Statistica Sinica **22** (2012), 1003-1020
doi:http://dx.doi.org/10.5705/ss.2010.267

VARIABLE SELECTION IN PARTLY LINEAR REGRESSION MODEL WITH DIVERGING DIMENSIONS FOR RIGHT CENSORED DATA

Shuangge Ma and Pang Du

Yale University and Virginia Tech

Abstract: Recent biomedical studies often measure two distinct sets of risk factors: low-dimensional clinical and environmental measurements, and high-dimensional gene expression measurements. For prognosis studies with right censored response variables, we propose a semiparametric regression model whose covariate effects have two parts: a nonparametric part for low-dimensional covariates, and a parametric part for high-dimensional covariates. A penalized variable selection approach is developed. The selection of parametric covariate effects is achieved using an iterated Lasso approach, for which we prove the selection consistency property. The nonparametric component is estimated using a sieve approach. An empirical model selection tool for the nonparametric component is derived based on the Kullback-Leibler geometry. Numerical studies show that the proposed approach has satisfactory performance. Application to a lymphoma study illustrates the proposed method.

Key words and phrases: Semiparametric regression, variable selection, right censored data, iterated Lasso.

1. Introduction

Consider regression models for right censored data. Let Ξ be the quantity of interest that is at risk of being censored from the right by a random variable Γ. One observes $\min(\Xi, \Gamma)$ or, often more conveniently, $Y = \min(T, C) \equiv \min(g_0(\Xi), g_0(\Gamma))$ for some known monotone transformation g_0 and indicator function $\delta = I_{[T \leq C]}$. Let $X \in \mathbb{R}^p$ and $Z \in \mathbb{R}^q$ be two sets of covariates related to T. Here the dimension p of X is often high and can even be allowed to diverge faster than the sample size n, whereas the dimension q of Z is usually low and can be considered as fixed. Given iid observations $\{(Y_i, \delta_i, X_i, Z_i),\ i = 1, \ldots, n\}$, we assume that the data can be described using the semiparametric regression model

$$T_i = \alpha + X_i^T \boldsymbol{\beta} + \eta(Z_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where α is the unknown intercept, $\boldsymbol{\beta}$ is an unknown coefficient vector, η is an unknown multivariate smooth function, and the ϵ_i's are iid random errors with an unknown distribution having mean 0 and unspecified finite variance σ^2. Since η is identifiable up to a constant, we adopt the constraint $\int \eta = 0$. Note that (1.1) includes the well-known accelerated failure time (AFT) model in survival analysis as a special case when g_0 is the logarithm function.

The data and model settings have been partly motivated by recent cancer prognosis studies. It is now commonly accepted that clinical and environmental risk factors do not have sufficient predictive power for cancer prognosis. Thus, in recent studies, two distinct sets of covariates are measured. The first set X represents high-dimensional genomic measurements such as microarray gene expressions or SNPs. The second set Z represents low-dimensional clinical and environmental risk factors. We refer to Ma and Huang (2007) for examples of such studies. With the high-dimensional X, it is of interest to identify a small subset that is associated with prognosis. For better interpretability and because of computational and theoretical limitations, the effect of X is usually modeled in a parametric way. With the low-dimensional Z, we adopt a more flexible nonparametric model, as many biological processes are nonlinear.

Variable selection for high-dimensional censored data has drawn much attention in the past decade. Various penalization procedures have been proposed assuming the Cox proportional hazards (PH) model. Examples include the LASSO in Tibshirani (1997), the SCAD in Cai et al. (2005), the adaptive LASSO in Zhang and Lu (2007) and Zou (2008), and the SIS in Fan, Feng, and Wu (2010). However, those models all assume a linear form of covariate effects in the relative risk. As an alternative to the PH model, the AFT model, as noted by Sir David R. Cox, is “in many ways more appealing because of its quite direct physical interpretation” (Reid (1994)). Under this direction, Johnson (2008) extended the SCAD procedures for selecting variables in an AFT model, but their model is a simplified parametric version of (1.1) with $\eta \equiv 0$. Zhang, Lu, and Wang (2010) further generalized these results to semiparametric transformation models with an unknown transformation function and linear covariate effects. In summary, the aforementioned variable selection procedures share the common limitation of assuming parametric covariate effects that may not be flexible enough in practice. Xie and Huang (2009) proposed the SCAD procedure for partially linear regression models with parametric covariates of diverging dimensions, but their model has limitations: it is for uncensored data, its nonparametric component is of one dimension, and the dimension of parametric covariates diverges in the order of $o(n^{1/2})$ which may not be appropriate for (e.g) gene expression studies with $p > n$. A recent work by Du, Ma, and Liang (2010) considered penalized variable selection procedures for PH models with semiparametric relative risk. Their approach allows more general nonparametric components but is limited to covariates of fixed dimensions. Johnson (2009) and Long et al. (2011) proposed regularized extensions to the rank estimation for partly linear AFT models that require a pre-specified stratification of nonparametric covariates. Johnson (2009) did not provide an estimate for the nonparametric component and focused on the case with fixed-dimension parametric covariates. Long et al. (2011) considered high-dimensional parametric covariates with $p > n$ but did not investigate the theoretical properties. Further, the simulations and data analysis in Long et al. (2011) only dealt with a one-dimensional nonparametric covariate effect, although the extension to additive nonparametric covariate effects was discussed. Our work may be innovative in that our model integrates the following: (i) It is a regression model for censored data that is semiparametric in two aspects: the error distribution is unspecified except for its zero mean which is the assumption of most existing semiparametric censored regression models, and our model allows flexible semiparametric covariate effects whose nonparametric part can contain multiple additive components. (ii) The dimensionality of parametric component can diverge in an exponential order of n, making it more appropriate for data with, for example, genomic measurements. (iii) Our approach provides a model selection tool for the nonparametric components.

There are several options for estimating censored regression models. Popular examples include the Buckley-James estimator (Buckley and James (1979)) and the rank-based estimator (Tsiatis (1990); Ying (1993); Wei, Ying, and Lin (1990)). However, the computational cost of these approaches can be too high for high-dimensional data. A more computationally feasible alternative is the weighted least squares approach (Stute (1993)), which is equivalent to inverse probability weighting. It involves the minimization of a weighted least squares objective function and has been used in AFT models with high-dimensional covariates by Huang, Ma, and Xie (2006).

We adopt LASSO-type penalties for variable selection with the parametric component. Compared with alternatives such as SCAD, bridge, and others, LASSO-type penalties are computationally easier. The selection properties of LASSO-type penalties with uncensored data have been established (Zhang and Huang (2008); Meinshausen and Buhlmann (2006)). The main conclusion is that LASSO is not selection consistent except under strong orthogonality conditions. A remedy for the inconsistent selection of LASSO is the adaptive LASSO (Zou (2006)) that requires a consistent initial estimate to compute the adaptive weights. When the dimensionality of covariates is low, the initial estimate can be easily constructed through simple linear regression. This is not feasible when the dimensionality of covariates is high. Motivated by this, we propose an iterated Lasso approach for semiparametric regression model with right censored data. Our approach uses the LASSO estimate as the initial estimate, which is l_2-estimation consistent even in the high-dimensional setting. Using the LASSO estimate as the initial estimate has also been suggested by Meinshausen and Buhlmann (2006), Meinshausen (2007), and Meier and Buhlmann (2008). We then use the initial estimate to construct weights and conduct a weighted LASSO estimation that has the selection consistency property. The nonparametric component η is estimated through a sieve approach (Schumaker (1981)). We also propose an empirical model selection approach for η derived from the Kullback-Leibler geometry (Gu (2004)).

