) method for the PO model under ultra high-dimensional settings. The average numbers of missing relevant and extra irrelevant variables are denoted as E^- and E^+, respectively.

Method	n	p	CP	E^-	E	E^+
(a) Cox proportional hazards model						
PVM	150	1,000	23	0.02 (0.14)	5.98 (0.14)	0.03 (0.22)
PVM	200	1,000	23	0 (0)	6 (0)	0.01 (0.10)
PVM	400	1,000	23	0 (0)	6 (0)	0 (0)
PVM	200	5,000	23	0.01 (0.10)	5.99 (0.10)	0 (0)
PVM	200	10,000	23	0 (0)	6 (0)	0 (0)
PVM	400	5,000	23	0 (0)	6 (0)	0 (0)
PVM	400	10,000	23	0 (0)	6 (0)	0 (0)
SIS	150	1,000	23	0.14 (0.64)	5.86 (0.64)	1.64 (1.43)
SIS	200	1,000	23	0 (0)	6 (0)	2.35 (1.60)
SIS	400	1,000	23	0 (0)	6 (0)	7.91 (3.57)
SIS	200	5,000	23	0.09 (0.64)	5.91 (0.64)	0.99 (3.15)
SIS	200	10,000	23	0.36 (1.11)	5.64 (1.11)	1.46 (5.27)
SIS	400	5,000	23	0 (0)	6 (0)	1.63 (1.38)
SIS	400	10,000	23	0 (0)	6 (0)	0.85 (0.90)
(b) Proportional odds model						
PVM	300	5,000	15	0 (0)	4 (0)	0.41 (0.71)
PVM	300	5,000	40	0.15 (0.39)	3.85 (0.39)	0.64 (0.98)
Song et al. (2014)	300	5,000	15	0 (0)	4 (0)	7.60 (4.70)
Song et al. (2014)	300	5,000	40	0 (0)	4 (0)	22.10 (7.80)

includes few irrelevant variables. In particular, for the PO model, the method of Song et al. (2014) chooses 22.1 irrelevant variables, on average, whereas the PVM selects 0.64 irrelevant variables, on average, under the $(n, p) = (300, 5{,}000)$ case. Detailed results for the performance of SIS under various settings is included in Appendix A3.

To examine the performance of the PVM for cases in which some important variables are marginally independent of the response variable, we also conducted simulations with settings that correspond to Cases 3 and 4 discussed in Fan, Yang and Wu (2010). In particular, for Case 3, the covariates $Z_1, \ldots, Z_p$ follow a multivariate Gaussian distribution. They follow a marginally $N(0, 1)$ distribution, with the correlation structure $corr(Z_i, Z_4) = 1/\sqrt{2}$ for all $i \neq 4$, and $corr(Z_i, Z_j) = 0.5$ if i and j are distinct elements of $\{1, \ldots, p\}\backslash\{4\}$. The covariates are configured as $\beta = (4, 4, 4, -6\sqrt{2}, 0, 0, \ldots, 0)$. The censoring rate is 30%. For Case 4, $Z_1, \ldots, Z_p$ are also multivariate Gaussian, each of which is marginally $N(0, 1)$ distributed, with correlation structure $corr(Z_i, Z_5) = 0$ for

Table 4. Comparison of performance between the PVM and Fan, Yang and Wu's (2010) approach under Cases 3 and 4.

Method	n	p	Fan, Yang and Wu (2010)	E^-	E	E^+
Cox proportional hazards model						
PVM	300	400	Case 3	0.23 (0.55)	3.77 (0.55)	0.84 (1.8)
PVM	300	400	Case 4	0.32 (0.55)	4.68 (0.55)	1.15 (1.74)
PVM	400	1,000	Case 3	0.14 (0.51)	3.86 (0.51)	0.13 (0.63)
PVM	400	1,000	Case 4	0.27 (0.62)	4.73 (0.62)	0.3 (1.2)
ISIS	300	400	Case 3	0 (0)	4 (0)	14.76 (3.91)
ISIS	300	400	Case 4	0 (0)	5 (0)	14.84 (4.03)
ISIS	400	1,000	Case 3	0 (0)	4 (0)	10.93 (3.77)
ISIS	400	1,000	Case 4	0 (0)	5 (0)	11.26 (3.32)

all $i \neq 5$, $corr(Z_i, Z_4) = 1/\sqrt{2}$ for all $i \notin \{4, 5\}$, and $corr(Z_i, Z_j) = 0.5$ if i and j are distinct elements of $\{1, \ldots, p\}\backslash\{4, 5\}$. The covariates are configured as $\beta = (4, 4, 4, -6\sqrt{2}, 4/3, 0, 0, \ldots, 0)$. The corresponding censoring rate is also around 30%, and similar performance is observed (see Table 4).

