Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2012, Article ID 478680, 14 pages
doi:10.1155/2012/478680

Review Article

High-Dimensional Cox Regression Analysis in Genetic Studies with Censored Survival Outcomes

Jinfeng Xu

Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546

Correspondence should be addressed to Jinfeng Xu, staxj@nus.edu.sg

Received 22 February 2012; Revised 21 May 2012; Accepted 26 May 2012

Academic Editor: Yongzhao Shao

With the advancement of high-throughput technologies, nowadays high-dimensional genomic and proteomic data are easy to obtain and have become ever increasingly important in unveiling the complex etiology of many diseases. While relating a large number of factors to a survival outcome through the Cox relative risk model, various techniques have been proposed in the literature. We review some recently developed methods for such analysis. For high-dimensional variable selection in the Cox model with parametric relative risk, we consider the univariate shrinkage method (US) using the lasso penalty and the penalized partial likelihood method using the folded penalties (PPL). The penalization methods are not restricted to the finite-dimensional case. For the high-dimensional ($p \to \infty$, $p \ll n$) or ultrahigh-dimensional case ($n \to \infty$, $n \ll p$), both the sure independence screening (SIS) method and the extended Bayesian information criterion (EBIC) can be further incorporated into the penalization methods for variable selection. We also consider the penalization method for the Cox model with semiparametric relative risk, and the modified partial least squares method for the Cox model. The comparison of different methods is discussed and numerical examples are provided for the illustration. Finally, areas of further research are presented.

1. Introduction

The modern high-throughput technologies offer the possibility of a powerful, genome-wide search for the genetic and environmental factors that have influential effects on diseases. The identification of such factors and the discernment of such a relationship can lead to better understanding of the causation of diseases and better predictive models. In the presence of a large number of covariates, it is very challenging to build a model which fully utilize all the information and excels in both parsimony and prediction accuracy. In classical settings where the number of covariates p is fixed and the sample size n is large, subset selection coupled with model selection criteria such as Akaike's information criterion (AIC) and

Bayesian information criterion (BIC) can be used to identify relevant variables or choose the best model with the optimal prediction accuracy. However, subset selection is inherently unstable because of its discreteness [1]. To overcome this drawback of subject selection, Tibshirani [2] proposed the least absolute shrinkage and selection operator (LASSO) for simultaneous coefficient estimation and variable selection. Fan and Li [3] further proposed the penalization method with the smoothly-clipped absolute deviation (SCAD) penalty and rigorously established its oracle properties. The optimal properties of the lasso or SCAD-based penalization methods are not restricted to the finite-dimensional case. In the high-dimensional case ($p \to \infty,\ p \ll n$), Fan and Peng [4] proved that the oracle properties are well retained. In the ultra high-dimensional case ($n \to \infty,\ n \ll p$), Fan and Lv [5] proposed the sure independence screening method (SIS) which first reduces dimensionality from high to a moderate scale that is below the sample size and then apply a penalization method. In a general asymptotic framework, the sure independence screening method is shown to fare well for even exponentially growing dimensionality. In high-dimensional or ultra high-dimensional situations, J. Chen and Z. Chen [6] proposed the extended Bayesian information criterion (EBIC) and established its selection consistency under mild conditions. The EBIC is further extended to the generalized linear model [7].

When the clinical outcome involves time to an event such as age at disease onset or time to cancer recurrence, the regression analysis is often conducted by the Cox relative risk model. The classical Cox model is only applicable to the situation where the number of subjects is much larger than the number of covariates. Thus, to accommodate the large p and small n scenario, some variable selection and dimension reduction techniques have to be implemented in a regression analysis. Recently, for variable selection in the Cox model, a number of approaches based on the efficient shrinkage method have been proposed and gained increased popularity. See, for example, LASSO [8], SCAD [9], and adaptive lasso [10, 11].

For high-dimensional variable selection in the Cox model with parametric relative risk, we review the univariate shrinkage method (US) [12] and the penalized partial likelihood approach [13]. The univariate shrinkage method [12] assumes the independence of the covariates in each risk set and the partial likelihood factors into a product. This leads to an attractive procedure which is univariate in its operation and most suitable for a high-dimensional variable selection setting. The variables are entered into the model based on the size of their Cox score statistics, and in nature the method is similar to univariate thresholding in linear regression and nearest shrunken centroids in classification. The univariate shrinakge method is applicable to the setting with an arbitrary number of variables but is less informative in identifying joint effects from multiple variables. The penalized partial likelihood approach [13] employs a class of folded-concave penalties to the Cox parametric relative risk model and strong oracle properties of non-concave penalized methods are established for nonpolynomial (NP) dimensional data. A coordinate-wise algorithm is used for finding the grid of solution paths. The penalized partial likelihood approach investigates joint effects from multiple variables and is applicable to both the finite-dimensional and high-dimensional cases. For the ultra high-dimensional case, some preliminary procedures such as the sure independence screening (SIS) method and the extended Bayesian information criterion (EBIC) can be used to reduce the number of variables to be moderately below the sample size before the penalized partial likelihood approach is formally adopted.

