Statistica Sinica Preprint No: SS-2017-0089	
Title	Pseudo value method for ultra high-dimensional semiparametric models with life-time data
Manuscript ID	SS-2017-0089
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0089
Complete List of Authors	Tony Sit Yue Xing Yongze Xu and Minggao Gu
Corresponding Author	Minggao Gu
E-mail	minggao@gmail.com
Notice: Accepted version subject to English editing.	

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Statistica Sinica

Pseudo value method for ultra high-dimensional semiparametric models with life-time data

Tony Sit, Yue Xing, Yongze Xu and Minggao Gu

Department of Statistics, The Chinese University of Hong Kong

Abstract: We develop a new procedure called "pseudo value method" (PVM) that can handle ultra high-dimensional variable selection problems for semiparametric survival models. While there has only been the sure independence screening (SIS)-type strategy for ultra high-dimensional life-time data so far in literature, the new 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 current problem of interest into a regular linear regression framework. Through this translation, all the existing and powerful techniques developed for the linear regression problems can be leveraged to tackle the new challenge at almost no extra cost.

Numerical performance of PVM has also demonstrated promising results: in addition to outperforming the (iterative) SIS for the Cox model, the new method can also accurately select the effective variables for probit, proportional odds and the AFT models, amongst which, to the best of our knowledge, has been studied under the ultra high-dimensional context on a case-by-case basis. Our unified

method was also applied to analyse Diffuse large-B-cell lymphoma data, which discovered genes that may be overlooked yet be influential. This finding itself can potentially be of scientific importance.

Key words and phrases: Semiparametric models, transformation models, accelerated failure time model, penalized log-marginal likelihood and 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 analysis attempts to model the dependence of the survival time T of a subject on his covariate variables $\mathbf{Z} = (Z_1, \ldots, Z_p)^\top$, where p indicates the dimensionality of the covariate space. With the rapid development in technologies, variable selection has been studied extensively since the mid 90's during which collection of vast volumes of data has become technically and economically feasible.

For life-time data, one popular class of semi-parametric models is the transformation models under which both the celebrated Cox (1972, 1975) proportional hazards model and the proportional odds model (Bennett 1983) are special cases. Given conventional high-dimensional datasets, variable selection for the Cox model is usually carried out via the parametric partial likelihood upon which penalties including the least absolute shrinkage and selection operator (LASSO, Tibshirani 1996), smoothly clipped absolute deviation (SCAD, Fan & Li 2001), least angle regression selection (LARS, Efron et al. 2004) and elastic net (Zou & Hastie 2005) are usually imposed.

For the Cox model, Li & Luan (2003) proposed a procedure that makes use of the reproducing kernel Hilbert spaces for inference on censored data; Gui & Li (2005) and Antoniadis et al.

(2010) studied the use of a threshold gradient descent regularization and the Dantzig selector, respectively, on the inference problems for the Cox model. For the proportional odds model, Lu & Zhang (2007) proposed an inference procedure via a penalized marginal likelihood of ranks, which was later extended in Li & Gu (2012) for a more general family of transformation models; see also Li et al. (2014).

Another attractive alternative to the Cox model is the accelerated failure time model (AFT) which directly relates the log of the survival time with the covariates; see Ying (1993), Jin et al. (2003) amongst others. This model offers a more straight-forward interpretation and is more appealing than the proportional hazards model in many aspects. There have been several variable selection methods: Huang & Harrington (2005) applied LASSO-type penalty on Buckley-James type estimator; a rank-based variable selection procedure for regular high-dimensional data was also studied by Cai et al. (2009), Xu et al. (2010) for the AFT model because, unlike the Cox model, partial likelihood is no longer available.

Microarray, proteomic and SNP data via bioimaging technology studies have induced a recent surge of interest in variable selection under an ultra high-dimensionality setting. Problems of interest involve an exponentially growing parameter space with respect to the sample size, *i.e.* $\log(p) = \mathcal{O}(n^{\alpha}),\ \alpha \in (0, 1/2)$. In the view of this new statistical challenge, Fan et al. (2010) applied the sure screening technique (SIS, see also Fan & Lv 2008) method to Cox's proportional hazard models; Song et al. (2014) applied SIS on censored rank data for the transformation models with the independent censoring assumption. A recent work of Khan & Shaw (2016) extended Stute (1993, 1996) weighted least squares formulation and applied a class of elastic net techniques to handle ultra high-dimensional data. For more general class of semiparametric transformation models, however, a unified procedure for other life-time data models under such an ultra high-dimension framework has been missing – the main difficulty remains the handling

of censored data. Unlike the special case of Cox model, where the partial likelihood can serve as an effective vehicle for inference, one may need to directly handle the likelihood or (rank-based) estimating equations in order to make the variable selection procedure possible.

More recently, Ing & Lai (2011) developed the orthogonal greedy algorithm (OGA), which is a stepwise regression modified for (ultra) high-dimensional linear regression model. 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 ensures that all the remaining variables become perpendicular to the selected variable(s). As a consequence, OGA tends to have lower false selection rates compared with the SIS counterpart in which case smaller correct models can be selected. To the best of our knowledge, there has not yet been any literature studying how this powerful tool in the arsenal of (ultra-) high dimensional sparse regression procedures be applied to time-to-event data.

