JP Journal of Biostatistics
Volume 3, Number 1, 2009, Pages 17-39

SURVIVAL PREDICTION WITH GENE EXPRESSION PROFILES

WENQING HE and GRACE Y. YI

Department of Statistical and Actuarial Sciences
University of Western Ontario
1151 Richmond Street North
London, Ontario
Canada N6A 5B7
e-mail: whe@stats.uwo.ca

Department of Statistics and Actuarial Science
University of Waterloo
200 University Avenue West
Waterloo, Ontario
Canada N2L 3G1
e-mail: yyi@uwaterloo.ca

Abstract

There is extensive research on prediction of various clinical phenotypes using gene expression profiles. Success has been demonstrated in molecular classification of different cancer types. However, relatively less attention has been paid to study the connection of gene expressions to time to event of patients such as time to tumour metastasis, an important problem in cancer research. One reason is that traditional survival analysis techniques may not be directly applicable in dealing

---

2000 Mathematics Subject Classification: 62N01, 62P10.

Keywords and phrases: Cox PH models, gene expression, microarray data, principal component analysis, Receive Operating Curve.

This research was supported by the Natural Sciences and Engineering Research Council of Canada.

Received August 23, 2008

with gene expression data, as typically the number of genes is much larger than the number of subjects. A primary objective of microarray studies is to identify informative or differentially expressed genes, and based upon them to make predictions on outcomes such as tumor type in cancer research. We develop methods for selecting survival relevant genes which may explain the time to event, and build prediction models, based on those genes, for the survival probability. Specifically, dimension reduction methods are invoked to pick out informative gene profiles that carry survival information. Cox proportional hazards models are utilized to conduct prediction, and the prediction accuracy is assessed by means of the Receive Operating Curve (ROC) method. Extensions to other survival models, such as accelerated failure time models, can be done along the same line. Simulation studies are conducted to evaluate the performance of the proposed methods under various conditions. A real microarray data set is analyzed using the proposed methods.

1. Introduction

In medical studies including cancer research, traditional methods focus on using clinical outcomes to study, for example, patient's survival information. Due to the fast advance of microarray technology, microarray data analysis has recently been receiving increasing interest in medical studies. Microarray technology is powerful in that it can simultaneously measure the expressions of thousands of genes and hence it becomes an increasingly common laboratory tool in biomedical and genomic research. This technology makes it possible to study clinical phenomenon such as tumor type identification and patient's survival prediction by means of gene profiles (e.g., West et al. [27], Dudoit et al. [7], van't Veer et al. [25]). It is believed that gene expression profiles, in combination with traditionally clinical information, may serve as a better tool to build prediction models on patient's survival probability. Due to the large variability in time to an event such as cancer recurrence or tumor metastasis among cancer patients, treating phenotypes as survival data which can be more effective than classifying the phenotypes as binary or categorical variable (Gui and Li [9]). It is of scientific interest to build prediction models of survival probability for cancer patients based upon informative genes, and this in turn may lead to novel approach to diagnosis and treatment. However, due to a number of challenges, traditional survival data analysis methods can not be applied directly to handle gene expression data. Collinearity and huge dimensionality, for instance, are typical features possessed by gene expression data that prevent direct application of survival data analysis. The number of genes is usually in thousands, but the number of subjects is a lot smaller. It is difficult or impossible to model those high dimensional data by applying usual survival models directly.

To employ survival analysis techniques, we need first to reduce the dimension of genes by using methods such as clustering discussed in von Heydebreck et al. [26] for which distinctive gene expressions are classified into different groups. However, this approach fails to use clinical information, and thereby it may lead to clusters that are not of survival relevance. To overcome this problem, authors including Nguyen and Rocke [19] and Park et al. [20] discussed the partial least squares method which accommodates the survival information by sequentially maximizing the covariance between the survival time and a linear combination of the genes. Recently, Li and Luan [18] proposed an L_2 penalized Cox proportional hazards model where kernel estimation is used to ease computational burden. Li and Li [17] proposed a dimension reduction strategy which combines principle components analysis and sliced inverse regression to identify useful linear combinations of genes. Bair et al. [2] explored the supervised principle component procedure which is similar to conventional principal components analysis except that it uses a subset of the predictors selected based on their association with the outcome. Zhao and Sun [28] used a modified correlation principal component regression to reduce the dimension and address the censoring in survival data. Among various dimension reduction methods, principle components analysis is perhaps the most popular one which provides the basis for many other modified dimension reduction methods. An overview of this method may be found in Jolliffe [15] and Chiaromonte and Martinelli [5].

These dimension reduction methods have enjoyed wide applications in microarray data analysis, and the focus mainly centers on building a prediction model for future patients based on the gene expression profiles and survival times of previous patients. Typically, the prediction model is formulated using the extracted linear combinations of all of the genes in the database, where those gene combinations are identified based on their variability in the gene expression levels and preservation of the phenotypic information (e.g., Li and Li [17]). However, from the biological point of view, only a small portion of genes is of predicting power for phenotypes. Including all or most of the genes in the predictive model may induce substantial noise and thereby lead to poor predictive performance. It is therefore important to identify individual informative genes that are of survival relevance and have to carry out prediction for future patient survival based on those genes. Furthermore, as a pleasant byproduct those selected genes may be used to fulfill other objectives in microarray studies. Identifying phenotypic relevant genes is of primary interest in many problems. There has been extensive research on this subject and various gene selection methods have been proposed in the literature (e.g., Chiaromonte and Martinelli [5], Dudoit et al. [7], He [11], Gui and Li [9]). Our proposed algorithm of survival relevant gene selection provides an addition to this topic.

In this paper we propose a three-step algorithm for survival relevant gene selection. Cox proportional hazards models are typically used to build prediction models for patients' survival probabilities using the selected gene profiles. The prediction model will be evaluated with the Receive Operating Curve (ROC) technique. The remainder is organized as follows. In Section 2 we present the notation and gene selection methodology. In Section 3 we establish the prediction model and describe the evaluation of its prediction accuracy. The performance of the proposed methods is assessed in Section 4 through simulation studies. In Section 5 we analyze a real microarray data set with the proposed methods. A general discussion is included in the last section.