The rest of the article is organized as follows. The estimation and variable selection procedure is described in Section 2. The selection consistency property is established. Numerical study, including simulation and analysis of a lymphoma prognosis study, is presented in Section 3. The article concludes with discussion in Section 4. Some technical details are provided in the Appendix.

2. Penalized Estimation and Variable Selection

2.1. Weighted least squares estimation

Let $Y_{(1)} \leq \cdots \leq Y_{(n)}$ be the order statistics of Y_i, $\delta_{(1)}, \ldots, \delta_{(n)}$ be the associated censoring indicators, and $(X_{(1)}, Z_{(1)}), \ldots, (X_{(n)}, Z_{(n)})$ be the associated covariates. Let F be the distribution function of T and $\hat{F}_n$ be its Kaplan-Meier estimator $\hat{F}_n(t) = \sum_{i=1}^n w_i I_{[Y_{(i)} \leq t]}$, where

$$w_1 = \frac{\delta_{(1)}}{n} \quad \text{and} \quad w_i = \frac{\delta_{(i)}}{n-i+1} \prod_{j=1}^{i-1} (\frac{n-j}{n-j+1})^{\delta_{(j)}}, i = 2, \ldots, n, \tag{2.1}$$

are the Kaplan-Meier weights (Stute (1993)). An equivalent set of weights, as shown in Huang, Ma, and Xie (2007), are the inverse probability weights $\tilde{w}_i = \delta_{(i)}/\hat{G}(Y_{(i)}-)$, where $\hat{G}$ is the Kaplan-Meier estimator of the survival function G of censoring time C and $\hat{G}(t-)$ is the left-hand limit of the function $\hat{G}$ at t. The weighted least squares loss function is

$$Q_n(\boldsymbol{\beta}, \eta) = \frac{1}{2} \sum_{i=1}^n w_i (Y_{(i)} - \alpha - X_{(i)}^T \boldsymbol{\beta} - \eta(Z_{(i)}))^2. \tag{2.2}$$

In this article, we make the reasonable assumption that η is continuously differentiable as most biological processes are smooth. We estimate η using a sieve approach. Let J be a roughness penalty and $\mathcal{H} = \{\eta : J(\eta) < \infty\}$. When η is of one dimension, an appropriate choice of J is $J(\eta) = \int (\eta''(z))^2 dz$ which yields the well-known reproducing kernel (RK) Hilbert space $\mathcal{H}$ defining cubic smoothing splines. Let $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$, $\mathcal{H}_J = \mathcal{H} \ominus \mathcal{N}_J$, and R_J be the reproducing kernel in $\mathcal{H}_J$. Consider the sieve $\mathcal{H}_n = \mathcal{N}_J \oplus \text{span}\{R_J(z_j, \cdot), j = 1, \ldots, r_n\}$, where r_n is a constant possibly increasing with n and $\{z_j : j = 1, \ldots, r_n\}$ is a random subset of $\{Z_i : i = 1, \ldots, n\}$. We choose $\mathcal{H}_n$ as our sieve since it can naturally incorporate multivariate functions through tensor product spline spaces.

Here we briefly describe an example of tensor product cubic spline space that is used in our numerical study. Consider the case of a bivariate continuous covariate $Z = (Z^{(1)}, Z^{(2)})$. For simplicity, assume that the domains for $Z^{(1)}$ and $Z^{(2)}$ are both $[0,1]$. Consider $\mathcal{H}^{(1)} = \mathcal{H}^{(2)} = W_2^2[0,1]$, where

$$W_2^2[0,1] = \left\{ f : \ f \text{ and } f' \text{ are absolutely continuous}, \ \int_0^1 (f'')^2 dz < \infty \right\} \tag{2.3}$$

is the cubic smoothing spline model space. $W_2^2[0,1]$ can be decomposed as

$$W_2^2[0,1] = \mathcal{H}_0 \oplus \mathcal{H}_1, \tag{2.4}$$

where $\mathcal{H}_0 = \text{span}\{1, k_1(z)\}$, $\mathcal{H}_1 = \{f : \int_0^1 f dz = \int_0^1 f' dz = 0, \ \int_0^1 (f'')^2 dz < \infty\}$, and $k_\nu(z) = B_\nu(z)/\nu!$ are the scaled Bernoulli polynomials. The RK for subspace $\mathcal{H}_1$ is $R_1(z_1, z_2) = k_2(z_1)k_2(z_2) - k_4(|z_1 - z_2|)$. Denote the decompositions corresponding to (2.4) for marginal spaces $\mathcal{H}^{(j)}$ as $\mathcal{H}^{(j)} = \mathcal{H}_0^{(j)} \oplus \mathcal{H}_1^{(j)}, j = 1, 2$. Taking the tensor product, one obtains the space $\mathcal{H} = \mathcal{H}^{(1)} \otimes \mathcal{H}^{(2)}$ with

$$\begin{aligned} J(f) = & \int_0^1 \left\{ \int_0^1 f_{22}(Z^{(1)}, Z^{(2)}) dZ^{(1)} \right\}^2 dZ^{(2)} + \int_0^1 \left\{ \int_0^1 f_{11}(Z^{(1)}, Z^{(2)}) dZ^{(2)} \right\}^2 dZ^{(1)} \\ & + \int_0^1 \left\{ \int_0^1 f_{122}(Z^{(1)}, Z^{(2)}) dZ^{(1)} \right\}^2 dZ^{(2)} + \int_0^1 \left\{ \int_0^1 f_{112}(Z^{(1)}, Z^{(2)}) dZ^{(2)} \right\}^2 dZ^{(1)} \\ & + \int_0^1 \int_0^1 \left\{ f_{1122}(Z^{(1)}, Z^{(2)}) \right\}^2 dZ^{(1)} dZ^{(2)}, \end{aligned}$$

where $f_{ij} = \frac{\partial^2 f}{\partial Z^{(i)} \partial Z^{(j)}}$, $f_{ijk} = \frac{\partial^3 f}{\partial Z^{(i)} \partial Z^{(j)} \partial Z^{(k)}}$, and $f_{ijkl} = \frac{\partial^4 f}{\partial Z^{(i)} \partial Z^{(j)} \partial Z^{(k)} \partial Z^{(l)}}$. Accordingly, $\mathcal{N}_J = \text{span}\{1, k_1(Z^{(1)}), k_1(Z^{(2)}), k_1(Z^{(1)})k_1(Z^{(2)})\}$ and

$$\begin{aligned} R_J(Z_1, Z_2) = & R_1(Z_1^{(1)}, Z_1^{(2)}) + R_1(Z_2^{(1)}, Z_2^{(2)}) + R_1(Z_1^{(1)}, Z_1^{(2)}) k_1(Z_2^{(1)}) k_1(Z_2^{(2)}) \\ & + k_1(Z_1^{(1)}) k_1(Z_1^{(2)}) R_1(Z_2^{(1)}, Z_2^{(2)}) + R_1(Z_1^{(1)}, Z_1^{(2)}) R_1(Z_2^{(1)}, Z_2^{(2)}). \end{aligned}$$

A RK Hilbert space can also be constructed for functions on a discrete domain. We refer to Chapter 2 of Gu (2002) for more details.