The key objective of this paper is to develop a general approach that utilizes the "pseudo-values" to bridge inference problems for survival data and those appear in conventional linear regression models under the ultra high-dimensional setting. Specifically, the pseudo-values can be regarded as a set of educated guesses for the response variables, some of which are not fully observable due to censoring. Our contribution in this paper is two-fold: Not only can the proposed method offer a solution to open challenges for modelling ultra high-dimensional life-time data, but the concept of pseudo-values also facilitates the use of existing variable selection machinery for linear models for more general model settings. The codes for simulations and numerical studies, composed in `MATLAB`, are also provided for interested users.

The remainder of the paper is organized as follows: Section 2 illustrates the pseudo-value

method (PVM) via two popular classes of models for life-time data, namely (i) the general transformation models and (ii) the accelerated failure time model. Sections 3 and 4 cover numerical simulations as well as numerical studies 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 the sequel, we let $(T, C, \mathbf{Z})$ denote the triplet for the survival time, the censoring time and their associated covariates, respectively. Under the conditional independence censoring mechanism, *i.e.* $C \perp T \mid \mathbf{Z}$, one 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 samples of the triplet $(\tilde{T}, \Delta, \mathbf{Z})$ that are 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, estimation of $\boldsymbol{\beta}$ in a regression problem is performed via (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 $p \times n$. Under the (ultra-) high dimensional setting, however, it is difficult to estimate the optimizer $\hat{\boldsymbol{\beta}}$ directly with respect to L^*. The key contribution of our methodology lies on an 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 can first obtain a set of values $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)^\top$ that maximizes L. Based upon these optimizers, estimation of $\hat{\boldsymbol{\beta}}$ can then be carried out subsequently. This particular set of optimizers $\hat{\mathbf{Y}}$ is regarded as the "pseudo values". It is noteworthy that the proposed method can significantly reduce the dimension of the problem of interest because $\dim(\hat{\mathbf{Y}}) = n \ll p$. Upon these "pseudo values", one can apply the orthogonal greedy

algorithm with high-dimensional information criterion (OGA; see Ing & Lai 2011) in the second stage of the inference without any need for further modification due to model-specific settings.

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

Stage 1: We first obtain an initial set of pseudo values which 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 various assumptions on the monotone transformation function $H(\cdot)$ and the residual ϵ for different models. Therefore, at this stage of optimization, it is natural to impose a restriction that the pseudo values should lie on the linear span of $\mathcal{Z}$, which is denoted as span($\mathcal{Z}$). In other words, we re-parametrize the parameters of interest so as to tackle challenges due to the high dimensionality of $\boldsymbol{\beta}$.

Stage 2: We introduce a penalty at this stage so as to reduce the dimension of the problem. Traditionally, as we see in LASSO, hard thresholding and SCAD, a penalty $p_\lambda(\cdot)$ is imposed on the regression coefficient $\boldsymbol{\beta}$. Due to the fact that our procedure first treats $\mathbf{Y}$ as the parameter of interest, we suggest the following conversion: Since $\mathbf{Y} = \mathcal{Z}^\top \boldsymbol{\beta}$, one 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 an 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 & Vandenberghe (2004). Results obtained in Stage 1 can be used as an initial set of value for this round of optimization. From another perspective, the introduction of this penalty is to adjust 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 LASSO, SCAD and adaptive LASSO can be applied to this method. Noteworthy, while the penalty applied here borrows the idea of penalizing the parameter of interest β, the main goal at this stage is to obtain the pseudo value upon which variable selection methods developed for linear models can be applied. In our simulation exercise in Section 3, we shall demonstrate that all these three approaches produce can produce similar and desirable results.

Stage 3: At the last stage, we can carry our variable selection procedure on $\boldsymbol{\beta}$ by solving a high-dimensional linear regression problem with $\hat{\mathbf{Y}}$ as the response variable, *i.e.* we estimate $Y = \mathbf{Z}^\top \boldsymbol{\beta} + \epsilon$. Any existing variable selection procedure that is found effective can be applied. After obtaining the active set, the set of $\boldsymbol{\beta}$'s that are non-zero, we can estimate the final regression coefficient via the classical procedures for low-dimensional settings.

In summary, the PVM procedure mainly aims at maximising the corresponding objective function via $\hat{\mathbf{Y}}$ with appropriate regularization penalty and linear constraints. The last stage is to apply OGA on the transformed $\hat{\mathbf{Y}}$ on the original covariates in the linear regression context so as to complete the final variable selection procedure. As shown in both Sections 3 and A2, we divide the algorithm in three stages so as to demonstrate clearly how much computation time is needed for each particular step. In the following subsections, we shall provide two generic examples in Sections 2.1 and 2.2 to illustrate specifically how this method can be used to tackle variable selection problems that appear in the general transformation models and the AFT model.