2. Framework and Methodology

Let the matrix $\boldsymbol{X} = [X_{ij}]$ denote a microarray data set of gene expressions, with rows being genes and columns being arrays (samples),

$i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, M$. Here $\boldsymbol{X}$ can be the relative log intensity ratios in cDNA array or the log intensity ratios of test samples versus a baseline sample in Affymetrix array, which have been processed by a certain normalization procedure to remove systematic background noise (e.g., He et al. [12]). Denote by $\boldsymbol{X}_{;j}$ the $N \times 1$ column vector of gene expressions for sample or subject j, and $\boldsymbol{X}_{i;}^{T}$ the $M \times 1$ vector of the realizations for the ith gene expression, where the superscript T represents the transpose of a vector or matrix. Let T_j and C_j be the survival and censoring times for subject j, respectively, and δ_j be the censoring indicator variable taking 1 if $T_j \le C_j$ and 0 otherwise. Let $t_j = \min(T_j, C_j)$ for $j = 1, 2, \ldots, M$. Independent censoring is assumed, i.e., the censoring process does not depend on the survival process, given the covariate process.

It is of primary interest to study the relationship between the survival times and the gene expression levels and based upon it to build a prediction model. The number of genes ranges from thousands to tens of thousands, but as believed in most microarray studies, the majority of those genes may not have informative expressions and thus do not have significant contribution to the survival information. Below we describe a three-step algorithm for selecting individual genes that are useful for prediction.

Step 1. Gene Screening

Single covariate Cox proportional hazards models are invoked to screen out those genes with no (or little) effects on survival. For a given $i = 1, 2, \ldots, N$, including the ith gene expression as the only covariate, we fit the Cox proportional hazards model with the hazard function

$$\lambda(t_j) = \lambda_0(t_j) \exp \lambda(\beta_i X_{ij}), \quad j = 1, 2, \ldots, M, \tag{1}$$

where β_i is the regression parameter, and $\lambda_0(t_j)$ represents the baseline hazard function that is left unspecified. The Cox partial loglikelihood (Cox [6])

$$\ell(\beta_i) = \sum_{j=1}^{M} \delta_j \left[\beta_i X_{ij} - \log\left(\sum_{l=1}^{M} Y_l(t_j) \exp(\beta_i X_{il}) \right) \right] \tag{2}$$

is employed to obtain the estimate $\hat{\beta}_i$ of the covariate coefficient β_i, where $Y_l(t) = I(t_l \geq t)$ is the at risk indicator and $I(\cdot)$ is the indicator function. The asymptotic normal distribution may be established for $\sqrt{M}(\hat{\beta}_i - \beta_i)$, $i = 1, 2, \ldots, N$, and based upon it, the test statistic can then be easily formulated to screen out those genes whose regression coefficients are not significant at a prespecified significance level. This method has been widely used for gene selection with survival data. The purpose of this step is to reduce the noise that would be induced by including a large number of non-informative genes. See Bair et al. [2], for example.

For screening purpose, an alternative may be employed. Note that gene expression may be represented by the relative expression between the test and reference samples such as in cDNA microarray studies, or the relative expression between the test sample and a common baseline sample in Affymetrix array. Informative genes with up-regulated (i.e., positive relative expression) or down-regulated (i.e., negative relative expression) expressions would have the measures farther away from zero, whereas non-informative genes have relative expressions close to zero. Furthermore, usual normalizations may shift the mean or median of the relative expressions within an array to be zero, but the mean or median across arrays for a single gene is not necessarily zero. Given these properties, we propose to use

$$f(i) = \hat{\beta}_i \overline{X}_i. \tag{3}$$

as a criterion to do screening. Here $\overline{X}_i. = M^{-1} \sum_{j=1}^{M} X_{ij}$. A large absolute value of $f(i)$ indicates that the survival contribution from gene i is different from the baseline survival information. Therefore, a threshold α_1 can be set to exclude genes with $|f(i)| < \alpha_1$, or a percentage (say, 90%) is specified such that a given proportion of genes is retained.

The use of $f(i)$ is motivated by the following consideration: for gene i, its expressions $X_{i1}, X_{i2}, \ldots, X_{iM}$ are usually assumed independently identically distributed with a normal distribution with mean μ_i and variance σ_i^2, and hence $\beta_i X_{ij} \sim N(\beta_i \mu_i, \beta_i^2 \sigma_i^2)$. To test the effect of gene i, one may test for the null hypothesis $H_o : \beta_i \mu_i = 0$. Based on the average of the realizations of gene i, we can use the test statistic $\hat{\beta}_i \overline{X}_{i\cdot}$ to test for H_o. This statistic is a special case of the compound covariate proposed in Tukey [24]. Such an idea was also discussed in Radmacher et al. [21] and Li and Gui [16].

We comment that issues related to multiple comparison may arise because thousands of comparisons are conducted in this step. It is a practical way to relax the significance level α_0 (or the threshold α_1) so that potentially survival relevant genes are not screened out.

Step 2. Gene Combination Selection

Let $\boldsymbol{X}^* = [X_{ij}^*]$ denote the new gene expression matrix containing the remaining potentially informative genes after Step 1, where i indexes these genes from 1 to m^*, say. After the screening step, the number m^* of the remaining genes may still be very large, and it may usually be much larger than the number of the observations. Moreover, the correlation among genes makes it difficult to select out the informative genes solely based on the first step. Now we invoke the principle components analysis method to further reduce the dimension of the gene expression variables. Extensions to other approaches such as supervised principle components procedure follow the same spirit.