Suppose $\{\phi_1, \ldots, \phi_m\}$ is a basis of $\mathcal{N}_J$. Then any function $\eta \in \mathcal{H}_n$ can be written as $\eta(\cdot) = \sum_{\nu=1}^m d_\nu \phi_\nu(\cdot) + \sum_{j=1}^n R_J(z_j, \cdot) \equiv \boldsymbol{\psi}^T(\cdot)\mathbf{b}$. We rewrite the objective function in (2.2) as $Q_n(\boldsymbol{\beta}, \mathbf{b}) = (1/2) \sum_{i=1}^n w_i (Y_{(i)} - \alpha - X_{(i)}^T \boldsymbol{\beta} - \boldsymbol{\psi}_{(i)}^T \mathbf{b})^2$, where $\boldsymbol{\psi}_{(i)} \equiv \boldsymbol{\psi}(Z_{(i)})$. We make the transformations $Y_{(i)}^* = \sqrt{w_i} \left(Y_{(i)} - \sum w_i Y_{(i)} / \sum w_i \right)$, $X_{(i)}^* = \sqrt{w_i} \left(X_{(i)} - \sum w_i X_{(i)} / \sum w_i \right)$ and $\boldsymbol{\psi}_{(i)}^* = \sqrt{w_i} \left(\boldsymbol{\psi}_{(i)} - \sum w_i \boldsymbol{\psi}_{(i)} / \sum w_i \right)$.

The objective function can then be rewritten as

$$Q_n(\boldsymbol{\beta}, \mathbf{b}) = \frac{1}{2}\sum_{i=1}^{n}(Y^*_{(i)} - X^{*T}_{(i)}\boldsymbol{\beta} - \boldsymbol{\psi}^{*T}_{(i)}\mathbf{b})^2. \tag{2.5}$$

2.2. Penalized variable selection

The proposed variable selection procedure consists of the following steps.

(S1) Initialize $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}) = \arg\min_{\boldsymbol{\beta},\mathbf{b}} Q_n(\boldsymbol{\beta}, \mathbf{b}) + \lambda_n \sum_j |\beta_j|$, where β_j is the j^{th} component of $\boldsymbol{\beta}$;

(S2) Compute $v_j = |\hat{\beta}_j|^{-\gamma}$ for a fixed $\gamma > 0$. Compute the adaptive Lasso estimate

$$(\hat{\boldsymbol{\beta}}, \hat{\mathbf{b}}) = \arg\min_{\boldsymbol{\beta},\mathbf{b}} Q_n(\boldsymbol{\beta}, \mathbf{b}) + \lambda_n \sum_j v_j|\beta_j|; \tag{2.6}$$

(S3) Repeat Step (S2) until convergence.

We adopt a sieve approach for the nonparametric covariate effects. Thus there is no need for additional constraints on $\mathbf{b}$ as smoothing spline estimation generally does. We borrow the basis functions of reproducing kernel Hilbert spaces, and the number of basis functions is taken to be much smaller than the sample size. With the Lasso penalty in (S1), the objective function is convex and can easily be minimized. In Section 2.4, we show that the Lasso can select all important covariates plus some false positives. This result justifies the validity of Lasso estimate as the initial estimate. In Steps (S2) and (S3), if $\hat{\beta}_j = 0$, then the corresponding covariate is taken out of penalized estimation. Section 2.4 shows that the one-step estimate after one iteration is selection consistent. However, our numerical study suggests that iterating until convergence may improve the finite sample property. Our experience shows that convergence can usually be achieved within a few iterations. The idea of improving consistency via iterated penalization is similar to that in Zou and Zhang (2009). The present study may be more complicated due to the presence of censoring and the nonparametric component η.

The proposed procedure involves computation of the (weighted) Lasso estimate that is implemented using the coordinate descent algorithm (Wu and Lange (2007)). The tuning parameter λ_n balances sparsity and goodness-of-fit, and can be chosen using V-fold cross validation. In Section 2.4, we provide conditions on λ_n under which the selection consistency holds.

2.3. Variable selection for nonparametric component

Even though the dimensionality of Z is low, it may still be of interest to identify components of $\eta(Z)$ that are not associated with the response variable. In this section, we derive a model selection procedure for the nonparametric component based on Kullback-Leibler geometry. For two estimates η_1 and η_2 of the true function η_0, their Kullback-Leibler distance reduces to

$$\mathrm{KL}(\eta_1, \eta_2) = \frac{1}{2n} \sum_{i=1}^{n} (\eta_1(Z_i) - \eta_2(Z_i))^2. \tag{2.7}$$

Suppose that the estimation of η_0 has been done in a space $\mathcal{H}_1$, but in fact $\eta_0 \in \mathcal{H}_2 \subset \mathcal{H}_1$. Let $\hat{\eta}$ be the estimate of η_0 in $\mathcal{H}_1$. Let $\tilde{\eta}$ be the Kullback-Leibler projection of $\hat{\eta}$ in $\mathcal{H}_2$, that is, the minimizer of $\mathrm{KL}(\hat{\eta}, \eta)$ for $\eta \in \mathcal{H}_2$, and let η_c be the estimate from the constant model. Set $\eta = \tilde{\eta} + \rho(\tilde{\eta} - \eta_c)$ for ρ real and $K(\rho) \equiv \mathrm{KL}(\hat{\eta}, \eta) = (1/2n) \sum_{i=1}^{n} (\hat{\eta}(Z_i) - (\tilde{\eta} + \rho(\tilde{\eta} - \eta_c))(Z_i))^2$. Differentiating $K(\rho)$ with respect to ρ and evaluating at $\rho = 0$, one has $(1/n) \sum_{i=1}^{n} (\hat{\eta}(Z_i) - \tilde{\eta}(Z_i))(\tilde{\eta}(Z_i) - \eta_c(Z_i)) = 0$. Straightforward calculation then yields

$$\mathrm{KL}(\hat{\eta}, \eta_c) = \mathrm{KL}(\hat{\eta}, \tilde{\eta}) + \mathrm{KL}(\tilde{\eta}, \eta_c).$$

Hence the ratio $\mathrm{KL}(\hat{\eta}, \tilde{\eta}) / \mathrm{KL}(\hat{\eta}, \eta_c)$ can be used to diagnose the feasibility of a reduced model $\eta \in \mathcal{H}_2$: the smaller the ratio, the more feasible the reduced model.