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

2.1 General Transformation Models

Under the general transformation models, it is assumed 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 $g^{-1}(\cdot) = 1 - F(\cdot) = \Pr\{\epsilon \geq \cdot\}$, one can rewrite (2.2) as $H(T) = \mathbf{Z}^\top \boldsymbol{\beta} + \epsilon$, which is the traditional transformation models; see Gu et al. (2005) for instance. 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 model (Chen & Little 2001).

Suppose we have n i.i.d. observations $\{(\tilde{T}_i, \mathbf{Z}_i, \Delta_i)\}_{i=1,\ldots,n}$. We define $k_n = \sum_{i=1}^n \Delta_i$ the total number of observed failure times. We also denote $\mathcal{R}_n^*$ as the partial ranking amongst 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 can define $\mathcal{S}_n$ as a set that is composed of all possible complete rankings $\mathcal{R}_n$ and

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

as a time set that is consistent with the order restriction $\mathcal{R}_n^*$. Here i_r denotes the label of the rth ordered observed failure and $\mathcal{L}_{i_r}$ is the set of labels of censored observations in $[T_{i_r}, T_{i_{r+1}})$

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

with $T_{i_0} = 0$ and $T_{i_{k_n+1}} = \infty$. It follows that the marginal likelihood can be rewritten as

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

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

As suggested in Gu & Kong (1998) and Gu et al. (2014), Monte Carlo method can be used to maximize (2.3). We assume that $\Phi(u,v)$ to be 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 $\mathbf{u} = (u_1, \ldots, u_n)^\top$,

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

where

$$H(Y_i, u_i) = \phi_v(1-u_i, Y_i)/\phi_u(1-u_i, Y_i)$$

with

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

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

To implement the pseudo value method, we first solve the following optimization problem without performing variable selection. Through this stage, we can obtain a preliminary estimates for our pseudo values that can be used in the next stage of our procedure. The maximum

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

likelihood estimates of $\mathbf{Y}$ given (2.3) can be obtained by maximising

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

where $\mathbf{H}_{\mathcal{Z}} = \mathcal{Z}(\mathcal{Z}\mathcal{Z}^\top)^+\mathcal{Z}^\top$ and 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

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

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

One difficulty involved in this optimization is that the integral is usually of very high dimension and the normalising constant in $p(\mathbf{u}, \mathbf{Y})$ defined in (2.5) has no closed analytic expression. The resulting computation of (2.6) cannot be solved trivially by using standard numerical methods. Here we adopt the MCMC-SA algorithm in the spirit of Gu & Kong (1998) and Gu et al. (2005) for our initial pseudo value estimation. The corresponding algorithm is described as follows:

Step 1: Choose positive integers λ, m, 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) for κ_1 times:

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

(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 $\Pi_{\mathbf{Y}^{(k-1)}}\{\boldsymbol{U}_{i-1}^{(k)}\}$. Construction of Markov transition probability is similar to the procedure discussed in Li & Gu (2012).

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

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

where

$$\begin{aligned}
\begin{bmatrix} \Delta \mathbf{Y}^{(k)} \\ \boldsymbol{\omega}^{(k)+} \end{bmatrix} &= \begin{bmatrix} -\boldsymbol{\Gamma}^{(k)} & (\mathbf{I} - \mathbf{H}_{\mathcal{Z}})^\top \\ \mathbf{I} - \mathbf{H}_{\mathcal{Z}} & \mathbf{0} \end{bmatrix}^{-1} \times \begin{bmatrix} -\bar{H}\{\hat{\mathbf{Y}}^{(k-1)}, \boldsymbol{U}^{(k)}\} \\ -(\mathbf{I} - \mathbf{H}_{\mathcal{Z}})\hat{\mathbf{Y}}^{(k-1)} \end{bmatrix} \\
\boldsymbol{\Gamma}^{(k)} &= \bar{\mathbf{I}}_0\{\mathbf{Y}^{(k-1)}, \boldsymbol{U}^{(k)}\} \\
\boldsymbol{\omega}^{(k)} &= \boldsymbol{\omega}^{(k-1)} + \boldsymbol{\nu}_k \left\{ \boldsymbol{\omega}^{(k)+} - \boldsymbol{\omega}^{(k-1)} \right\}
\end{aligned}$$

with

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

At the end of this stage, we can 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

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

constraint on $\mathbf{Y}$:

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

Similar to Step 1, we propose an MCMC-SA algorithm to solve (2.7). Different from (2.6), we have to tackle 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 challenge can be solved by adopting Fan & Li (2001) and Fan et al. (2010) approach and applying a local quadratic approximation to the objective function. Since the iterative update procedure is similar to Step 1, description of the exact algorithm will be deferred to Appendix A1.