Principle components can be obtained through the singular value decomposition of the matrix $\boldsymbol{X}^*$ (Hastie et al. [10]). The decomposition of $\boldsymbol{X}^*$ is given by

$$\boldsymbol{X}^* = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T,$$

where $\boldsymbol{D}$ is the $M \times M$ diagonal matrix with diagonal elements being ordered singular values $d_1 \geq d_2 \geq \cdots \geq d_M \geq 0$ of $\boldsymbol{X}^*$, i.e., they are the square roots of the eigenvalues of $\boldsymbol{X}^{*T}\boldsymbol{X}^*$. Here $\boldsymbol{U}$ and $\boldsymbol{V}$ are, respectively, the $m^* \times M$ and $M \times M$ matrices having $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_M$ and $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_M$, where $\boldsymbol{I}_M$ is the $M \times M$ unit matrix. Letting

$$\boldsymbol{V}^T = \boldsymbol{D}^{-1}\boldsymbol{U}^T\boldsymbol{X}^*$$

gives that the column vectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_M$ of $\boldsymbol{V}$ are the linear combinations of the row vectors of $\boldsymbol{X}^*$ with coefficients $\boldsymbol{D}^{-1}\boldsymbol{U}^T = \boldsymbol{A}$, say, and they form the principle components of $\boldsymbol{X}^*$. Note that $\boldsymbol{A}$ is an $M \times m^*$ matrix with orthogonal row vectors, i.e., $\boldsymbol{A}\boldsymbol{A}^T$ is an $M \times M$ diagonal matrix. As the singular value d_j is the square root of the variance that the associated principle component can explain, thus we further select a number of principle components so that the associated total variance can be explained over a threshold $\gamma = 0.8$, say. Let m be the label such that

$$\frac{d_1^2 + \cdots + d_m^2}{d_1^2 + \cdots + d_M^2} \geq \gamma,$$

and $\widetilde{\boldsymbol{A}}$ be the submatrix of the first m rows of $\boldsymbol{A}$. Define $\boldsymbol{Z} = \widetilde{\boldsymbol{A}}\boldsymbol{X}^*$. This $m \times M$ matrix consists of the expressions of gene combinations that are important in the sense that they explain a large proportion of the total variation of gene expressions. To exclude possible gene combinations that are not of survival relevance, we may repeat Step 1 for matrix $\boldsymbol{Z}$ to pick out n survival relevant combinations and form a new matrix $\boldsymbol{Z}^* = [Z_{ij}^*]$ with $n{\times}M$ dimension, say. Denote by $\boldsymbol{A}^*{=}(\boldsymbol{a}_1^*,\ldots,\boldsymbol{a}_{m^*}^*)$ the corresponding $n \times m^*$ submatrix of $\widetilde{\boldsymbol{A}}$, where $\boldsymbol{a}_i^* = (a_{1i}^*, a_{2i}^*, \ldots, a_{ni}^*)^T$ is an $n \times 1$ column vector, $i = 1, 2, \ldots, m^*$. As a result, $\boldsymbol{Z}^*{=}\boldsymbol{A}^*\boldsymbol{X}^*$.

Step 3. Individual Gene Selection

The selected gene combinations $Z_{1j}^*, Z_{2j}^*, \ldots, Z_{nj}^*$ (with $n < m$) may be used to build the prediction model using the Cox proportional hazard function

$$\lambda(t_j) = \lambda_0(t_j)\exp\left(\sum_{k=1}^{n}\beta_k^* Z_{kj}^*\right), \quad j = 1, 2, ..., M, \tag{4}$$

where $\beta_1^*, ..., \beta_n^*$ are the regression coefficients. The partial log-likelihood can be employed to estimate $\boldsymbol{\beta}^* = (\beta_1^*, ..., \beta_n^*)^T$, and let $\hat{\boldsymbol{\beta}}^*$ be the resulting estimator of $\boldsymbol{\beta}^*$. Though those selected gene combinations explain a major proportion of variation, their biological meaning is not clear as the expression combinations of Z_{kj}^*'s combine all or most of the gene expressions on an array for which survival irrelevant genes are also included. It is desirable to identify individual genes that are associated with the survival time. If a model based on individual survival relevant genes (often of a considerably small number) can be built for prediction, then a lot cheaper methods can be used to solely measure those small number of genes for patients rather than to measure all the gene expressions. To this end we propose the following step.

In model (4), we may explicitly present the initial gene expressions $\boldsymbol{X}^*$ by using

$$Z_{kj}^* = \sum_{i=1}^{m^*} a_{ki}^* X_{ij}^*.$$

Therefore,

$$\sum_{k=1}^{n}\beta_k^* Z_{kj}^* = \sum_{i=1}^{m^*}\left(\sum_{k=1}^{n}\beta_k^* a_{ki}^*\right)X_{ij}^*,$$

that is, the coefficient related to original gene i is $\sum_{k=1}^{n}\beta_k^* a_{ki}^*$. Let $\beta_i^{**} = \sum_{k=1}^{n}\beta_k^* a_{ki}^*$. Then model (4) may be written as

$$\lambda(t_j) = \lambda_0(t_j)\exp\left(\sum_{i=1}^{m^*}\beta_i^{**} X_{ij}^*\right), \quad j = 1, 2, ..., M, \tag{5}$$

with m^* original genes as predictors. However, model (5) can not be directly used to conduct prediction because the number m^* of the predictors X^*_{ij}'s could be greater than the number M of the subjects. To detect the effects of genes X^*_{ij}'s we need to test $H_o : \beta^{**}_i = 0$ for each $i = 1, 2, \ldots, m^*$. In the same spirit as (3) in Step 1, we employ

$$f^*(i) = \hat{\beta}^{**}_i \overline{X}^*_{i\cdot}$$

to choose survival relevant individual genes from $\boldsymbol{X}^*$, where $\overline{X}^*_{i\cdot} = M^{-1}\sum_{j=1}^{M} X^*_{ij}$, and $\hat{\beta}^{**}_i = \sum_{k=1}^{n} \hat{\beta}^*_k a^*_{ki}$. More specifically, for a given threshold α^*, we proceed through initial gene expression matrix $\boldsymbol{X}^*$, and include gene i in the final set of selected genes if $|f^*(i)| > \alpha^*$. Let $\{\boldsymbol{X}^*_{i;}, i = 1, 2, \ldots, N^*\}$ denote the gene expression of those selected genes. The value of α^* can be chosen based on the feature of individual data set. In principle, we can take α^* such that $N^* < M$ and the proportion N^*/m^* of survival relevant genes is smaller than a desired percentage.