2.4. Asymptotic properties

For fixed $\boldsymbol{\beta}$, $\hat{\mathbf{b}}$ satisfies $\sum_{i=1}^{n} \boldsymbol{\psi}^*_{(i)} (Y^*_{(i)} - X^{*T}_{(i)} \boldsymbol{\beta} - \boldsymbol{\psi}^{*T}_{(i)} \hat{\mathbf{b}}) = 0$. That is,

$$\hat{\mathbf{b}} = \left(\sum_{i=1}^{n} \boldsymbol{\psi}^*_{(i)} \boldsymbol{\psi}^{*T}_{(i)} \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{\psi}^*_{(i)} (Y^*_{(i)} - X^{*T}_{(i)} \boldsymbol{\beta}) \right).$$

Let $P_i = \boldsymbol{\psi}^{*T}_{(i)} \left(\sum_{i=1}^{n} \boldsymbol{\psi}^*_{(i)} \boldsymbol{\psi}^{*T}_{(i)} \right)^{-1} \boldsymbol{\psi}^*_{(i)}$ be the projection. The objective function (2.5) can be rewritten as

$$Q_n(\boldsymbol{\beta}) = \frac{1}{2} \sum_{i=1}^{n} ((I - P_i)(Y^*_{(i)} - X^{*T}_{(i)} \boldsymbol{\beta}))^2 = \frac{1}{2} \sum_{i=1}^{n} (\tilde{Y}^*_{(i)} - \tilde{X}^{*T}_{(i)} \boldsymbol{\beta})^2. \tag{2.8}$$

Let $\tilde{Y} = (\tilde{Y}^*_{(1)}, \ldots, \tilde{Y}^*_{(n)})^T$ and $\tilde{X}$ be the $n \times p$ matrix consisting of row vectors $\tilde{X}^{*T}_{(1)}, \ldots, \tilde{X}^{*T}_{(n)}$. Let $\tilde{X}_1, \ldots, \tilde{X}_p$ be the p columns of $\tilde{X}$. Let $W = \mathrm{diag}(nw_1, \ldots, nw_n)$ be a $n \times n$ diagonal matrix. For $A \subset \{1, \ldots, p\}$, let $\tilde{X}_A = (\tilde{X} : j \in A)$ be the matrix with columns $\tilde{X}_j$s for $j \in A$. Write $\Sigma_A = \tilde{X}_A^T W \tilde{X}_A / n$ and denote the cardinality of A by $|A|$. Let $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^T$ be the unknown true regression coefficients. Let $A_1 = \{j : \beta_{0j} \neq 0\}$ be the set of nonzero regression coefficients and $|A_1| = p_1$. For $i = 1, \ldots, n$, let $\tau_i = w_i \epsilon_{(i)} \equiv w_i(\tilde{Y}^*_{(i)} - \tilde{X}^{*T}_{(i)} \boldsymbol{\beta}_0)$ and $\xi_j = \sum_{i=1}^n \tilde{X}^*_{ij} \tau_i, 1 \leq j \leq p$. We assume the following.

(A1) p_1 is finite.

(A2) (a) $\{(Y_i, \delta_i, X_i, Z_i), i = 1 \ldots n\}$ are iid. (b) The random errors $\epsilon_1, \ldots, \epsilon_n$ are iid with mean 0 and finite variance σ^2, and there exist $K_1, K_2 > 0$ such that $P(|\epsilon_i| > u) \leq K_2 \exp(-K_1 u^2)$ for all $u > 0$.

(A3) (a) The distributions of ξ_j's are subgaussian. (b) There exists $M > 0$ such that $|X_{ij}|, |Z_{ij}| \leq M$.

(A4) Matrix $\tilde{X}$ satisfies the sparse Riesz condition (SRC) with rank p_1^*: there exist constants $0 < c_* < c^* < \infty$ such that, for $C = c^*/c_*$ and $p_1^* = (3 + 4C)p_1$, with probability converging to 1, $c_* \leq \nu^T \Sigma_A \nu / ||\nu||^2 \leq c^*$, for any A with $|A| = p_1^*$ and $\nu \in \mathbb{R}^{p_1^*}$. Here $|| \cdot ||$ is the ordinary l_2 norm.

The model is sparse under (A1), reasonable in genomic studies where the number of genes profiled can be large, but only a very small number of genes are associated with the response variables. The subgaussian assumption (A2) is commonly made in high dimensional data analysis but can be weakened at the price of a smaller p. The subgaussian property in (A3) is required for Theorem 1; It can be ensured by the boundedness of covariates X and Z plus certain other conditions. For example, a sufficient condition can be boundedness of random errors. Another sufficient condition that leads to subgaussian ξ_j's is that $w_i \leq c/n$ for some constant $c > 1$. This happens, for example, when $\delta_{(j)} = 1$ for all $j > n/(k_c)$ with $k_c = c/(c-1)$. The latter can be achieved when $C \leq \tau_0$ for a constant τ_0 and $P\{(X, Z) : P(T > \tau_0 | X, Z) = 1\} > 0$. In Huang and Ma (2010) with parametric AFT models, to achieve the subgaussian property of ξ_j, it is assumed that the errors $(\epsilon_1, \ldots, \epsilon_n)$ are independent of the weights $(w_1, \ldots, w_n)$. Although reasonable arguments have been provided in Huang and Ma (2010), it is worth noting that the weights are estimates generated from data. Other sufficient conditions for (A3) are certainly possible. The SRC condition is proposed in Zhang and Huang (2008). It guarantees that all eigenvalues of any $d \times d$ submatrix of $\tilde{X}^T W \tilde{X}/n$ with $d \leq p_1^*$ lie between c_* and c^*. That is, any model with dimensionality no greater than p_1^* is identifiable. In this study, conditions on the basis functions of η are built in the conditions on $\tilde{X}$. The presence of censoring brings considerable difficulty, which makes it hard to "separate" conditions on η as in Xie and Huang (2009). With a fixed dimensionality and correlation structure for X, the SRC condition in (A4) needs to be checked following Zhang and Huang (2008).

2.4.1. The initial estimate

We first investigate the Lasso estimate computed in (S1) of the proposed procedure. The estimate is $\tilde{\boldsymbol{\beta}} = \arg\min Q_n(\boldsymbol{\beta}) + \lambda_n \sum_j |\beta_j|$. Define $\tilde{A}_1 = \{j : \tilde{\beta}_j \neq 0\}$ as the set of nonzero Lasso estimate coefficients.

Theorem 1. *Suppose that* (A1)−(A4) *hold and* $\lambda_n/\sqrt{n\log(p)}$ *is bounded away from zero. Then*

(a) *with probability converging to 1,* $|\tilde{A}_1| \leq (2+4C)p_1$*;*

(b) *if* $\lambda_n/n \to 0$ *and* $\log(p)/n \to 0$*, then with probability converging to 1, all components of* X *with nonzero coefficients are selected;*

(c) $||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2 \leq 16\lambda_n^2 p_1/(n^2 c_*^2) + O\left(|\tilde{A}_1|\log(p)/(nc_*^2)\right)$*, and if* $\lambda_n = O(\sqrt{n\log(p)})$*, then* $||\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}||^2 = O(\log(p)/n)$*.*

Theorem 1 suggests that, with a high probability, all true positives are selected. Thus, the Lasso estimate serves well as the initial estimate. In addition, the Lasso estimate is *estimation consistent*, a desired property for the initial estimate of the adaptive Lasso.