Finally, we also considered a high-dimensional setting for the AFT model similar to that studied in Khan and Shaw (2016). Specifically, we set $(n, p) = (100, 120)$, with the first 20 coefficients for $\boldsymbol{\beta}$ set to four, and the remaining coefficients chosen to be zero. The covariates were generated as Z from Uniform(0, 1), with correlations 0 and $0.5^{|i-j|}$ for the two separate cases, and the error following a standard normal distribution. The censoring time was generated using the log-normal distribution $\exp\{N(\sqrt{2}c_0, 2)\}$, where c_0 was calculated analytically to produce the chosen censoring rate of 30% or 50%. According to the results shown in Table 5, where p_γ refers to significant variables and $p - p_\gamma$ represents non-relevant variables, the PVM produces the highest net selection accuracy, which we define as the percentage of relevant covariates selected minus that of nonrelevant variables chosen. The performance is more distinct for the nondependent case. This can be explained by the fact that OGA employs orthogonal projections to determine the next immediate relevant covariate. In cases where the linear dependence is strong amongst the variables, the OGA approach may choose the next linearly independent covariate that is not explained by the previously selected variables.

To conclude this section, we provide numerical results for cases with censoring times that depend on covariates. Specifically, we consider four models, namely, the Cox, probit, PO and AFT, with i.i.d. covariates Z_{ij} generated from Uniform $(0, 1)$. The covariates are configured as $\beta = (4, -4, 4, -4, 4, 0, \ldots, 0)$.

Table 5. Comparison of performance between the PVM and Khan and Shaw's (2016) approach under ultra high-dimensional settings

CP	Methods	Parameters	$r_{ij}=0$	$r_{ij}=0.5^{\lvert i-j\rvert}$
30	PVM	p_γ	91.9	43.4
		$p-p_\gamma$	4.6	1.0
	AEnet	p_γ	84.8	50.7
		$p-p_\gamma$	9.4	21.1
	AEnetCC	p_γ	89.8	62.0
		$p-p_\gamma$	14.3	40.3
	WEnet	p_γ	80.0	43.3
		$p-p_\gamma$	12.3	1.4
	WEnetCC	p_γ	87.3	61.8
		$p-p_\gamma$	12.4	11.0
50	PVM	p_γ	68.5	32.0
		$p-p_\gamma$	5.1	0.7
	AEnet	p_γ	68.1	50.7
		$p-p_\gamma$	8.8	26.0
	AEnetCC	p_γ	76.4	57.2
		$p-p_\gamma$	23.6	37.9
	WEnet	p_γ	56.8	28.4
		$p-p_\gamma$	10.6	1.8
	WEnetCC	p_γ	75.4	55.5
		$p-p_\gamma$	21.1	12.4

The censoring times are generated from an $\exp(6Z_1)$ distribution. The results are presented in Table 6. Again, our proposed procedure yields acceptable results, with few cases of over-selection.

4. Data Analyses

We examine two data sets in this section, namely, Stanford heart transplant data (Miller and Halpern (1982)) and diffuse large-B-cell lymphoma data (Rosenwald et al. (2002)).

4.1. Stanford heart transplant data

To demonstrate the PVM's performance for a data set with a regular dimension, we first present an analysis of a classical data set for the Cox proportional hazards model, namely, the Stanford heart transplant data collected in February 1980. The data set contains 157 observations of the following four variables: (i) the survival days for each patient, (ii) the censoring indicator, (iii) the age at time of first transplant, and (iv) the mismatch score. Because we need not

Table 6. Comparison of performance between the PVM and relevant existing methods on the (a) Cox proportional hazards model, (b) probit model, (c) PO model and (d) AFT model under dependent censoring. Frequency, in 100 simulations, of including all relevant variables (Correct), selecting exactly the relevant variable (E), selecting all relevant variables and i irrelevant variables ($E+i$), and selecting some relevant variables with i relevant variables omitted ($E-i$). The notation 3^+ denotes the results with at least three irrelevant variables selected. The column "Correct" specifies the number of cases where all relevant variables are selected.