3. Survival Prediction Model

Using the selected genes $\{\boldsymbol{X}^*_{i;}, i = 1, 2, \ldots, N^*\}$, we build a prediction model by specifying the hazard function as

$$\lambda(t_j) = \lambda_0(t_j)\exp\left(\sum_{i=1}^{N^*} \tilde{\beta}_i X^*_{ij}\right), \quad j = 1, 2, \ldots, M, \tag{6}$$

where $\tilde{\beta}_i$'s are the regression coefficients corresponding to individually selected genes, and $\lambda_0(t_j)$ is the baseline hazard function.

Let $\hat{\tilde{\beta}}_i$'s be the estimates of $\tilde{\beta}_i$'s and $\hat{\Lambda}_0(t)$ be the Breslow estimate for the baseline cumulative hazard function $\Lambda_0(t)$. Then the estimate of the survival probability at a given time t is given by

$$\hat{S}(t) = \exp\left\{-\hat{\Lambda}_0(t)\exp\left(\sum_{i=1}^{N^*}\hat{\tilde{\beta}}_i x_i^*\right)\right\},$$

where $\{x_1^*, x_2^*, \ldots, x_{N^*}^*\}$ are the expressions of genes obtained by the algorithm in Section 2 for a subject to be predicted.

In addition to building a useful prediction model, we also concern about how the prediction model (6) performs. It is desirable to develop a tool to measure the prediction accuracy of model (6). In classification problems, for example, it is often a concern to estimate the classification error. The cross-validated misclassification rate is often used for this purpose. With linear regression analysis we commonly use the mean squared residuals as a criterion to assess models. For survival models, however, those criteria can not be applied due to the complications caused by censoring. If we use only the observed survival times for constructing a measure of prediction power, we essentially loss much information that is conveyed by censored data. Particularly for survival data with a large portion of censoring, mean squared errors can provide very poor evaluation for a survival prediction model. Heagerty et al. [14] proposed a time-dependent Receive Operation Curve (ROC) for censoring data. The Area Under the Curve (AUC) is utilized to assess survival prediction models (Li and Gui [16]). Here we adapt the discussion in Li and Gui [16] to assess the accuracy of the prediction model (6).

In model (6), the covariate effects $\sum_{i=1}^{N^*}\tilde{\beta}_i X_{ij}^*$ have decreasing impact on the survival probability. That is, the larger the value of $\sum_{i=1}^{N^*}\tilde{\beta}_i X_{ij}^*$, the smaller the value of the survival probability. Denote $g(\boldsymbol{X}_{;j}^*) = \sum_{i=1}^{N^*}\hat{\tilde{\beta}}_i X_{ij}^*$. Let $D_j(t)$ be the event status of subject j at time t, i.e., $D_j(t) = 1$ if the event T_j occurs at time t, and 0 otherwise. For a constant c and a time point t, define

$$\text{Sensitivity}\,(c, t) = \Pr(g(\boldsymbol{X}_{;j}^*) > c \mid D_j(t) = 1),$$

$$\text{Specificity}\,(c,\,t) = \Pr(g(\boldsymbol{X}^*_{;j}) \leq c \mid D_j(t) = 0). \tag{7}$$

Those quantities can be estimated through the nearest neighbor estimation of the bivariate distribution of c and t (Akritas [1]).

For a given time point t, the plot of estimated *sensitivity* against 1-*specificity* with cutoff point c varying gives the ROC curve, and the AUC as the area under the ROC curve may then be calculated. AUC values range from 0 to 1, and a larger AUC at time t suggests a better predictability at time t.

4. Simulation Studies

We conduct simulation studies to evaluate the performance of the proposed methods. 100 simulations are run for each configuration. In practice, a microarray usually has tens of thousands of genes with a small portion of informative genes. But for the purposes of maintaining computation ease and getting a direct insight into the proposed methods, we consider a setting with $N = 3000$ genes and $M = 100$ samples, and keep the proportion of informative genes in an array to be similar to real situations. Namely, about 1% of genes, or 30 genes here, are set to be of survival relevance. To facilitate possible heterogeneity we generate informative genes from two different distributions.

To be specific, we partition the gene expression matrix $\boldsymbol{X} = [X_{ij}]_{3000\times 100}$ as

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}^{11} & \boldsymbol{X}^{12} \\ \boldsymbol{X}^{21} & \boldsymbol{X}^{22} \\ \boldsymbol{X}^{31} & \boldsymbol{X}^{32} \end{bmatrix},$$

where the submatrices $[\boldsymbol{X}^{11}, \boldsymbol{X}^{12}]$ and $[\boldsymbol{X}^{21}, \boldsymbol{X}^{22}]$, respectively, represent each of the two groups of informative genes, each of dimension 15×100. The submatrices $\boldsymbol{X}^{l1}$ and $\boldsymbol{X}^{l2}$, $l = 1, 2$, are introduced to divide 100 samples into two different categories, where $\boldsymbol{X}^{l1}$ is of dimension 15×40 and $\boldsymbol{X}^{l2}$ is of dimension 15×60. As a result, the remaining matrices $\boldsymbol{X}^{31}$ and $\boldsymbol{X}^{32}$ represent 2970 non-informative gene expressions, corresponding to the first 40 and the last 60 samples, respectively.