2.4.2. The iterated estimate

We now investigate properties of $\hat{\boldsymbol{\beta}} = \arg\min Q_n(\boldsymbol{\beta}) + \lambda_n \sum_j v_j|\beta_j|$, the adaptive Lasso estimate defined in (S2). For a vector $u = (u_1, \ldots, u_p)$, let $\text{sign}(u) = (\text{sign}(u_1), \ldots, \text{sign}(u_p))$, where $\text{sign}(u_i) = 1, 0, -1$ if $u_i > 0, = 0, < 0$.

Theorem 2. *Suppose that* (A1)−(A4) *hold, that* $\log(p)/n \to 0$ *and* $\lambda_n = O(\sqrt{n\log(p)})$*. Then* $P(sign(\hat{\boldsymbol{\beta}}) = sign(\boldsymbol{\beta}_0)) \to 1$.

Theorem 2 suggests that the one-step adaptive Lasso estimate computed in (S2) is selection consistent. Following a similar strategy, it can be proved that any finite-step estimate is selection consistent. Under conditions described above, Theorems 1 and 2 hold if $\log(p)/n \to 0$. Thus, the proposed approach can accommodate $p = \exp(o(n))$, that is, very high-dimensional data.

3. Numerical Study

3.1. Simulation

We simulated data from the AFT model such that $g_0 = \log(\cdot)$. Let $\mathcal{W}(a,b)$ denote the Weibull distribution with shape parameter a and scale parameter b. The failure time Ξ was generated from $\mathcal{W}(4, \exp(\mu_0(x,z)))$ where $\mu_0(x,z) =$

$x^T\boldsymbol{\beta} + \eta(z)$. The censoring time Γ was generated from an exponential distribution whose parameter was adjusted to yield a censoring rate about 30%. The dimensionality p of covariate X was either 200 or 500. X_i's were independently generated from the multivariate normal distribution with zero mean and $\text{cov}(X_i, X_j) = 0.112 \cdot 0.5^{|i-j|}$. The first 15 entries of the coefficient vector $\boldsymbol{\beta}$ were $(1.0, 0.5, 0.9, 0.7, 1.0, 0.7, 0.9, 0.5, 0.6, 0.7, 0.6, 0.9, 1.0, 1.0, 0.6)$, and the rest of the entries were zero. The covariate $Z = (Z_1, Z_2, Z_3)$ had three dimensions, and each component was simulated from the uniform distribution on $[0, 1]$. We took $\eta(z) = \eta_1(z_1) + \eta_2(z_2) + \eta_3(z_3)$, with $\eta_1(z) = 0.5\sin(2\pi z - \pi/2)$, $\eta_2(z) = 2(z - 0.4)^2 + 2.28e^{-z} - 1.628$, and $\eta_3(z) = z - 0.5$. Note that all η_js integrate to zero to make model identifiable.

For a prediction procedure $\mathcal{M}$ and the estimator $(\hat{\boldsymbol{\beta}}_{\mathcal{M}}, \hat{\eta}_{\mathcal{M}})$ obtained from the procedure, an appropriate measure of prediction performance is the empirical prediction error $\text{PE}(\hat{\boldsymbol{\beta}}_{\mathcal{M}}, \hat{\eta}_{\mathcal{M}}) = (1/N)\sum_{i=1}^{N} w_{i,0}(Y_{(i),0} - X_{(i),0}^T\hat{\boldsymbol{\beta}}_{\mathcal{M}} - \hat{\eta}_{\mathcal{M}}(Z_{(i),0}))^2$. Here $w_{i,0}$ and $(Y_{(i),0}, X_{(i),0}, Z_{(i),0})$ are, respectively, the Kaplan-Meier weights and ordered statistics for a test data set $\{(Y_{i,0}, \delta_{i,0}, X_{i,0}, Z_{i,0}) : i = 1, \ldots, N\}$ independently generated from the true model. The relative model error (RPE) of $\mathcal{M}_1$ versus $\mathcal{M}_2$ is defined as the ratio $\text{PE}(\hat{\boldsymbol{\beta}}_{\mathcal{M}_1}, \hat{\eta}_{\mathcal{M}_1}) / \text{PE}(\hat{\boldsymbol{\beta}}_{\mathcal{M}_2}, \hat{\eta}_{\mathcal{M}_2})$. The procedure $\mathcal{M}_0$ with complete oracle is used as a benchmark. In $\mathcal{M}_0$, $(X_1, \ldots, X_{15}, Z_1, Z_2, Z_3)$ are known to be the only contributing covariates, the exact form of η_0 is known, and the only parameters to be estimated are the coefficients of $X_1, \ldots, X_{15}$. Note that $\mathcal{M}_0$ can be implemented only in simulation and is unrealistic in practice. We compared performance of the following procedures through their RPEs versus $\mathcal{M}_0$.

$\mathcal{M}_A$: The procedure with partial oracle and estimated η_0. That is, $(X_1, \ldots, X_{15}, Z_1, Z_2, Z_3)$ are known to be the only contributing covariates, but the form of η_0 is unknown. η_0 is estimated together with the coefficients for $(X_1, \ldots, X_{15})$ using the penalized weighted least squares approach defined in (2.6), with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{15})^T$.

$\mathcal{M}_B$: The procedure with partial oracle and misspecified η_0. Here, $(X_1, \ldots, X_{15}, Z_1, Z_2, Z_3)$ are known to be the only contributing covariates, but η_0 is misspecified to be of the parametric form $\eta_0(W) = Z^T\boldsymbol{\beta}_Z$ and $\boldsymbol{\beta}_Z$ is estimated together with the coefficients for $(X_1, \ldots, X_{15})$.

$\mathcal{M}_C$: The procedure with all the covariate effects assumed to be of linear form, and an iterated Lasso procedure is applied to all coefficients.

$\mathcal{M}_D$: The procedure ignoring nonparametric covariates (Z_1, Z_2, Z_3). That is, $\eta \equiv 0$, and one applies the iterated Lasso procedure to $\boldsymbol{\beta}$.

$\mathcal{M}_E$: The proposed iterated Lasso procedure with a nonparametric additive model assumed for η.

Procedures $\mathcal{M}_A$ and $\mathcal{M}_B$ have a partial oracle property and are unrealistic in practice. Procedure $\mathcal{M}_B$ has a misspecified covariate effect. This procedure is of interest as some studies model effects of low-dimensional covariates linearly. Procedures $\mathcal{M}_C$ and $\mathcal{M}_D$ misspecify the effects of (Z_1, Z_2, Z_3), one to be linear and the other to be zero. We intend to show that it is important to properly specify the effects of low-dimensional covariates. In procedures $\mathcal{M}_C$, $\mathcal{M}_D$ and $\mathcal{M}_E$, five-fold cross validation was used to select the λ_n that minimizes the mean sum of squared prediction errors on a common grid of $\log(\lambda_n) = -3$ to 3 by 0.1. The tensor product cubic spline basis functions described in Section 2.1, with $r_n = 5$, were used in the estimation of nonparametric effects in $\mathcal{M}_E$.

