Title
Partial Cox Regression Analysis for High-Dimensional Microarray Gene Expression Data

Authors
Li, Hongzhe
Gui, Jiang

Publication Date
2004-03-01

Peer reviewed

BIOINFORMATICS

Vol. 00 no. 00 2004
Pages 1–8

Partial Cox Regression Analysis for High-Dimensional Microarray Gene Expression Data

Hongzhe Li and Jiang Gui

Rowe Program in Human Genetics and Department of Statistics, University of California, Davis, CA 95616,USA

ABSTRACT

Motivation: An important application of microarray technology is to predict various clinical phenotypes based on the gene expression profile. Success has been demonstrated in molecular classification of cancer in which different types of cancer serve as categorical outcome variable. However, there has been less research in linking gene expression profile to censored survival outcome such as patients' overall survival time or time to cancer relapse. In this paper, we develop a partial Cox regression method for constructing mutually uncorrelated components based on microarray gene expression data for predicting the survival of future patients.

Results: The proposed partial Cox regression method involves constructing predictive components by repeated least square fitting of residuals and Cox regression fitting. The key difference from the standard principal components Cox regression analysis is that in constructing the predictive components, our method utilizes the observed survival/censoring information. We also propose to apply the time dependent receiver operating characteristic curve analysis to evaluate the results. We applied our methods to a publicly available data set of diffuse large B-cell lymphoma. The results indicated that combining the partial Cox regression method with principal components analysis results in parsimonious model with fewer components and better predictive performance. We conclude that the proposed partial Cox regression method can be very useful in building a parsimonious predictive model that can accurately predict the survival of future patients based on the gene expression profile and survival times of previous patients.

Availability: R codes are available upon request.

Contact: hli@ucdavis.edu

INTRODUCTION

DNA microarray technology permits simultaneous measurements of expression levels for thousands of genes, which offers the possibility of a powerful, genome-wide approach to the genetic basis of different types of tumors. The genome-wide expression profiles can be used for molecular classification of cancers, for studying varying levels of drug responses in the area of pharmacogenomics and for predicting different patients' clinical outcomes. The problem of cancer class prediction using the gene expression data, which can be formulated as predicting binary or multi-category outcomes, has been studied extensively and has been demonstrated great promise in recent years (Golub *et al.*, 1999; Alon *et al.*, 1999; Garber *et al.*, 2001; Rosenwald *et al*., 2002). However, there has been less development in relating gene expression profiles to censored survival phenotypes such as time to cancer recurrence or death due to cancer. Due to large variability in time to cancer recurrence among cancer patients, studying possibly censored survival phenotypes can be more informative than treating the phenotypes as binary or categorical variables.

From a statistical perspective, one challenge to studying time to event outcome results from right censoring during patient followup, since some patients may still be event-free. These patients are termed right-censored, and for these patients, we only know that the time to event is greater than the time of last followup. An additional challenge is in the microarray gene expression data itself. Microarray gene expression data is often measured with great deal of background, irrelevant readings, and the sample size of tissues or patients is usually very small compared to the number of genes measured by expression arrays. In addition, there is a potential of high collinearity of the gene expression levels among many genes. Censoring of patients proves difficult when compared to binary or continuous phenotypes. A frequent approach to relating gene expression profiles to survival phenotypes is to first group tumor samples into several clusters based on gene expression patterns across many genes, and then to use the Kaplan-Meier (KM) curve or the log-rank test to indicate whether there is a difference in survival time among different tumor groups. Another approach is to cluster genes first based on their expression across different samples, and use the sample averages of the gene expression levels in a Cox model (Cox, 1972) for survival outcome. Both approaches suffer the

drawback that the phenotype information is completely ignored in the clustering step and therefore may result in loss of efficiency. Additionally, results will potentially be sensitive to the clustering algorithm and distance metrics employed, as well as the number of clusters chosen.