Model	n	p	E	$E+1$	$E+2$	Correct	$E-1$	$E-2$
Cox	400	10,000	99	1	0	100	0	0
Probit	400	10,000	100	0	0	100	0	0
PO	400	10,000	99	1	0	100	0	0
AFT	400	10,000	100	0	0	100	0	0

Table 7. Analysis of Stanford heart transplant data using the PVM

Estimator	Age($\hat{\beta}_1$)	Mismatch score ($\hat{\beta}_2$)
Original Cox	0.030	0.167
Cox with PVM	0.030	0.157

carry out a dimension-reduction procedure here, we only execute the first stage of our algorithm. Compared with the result obtained from the original Cox regression model, our estimates are virtually the same; see Table 7. This verifies the performance of the PVM for low-dimensional cases.

4.2. Diffuse large-B-cell lymphoma studies

Here, we analyze a type of diffuse large-B-cell lymphoma, which is the most common type of lymphoma in adults, and can be cured by chemotherapy for only 35 to 40 percent of patients. Rosenwald et al. (2002) examined whether gene-expression profiles of the lymphoma of interest can be used to predict the outcome of chemotherapy using a multivariate Cox proportional hazards model. Biopsy samples of diffuse large-B-cell lymphoma from 240 patients were examined for 7,399 gene expressions using DNA microarrays and were analyzed for genomic abnormalities. The data set is available to the public at http://llmpp.nih.gov/DLBCL/. Of the 240 samples, 138 patients died (57% of the patients recruited) during the follow-ups, with a median death time of 2.8 years. The median age of the patients was 63 years, and 56 percent were men. According to the Ann Arbor classification, 15 percent of patients had stage I disease, 31 percent had stage II, 20 percent had stage III, and 34 percent had stage IV. The Kaplan−Meier plot of the overall survival for this data set is shown

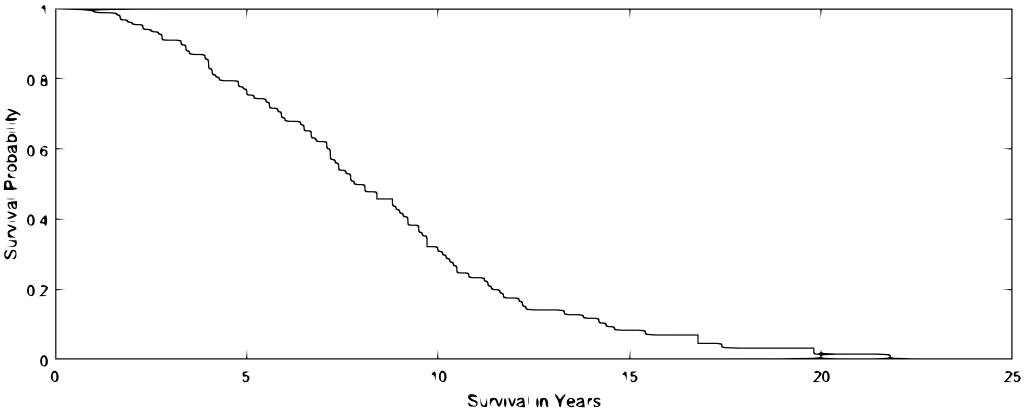

Figure 1. Kaplan−Meier plot of overall survival of patients for large-b-cell lymphoma data.

in Figure 1.

To analyze this data set, we adopted the same initial setup as that shown in our simulation studies. Using the pseudo values $\hat{Y}$ obtained in the first two stages of our procedure, we performed the variable selection using the OGA with the HDIC as the model selection criterion. The genes selected are listed in Table 8. In particular, the genes with asterisks were identified as significant classes of genes in Rosenwald et al. (2002), namely, the Germinal-center B-cell signature, MHC class-II signature, and Lymph-node signature. It is natural for our method to select the three genes under the Cox model, because Rosenwald et al. (2002) adopted the same model in their analysis. The additional genes selected under the various models are worthy of attention because they are the influential variables most likely to be overlooked.