We used $n = 100, 200$ and $p = 200, 500$. In studies with microarray measurements, a large number of covariates are measured. But it is commonly accepted that only a small number of covariates are associated with the response variables. Recent studies on marginal screening show that it is possible to reduce the number of covariates to a few hundred via screening. Thus the scenario considered here for simulation is reasonable. For each combination, we simulated 500 data replicates and computed the following: the mean and standard deviation of the 500 RPEs of the complete oracle procedure $\mathcal{M}_0$ versus procedures $\mathcal{M}_A$ to $\mathcal{M}_E$, the proportion of being selected for each of the 15 nonzero coefficients, and the average number of noise variables selected. The prediction and variable selection results are summarized, respectively, in Tables 1 and 2. From Table 1, we can see that although the proposed procedure $\mathcal{M}_E$, as expected, did not predict as well as the two partial oracle procedures $\mathcal{M}_A$ and $\mathcal{M}_B$, it had significantly better prediction performance than did procedures $\mathcal{M}_C$ and $\mathcal{M}_D$ where Z-covariate effects were misspecified. For variable selection performance, only procedures $\mathcal{M}_C$ to $\mathcal{M}_E$ are relevant. Table 2 shows competitive performance of the three procedures in selecting the signal variables. This competitiveness suggests that misspecifying nonlinear effects that are independent of linear effects may not have a dramatic impact on variable selection of linear covariates. The proposed procedure slightly outperformed the others by consistently including fewer noise variables. Both prediction and variable selection clearly improved as n increased or p decreased.

To evaluate estimation of the nonparametric part, Figure 1 shows the top 10%, 50%, and 90% function estimates against their corresponding true functions. Here the estimates were ranked according to mean integrated square error (MISE) of the estimate $\hat{\eta}$ against the true nonparametric function. We can see that the proposed approach provides reasonable estimates of the nonparametric

Table 1. Prediction performance comparison by the means and standard deviations (in the brackets) of the RPEs computed from 500 replicates.

n	p	$\mathcal{M}_A$	$\mathcal{M}_B$	$\mathcal{M}_C$	$\mathcal{M}_D$	$\mathcal{M}_E$
100	200	0.576(0.117)	0.367(0.050)	0.167(0.040)	0.158(0.033)	0.206(0.057)
	500	0.536(0.117)	0.366(0.045)	0.128(0.031)	0.125(0.029)	0.167(0.052)
200	200	0.735(0.083)	0.374(0.033)	0.200(0.037)	0.181(0.026)	0.264(0.054)
	500	0.725(0.102)	0.373(0.031)	0.157(0.030)	0.147(0.022)	0.222(0.050)

Table 2. Variable selection frequencies for parametric components. Values are the average numbers of selection for signal variables and average total numbers of selected noise variables computed from 500 replicates.

	Signal Variables (with values of β_js)															
	1.0	0.5	0.9	0.7	1.0	0.7	0.9	0.5	0.6	0.7	0.6	0.9	1.0	1.0	0.6	**Noise**
	$n = 100, p = 200$															
$\mathcal{M}_C$	0.68	0.43	0.69	0.49	0.74	0.49	0.71	0.37	0.48	0.54	0.52	0.66	0.74	0.74	0.34	2.76
$\mathcal{M}_D$	0.65	0.43	0.67	0.49	0.73	0.51	0.68	0.37	0.47	0.51	0.50	0.63	0.72	0.71	0.34	3.00
$\mathcal{M}_E$	0.67	0.41	0.69	0.49	0.74	0.52	0.69	0.37	0.48	0.51	0.53	0.63	0.74	0.73	0.34	2.37
	$n = 100, p = 500$															
$\mathcal{M}_C$	0.61	0.37	0.65	0.51	0.70	0.49	0.63	0.39	0.39	0.46	0.47	0.65	0.68	0.71	0.28	3.57
$\mathcal{M}_D$	0.58	0.38	0.64	0.54	0.70	0.47	0.63	0.41	0.39	0.49	0.46	0.65	0.70	0.68	0.28	3.62
$\mathcal{M}_E$	0.61	0.37	0.64	0.54	0.72	0.48	0.64	0.36	0.40	0.47	0.47	0.64	0.71	0.71	0.28	3.12
	$n = 200, p = 200$															
$\mathcal{M}_C$	0.93	0.33	0.89	0.60	0.94	0.60	0.86	0.43	0.53	0.76	0.49	0.87	0.89	0.92	0.41	1.27
$\mathcal{M}_D$	0.90	0.36	0.83	0.62	0.91	0.57	0.83	0.44	0.50	0.71	0.48	0.84	0.87	0.89	0.39	0.31
$\mathcal{M}_E$	0.92	0.33	0.85	0.62	0.92	0.58	0.84	0.42	0.51	0.72	0.49	0.85	0.89	0.91	0.38	0.22
	$n = 200, p = 500$															
$\mathcal{M}_C$	0.86	0.36	0.83	0.60	0.88	0.57	0.86	0.38	0.57	0.63	0.50	0.80	0.85	0.86	0.36	1.25
$\mathcal{M}_D$	0.85	0.38	0.81	0.57	0.85	0.54	0.83	0.40	0.56	0.61	0.49	0.78	0.83	0.85	0.37	0.88
$\mathcal{M}_E$	0.86	0.37	0.82	0.58	0.87	0.53	0.83	0.39	0.57	0.60	0.50	0.80	0.83	0.87	0.36	0.70

covariate effects. In some plots, we see a shrinkage towards zero; this is reasonable considering the connection between $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{b}}$, and the shrinkage nature of penalized estimates.

We also conducted simulations to evaluate the model selection procedure for the nonparametric part. Z_1, Z_2, and Z_3 were independent Uniform(0,1). We considered two scenarios for the true nonparametric part: (i) a nonparametric bivariate additive model $\eta(z) = \eta_1(z_1) + \eta_2(z_2)$, or (z_1, z_2) using shorthand notation, and (ii) a nonparametric additive model with three covariates $\eta(z) = \eta_1(z_1) + \eta_2(z_2) + \eta_3(z_3)$, or (z_1, z_2, z_3). In both scenarios, the fitted models were the nonparametric additive model with all three covariates, and the ratios $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c)$ for the projections to the bivariate models (z_2, z_3), (z_1, z_3), and (z_1, z_2) were computed. We claim that a reduced model is feasible when the ratio $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c) < 0.05$. Note that each of these three reduced models