Let $\boldsymbol{X}_{;j}^{ll'}$ be the column vector of the submatrix $\boldsymbol{X}^{ll'}$, $l, l' = 1, 2,$ corresponding to the jth sample. For each j we simulate $\boldsymbol{X}_{;j}^{ll'}$ from a multivariate normal distribution $\mathrm{MVN}(\boldsymbol{\mu}^{ll'}, \boldsymbol{\Sigma}),$ where

$$\boldsymbol{\mu}^{11} = 0.8 \cdot \mathbf{1}, \quad \boldsymbol{\mu}^{12} = 0.1 \cdot \mathbf{1},$$

$$\boldsymbol{\mu}^{21} = -0.1 \cdot \mathbf{1}, \quad \boldsymbol{\mu}^{22} = 0.2 \cdot \mathbf{1},$$

and $\mathbf{1}$ is the 15×1 unit vector. The variance matrix $\boldsymbol{\Sigma}$ is specified as $\boldsymbol{\Sigma} = [v_{ij}]$ with $v_{ii} = \sigma^2$ and $v_{ii} = \rho\sigma^2$ for $i \neq j$ to facilitate exchangeable correlation among gene expressions. Set $\sigma = 0.5.$ Non-informative gene expression is simulated as $X_{ij}^{3l} \sim N(0, 0.2^2),$ where X_{ij}^{3l} is the (i, j) component of $\boldsymbol{X}^{3l},$ $l = 1, 2.$ Survival times are simulated using a Cox proportional hazards model with the informative gene expressions included as covariates, where the baseline hazard function is specified as a constant. That is, an exponential model is employed to generate the survival time for patient $j = 1, 2, \ldots, 100,$

$$T_j = 100(-\log u_j) \exp\left(-\sum_{i=1}^{30} \beta_i X_{ij} \right), \tag{8}$$

with $\beta_i = -0.4$ for $i = 1, \ldots, 15$ and $\beta_i = 0.4$ for $i = 16, \ldots, 30,$ where u_j is generated from the uniform distribution $U[0, 1].$ We consider fixed censoring times $C_j = 70$ for $j = 1, 2, \ldots, 100.$

To illustrate how the proposed three-step algorithm in Section 2 may be applied to select survival relevant genes, we report the results for a case with ρ specified as 0. Various other values of ρ may be considered in a similar manner but not reported here. In the first step we choose a threshold such that 5% of genes pass this screening step. In the second step we pick a number of principle components so that 80% of the total variation can be explained, and a threshold 1.0 is used to exclude survival irrelevant gene combinations $\boldsymbol{Z}_{i;}$. In Step 3, we set the threshold value α^* as 0.25, 0.20, 0.15 and 0.10, respectively, to control different proportions of selected genes. Table 1 reports on the number of selected genes, the average number of true positives in selected genes and the false discovery rate, along with the empirical standard errors reported in the brackets. It is seen that when the threshold α^* is small, say, 0.10 or 0.15, almost all selected genes are truly survival relevant, and no irrelevant genes are selected. However, some survival relevant genes are screened out at Step 1 already, and thus are not included in the final list. Typically about 50% relevant genes are missed in the case of the threshold 0.10. As the threshold value α^* increases, more relevant genes are selected, but we can see that more irrelevant genes are also being selected as well. It appears that the threshold 0.20 gives the optimal result with about 92% truly relevant genes being selected and falsely selected irrelevant genes being kept at a reasonably low proportion. This finding suggests that the selection based on the three steps in Section 2 can provide reasonable results, provided the threshold values are properly chosen.

Table 1. Simulation results: selected genes by the three-step algorithm in Section 2

Threshold	Number of Selected genes	Number of positives	True Positive rate[1]	False discovery rate[2]
0.25	37	28.740 (0.645)	0.958 (0.022)	0.223 (0.017)
0.20	31	27.630 (1.454)	0.921 (0.048)	0.109 (0.047)
0.15	23	22.650 (0.880)	0.755 (0.029)	0.015 (0.038)
0.10	15	14.990 (0.001)	0.500 (0.003)	0.000 (0.007)

$$^{1}\text{True postive rate} = \frac{\text{Selected positive}}{\text{True positive}}$$

$$^{2}\text{False discovery rate} = \frac{\text{Selected false positive}}{\text{True selected}}$$

Our second objective here is to assess the power of the prediction models that are built based on individual survival relevant genes that are selected from the three-step algorithm in Section 2. Here we explicitly consider the cases with $\rho = 0,\ 0.5,$ and 0.9 to reflect independence, mild and strong correlation among gene expressions. The prediction models are applied to predict the survival probability at given time points $5, 10, \ldots, 70,$ and the AUC at each time point is calculated. Figure 1 displays the plots of the AUC values against the chosen time points. It is apparent that the AUCs are between 0.8 and 0.9 for the scenarios with correlation ρ ranging from 0 to 0.9. This demonstrates that the prediction model using the selected individual genes performs reasonably well.