Gui and Li (2005) studied the same data set using the LARS method on the $\mathcal{L}_1$-penalized Cox model. Amongst the four genes selected by their method, Gui and Li (2005) also regarded the Germinal-center B-cell signature and MHC class II signature as influential genes. Whereas genes that belong to the Lymph-node signature were also selected in Gui and Li (2005), they selected `LC_29222` and `X59812` in their final conclusion.

To compare the conclusions from the different approaches, we also performed a likelihood ratio test to compare our selection with that of Gui and Li (2005) based on the Cox model. As reported, the PVM chose six variables and the corresponding log-likelihood is −387.13; Gui and Li (2005) selected ten variables with a log-likelihood of −386.23. Combining these two results, we constructed

Table 8. Rosenwald et al. (2002), diffuse large B-cell lymphoma studies data: Significant genes selected under Cox, probit, and AFT models with respect to the survival time. Genes marked with asterisks (*) and daggers (†) correspond to genes that were also selected in Rosenwald et al. (2002) and Gui and Li (2005), respectively.

Model	GenBank IDs of the selected genes					
Cox	X00452*,†	AA805575*,†	X14420*	U14791	M63438	X90858
Probit	X00452*,†	AA805575*,†	AI540204	U64197		
AFT	X14420*	AA805575*,†				

a test that covered the union of the 14 variables selected. The corresponding new model (Model C_1) has a log-likelihood of -383.60. Here, we define $D_{1,2}$ as twice the difference between the log-likelihoods, which means that for two models M_1 and M_2, $D_{1,2} = 2 \times$ (log-likelihood for M_1 − log-likelihood for M_2). The statistics $D_{\text{PVM, } C_1}$ and $D_{\text{Gui and Li (2005), } C_1}$ are found to be -7.1 and -5.3 with 8 and 4 degrees of freedom, respectively. Neither of these two models was found to be statistically different from the full model. In other words, both models are statistically equivalent; however, the model obtained from the PVM is more parsimonious. To make the comparison easier to understand, we also chose the best ten variables based on the PVM, which yielded a log-likelihood of -383.56. With the model that involves 10 variables chosen by Gui and Li (2005), the corresponding log-likelihood is -386.23. Thus, the model chosen by the PVM achieves a higher (log-)likelihood with the same number of variables selected.

5. Discussion

We have introduced an innovative PVM with applications to ultra high-dimensional lifetime data. These pseudo values can be regarded as a set of educated guesses for the response variables, some of which are not fully observable owing to censoring. Our numerical results have demonstrated the promising performance of the PVM. That is, the model identifies the relevant variables, while minimizing the number of irrelevant variables under statistically challenging settings with $n \ll p$.

Although many procedures have been designed to address variable selection problems, most have been developed in a linear regression context. To implement these ideas in survival models, we have to rely heavily on the (pseudo-/partial-) likelihood upon which a penalty can be applied. As a result, it is not trivial to incorporate a SIS (Fan and Lv (2008)) component into these semiparametric

models because the likelihood cannot be easily calculated. One main contribution of our method is that it bridges the gap between the tools developed for linear models and semiparametric survival models such that ultra high-dimensional variables and censored data can be handled properly.

Finally, note that the PVM is a generic approach, and so is not restricted to the two classes of models studied here. Other models, such as general linear models and quantile regression models, can also be handled using a similar procedure. It is expected that this method will be effective for a wide range of regression-type problems. The development for such models is left to future research.

Supplementary Materials

The Supplementary Material includes the algorithm adopted in Step 2 of our proposed method for the general transformation models, additional details on the computation time of our proposal, and further numerical results for Cox's proportional hazards model using SIS. The proofs for the theorems presented in Section 2.3 are also included here.

Acknowledgment

The first author acknowledges the financial support of Hong Kong Research Grant Council Research Grants ECS-24300514 and GRF-14317716.

References

Antoniadis, A., Fryzlewicz, P. and Letue, F. (2010). The dantzig selector in cox's proportional hazads model. *Scandinavian Journal of Statistics* **37**, 531–552.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine* **2**, 273–277.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

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

Chen, H. Y. and Little, R. J. (2001). A profile conditional likelihood approach for the semiparametric transformation regression model with missing covariates. *Lifetime Data Analysis* **7**, 207–224.