Step 3: After the first two steps of our proposed algorithm, we obtained $\hat{\mathbf{Y}}$, which are the pseudo values for the true $\mathbf{Y} = \mathcal{Z}^\top \boldsymbol{\beta}$. In other words, the transformation of a high-dimensional semi-parametric problem to a high-dimensional linear regression problem has been completed. The remaining problem is to estimate the effective regression parameter. This can be done by existing variable selection method designed under the context of a high-dimensional linear regression model: $\hat{\mathbf{Y}} = \mathcal{Z}^\top \boldsymbol{\beta} + \epsilon$. Amongst all the existing methods, our experience reveals that OGA can offer fast and accurate results. After choosing the effective regression coefficients, say $\hat{\boldsymbol{\beta}}^*$, we only need to use the corresponding selected variable $\mathbf{Z}^*$ as a new input for Gu et al. (2005) algorithm.

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

2.2 The Accelerated Failure Time (AFT) Model

The pseudo value method can also be applied for ultra high-dimensional problems under the accelerated failure time (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 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 pseudo value method for the AFT model under high-dimensional setting, we rewrite $e_i(Y_i)$ as $e_i^*(\boldsymbol{\beta})$, *i.e.* $e_i(Y_i) = \log(\tilde{T}_i) - Y_i$ for $i = 1, \ldots, n$. Similar 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 l-th column vector of $\mathbf{I}$ for $l = 1, \ldots, n$ and M is a large constant. Since the gradient of $L(\mathbf{Y})$ in this case is not differentiable, to implement the pseudo value method that is similar to the procedure introduced in Section 2.1, we use an identity matrix to approximate its gradient, *i.e.* we replace $\boldsymbol{\Gamma}$ by $\mathbf{I}$ in the aforementioned MCMC-SA algorithm.

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

2.3 Asymptotic Properties

We adopt the conditions required in Ing & Lai (2011) that can ensure the convergence for OGA on linear regression models in 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 $\lambda = O(\frac{\log p_n}{n})$.

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

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

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

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

(C6) Define J be an index set of selected attributes, $\Gamma(J) = \mathbb{E}(z(J)z(J)^T)$, $g_i(J) = \mathbb{E}(z_i z(J))$, then

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

Remark 1. *Since different models, penalties and attributed chosen all relate to the possible bound of p_n, we just write it as $f(n)$ for generality in (C1).*

Remark 2. *For 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 trans-*

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

formation models, by Klaassen et al. (2017), 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 AFT model, using the method described in Xu et al. (2010), we have* $\tilde{\beta}_i$ *is* $n^{-1/2}$*-consistent for* $i = 1, ..., s$ *and* $\tilde{\beta}_i = \beta_i = 0$ *for* $i = s+1, ..., p_n$. *If* p_n *increases with* n, *we have*

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

We present the following two theorems that provide justification for the proposed procedure. The corresponding proofs will be deferred to the supplementary materials.

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* $N_n = \{1 \leq j \leq p_n : \beta_j \neq 0\}$ *denotes the set of relevant input variables.*

Recall the high-dimensional information criterion (HDIC) introduced in Ing & 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 HDIC through OGA, we have the following result:

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

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

where

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

and

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

Furthermore

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

Remark 3. *Theorems are modified results based on Theorems 3, 4 and 5 of Ing & Lai (2011) under the PVM framework, respectively. These results guarantee that all the relevant variables can be selected by 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 PVM, the use of OGA can select substantially fewer variables amongst which almost all are relevant ones; 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. Three special cases for the transformation models, namely Cox's proportional hazards model, the probit model and the proportional odds (PO) model were considered. In addition, we also include our results for the AFT model under various settings.

In all of the following examples, after obtaining the pseudo values $\hat{\mathbf{Y}}$ in stage two, we applied the OGA method with high-dimensional information criterion (HDIC) as the selection criterion to perform variable selection on the regression parameter $\boldsymbol{\beta}$. We only report our selection results here since the final estimation performance is entirely determined by the classical low-dimensional regression procedure. All the simulation results were obtained by using a standard desktop equipped with an i7-2600 3.40GHz CPU and 8.00Gb RAM. Noteworthy, although PVM utilizes MCMC-SA optimization procedure, the mean computation time needed for model selection based on one dataset with sample size 400 yet 5000 covariates is merely around 3,485

seconds (58 minutes). The computation burden for our proposed method is much less demanding than it appears, especially when parallel computing is employed. We shall relegate the details about the computation time in A2 of the Appendix; all the simulation are tabulated in Tables 1 – 5 in the main text.

For the transformation models, similar to Case 5 of the simulation study discussed in Fan et al. (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 all to be 1 and other elements to be 0.5. We set $p = 1000$ and $n = 150, 200$ and 400 for each model. Survival time T's were generated from by model (2.2), with $H^{-1}(\cdot)$ in $\Phi(u, v, w) = H^{-1}\{H(u) + v^\top w\}$ taking standard extreme value survival function, standard standard normal survival function and standard logistic survival function for Cox's proportional hazards model, probit model and proportional odds models, respectively. The true regression parameters are set to be the values considered in Fan et al. (2010), *i.e.* $\boldsymbol{\beta}_0 = (-1.5140, 1.2799, -1.5307, 1.5164, -1.3020, 1.15833, 0, \ldots, 0)^\top$, which was a p-column vector with only the first six elements non-zero.