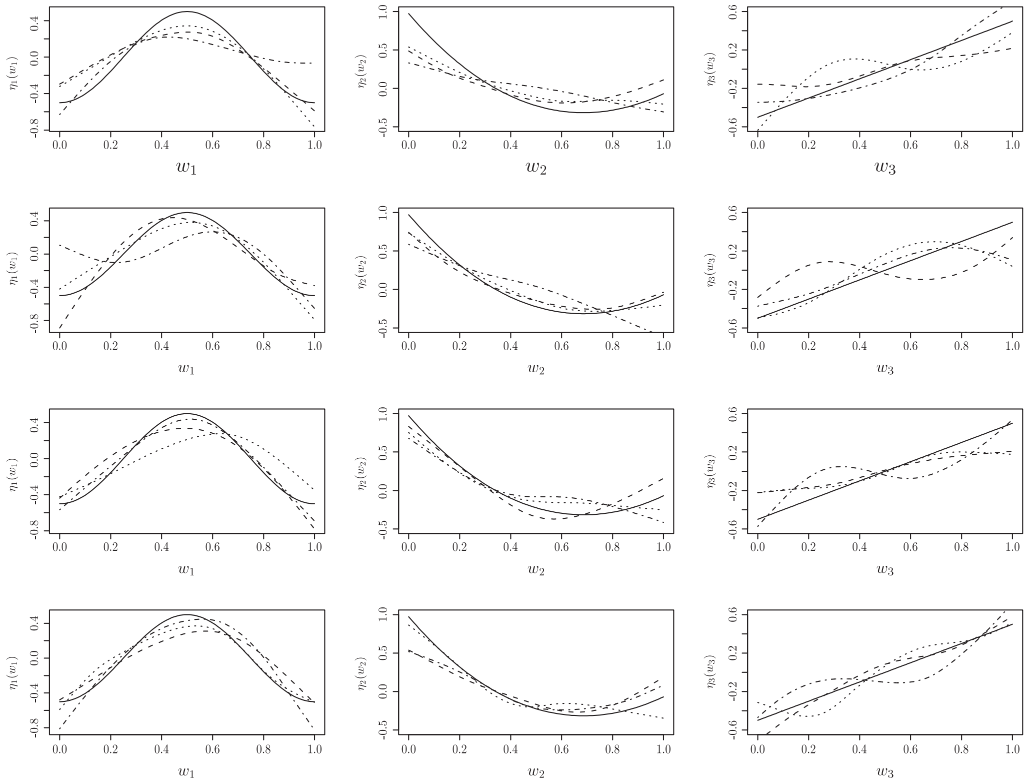

Figure 1. Estimates for nonparametric components. From top to bottom: $(n, p) = (100, 200), (100, 500), (200, 200), (200, 500)$. Solid lines are true functions, and dashed, dotted and dot-dashed lines are, respectively, the top 10%, 50%, 90% estimates ranked by MISE.

drops one covariate from the full additive model. If none of these reduced models is feasible, then the full additive model is kept as the final model. The results are summarized in Table 3. The procedure was very successful in keeping the signal variables in all the simulations. This resulted in very low percentages of under-fitted final models, defined as models missing any signal variable. On the other hand, the procedure seemed to be conservative in that it included the noise variable at times. The selection performance clearly improved as the sample size increased, but seemed to be less affected by the total number of parametric covariates in the model.

3.2. Analysis of mantle cell lymphoma data

Rosenwald et al. (2003) reported a gene expression profiling study of mantle cell lymphoma (MCL) prognosis. Among 101 untreated patients with no history of previously diagnosed lymphoma, 92 were classified as having MCL based on

Table 3. Performance of model selection for the nonparametric part (500 replicates). Under-fit means missing at least one signal Z_j, correct-fit means a match of selected Z_j's to true signal Z_j's, and over-fit in the (Z_1, Z_2) true model case means that all three Z_j's are selected.

		Proportion of Selecting			**Proportion of**		
n	p	Z_1	Z_2	Z_3	**Under-fit**	**Correct-fit**	**Over-fit**
		True model: $\eta(z) = \eta_1(z_1) + \eta_2(z_2)$					
100	200	0.996	0.992	0.712	0.012	0.284	0.704
	500	0.990	1.000	0.727	0.010	0.273	0.717
200	200	1.000	1.000	0.396	0.000	0.604	0.396
	500	1.000	1.000	0.374	0.012	0.626	0.374
		True model: $\eta(z) = \eta_1(z_1) + \eta_2(z_2) + \eta_3(z_3)$					
100	200	0.988	0.994	0.964	0.054	0.946	-
	500	0.980	0.994	0.960	0.066	0.934	-
200	200	1.000	1.000	0.996	0.004	0.996	-
	500	1.000	1.000	0.994	0.006	0.994	-

established morphologic and immunophenotypic criteria. During the followup, 64 patients died of MCL, and the other 28 patients were censored. The median survival time was 2.8 years. This dataset contains two distinct sets of covariates. The first contains five clinical covariates: BMI expression (Z_1), cyclinD-1 taqman result (Z_2), indicator of INK/ARF deletion (Z_3), indicator of ATM deletion (Z_4), and indicator of P-53 deletion (Z_5). The second set contains the expressions of 8810 genes. Lymphochip DNA microarrays were used to quantify mRNA expression in the lymphoma samples from the 92 patients. After removing 7 subjects with missing values for Z_3 to Z_5, we were left with 85 subjects.

With gene expressions, we first conducted unsupervised screening. We computed the interquartile ranges of all gene expressions and removed genes with interquartile ranges smaller than their first quartiles. Furthermore, since genes with higher variations are usually of higher interest, we selected 200 and 500 genes with the highest variations. We then rescaled gene expressions to have mean zero and variance one.

In our analysis, we started with the model with all gene expression covariate effects linear and all clinical covariate effects nonlinear. An additive model with all five clinical covariates was used for the nonparametric part. A backward application of the model selection procedure described in Section 2.3 reduced to the same additive model with Z_1 and Z_2 in the analysis with 200 and 500 genes. The clinical covariates were removed in the order Z_3, Z_4, and Z_5, with the corresponding ratios $\mathrm{KL}(\hat{\eta}, \tilde{\eta})/\mathrm{KL}(\hat{\eta}, \eta_c)$ of 0.002, 0.040, and 0.048 for the estimation with 500 genes, and 0.010, 0.017, and 0.046 for the estimation with 200 genes.

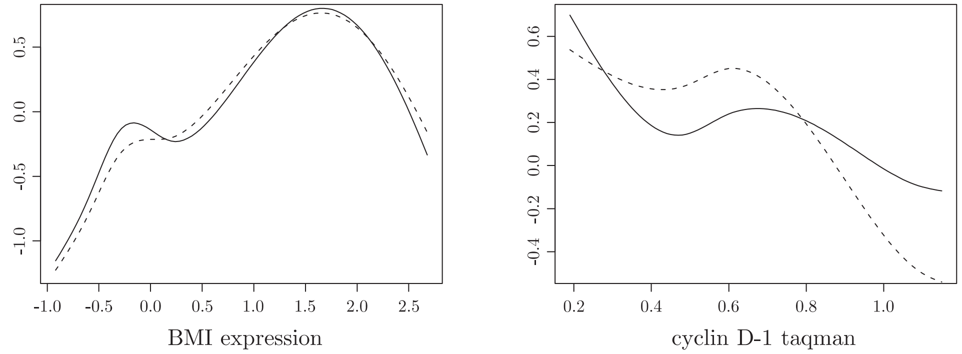

Figure 2. Estimates of nonparametric covariate effects for MCL data. Solid lines are estimates with 500 genes, and dashed lines are estimates with 200 genes.