Perhaps the most developed technique in relating gene expression profiles to phenotypes is the gene harvesting procedure of Hastie *et al.* (2001). This is a forward stepwise regression method that can be applied to a spectrum of outcome types, including survival data, in which case the stepwise regression corresponds to a stepwise Cox model. The central strategy of gene harvesting, and what distinguishes it from conventional forward stepwise techniques, is to initially cluster all genes via hierarchical clustering, and then to consider the average expression profiles from all of the clusters in the resulting dendrogram as additional covariates (beyond the individual gene expression profiles). The number of terms retained is determined by cross-validation. By using clusters as covariates, selection of correlated sets of genes is favored, which in turn potentially reduces overfitting. However, gene harvesting is sensitive to clustering procedure specifications and more importantly as demonstrated by Segal *et al.* (2003), gene harvesting admits artifactual solutions. These arise as a result of the nature and extent of the basis expansion represented by the additional covariates in the typically small sample size settings.

Another approach to dealing with the problem of high-dimensionality and multi-collinearity is through penalized maximum partial likelihood estimation. Li and Luan (2003) developed a penalized estimation procedure for the Cox model using kernels. The procedure is in fact reduced to the L_2 penalized estimation of the standard Cox model with linear predictors when the inner product kernel is used. However, the paper did not provide a formal practical procedure for choosing different kernels or the corresponding tuning parameters.

Partial least squares (PLS) (Wold, 1966) is a method of constructing linear regression equations by constructing new explanatory variables or factors or components using linear combinations of the original variables. The methods can be effectively applied to the settings where the number of explanatory variables is very large (Wold, 1966; Garthwaite, 1994). Different from the principal components (PC) analysis, this method makes use of the response variable in constructing the latent components. The method identifies linear combinations of the original variables as predictors and uses these linear components in the standard regression analysis. Nguyen and Rock (2002) applied the standard PLS methods of Wold (1966) directly to survival data and used the resulted PLS components in the Cox model for predicting survival time. The approach did not really generalize the linear PLS to censored survival data, but applied it directly. However, such direct application of the Wold algorithm to survival data is questionable and indeed does not seem reasonable since it treats both time to event and time to censoring as the same in the linear PLS procedure. Park *et al.* (2002) reformulated the Cox model as a Poisson regression and applied the formulation of PLS of Marx (1996) for the generalized linear models to derive the PLS components. However, such reformulation introduces many additional nuisance parameters and when the number of covariates is large, the algorithm may fail to converge. In addition, Park *et al.* (2002) did not evaluate how well the model predicts the survival of a future patient.

In this paper, we propose a different extension of PLS to the censored survival data in the framework of the Cox model by providing a parallel algorithm for constructing the latent components. The algorithm involves constructing predictive components by repeated least square fitting of residuals and Cox regression fitting. These components can then be used in the Cox model for building a useful predictive model for survival. We call this method the partial Cox regression (PCR) method. In addition, we propose to employ the time dependent receiver operating characteristic (ROC) curve (Heagerty *et al.*, 2000) to assess how well the model predicts the survival. The rest of the paper is organized as follows: we first present the PCR methods for constructing the predictive components for the Cox model. We then apply the methods to analysis of the diffuse large B-cell lymphoma (DLBCL) survival data set of Rosenwald *et al.* (2002) and compare their performance in prediction by splitting the data into training and testing sets and by using the concept of the time-dependent ROC curves. We conclude the paper with a discussion of the results presented in this paper.

METHODS AND RESULTS

An algorithm for constructing the partial Cox regression model

Suppose that we have a sample size of n from which to estimate the relationship between the survival time and the gene expression levels $X_1, \cdots, X_p$ of p genes. Due to censoring, for $i = 1, \cdots, n$, the ith datum in the sample is denoted by $(t_i, \delta_i, x_{i1}, x_{i2}, \cdots, x_{ip})$, where δ_i is the censoring indicator and t_i is the survival time if $\delta_i = 1$ or censoring time if $\delta_i = 0$, and $\{x_{i1}, x_{i2}, \cdots, x_{ip}\}'$ is the vector of the gene expression level of p genes for the ith sample. Let $x_j = \{x_{1j}, x_{2j}, \cdots, x_{nj}\}'$ be the vector of gene expression levels of the jth gene over n samples. Our aim is to build the following Cox regression model for the hazard of cancer recurrence or death at time t

$$\begin{aligned} \lambda(t) &= \lambda_0(t)\exp(\beta_1 T_1 + \beta_2 T_2 + \cdots + \beta_k T_k) \\ &= \lambda_0(t)\exp(f(X)), \end{aligned} \quad (1)$$

where each component T_k and the risk score function $f(X)$ is a linear combination of $X = \{X_1, X_2, \cdots, X_p\}$. In this model, $\lambda_0(t)$ is an unspecified baseline hazard function.