The aforementioned two methods both adopt the Cox parametric relative risk model for the covariance analysis. In practice, the parametric form of the relative risk model is quite

restrictive and may not be tenable. In Section 3, we review a penalization method in the Cox model with semiparametric relative risk approach [14]. The relative risk is assumed to partially linear with one parametric component and one nonparametric component. Two penalties are applied sequentially to simultaneously estimate the parameters and select variables for both the parametric and the nonparametric parts. The semiparametric relative risk model greatly relaxes the restrictive assumption of the classical Cox model and facilitate its use in exploratory data analysis. Although the method is proposed for the finite-dimensional setting, it is straightforward to be extended to the high dimensional and ultra high-dimensional situations the same as the penalization method for the Cox model with parametric relative risk.

In Section 4, we review a modified partial least squares method for dimension reduction in the Cox regression approach [15] which provides another alternative approach to dealing with the problem of high-dimensionality. By mimicking the partial least squares in the linear model, it first constructs the components which are linear combinations of original covariates. By sequentially determining the components and using the cross-validation to select the number of components, a parsimonious model with good predictive accuracy can be obtained.

In Section 5, we discuss the comparison of different methods and numerical examples are provided for the illustration. Finally, several important problems for future research are also presented in Section 6.

2. The Penalization Methods for the Cox Model with Parametric Relative Risk

2.1. The Cox Model with Parametric Relative Risk

We consider the setting where the time to event is subject to right censoring and the observations consist of $\{Y_i = T_i \wedge C_i,\ \delta_i = I(T_i \leq C_i),\ Z_i,\ i = 1, \ldots, n\}$, where T_i is the survival time, C_i the censoring time, and Z_i is the p-dimensional vector of covariates. The Cox relative risk model assumes that the conditional hazard function of T given the covariates $Z = z$ takes the following form:

$$\lambda(t \mid Z = z) = \lambda_0(t) \exp\left(\beta_0^T Z\right), \tag{2.1}$$

where $\lambda_0(t)$ is the unknown baseline hazard function and β_0 is the unknown vector of coefficients. The influential effects that the covariates might have on the time T_i are examined by the relative risk. The unknown coefficient vector β_0 is estimated by maximizing the partial likelihood function

$$L(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta^T Z_i)}{\sum_{j \in R_i} \exp(\beta^T Z_j)} \right\}^{\delta_i}, \tag{2.2}$$

or equivalently, the log partial likelihood function

$$\ell(\beta) = \sum_{i=1}^{n} \delta_i \left\{ \beta^T Z_i - \log \left[\sum_{j \in R_i} \exp\left(\beta^T Z_j\right) \right] \right\}, \tag{2.3}$$

where $R_i = \{j : Y_j \geq Y_i\}$. As in the least squares estimation, the estimation of β from the partial likelihood function requires the sample size n is much larger than the dimension of the covariate vector p. In practice, a marginal approach is often adopted which includes one covariate at a time and maximizes

$$L_k(\beta_k) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta_k Z_{ik})}{\sum_{j \in R_i} \exp(\beta_k Z_{jk})} \right\}^{\delta_i}, \tag{2.4}$$

or

$$\ell_k(\beta_k) = \sum_{i=1}^{n} \delta_i \left\{ \beta_k Z_{ik} - \log \left[\sum_{j \in R_i} \exp(\beta_k Z_{jk}) \right] \right\}, \tag{2.5}$$

for $k = 1, \ldots, p$.

2.2. The Univariate Shrinkage Method

To identify the variables which are associated with T, multiple testing procedures will be used to make valid statistical inferences. However, Tibshirani [12] looks at the problem from another perspective. Since the maximizer of the partial likelihood is not unique when $n \ll p$, he proposes the regularized partial likelihood approach by using the lasso penalty as follows:

$$J(\beta) = \ell(\beta) - \lambda \sum_{k=1}^{p} |\beta_k|. \tag{2.6}$$

By assuming that both conditionally on each risk set, and marginally, the covariates are independent of one another, and using Bayes's theorem, Tibshirani [12] shows that the log partial likelihood function

$$\ell(\beta) = \text{constant} + \sum_{k=1}^{p} \ell_k(\beta_k). \tag{2.7}$$

The regularized partial likelihood function is

$$J(\beta) = \text{constant} + \sum_{k=1}^{p} \ell_k(\beta_k) - \lambda \sum_{k=1}^{p} |\beta_k|, \tag{2.8}$$