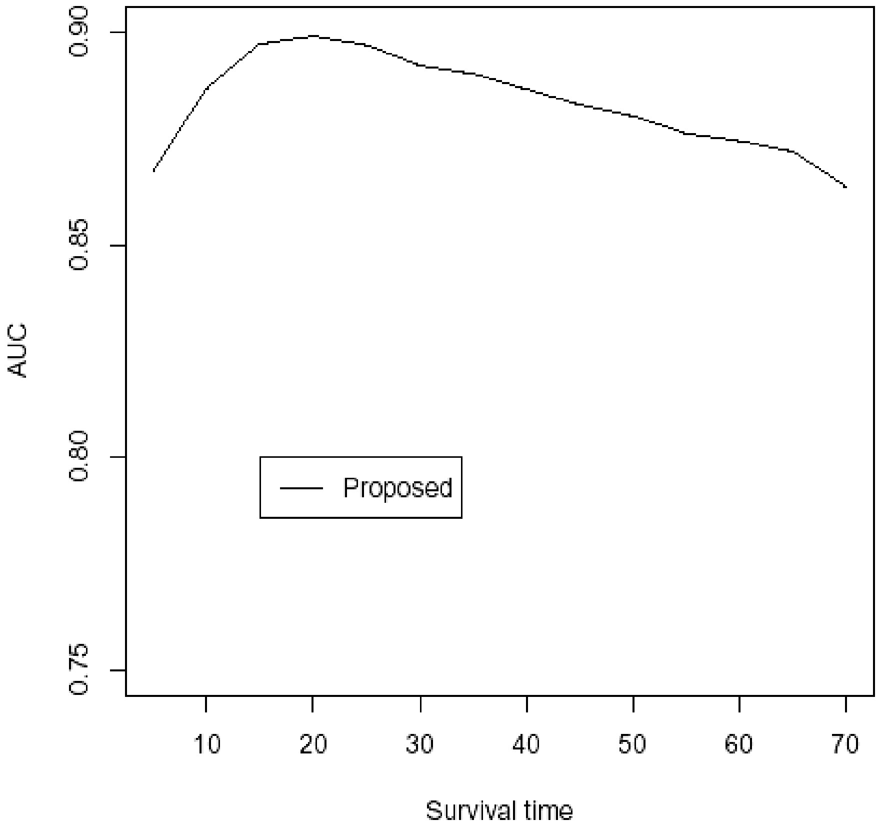

Figure 1. The AUC values against survival times from the simulation study.

To see how the performance of the proposed method is affected by the correlation ρ among gene expressions, we report the AUC values in contrast to ρ values in Figure 2. It seems that the method is not very much influenced by the strength of the correlation $\rho,$ as the AUC values are fairly stable over the long range of time points from 5 to 70. It is not surprising that the AUC values tend to decrease as the prediction time gets large.

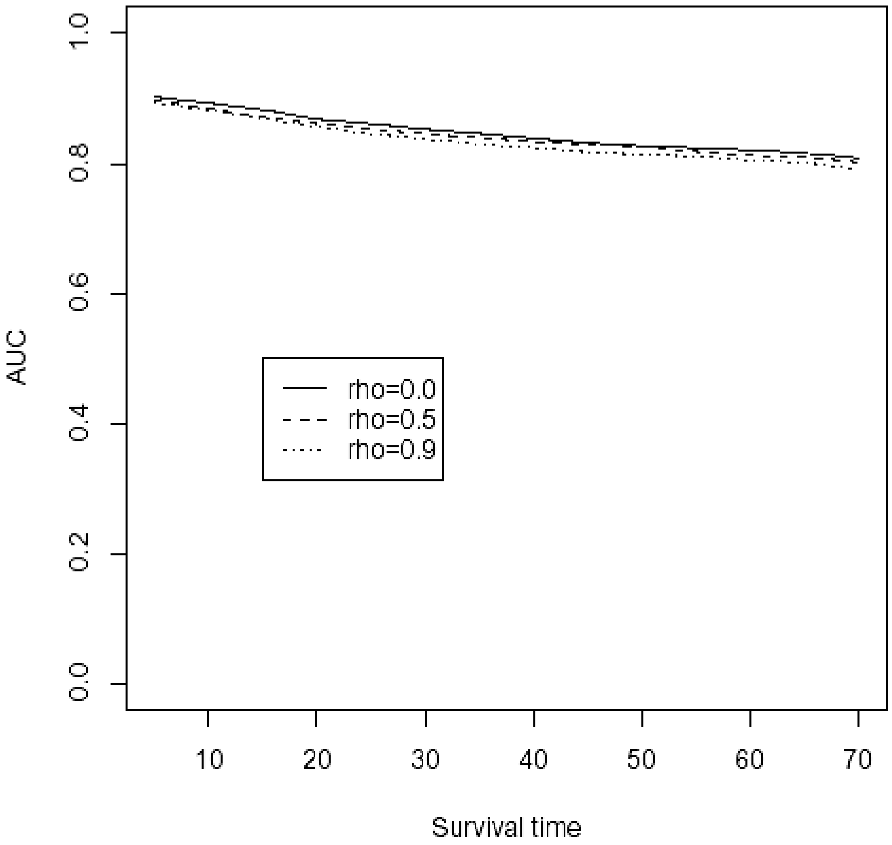

Figure 2. The AUC values for simulated data with different correlations.

5. An Example

We apply the proposed methods to analyze a microarray data set discussed in Rosenwald et al. [22]. The data set consists of 240 biopsy samples from patients with diffuse large B-cell lymphoma (DLBCL), the most common lymphoma in adults, with the gene expression measured for 7399 genes after some preprocessing procedures. The outcomes include the survival information of the patients, either observed death times or right censored times. Survival times vary in a range from 0 to 21.8 years with median 7.3 years. About 43% patients are still alive at the end of the study.

A well-established predictor of survival in DLBCL is the International Prognostic Index (IPI) which is based on five clinical variables (i.e., age, tumor stage, serum lactate dehydrogenase concentration, performance status, and the number of extranodal disease sites). But it has been found that the outcome in patients with DLBCL who have identical IPI values may vary considerably. As a result, Rosenwald et al. [22] hypothesized that gene-expression profiles of DLBCL could be used independently of the IPI factor to predict patients' survival after their chemotherapy. Many authors have analyzed survival times based on gene expression profiles, and their findings suggested it is possible to characterize patients' survival based on gene expression data (Sinisi et al. [23]).

Here our objectives are to (1) select individual survival relevant genes that may be of prime interest for other microarray studies; and (2) build prediction models using the methods described in Section 3. For the first screening step we set $\alpha_0 = 0.05$, and this leads to 367 genes that are potentially useful for survival prediction. A threshold of $\gamma = 0.8$ is used for the second step to select principle components, and the threshold is set to be 1 for repeating Step 1 for the matrix $\boldsymbol{Z}^*$. Set a value for α^* such that 20% genes are selected at the third step, and this results in 73 genes as survival relevant genes. We consider two methods here to build prediction models. With Method 1 we formulate a prediction model by entering those selected genes in hazard function (6). To accommodate possible correlation of the screened out genes with those 73 genes, we add back those excluded genes if they have a correlation with one of the 73 genes higher than 0.7. It turns out that only one screened out gene is added back. We use those 74 genes to build a prediction model by using hazard function form (6), and this is Method 2. To compare the performance of these methods, we work on the ROC curve for the time point $t = 10$ and the AUC values for a given sequence of time points ranging from 0 to 22. The results are displayed in Figures 3 and 4, respectively. It can be seen that, from Figure 3, the shape of the ROC curves obtained from Methods 1 and 2 is very similar.