In Step 1 of the MCMC-SA procedure, as adopted in Gu et al. (2005), m was chosen to be 50. Different penalties were applied for the three cases so as 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 proportional odds model, adaptive LASSO penalty with $\gamma_k = k^{-1/6}$ was imposed. Our numerical experience reveals that the choice of the penalty functions does not affect the selection result significantly. The corresponding tuning parameter can be determined by using GCV specified in Li & Gu (2012). The censoring times were generated from exponential distribution with mean 10, which yielded an average censoring rate of 25%. As presented in Table 1, the pseudo value method performed well for the transformation

models.

For the AFT model, we mimicked the example investigated in Xu et al. (2010) except that an ultra high-dimensional setting ($p = 1000, 5000$ and 10000) was considered. 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; $\mathbf{Z}$'s were generated from a multivariate Gaussian distribution $N(0, V_2)$ with V_2 as a $p \times p$ matrix with its (i, j)'s element $= 0.5^{|i-j|}$.

Table 1 summarizes the performance of PVM under ultra high-dimensional settings. For cases with relatively small sample sizes, say $n = 200$, PVM can always discover all the relevant variables with occasionally one extra variable selected. This sample size is particularly chosen so as to represent a similar situation as considered in our data analysis. It can also be seen from Table 1 that in most cases, PVM can select all relevant variables with occasionally few irrelevant chosen. Under the setting of $n = 400$ with $p = 5000$ or 10000, PVM always selects exactly all the relevant variables.

We present a comparison between PVM and some existing methods for the transformation models (Li et al. 2014) and the AFT model (Xu et al. 2010) under the regular high-dimensional setting with $p < n$. The corresponding results are summarized in Table 2. As we can see from the table, for transformation models, PVM outperforms Li et al. (2014)'s method, especially for small sample cases where $n = 150$. With the exception of proportional odds model, PVM never omits any relevant variables whereas Li et al. (2014)'s approach tend to omit one to two relevant variables. In addition, for cases where all relevant variables are selected, PVM also produces a smaller active set as we can see from columns $4-7$ in Table 2. A similar, yet more distinctive, pattern can also be found in the comparison between PVM and Xu et al. (2010)'s method.

Under the ultra high-dimensional settings, we also compare PVM with SIS (Fan et al. 2010)

and Song et al. (2014)'s method for Cox's proportional hazards model and proportional odds model, respectively. For Cox's model, we adopted the same setting as shown in Table 1 except that $n < p$. For proportional odds model, we change the variance matrix to be a $p \times p$ matrix with its (i, j)'s element $= 0.5^{|i-j|}$. The regression parameters $\boldsymbol{\beta}_0$ follow Song et al. (2014)'s setting, which are to be $(-1, -0.9, 0, 0, 0, 0, 0, 0, 0.8, 1, 0, \ldots, 0, 0)$ with $H(t) = \log\{0.5(e^{2t} - 1)\}$. The censoring times were generated from uniform distributions Uniform$(0, 5.8)$ and Uniform$(0, 1.9)$ to achieve the censoring rates of 15% and 40%, respectively. Table 3 presents the results. It is noteworthy 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 so that all the relevant variables can be included. There is, therefore, no circumstances under which any single relevant variables is missing. That explains why the E^- column in Table 3 is always 0 for these two methods. We also summarize the mean numbers of extra irrelevant variables E^+. The figures in Table 3 reveal that PVM offers a high chance to select important variables with few irrelevant ones. In particular, for the proportional odds model, Song et al. (2014)'s proposal, on average, chooses 22.1 immaterial variables whereas PVM merely selects a mean of 0.64 extra unimportant one under the $(n, p) = (300, 5000)$ case. A more extensive table in which details about the performances of SIS under various settings is also included in the Appendix; see A3.

To examine the performance of 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 et al. (2010). In particular, for Case 3, covariates $Z_1, \cdots, Z_p$ are multivariate Gaussian distributed. They follow marginally $N(0, 1)$, and 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 $\{1, \cdots, p\} \backslash \{4\}$. $\beta = (4, 4, 4, -6\sqrt{2}, 0, 0, \cdots, 0)$. The censoring

rate is 30%. For Case 4, $Z_1, \cdots, Z_p$ are also multivariate Gaussian, each of which is marginally $N(0,1)$ distributed, and with correlation structure $corr(Z_i, Z_5) = 0$ for 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 $\{1, \cdots, p\} \backslash \{4,5\}$. Covariates are configured as $\beta = (4, 4, 4, -6\sqrt{2}, 4/3, 0, 0, \cdots, 0)$. The corresponding censoring rate is also around 30%. Similar performance is observed and is summarized in Table 4.