Table 1: Simulation results for variable selection in the Cox model with parametric relative risk.

p	Method	MSE	FDR	PSR	MMS
250	US	1.62	0.06	0.58	3.97
	PS	0.92	0.16	0.76	6.55
	EBIC	0.96	0.10	0.69	4.70
500	US	1.71	0.07	0.55	3.95
	PS	1.01	0.18	0.74	6.63
	EBIC	1.06	0.11	0.66	4.73
1000	US	1.84	0.08	0.52	3.91
	PS	1.12	0.20	0.70	6.69
	EBIC	1.15	0.13	0.63	4.78

Table 2: Results for the microarray lung cancer dataset.

Method	Number of selected genes	Median P value ($\times 10^{-4}$)
US	5	10.06
PS	13	0.064
EBIC	8	0.082

and results in the the Cox univariate shrinkage (CUS) estimator which maximizes the penalized function. Since the maximization is a set of one-dimensional maximization $\ell_k(\beta_k) - \lambda|\beta_k|,\ k = 1,\ldots,p$, for a range of λ, we can fairly easily get the penalized estimates $\widehat{\beta}_k$. Actually, the entire paths of the regularization estimates can be obtained. It can also be shown that

$$\widehat{\beta}_k \neq 0 \Longleftrightarrow \frac{|U_k|}{\sqrt{V_k}} > \lambda, \tag{2.9}$$

where U_k and V_k are the gradient of the (unpenalized) log-partial likelihood and the (negative) observed Fisher information. This is similar to soft/hard thresholding. Hence, the Cox univariate shrinkage method ranks all the covariates based on the Cox score statistic. As the Cox score is often used for determining the univariate significance of covariates, the results have easy interpretation. The tuning parameter λ can be selected by cross-validation as in Verweij and van Houwelingen [16] or directly determined as in Donoho and Johnstone [17]. The Cox univariate shrinkage method presents a numerically convenient approach for high-dimensional variable selection in the Cox model. In the literature, the modified shooting algorithm [10] and the least squares approximation based algorithm [11] both yield the entire solution paths, but only when n is much larger than p.

One drawback of the Cox univariate shrinkage procedure is that the variables enter into the model based on their univariate Cox scores. Thus, when two predictors are both strongly predictive and highly correlated with each other, both will appear in the model. In that case, it may be more desirable to just include one of them for parsimony. This can be done using preconditioning [18] as is demonstrated by Tibshirani [12].

2.3. The Penalized Partial Likelihood Method

The penalized partial likelihood estimation with noncave penalties has been extensively studied by Fan and Li [9] for the case where the sample size n is much larger than the dimension of Z Bradic et al. [13] considered the folded penalties for the penalized partial likelihood estimation when the dimension of Z is nonpolynomial (NP). The folded penalties include the smoothly clipped absolute deviation (SCAD) and the minimax concave penalty (MCP) as special cases. The penalized log partial likelihood becomes

$$\ell(\beta) - \lambda_n \sum_{k=1}^{p} p_{\lambda_n}(|\beta_k|), \tag{2.10}$$

where $p_{\lambda_n}(\cdot)$ is a penalty function and λ_n is a nonnegative tuning parameter. For a class of folded penalties, by clarifying the identification problem of the penalized partial likelihood estimates and deriving a large deviation result for divergence of a martingale from its compensator, Bradic et al. [13] establish the strong oracle properties for the penalized estimates. Note that their results also hold for the lasso penalty. The strong oracle property indicates that as both n and p goes to ∞, with probability tending to 1, the penalized estimator behaves as if the true relevant variables in the model were known. This is different from the classical notion of oracle which just requires that the estimator behaves like the oracle rather than an actual oracle itself. The strong oracle property implies the classical oracle property of Fan and Li [9] and sign consistency of Bickel et al. [19]. This tighter notion of an oracle property was first mentioned in Kim et al. [20] for the SCAD estimator of the linear model with polynomial dimensionality and then extended by Bradic et al. [21] to the penalized M-estimators under the ultrahigh dimensionality setting. Bradic et al. [13] further extended it to the Cox model by employing sophisticated techniques dealing with martingale and censoring structures.