Figure 4 presents the change of the AUC with respect to the predicting times, corresponding to each method. It is seen that the AUCs for Methods 1 and 2 are almost the same, though Method 2 yields a bit higher AUCs, indicating a slightly better prediction. The trend of the AUC changes is the same for the two methods. The AUCs are about 0.8 before 15 years, which suggests that the survival prediction before 15 years is very good. The AUCs decrease to about 0.6 at 17 years and then to about 0.5 at 22 years, which pronounces the prediction after 15 years is not very reliable. This may not be surprising because only 7 samples after 15 years are observed, and among them only one observed failure at 16.9 years while the other 6 samples are all censored.

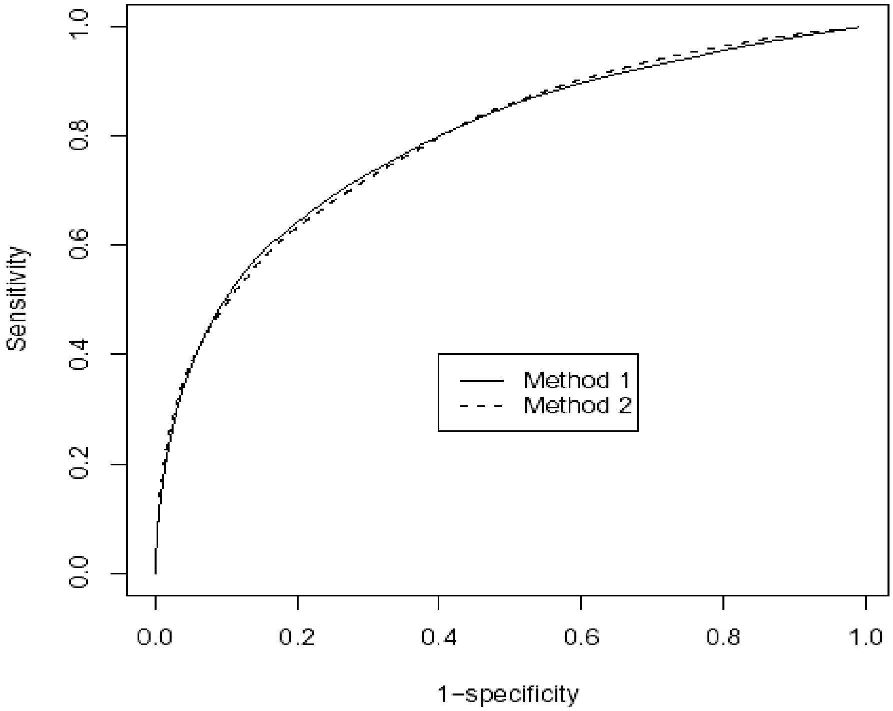

Figure 3. The ROC curves for survival prediction at time $t = 10$ for DLBCL data.

6. Discussion

In this paper, we develop a method to select survival relevant genes and based upon them build survival prediction models. The genes used in the survival model are both survival relevant and biologically interpretable, and this is a typically appealing feature of the proposed method. Many existing methods focus on building prediction models with the linear combinations of genes, but our methods here, in addition to building a prediction model for patient's survival probability, are able to identify individual survival relevant genes, which is of interest in its own right. This feature distinguishes the proposed methods from other approaches such as the Supervised Principle Component (SPC) method in Bair et al. [2] for which the linear combinations of a subset of genes are incorporated in the survival prediction model. Simulation studies demonstrate that the proposed methods perform generally well under various situations.

In this paper we employ Cox proportional hazards models to relate gene profiles to patient survival. Extensions to other survival models such as accelerated failure time (AFT) models are straightforward. Such models have transparent interpretation of the covariate effects on the survival information (e.g., He and Lawless [13]).

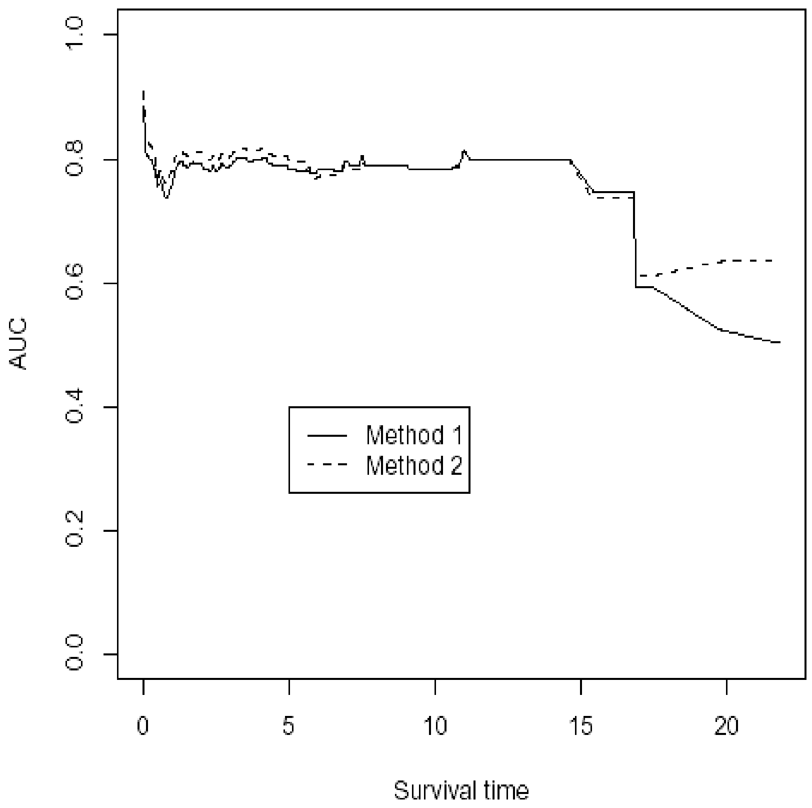

Figure 4. The AUC values for predicting survival at distinct time points of DLBCL data.