Finally, we also considered another high dimensional setting for the AFT model that is similar to that studied in Khan & Shaw (2016). Specifically, we set $(n, p) = (100, 120)$ with the first 20 coefficients for $\boldsymbol{\beta}$'s are specified to be 4 whereas the remaining coefficients of $\boldsymbol{\beta}$ are chosen to be zero. Covariates were generated as Z follows Uniform$(0,1)$ with correlations 0 and $0.5^{|i-j|}$ for two separate cases while the error is standard normally distributed. 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 to be 30% or 50%. According to the results shown in Table 5, with p_γ refers to significant variables and $p - p_\gamma$ represents non-relevant ones, PVM produces the highest net selection accuracy, which we define as the percentage of the relevant covariates selected minus that of the non-relevant variables chosen. The performance is more distinct for non-dependent case. This can be explained by the fact that OGA employs orthogonal projections in order to seek the next immediate relevant covariate. Under cases where 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 end this section, we also provide numerical results for cases with censoring times that depend on covariates. Specifically, we consider four aforementioned models, namely, Cox, probit, proportional odds and AFT, with covariates i.i.d. Z_{ij} generated from Uniform $(0,1)$. $\beta = (4, -4, 4, -4, 4, 0, \cdots, 0)$. Censoring times are generated from $\exp(6Z_1)$ distribution. The results are presented in Table 6. Again, our proposed procedure yields decent results with merely few

cases of over-selection.

Statistica Sinica

Table 1: Performance between PVM on (a) Cox proportional hazards model, (b) probit model, (c) proportional odds model and (d) the accelerated failure time (AFT) model under ultra -high dimensional settings, *i.e.* $n \ll p$. Frequency, in 100 simulations, of including all relevant variables (Correct), of selecting exactly the relevant variable (E), of selecting all relevant variables and i irrelevant variables ($E+i$), and of selecting some relevant variables with i relevant ones omitted ($E-i$). The column "Correct" specifies the number of cases where all the 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	1000	96	1	1	98	2	0	150	1000	97	3	0	100	0	0
200	1000	99	1	0	100	0	0	200	1000	100	0	0	100	0	0
400	1000	100	0	0	100	0	0	400	1000	100	0	0	100	0	0
200	5000	99	0	0	99	1	0	200	5000	99	0	0	99	1	0
200	10000	100	0	0	100	0	0	200	10000	100	0	0	100	0	0
400	5000	100	0	0	100	0	0	400	5000	100	0	0	100	0	0
400	10000	100	0	0	100	0	0	400	10000	100	0	0	100	0	0
(c) Proportional odds Model								(d) The Accelerated failure time (AFT) model							
300	1000	96	4	0	100	0	0	200	1000	100	0	0	100	0	0
400	1000	100	0	0	100	0	0	400	1000	100	0	0	100	0	0
300	5000	99	0	0	99	1	0	200	5000	98	0	0	98	2	0
300	10000	94	0	0	94	4	2	200	10000	100	0	0	100	0	0
400	5000	100	0	0	100	0	0	400	5000	100	0	0	100	0	0
400	10000	100	0	0	100	0	0	400	10000	100	0	0	100	0	0

Table 2: Comparison of performance between PVM and relevant existing methods on (a) Cox proportional hazards model, (b) probit model, (c) proportional odds model and (d) the accelerated failure time (AFT) model under traditional high-dimensional settings, *i.e.* $n > p$, $p \approx n$. Frequency, in 100 simulations, of including all relevant variables (Correct), of selecting exactly the relevant variable (E), of selecting all relevant variables and i irrelevant variables ($E+i$), and of 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 the 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

Continued on next page

Table 2 – continued from previous page

Method	n	p	E	$E+1$	$E+2$	$E+3^+$	Correct	$E-1$	$E-2$	$E-1+1$	$E-2+3$	$E-3+2$
(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 et al. (2010)	150	100	64	23	9	3	99	1	0	0	0	0
Xu et al. (2010)	200	150	67	23	9	0	99	1	0	0	0	0
Xu et al. (2010)	400	300	82	16	2	0	100	0	0	0	0	0

Statistica Sinica

Table 3: Comparison of performance between PVM and (a) SIS for Cox proportional hazards model and (b) Song et al. (2014)'s method for proportional odds 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	1000	23	0.02 (0.14)	5.98 (0.14)	0.03 (0.22)
PVM	200	1000	23	0 (0)	6 (0)	0.01 (0.10)
PVM	400	1000	23	0 (0)	6 (0)	0 (0)
PVM	200	5000	23	0.01 (0.10)	5.99 (0.10)	0 (0)
PVM	200	10000	23	0 (0)	6 (0)	0 (0)
PVM	400	5000	23	0 (0)	6 (0)	0 (0)
PVM	400	10000	23	0 (0)	6 (0)	0 (0)
SIS	150	1000	23	0.14 (0.64)	5.86 (0.64)	1.64 (1.43)
SIS	200	1000	23	0 (0)	6 (0)	2.35 (1.60)
SIS	400	1000	23	0 (0)	6 (0)	7.91 (3.57)
SIS	200	5000	23	0.09 (0.64)	5.91 (0.64)	0.99 (3.15)
SIS	200	10000	23	0.36 (1.11)	5.64 (1.11)	1.46 (5.27)
SIS	400	5000	23	0 (0)	6 (0)	1.63 (1.38)
SIS	400	10000	23	0 (0)	6 (0)	0.85 (0.90)
(b) Proportional odds model						
PVM	300	5000	15	0 (0)	4 (0)	0.41 (0.71)
PVM	300	5000	40	0.15 (0.39)	3.85 (0.39)	0.64 (0.98)
Song et al. (2014)	300	5000	15	0 (0)	4 (0)	7.60 (4.70)
Song et al. (2014)	300	5000	40	0 (0)	4 (0)	22.10 (7.80)

4 Data Analyses

We studied two datasets in this section, namely Stanford heart transplant data (Miller & Halpern 1982) and Diffuse large-B-cell lymphoma data (Rosenwald et al. 2002).