After fitting the semiparametric models with the effects of Z_1 and Z_2 nonparametric, the analysis with 500 genes selected two genes (with estimated coefficients -0.765 and -0.022), and the analysis with 200 genes selected five genes (with estimated coefficients -0.139, 0.516, -0.312, -0.046, and -0.027). The two sets of identified genes have no overlap, not surprising considering that multiple sets of genes may have equal predictive power and the extremely noisy nature of gene expression data. The estimates of the nonparametric part are plotted in Figure 2. Overall the effect of BMI expression has a bell shape. The effect of cyclin D-1 has an overall decreasing trend, but the effect is not monotone. Previous studies, such as Rosenwald et al. (2003), have analyzed this dataset, but we may be the first to observe nonlinear trends that may provide further insights into the biological mechanisms.

The same dataset is also analyzed in Huang and Ma (2010) with the clinical covariates ignored. Comparing the two studies shows that our approach, which accommodates the nonlinear effects of clinical covariates, identifies fewer genes. This is reasonable; with the clinical covariates explaining part of the variation in response, fewer genes are needed. Compared with gene expressions, clinical covariates can be easier to measure and have more lucid interpretations. Thus, a model with fewer gene expressions may be preferred in practice.

4. Discussion

In this study, we consider variable selection for semiparametric high-dimensional censored regression model which includes the AFT model as a special case. In for example cancer prognosis studies, both the low-dimensional clinical/environmental covariates and the high-dimensional genomic covariates have been shown to have predictive power. The semiparametric model we propose provides a useful tool for analyzing such data.

We propose an iterated LASSO approach for variable selection with the parametric component. It is possible to extend the iterated approach with other types of penalties, for example the SCAD and elastic net. The LASSO penalty is preferred here because of its computational simplicity. We establish that using the LASSO initial estimate will not miss any important covariates. In practice if there is concern over missing important covariates at the first step, a tuning parameter slightly smaller than the one selected by cross validation may be used. When there exist extremely high correlations among covariates, the SRC condition may be violated and the LASSO approach may miss important covariates. We conjecture that an iterated Elastic Net procedure, extending Zou and Zhang (2009), may ameliorate the problem, but such an extension is beyond the scope of this paper.

The nonparametric component is estimated using a sieve approach. As a limitation of this study, because of the complexity introduced by censoring, we are unable to "separate" the conditions on the basis functions. Rather, they are built in the SRC condition. This condition needs to be checked on a case-by-case basis, following Zhang and Huang (2008), if different basis functions are adopted. There are many publications on choosing the basis functions. Because of the high dimensionality of X, we recommend that a small number of basis functions be used for the nonparametric part.

We also propose a Kullback-Leibler geometry-based approach for model selection in the nonparametric component. It was motivated by similar concepts in simple linear regression models. This procedure essentially resembles a hypothesis testing where both reduced and complete models belong to infinite dimensional model spaces. Theoretical investigation of such a testing problem in nonparametric function estimation is notoriously challenging. To the best of our knowledge, there is still no satisfactory solution. We leave the theoretical investigation of this procedure as an open problem.

Acknowledgement

We would like to thank the Editor, an associate editor, and two referees for careful review and insightful comments. Ma's research has been supported by grants DMS 0904181 from NSF and CA142774 from NIH, and Du's research has been supported by grant DMS 1007126 from NSF.

References

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

Cai, J., Fan, J., Li, R. and Zhou, H. (2005). Variable selection for multivariate failure time data. *Biometrika* **92**, 303-316.

Du, P., Ma, S. and Liang, H. (2010). Penalized variable selection procedure for Cox models with semiparametric relative risk. *Ann. Statist.* **38**, 2092-2117.

Fan, J., Feng, Y. and Wu, Y. (2010). High-dimensional variable selection for Cox's proportional hazards model. In *Borrowing Strengh: Theory Powering Applications – A Festschrift for Lawrence D. Brown* (Edited by J. O. Berger, T. T. Cai, and I. M. Johnstone), 70-86. Institute of Mathematical Statistics.

Gu, C. (2002). *Smoothing Spline ANOVA Models.* Springer, New York.

Gu, C. (2004). Model diagnostics for smoothing spline ANOVA models. *Canadian J. Statist.* **32**, 347-358.

Huang, J. and Ma, S. (2010). Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Anal.* **16**, 176-195.

Huang, J., Ma, S. and Xie, H. (2006). Regularized estimation in the accelerated failure time model with high dimensional covariates. *Biometrics* **62**, 813-820.

Huang, J., Ma, S. and Xie, H. (2007). Least absolute deviations estimation for the accelerated failure time model. *Statist. Sinica* **17**, 1533-1548.

Huang, J., Ma, S. and Zhang, C. (2008). Adaptive LASSO for high dimensional regression models. *Statist. Sinica* **18**, 1603-1618.

Johnson, B. A. (2008). Variable selection in semiparametric linear regression with censored data. *J. Roy. Statist. Soc. Ser. B* **70**, 351-370.

Johnson, B. A. (2009). Rank-based estimation in the l_1-regularized partly linear model for censored outcomes with application to integrated analyses of clinical predictors and gene expression data. *Biostatistics* **10**, 659-666.

Long, Q., Chung, M., Moreno, C. and Johnson, B. A. (2011). Risk prediction for cancer recurrence through regularized estimation with simultaneous adjustment for nonlinear clinical effects. *Ann. Appl. Stat.* In press.

Ma, S. and Huang, J. (2007). Combining clinical and genomic covariates via Cov-TGDR. *Cancer Informatics* **3**, 381-388.

Meier, L. and Buhlmann, P. (2008). Discussion of "one-step sparse estimates in non-concave penalized likelihood models" by Zou and Li. *Ann. Statist.* **36**, 1534-1541.

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

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

Reid, N. (1994). A conversation with Sir David Cox. *Statist. Sci.* **9**, 439-455.

Rosenwald, A., Wright, G., Wiestner, A., Chan, W. et al. (2003). The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell* **3**, 185-197.

Schumaker, L. (1981). *Spline Functions: Basic Theory.* Wiley, New York.

Stute, W. (1993). Consistent estimation under random censorship when covariables are present. *J. Multivariate Anal.* **45**, 89-103.

Tibshirani, R. (1997). The LASSO method for variable selection in the Cox model. *Statist. in Med.* **16**, 385-395.

Tsiatis, A. (1990). Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.* **18**, 354-372.

Wei, L., Ying, Z. and Lin, D. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika* **77**, 845-851.

Wu, T. and Lange, K. (2007). Coordinate descent procedures for LASSO penalized regression. *Ann. Appl. Statist.* **2**, 224-244.

Xie, H. and Huang, J. (2009). Scad-penalized regression in high-dimensional partially linear models. *Ann. Statist.* **37**, 673-696.

Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *Ann. Statist.* **21**, 76-99.

Zhang, C. H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high dimensional linear regression. *Ann. Statist.* **36**, 1567-1594.

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

Zhang, H. H., Lu, W. and Wang, H. (2010). On sparse estimation for semiparametric linear transformation models. *J. Multivariate Anal.* **101**, 1594-1606.

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

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

Zou, H. and Zhang, H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.* **37**, 201-221.

School Public Health, Yale University, New Haven, CT 06520, U.S.A.

E-mail: shuangge.ma@yale.edu

Department of Statistics, Virginia Tech, Blacksburg, VA 24061, U.S.A.

E-mail: pangdu@vt.edu

(Received November 2010; accepted May 2011)