Analogous to the Cox univariate shrinkage method, the penalized Cox relative risk method [13] proposes a coordinate wise algorithm which is especially attractive for the situation of $p \gg n$ and have been previously studied for linear and generalized linear models [5, 22, 23]. Since the coordinate-wise maximization algorithm in each iteration provides limits that are stationary points of the overall optimization, each output of the iterative coordinate ascent algorithm (ICA), Bradic et al. [13] propose gives a stationary point.

For each iteration, sequentially for $k = 1, \ldots, p$, by the partial quadratic approximation of $\ell(\beta)$ at the current estimate along the k-th coordinate while fixing the other coordinates, the k-th coordinate of the estimate is updated by maximizing the univariate penalized likelihood. Due to the univariate nature, the problem can be solved analytically, avoiding the challenges of nonconcave optimization. It updates each coordinate if the maximizer of the penalized univariate optimization increases the penalized objective function as well. The algorithm stops when two values of the penalized objective function are not different by more than a small threshold value.

Although both the iterative coordinate ascent algorithm (ICA) and the univariate shrinkage method exploit the convenient of univariate optimization, the univariate shrinkage method separates the coefficient estimates while in the iterative coordinate ascent algorithm, the coefficient estimates, are still related to one another in the iterative updating. These two methods also require different conditions. The univariate shrinkage method assumes that both conditionally on each risk set, and marginally, the covariates are independent of

one another while the penalized Cox relative risk method assumes the conditions on the folded-concave penalties, the sparsity level, the dimensionality of the covariate vector, and the magnitude of the tuning parameter λ_n.

2.4. The Penalized Partial Likelihood Approach for the Ultra High-Dimensional Case

While the univariate shrinkage method is applicable to an arbitrary dimensionality, the penalized partial likelihood requires that the sample size is larger than the number of variables. Thus, to apply the penalized partial likelihood approach to the ultra high-dimensional case, a preliminary screening procedure is needed. Fan and Lv [5] proposed the sure independence screening procedure which first shrinks the full model $1, \ldots, p$ straightforwardly and accurately down to a submodel with size $d = o(n)$. Thus, the original problem of estimating the sparse p-vector β reduces to estimating a sparse d-vector that is based on the now much smaller submodel. The penalized partial likelihood method in Section 2.3 can then be applied to the submodel. Fan and Lv [5] proved the sure independence screening method has optimal theoretical properties for even exponentially growing dimensionality.

For small n large P problems, the traditional model selection criteria such as AIC, BIC, and cross-validation choose too many features. To overcome the difficulties, J. Chen and Z. Chen [6] developed a family of extended Bayes' information criteria (EBIC). The EBIC is shown to be consistent with nice finite sample properties in both the linear model [6] and the generalized linear model [7]. For any subset model $s \subset \{1, 2, \ldots, p\}$, denote its size by $\nu(s)$. Let $\widehat{\beta}(s)$ be the maximum partial likelihood estimate corresponding to the subset model s. The extended Bayesian information criterion is defined as

$$-2\ell\left(\widehat{\beta}(s)\right) + \nu(s) \log n + 2\nu(s)\gamma \log p, \tag{2.11}$$

where γ is a prespecified constant and can be chosen to be 0.5 as suggested by J. Chen and Z. Chen [7]. Optimal theoretical properties such as selection consistency of the EBIC have been rigorously obtained by J. Chen and Z. Chen [6] for the linear model and by J. Chen and Z. Chen [7] for the generalized linear model. The EBIC can be appealingly applied to the Cox model and it is worthwhile to further investigate its theoretical properties in the Cox model which has not yet been addressed in the literature.

3. The Penalization Method for the Cox Model with Semiparametric Relative Risk

The Cox relative risk model is sometimes too restrictive in examining the covariate effects. It seems implausible that the linearity assumption holds in the presence of a large number of predictors. Intuitively, at least for some of them, the linearity assumption might be violated and the modeling of covariate effects via the parametric relative risk model might lead to erroneous results. On the other hand, there are two objectives in the high-dimensional regression analysis of genetic studies with censored survival outcomes, we not only want to identify the predictor variables which are associated with the time but also to discern

such a relationship if there does exist an association. Therefore, it is worth looking at other alternative survival models in examining the covariate effects.