4.1 Stanford heart transplant data

To demonstrate PVM's performance for a dataset with a regular dimension, we first present an analysis of a classical dataset for Cox proportional hazards model, namely, Stanford heart

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Table 4: Comparison of performance between PVM and Fan et al. (2010)'s approach under Cases 3 and 4 settings

Method	n	p	Fan et al. (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	1000	Case 3	0.14 (0.51)	3.86 (0.51)	0.13 (0.63)
PVM	400	1000	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	1000	Case 3	0 (0)	4 (0)	10.93 (3.77)
ISIS	400	1000	Case 4	0 (0)	5 (0)	11.26 (3.32)

Table 5: Comparison of performance between PVM and Khan & Shaw (2016)'s 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

transplant data collected in February 1980. The dataset contains 157 observations of the following 4 variables: (i) the survival days for each patient, (ii) the censoring indicator, (iii) the

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Table 6: Comparison of performance between PVM and relevant existing methods on (a) Cox proportional hazards model, (b) probit model, (c) proportional odds model and (d) the accelerated failure time (AFT) model under dependent censoring. Frequency, in 100 simulations, of including all relevant variables (Correct), of selecting exactly the relevant variable (E), of selecting all relevant variables and i irrelevant variables ($E+i$), and of 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 the relevant variables are selected.

Model	n	p	E	$E+1$	$E+2$	Correct	$E-1$	$E-2$
Cox	400	10000	99	1	0	100	0	0
Probit	400	10000	100	0	0	100	0	0
PO	400	10000	99	1	0	100	0	0
AFT	400	10000	100	0	0	100	0	0

age at time of first transplant and (iv) the mismatch score. Since we need not carry out any 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 PVM for low-dimensional cases.

Table 7: Analysis of Stanford heart transplant data using the pseudo value method

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

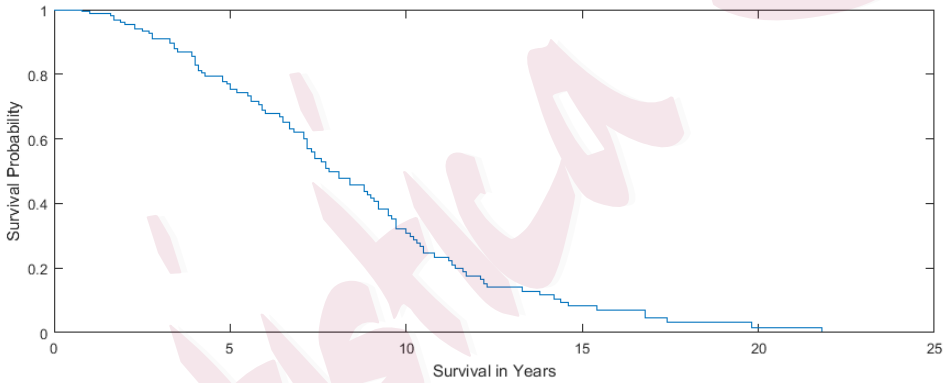
4.2 Diffuse large B-cell lymphoma studies

We also analysed a type of diffuse large-B-cell lymphoma, which is the most common type of lymphoma in adults that be cured by chemotherapy for only 35 to 40 percent of patients. In Rosenwald et al. (2002), the authors examined if gene-expression profiles of the type lymphoma of interest can be used to predict the outcome of chemotherapy via a multivariate Cox proportional hazards model. Biopsy samples of diffuse large-B-cell lymphoma from 240 patients were examined for 7,399 gene expression with the use of DNA microarrays and were analysed for ge-

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

nomic abnormalities. The dataset is available to the public at http://llmpp.nih.gov/DLBCL/. Among these 240 samples, 138 patient died (which accounts for 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 dataset is shown as follows:

Figure 1: Kaplan-Meier Plot of Overall Survival of Patients for Large-B-Cell Lymphoma Data



To analyse this dataset, we adopted the same initial setup as shown in our simulation studies. By using the pseudo values $\hat{Y}$ obtained in the first two stages of our procedure, we performed the variable selection via OGA with HDIC as the model selection criterion. The genes selected are listed in Table 8. In particular, the genes with asterisks were concluded as significant classes of genes in Rosenwald et al. (2002), namely, 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 various models are worth some attention

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

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 & 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*,†				

because they can be most likely overlooked influential variables.