One may note that in building prediction models, only gene profiles are included as covariates. This is driven by that gene profiles are regarded as of most importance and interest in predicting survival. However, one may also wish, in some situations, to control clinical covariates such as age, gender, and other health information when building prediction models. A possible option is to use the partially linear regression (Carroll et al. [4]) form $\boldsymbol{\beta}'\boldsymbol{x} + \eta(\boldsymbol{\alpha}'\boldsymbol{z})$ to specify the hazard function, i.e., let

$$\lambda(t) = \lambda_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x} + \eta(\boldsymbol{\alpha}'\boldsymbol{z})), \tag{9}$$

where $\lambda_0(t)$ is the baseline hazard function, $\boldsymbol{x}$ represents the gene profiles that are of interest, and $\boldsymbol{z}$ includes clinical covariates that are of less importance. This specification generalizes traditional Cox proportional hazards models by including an unknown smooth function $\eta(\cdot)$. It would be interesting to conduct survival prediction by employing (9).

References

[1] M. G. Akritas, Nearest neighbor estimation of bivariate distribution under random censoring, Ann. Statist. 22 (1994), 1299-1327.

[2] E. Bair, T. Hastie, D. Paul and R. Tibshirani, Prediction by supervised principle components, J. Amer. Statist. Assoc. 101 (2006), 119-137.

[3] E. Bair and R. Tibshirani, Semi-supervised methods to predict patient survival from gene expression data, PLos Biology 2 (2004), 411-522.

[4] R. J. Carroll, J. Fan, I. Gijbels and M. P. Wand, Generalized partially linear single-index models, J. Amer. Statist. Assoc. 92 (1997), 477-489.

[5] F. Chiaromonte and J. Martinelli, Dimension reduction strategies for analyzing global gene expression data with a response, Math. Biosci. 176 (2002), 123-144.

[6] D. J. Cox, Partial likelihood, Biometrika 62 (1975), 269-276.

[7] S. Dudoit, J. Fridlyand and T. P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, J. Amer. Statist. Assoc. 97 (2002), 77-87.

[8] A. Gould and J. F. Lawless, Consistency and efficiency of regression coefficient estimates in location-scale models, Biometrika 75 (1988), 535-540.

[9] J. Gui and H. Li, Penalized Cox regression analysis in the high dimensional and low-sample size setting, with applications to microarray gene expression data, Bioinformatics 21 (2005), 3001-3008.

[10] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer-Verlag, New York, 2001.

[11] W. He, A spline function approach for detecting differentially expressed genes in microarray data analysis, Bioinformatics 20 (2004), 2954-2963.

[12] W. He, S. B. Bull, N. Gokgoz, I. Andrulis and J. Wunder, Application of reliability coefficients in cDNA microarray data analysis, Statist. Med. 25 (2006), 1051-1066.

[13] W. He and J. F. Lawless, Bivariate location-scale models for regression analysis, with applications to lifetime data, J. Roy. Statist. Soc., Ser. B Methodol. 67 (2005), 63-78.

[14] P. J. Heagerty, T. Lumley and M. Pepe, Time dependent ROC curves for censored survival data and a diagnostic marker, Biometrics 56 (2000), 337-344.

[15] I. T. Jolliffe, Principle Component Analysis, 2nd Edition, Springer, 2002.

[16] H. Li and J. Gui, Partial Cox regression analysis for high-dimensional microarray gene expression data, Bioinformatics 20 Suppl (2004), i208-i215.

[17] L. Li and H. Li, Dimension reduction methods for microarrays with application to censored survival data, Bioinformatics 20 (2004), 3406-3412.

[18] H. Li and Y. Luan, Kernel Cox regression models for linking gene expression profiles to censored survival data, Pacific Symposium of Biocomputation 8 (2003), 65-76.

[19] D. Nguyen and D. M. Rocke, Partial least squares proportional hazard regression for application to DNA microarray data, Bioinformatics 18 (2002), 1625-1632.

[20] P. J. Park, L. Tian and I. S. Kohane, Linking expression data with patient survival times using partial least squares, Bioinformatics 18 (2002), S12-S127.

[21] M. D. Radmacher, L. M. McShane and R. Simon, A paradigm for class prediction using gene expression profiles, J. Comput. Biol. 9 (2002), 505-511.

[22] A. Rosenwald, G. Wright, W. C. Chan, J. M. Connors, E. Campo, R. I. Fisher, R. D. Gascoyne, H. K. Muller-Hermelink, E. B. Smeland and L. M. Staudt, The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma, New England J. Med. 346 (2002), 1937-1947.

[23] S. E. Sinisi, R. Neugebauer and M. K. van der Laan, Cross-validated bagged prediction of survival, Statist. Appl. Gene. Mole. Bio. 5 (2006), Article 12.

[24] J. W. Tukey, Tightening the clinical trials, Controlled Clinical Trials 14 (1993), 266-285.

[25] L. J. van't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse et al., Gene expression profiling predicts clinical outcome of breast cancer, Nature 415 (2002), 530-536.

[26] A. von Heydebreck, W. Huber, A. Poustka and M. Vingron, Identifying splits with clear separation: a new class discovery method for gene expression data, Bioinformatics 17(Suppl I) (2001), S107-S114.

[27] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, Jr. J. A. Olson, J. R. Marks and J. R. Nevins, Predicting the clinical status of human breast cancer by using gene expression profiles, Proc. National Acad. Sci. 98 (2001), 11462-11467.

[28] Q. Zhao and J. Sun, Cox survival analysis of microarray gene expression data using correlation principal component regression, Statist. Appl. Gene. Mole. Biol. 6(1) (2007), Article 16.

■