Du et al. [14] proposed the penalized method for the Cox model with semiparametric relative risk model. Let $Z^T = (U^T, W^T)$, where U and W are the subvectors of Z with dimensions $d = p - q$ and q, respectively. Instead of (2.1), they assume that

$$\lambda(t \mid Z = z) = \lambda_0(t) \exp\left[\beta_0^T U + \eta(W)\right], \tag{3.1}$$

where $\eta(w) = \eta(w_1, \ldots, w_q)$ is an unknown multivariate smooth function. The model assumes the additivity of the effects of U and W and only the effect of U is postulated to be linear. The effect of W can be of any form. This greatly enhances the flexibility and facilitates more robust investigation of the covariate effects across a large number of genetic and environment factors. Similarly, the log partial likelihood is

$$\ell(\beta, \eta) = \sum_{i=1}^{n} \delta_i \left\{ \beta^T U_i + \eta(W_i) - \log\left[\sum_{j \in R_i} \exp\left(\beta^T U_j + \eta(W_j)\right) \right] \right\}. \tag{3.2}$$

Du et al. [14] proposed two penalties for the model (3.1), one penalty for the roughness of the function η and the other penalty for simultaneous coefficient estimation and variable selection. The estimation iterates between the estimation of η given an initial estimator of β and the estimation of β given an initial estimator of η. Given an estimate $\widehat{\beta}$ of β, η is estimated by maximizing

$$\ell\left(\widehat{\beta}, \eta\right) - \lambda J(\eta), \tag{3.3}$$

where J is a roughness penalty specifying the smoothness of η, and $\lambda > 0$ is a smoothing parameter controlling the tradeoff. A popular choice for J is the L_2-penalty which yields tensor product cubic splines for multivariate W. Given an estimate $\widehat{\eta}$ of η, β can be estimated by

$$\ell(\beta, \widehat{\eta}) - \sum_{k=1}^{d} p_{\theta_n}(|\beta_k|), \tag{3.4}$$

where $p_{\theta_n}(\cdot)$ is the SCAD penalty function and θ_n is the tuning parameter. In its numerical implementation, the SCAD penalty is approximated by a one-step approximation which transforms the SCAD penalty problem to a LASSO-type optimization, where the celebrated LARS algorithm [24] can be readily used to yield the entire solution path.

The algorithm converges quickly within a few iterations. In this approach, the SCAD penalty facilitates the simultaneous coefficient estimation and variable selection in the parametric component of relative risk model. As the multivariate smooth function W also involves multiple predictor variables, it is therefore necessary to identify the correct structure of η and relevant variables in W too. Taking care of variable selection for the parametric components, we still need an approach to assess the structure of the nonparametric components. By transforming the profile partial likelihood to a density estimation problem

with biased sampling, Du et al. [14] further derive a model selection tool based on the Kullback-Leibler geometry for the nonparametric component η. Specifically, a quantity based on the ratio of two Kullback-Leibler distances can be used to diagnose the feasibility of a reduced model η, the smaller the ratio is, the more feasible the reduced model is. Thus, the penalized Cox semiparametric relative risk approach provides a flexible tool for identifying relevant variables in both the parametric and nonparametric components.

4. The Modified Partial Least Squares Method for Dimension Reduction in the Cox Model

Partial least squares (PLS) [25] is a classical dimension reduction method of dealing with a large number of covariates. By constructing new variables which are linear combination of the original variables, it fully utilizes the information and a proper regression analysis can be conducted using the new variables. Different from the principal components (PCs) analysis, partial least squares utilizes the information contained in both the response variable and the predictor variables to construct new variables. This complicates its direct application to censored survival data since the response variable is subject to right censoring. Nguyen and Rocke [26] applied the standard PLS methods of Wold [25] directly to survival data and used the resulted PLS components in the Cox model for predicting survival time. Since the approach did not take into account that some of the survival time are censored and not exactly the underlying time to event, the resulting components are questionable and may induce bias. Alternatively, by reformulating the Cox model into a generalized linear model, Park et al. [27] applied the formulation of PLS of Marx [28] to derive the PLS components. Despite its validity, the introduction of many additional nuisance parameters in the reformulation makes the algorithm fail to converge when the number of covariates is large.

Li and Jiang [15] proposed a modified partial least squares method for the Cox model by constructing the components based on repeated least square fitting of residuals and Cox regression fitting. Let $w_{ij} \propto \mathrm{var}(V_{ij})$ be the weights and $\sum_{i=1}^{n} w_{ij} = 1$. First, let $X_j = (Z_{1j}, \ldots, Z_{nj})^T$ and define

$$V_{1j} = X_j - z_{\cdot j}\mathbf{1}, \tag{4.1}$$

where $z_{\cdot j} = (1/n)\sum_{i=1}^{n} Z_{ij}$, and $\mathbf{1}$ is an n-dimensional vector of all elements 1. After fitting the Cox model with one covariate at one time, we obtain the maximize partial likelihood estimate $\widehat{\beta}_{1j}$ for the predictor variable V_{1j}, $j = 1, \ldots, p$. Combining these estimates, we get the first component