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

To compare the conclusions yielded from different approaches, we also performed a likelihood ratio test to compare our selection with Gui & Li (2005) based on the Cox model. As reported, PVM chose 6 variables and the corresponding model's log-likelihood is -387.13; Gui & Li (2005) selected 10 variables with log-likelihood equals -386.23. Combining these two results, we constructed a test that covered the union of the 14 variables selected. The corresponding new model (Model C_1) has the log-likelihood of -383.60. Here we define $D_{1,2}$ as twice the difference in the log-likelihoods, which means 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 $D_{\text{Gui \& Li (2005)},\ C_1}$ are found to be -7.1 and -5.3 with 8 and 4 degrees of freedom, respectively. Neither of these two models was found statistically different from the full model. In other words, both models are statistically equivalent; the model obtained via PVM, however, is more parsimonious. To make the comparison easier to understand, we also chose the best 10 variables based on PVM, which

yielded the log-likelihood of -383.56. With the model that involves 10 variables chosen by Gui & Li (2005), the corresponding log-likelihood equals -386.23. It also reflects that the model chosen by PVM achieves higher (log-)likelihood with the same number of variables selected.

5 Discussion

In this paper, we introduce an innovative pseudo value method (PVM) with applications on ultra high-dimensional life-time data. These pseudo-values can be regarded as a set of educated guesses for the response variables, some of which are not fully observable due to censoring. Our numerical results have demonstrated the promising performance of PVM: It can identify the relevant variables while minimizing the number of irrelevant ones under statistically challenging settings with $n \ll p$.

Although there has been a wealth of procedures that are designed to tackle variable selection problems, most of them are developed under the linear regression context. To implement these ideas on survival models, one has to rely heavily on the (pseudo-/partial-) likelihood upon which a penalty can be applied. As a consequence, it is not trivial to incorporate SIS (Fan & Lv 2008) component to these semiparametric models whose likelihood cannot be easily calculated. One main contribution of our method is that it bridges the gap between these powerful results developed for linear models with semiparametric survival models so that challenges due to ultra high-dimensional variables and censored data can be properly handled.

As a final remark, it is noteworthy that PVM is a generic approach and by no means it should be restricted to only the two classes of models studied. Other models like general linear models (GLM) and quantile regression models can also be handled using a similar procedure. It is expected that this method remains effective for a wide range of regression-type problems.

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Development for these models will be considered in future research.

Supplementary Materials

The supplementary materials include the algorithm adopted in Step 2 of our proposed method for the general transformation models, additional details on the computation time of our proposal as well as further numerical results on for Cox's proportional hazards model using SIS. Proofs for theorems presented in Section 2.3 will also be presented.

References

Antoniadis, A., Fryzlewicz, P. & Letue, F. (2010), 'The dantzig selector in cox's proportional hazads model', *Scandinavian Journal of Statistics* **37**, 531–52.

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

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

Cai, T., Huang, J. & Tian, L. (2009), 'Regularized estimation for the accelerated failure time model', *Biometrics* pp. 394–404.

Chen, H. Y. & Little, R. J. (2001), 'A profile conditional likelihood approach for the semiparametric transformation regression model with missing covariates', *Lifetime Data Analysis* **7**, 207–24.

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

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Cox, D. (1975), 'Partial likelihood', *Biometrika* **62**, 269–76.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. J. (2004), 'Least angle regression', *The Annals of Statistics* **32**, 407–99.

Fan, J. & Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of American Statistical Association* **96**, 1348–60.

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

Fan, J., Yang, F. & 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. G., Sun, L. & Zuo, G. (2005), 'A baseline-free procedure for transformation models under interval censorship', *Lifetime Data Analysis* **11**, 473–88.

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

Gu, M. & 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**, 7270–74.

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

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

21, 3001–8.

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

Huang, J. & 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. & Zhang, C.-H. (2013), 'Oracle inequalities for the lasso in the cox model', *Annals of statistics* **41**(3), 1142.

Ing, C.-K. & 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. & Ying, Z. (2003), 'Rank-based inference for accelerated failure time model', *Biometrika* **90**, 341–53.

Khan, M. H. R. & Shaw, J. E. H. (2016), 'Variable selection for survival data with a class of adaptive elastic net techniques', *Statistics and Computing* **26**, 725–41.

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

Li, H. & 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. & Gu, M. (2012), 'Adaptive lasso for general transformation modelds with right censored data', *Computational Statistics and Data Analysis* **56**(2583–97).

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

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

Lu, W. & Zhang, H. (2007), 'Variable selection for proportional odds model', *Statistics in Medicine* **26**, 3771–81.

Miller, R. & Halpern, J. (1982), 'Regression with censored data', *Biometrika* **69**, 521–31.

Prekopa, A. (1973), 'On logarithmic convace measures and functions', *Acta Scientiarum Mathematicarum* **34**, 335–43.

Rosenwald, A., Wright, G., Chan, W., Connors, J., Campo, E., Fisher, R., Gascoyne, R., Muller-Hermelink, H., Smeland, E. & 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**, 1937–47.

Song, R., Lu, W., Ma, S. & 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**, 461–71.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society, Series B* **58**, 267–88.

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Xu, J., Leng, C. & Ying, Z. (2010), 'Rank-based variable selection with censored data', *Statistics and Computing* **20**, 165–76.

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

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society, Series B* **67**, 301–20.

Statistica Sinica