Cox, D. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **34**, 187–220.

Cox, D. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. J. (2004). Least angle regression. *The Annals of Statistics* **32**, 407–499.

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

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

Fan, J., Yang, F. and Wu, Y. (2010). High-dimensional variable selection for cox's proportional hazards model. *IMS Collections, Borrowing Strength: Theory Powering Applications - A Festschrift for Lawrence D. Brown* **6**, 70–86.

Gu, M. and Kong, F. H. (1998). A stochastic approximation algorithm with markov chian monte carlo method for imcomplete data estimation problems. *Proceedings of the National Academy of Sciences* **95**, 7270–7274.

Gu, M. G., Sun, L. and Zuo, G. (2005). A baseline-free procedure for transformation models under interval censorship. *Lifetime Data Analysis* **11**, 473–488.

Gu, M. G., Wu, Y. and Huang, B. (2014). Partial marginal likelihood estimation for general transformation models. *Journal of Multivariate Analysis* **123**, 1–18.

Gui, J. and Li, H. (2005). Penalized cox regression analysis in the high-dimensional and low-sample size settings, with applications to microarray gene expression data. *Bioinformatics* **21**, 3001–3008.

Hsieh, F. (2001). On heteroscedastic hazards regression models: Theory and application. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **63**, 63–79.

Huang, J. and Harrington, D. (2005). Iterative partial least squares with right censored data analysis: a comparison to other dimension reduction techniques. *Biometrics* **61**, 17–24.

Huang, J., Sun, T., Ying, Z., Yu, Y. and Zhang, C.-H. (2013). Oracle inequalities for the lasso in the cox model. *The Annals of Statistics* **41**, 1142.

Ing, C.-K. and Lai, T. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statisica Sinica* **21**, 1473–1513.

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

Khan, M. H. R. and Shaw, J. E. H. (2016). Variable selection for survival data with a class of adaptive elastic net techniques. *Statistics and Computing* **26**, 725–741.

Klaassen, S., Kueck, J. and Spindler, M. (2017). Transformation models in high-dimensions. *arXiv preprint arXiv:1712.07364*.

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

Li, J. and Gu, M. (2012). Adaptive lasso for general transformation modelds with right censored data. *Computational Statistics and Data Analysis* **56**, 2583–2597.

Li, J., Gu, M., Zhang, R. and Lian, H. (2014). Variable selection for general transformation models with ranking data. *Statistics* **48**, 81–100.

Lu, W. and Zhang, H. (2007). Variable selection for proportional odds model. *Statistics in Medicine* **26**, 3771–3781.

Miller, R. and Halpern, J. (1982). Regression with censored data. *Biometrika* **69**, 521–531.

Prekopa, A. (1973). On logarithmic convace measures and functions. *Acta Scientiarum Mathematicarum* **34**, 335–343.

Rosenwald, A., Wright, G., Chan, W., Connors, J., Campo, E., Fisher, R., Gascoyne, R.,

Muller-Hermelink, H., Smeland, E. and Staudt, L. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *The New England Journal of Medicine* **346**, 1937–1947.

Song, R., Lu, W., Ma, S. and Jeng, X. (2014). Censored rank independence screening for high-dimensional survival data. *Biometrika* **101**, 799–814.

Stute, W. (1993). Consistent estimation under random censorship when covariables are available. *Journal of Multivariate Analysis* **45**, 89–103.

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

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

Xu, J., Leng, C. and Ying, Z. (2010). Rank-based variable selection with censored data. *Statistics and Computing* **20**, 165–176.

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

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

Department of Statistics, The Chinese University of Hong Kong, Ma Liu Shui, Hong Kong.

E-mail: tonysit@sta.cuhk.edu.hk

Department of Statistics, Purdue University, West Lafayette, IN 47907, USA.

E-mail: xing49@purdue.edu

Department of Statistics, The Chinese University of Hong Kong, Ma Liu Shui, Hong Kong.

E-mail: colexyz@163.com

Department of Statistics, The Chinese University of Hong Kong, Ma Liu Shui, Hong Kong.

E-mail: minggao@sta.cuhk.edu.hk

(Received February 2017; accepted February 2018)