$$T_1 = \sum_{j=1}^{p} w_{1j}\widehat{\beta}_{1j}V_{1j}. \tag{4.2}$$

The information in X that is not in T_1 can be written as the residuals of regressing $V_{1,j}$ on T_1

$$V_{2,j} = V_{1,j} - \frac{T_1^T V_{1j}}{T_1^T T_1}T_1. \tag{4.3}$$

By performing the Cox regression analysis with T_1 and V_{1j} (one j at a time), we obtain the maximized partial likelihood estimates $\widehat{\beta}_{2j}$ and consequently get the second component

$$T_2 = \sum_{j=1}^{p} w_{2j}\widehat{\beta}_{2j}V_{2j}. \tag{4.4}$$

This procedure extends iteratively in a natural way to give component $T_2, \ldots, T_K$, where the maximum value of K is the sample size n. Specifically, suppose that T_i has just been constructed, and to construct T_{i+1}, we first regress V_{ij} against T_i and denote the residual as $V_{(i+1),j}$, which can be written as

$$V_{(i+1),j} = V_{i,j} - \frac{T_i^T V_{ij}}{T_i^T T_i} T_i. \tag{4.5}$$

Then we fit the Cox relative risk model

$$\lambda(t \mid Z) = \lambda_0(t)\exp\left(\beta_1 T_1 + \cdots + \beta_i T_i + \beta_{(i+1),j}V_{(i+1),j}\right), \tag{4.6}$$

and obtain the maximum partial likelihood estimates $\widehat{\beta}_{(i+1),j}$ and

$$T_{i+1} = \sum_{j=1}^{p} w_{(i+1)j}\widehat{\beta}_{(i+1)j}V_{(i+1)j}. \tag{4.7}$$

With the components $T_1, \ldots, T_K$, a standard Cox regression model can be fitted and the risk score can be obtained as

$$\widehat{\beta}_1 T_1 + \cdots + \widehat{\beta}_K T_K, \tag{4.8}$$

where $\widehat{\beta}_j,\ j = 1, \ldots, K$ is the maximum partial likelihood estimate of β_j when we fit the Cox relative risk model

$$\lambda(t \mid Z) = \lambda_0(t)\exp\left(\beta_1 T_1 + \cdots + \beta_K T_K\right). \tag{4.9}$$

This can then be used for estimating the hazard function for future samples on the basis of their X values. By examining the coefficients of X values in the final model with K components, one can rank the covariate effects by the risk score. The number of K can be chosen by applying the cross-validation.

5. Comparison of Different Methods and Numerical Examples

5.1. Comparison Using Survival Prediction

In the previous sections, four different approaches have been used to identify the relevant factors which have influential effects on the survival time. In practice, it would be important

and interesting to compare different methods which can be done by using certain measure of survival prediction. To assess the performance of the methods, the data set is first divided into the training sample and the test sample randomly. For example, a v-fold cross-validation divide the sample into v parts randomly. One part is retained as test set while the rest $v-1$ folds are used as the training set. The training sample gives the estimated risk score for a given model (method) and then used in test sample for prediction. There are many measures for survival prediction. One of them mimics the random clinical trial in assigning the test sample into two groups—one "good" group and one "bad" group. Whether a subject in the test sample falls into a good group or a bad group depends on whether his/her risk score is smaller than a threshold value for the risk score. The log-rank test can then be used to test the hypothesis that there is no difference between the two groups. The smaller the P value the resulting log-rank test has, the better predictive power the estimated risk score has, which translates into the better performance of the method/model. The dataset is randomly split into training and test sample and hence for a large number of replications, the comparison of different methods can be made by looking at the summary of the P-values of the log-rank test, say, the median.

The disadvantage of the log-rank test is that the subjects are only assigned to two groups and the risk score is only utilized in comparing with a threshold value. The information contained in the risk score which is continuous is not fully utilized for survival prediction. Alternatively, we can fit a Cox regression for the test sample using the risk score estimated from the training sample as a single covariate. The predictive power of the estimated risk score can be indicated by the significance of the risk score covariate in the fitted Cox regression model for the test sample. Again, the obtained P values using different methods in a large number of replications can help us assess their performance in terms of survival prediction.

5.2. Simulation Studies

We conduct simulation studies to compare different methods. As a simple illustration, we focus on the univariate shrinkage method (US) and the penalized shrinkage method (PS) reviewed in Section 2. We set the sample size $n = 500$ and the number of covariates $p = 250, 500$, and 1000, respectively. The covariates are jointly normally distributed with equal correlation coefficient $\rho = 0.5$. The first six covariates are the only relevant variables with $\beta_1 = \beta_3 = \beta_5 = 1$ and $\beta_2 = \beta_4 = \beta_6 = -1$. The baseline hazard function in (2.1) is set to be constant 1 and the censoring time is generated from the $niform(0, \tau)$, where τ is chosen to yield the censoring proportion 30%. For the univariate shrinkage method, the top ranked variables with significance at 0.05 after Bonferroni's correction will be selected. For the penalized shrinkage method, the sure independence screening procedure preselects $n/(4 \log n) = 20$ and the penalized partial likelihood method is then applied to obtain the final model. As a third method, we directly use the EBIC to select a subset model. We report the median squared estimation error (MSE) and the squared estimation error is defined as

$$\sum_{j=1}^{p} \left| \widehat{\beta}_j - \beta_j \right|^2. \tag{5.1}$$

We also report the average number of selected variables (MMS), the average positive selection, and false discovery rates (PSR and FDR), where

$$\begin{aligned} \text{PSR} &= \frac{\sum_{j=1}^{N} \nu\left(s_j^* \bigcap s_0\right)}{N\nu(s_0)}, \\ \text{FDR} &= \frac{\sum_{j=1}^{N} \nu\left(s_j^* / s_0\right)}{\sum_{j=1}^{N} \nu\left(s_j^*\right)}, \end{aligned} \tag{5.2}$$

$N = 200$ is the number of replications, s_0 denotes the true model, and s_j^* denotes the selected model in the jth replication. The simulation results are summarized in Table 1. From Table 1, we can see that both the PS and the EBIC perform better than the US. Compared with the EBIC, the PS selects slightly more variables and has relatively larger FDRs and PSRs.

5.3. A Real Example

We analyzed microarray data by the lung cancer dataset from Beer et al. [29]. The dataset consists of gene expressions of 4966 genes for 83 patients. The patients were classified according to the progression of the disease. Sixty four patients were classified as stage I. Nineteen patients were classified as stage III. For each of the 83 patients, the survival time as well as the censoring status is available. Other covariate variables in addition to the gene expressions are age, gender, and smoking status. Our aim is to study the association of survival time with the gene expressions adjusting for the effects of the other covariates via the Cox model with parametric relative risk. The US, PS, and EBIC are used to select variables. We divide the 83 patients into two groups by randomly assigning 32 of the 64 stage I and 9 of 19 stage III patients to the training group and the remaining patients to the test group. By adjusting for the covariate (gender, age, smoking) effects, we fit the Cox model with the selected genes and construct a risk index. The 50th percentile of the risk index from the training group is employed as the threshold. We then apply the threshold to test dataset to define the low-risk and high-risk groups. To assess the predictability of the so-defined discriminant criterion, we perform a log-rank test of the difference of survivals of the two groups defined by the risk index. If the survival times of the two groups can be well separated (measured by the P-value of the log-rank test), then the method has a better predictability. We therefore use the resulting median P-value (among 1000 random splitting data into training and test sets) as the measure of prediction accuracy of different methods. The results are summarized in Table 2. It is shown that the PS and EBIC have comparable predictability which is much better than the US.

6. Further Work

We review in this paper some recently developed methods for high-dimensional regression analysis in genetic studies with censored survival outcomes. The identification of relevant variable that have the influential effects on the survival time leads to a better understanding of disease and gene/environment association for many complex diseases. Although the Cox model is widely used to examine the covariate effects through the relative risk, the

proportional hazard assumption may be violated in practice, for example, when there are long-term survivors. In some situations, other alternative models such as the additive risks model, the proportional odds model, or more generally the semiparametric transformation models may fare better. Furthermore, as we discussed before, the linearity assumption may not be tenable either. It would be interesting to develop parallel methodologies in these alternative models.

Although the Cox semiparametric relative risk relaxes the assumption to some extent, the classification of the covariates into the parametric component (with linearity assumption) and the nonparametric component (without linearity assumption) is challenging and unsolved. The problem would be more difficult when both the proportional hazards assumption and the linearity assumption are violated. In the presence of a large number of genetic and environment factors, undoubtedly we have to make necessary assumptions on the underlying structure to proceed. It is worth investigating that how the nonproportionality and the linearity assumptions alone or jointly with each other impact on the high-dimensional regression analysis. In particular, how sensitive the identification of relevant variables is to the misspecification of the model and whether there are other good structures to be postulated which have appealing properties and are most suitable for the high-dimensional regression analysis.

It is also worthy to note that for model selection, there are two different purposes. One is selection consistency such as the oracle properties. The other is the prediction accuracy. While the prediction accuracy can be well assessed by cross-validation, the selection consistency should be assessed by using the FDRs and PSRs.

Acknowledgments

The authors are very grateful to Professor Yongzhao Shao and three anonymous references for many helpful comments which improved the presentation of the paper. This work was supported by the grant from National University of Singapore (R-155-000-112-112).

References

[1] L. Breiman, "Heuristics of instability and stabilization in model selection," *Annals of Statistics*, vol. 24, pp. 2350–2383, 1996.
[2] R. Tibshirani, "Regression shrinkage and selection via the lass," *Journal of the Royal Statistical Society B*, vol. 58, pp. 267–288, 1996.
[3] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.
[4] J. Fan and H. Peng, "Nonconcave penalized likelihood with a diverging number of parameters," *The Annals of Statistics*, vol. 32, no. 3, pp. 928–961, 2004.
[5] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society B*, vol. 70, no. 5, pp. 849–911, 2008.
[6] J. Chen and Z. Chen, "Extended Bayesian information criteria for model selection with large model spaces," *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008.
[7] J. Chen and Z. Chen, "Extended BIC for small-n-large-P sparse GLM," *Statistica Sinica*. In press.
[8] R. Tibshirani, "The Lasso method for variable selection in the cox model," *Statistics in Medicine*, vol. 16, pp. 385–395, 1997.
[9] J. Fan and R. Li, "Variable selection for Cox's proportional hazards model and frailty model," *The Annals of Statistics*, vol. 30, no. 1, pp. 74–99, 2002.
[10] H. H. Zhang and W. Lu, "Adaptive Lasso for Cox's proportional hazards model," *Biometrika*, vol. 94, no. 3, pp. 691–703, 2007.

[11] H. Zou, "A note on path-based variable selection in the penalized proportional hazards model," *Biometrika*, vol. 95, no. 1, pp. 241–247, 2008.
[12] R. J. Tibshirani, "Univariate shrinkage in the Cox model for high dimensional data," *Statistical Applications in Genetics and Molecular Biology*, vol. 8, pp. 3498–3528, 2009.
[13] J. Bradic, J. Fan, and J. Jiang, "Regularization for Cox's proportional hazards model with NP-dimensionality," vol. 39, no. 6, pp. 3092–3120, 2011.
[14] P. Du, S. Ma, and H. Liang, "Penalized variable selection procedure for Cox models with semiparametric relative risk," *The Annals of Statistics*, vol. 38, no. 4, pp. 2092–2117, 2010.
[15] H. Li and G. Jiang, "Partial Cox regression analysis for high-dimensional microarray gene expression data," *Bioinformatics*, vol. 20, 1, pp. i208–i215, 2004.
[16] P. Verweij and H. van Houwelingen, "Cross-validation in survival analysis," *Statistics in Medicine*, vol. 12, pp. 2305–2314, 1993.
[17] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.
[18] D. Paul, E. Bair, T. Hastie, and R. Tibshirani, "'Preconditioning' for feature selection and regression in high-dimensional problems," *The Annals of Statistics*, vol. 36, no. 4, pp. 1595–1618, 2008.
[19] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, "Simultaneous analysis of LASSO and Dantzig selector," *Annals of Statistics*, vol. 37, pp. 1705–1732, 2009.
[20] Y. Kim, H. Choi, and H.-S. Oh, "Smoothly clipped absolute deviation on high dimensions," *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1665–1673, 2008.
[21] J. Bradic, J. Fan, and W. Wang, "Penalized composite quasi-likelihood for ultrahigh-dimensional variable selection," *Journal of Royal Statistical Society Series B*. In press.
[22] T. T. Wu and K. Lange, "Coordinate descent algorithms for lasso penalized regression," *The Annals of Applied Statistics*, vol. 2, no. 1, pp. 224–244, 2008.
[23] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, pp. 1–22, 2010.
[24] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004, With discussion, and a rejoinder by the authors.
[25] H. Wold, "Estimation of principal components and related models by iterative least squares," in *Multivariate Analysis*, P. R. Krishaiaah, Ed., pp. 391–420, Academic Press, New York, NY, USA, 1966.
[26] D. Nguyen and D. M. Rocke, "Partial least squares proportional hazard regression for application to DNA microarray data," *Bioinformatics*, vol. 18, pp. 1625–1632, 2002.
[27] P. J. Park, L. Tian, and I. S. Kohane, "Linking expression data with patient survival times using partial least squares," *Bioinformatics*, vol. 18, pp. S120–S127, 2002.
[28] B. D. Marx, "Iteratively reweighted partial least squares estimation for generalized linear regression," *Technometrics*, vol. 38, pp. 374–381, 1996.
[29] D. G. Beer, S. L. Kardia, C. C. Huang et al., "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature Medicine*, vol. 8, pp. 816–824, 2002.