Following the idea of PLS (Garthwaite, 1994), we adopts the principle that when considering the relationship between

the hazard and some specified X variable, other X variables are not allowed to influence the relationship directly but are only allowed to influence it through the components T_k. Particularly, we propose to develop the following procedure to determine the components sequentially. To construct the first component, first define

$$V_{1j} = X_j - \bar{x}_{.j} \tag{2}$$

where $\bar{x}_{.j} = 1/n \sum_{i=1}^{n} x_{ij}$. The vector of the sample values of V_{1j} is $v_{1j} = \{v_{11j}, \cdots, v_{n1j}\} = x_j - \bar{x}_{.j}$ and therefore the sample mean of V_{1j} is zero. Then for each gene j, we fit the following Cox model

$$\lambda(t) = \lambda_0(t) \exp(\beta_{1j} V_{1j})$$

based on the sample value of V_{1j} and obtain the maximum partial likelihood estimate (MPLE) of β_{1j}, denoted by $\hat{\beta}_{1j}$. Then each $\hat{\beta}_{1j} V_{1j}$ provides an estimate of the log relative hazard in the hazard function. To reconcile these estimates, we set T_1 equal to the weighted average, so

$$T_1 = \sum_{j=1}^{p} w_{1j} \hat{\beta}_{1j} V_{1j}, \tag{3}$$

where w_{1j} is a weight with $\sum w_{1j} = 1$. It is easy to see that the sample mean of T_1 is also zero. Note that T_1 is a special type of compound covariates advocated by Tukey (1993) in a clinical trial setting when there are many covariates.

Note that the X variables potentially contain further useful information for predicting the risk of recurrence or survival. So one should not stop at the T_1 step. The information in X_j that is not in T_1 may be estimated by residuals from a regression of V_{1j} (equivalently, X_j) on T_1, and denote the residual as V_{2j}, which can be written as

$$V_{2j} = V_{1j} - \frac{V_{1j}^{'} T_1}{T_1^{'} T_1} T_1. \tag{4}$$

Similarly, the contribution of the residual information in V_{1j} to the variability in the risk of recurrence or death after adjusting T_1 can be estimated by performing the following Cox regression analysis,

$$\lambda(t) = \lambda_0(t) \exp(\beta_1 T_1 + \beta_{2j} V_{2j})$$

and denote the MPLE of β_{2j} as $\hat{\beta}_{2j}$. Then each $\hat{\beta}_{2j} V_{2j}$ provides an estimate of the log relative hazard after adjusting for T_1. We define the next component T_2 as

$$T_2 = \sum_{j=1}^{p} w_{2j} \hat{\beta}_{2j} V_{2j}, \tag{5}$$

where w_{2j}s are the weights. Note here the sample mean of T_2 is zero.

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

$$V_{(i+1)j} = V_{ij} - \frac{V_{ij}^{'} T_i}{T_i^{'} T_i} T_i. \tag{6}$$

Then perform the following Cox regression analysis for each j,

$$\lambda(t) = \lambda_0(t) \exp(\beta_1 T_1 + \beta_2 T_2 + \cdots, \beta_i T_i + \beta_{(i+1)j} V_{(i+1)j})$$

and denote the MPLE of $\beta_{(i+1)j}$ as $\hat{\beta}_{(i+1)j}$. We then construct T_{i+1} as

$$T_{i+1} = \sum_{j=1}^{p} w_{(i+1)j} \hat{\beta}_{(i+1)j} V_{(i+1)j}, \tag{7}$$

where $w_{(i+1)j}$s are the weights.

We call the above procedure the partial Cox regression (PCR) and the components $T_1, T_2, \cdots$ constructed the PCR components. It is easy to verify that these components are mutually uncorrelated with sample mean of zero. In addition, it is also easy to verify that the procedure gives precisely the PLS in the case of linear model and when the simple linear regression is used in place of the Cox regression. In constructing the PCR weights, we let $w_{ij} \propto var(V_{ij})$, i.e., variables with large variance are given larger weights. This weight was also suggested in the PLS literatures (e.g., Garthwaite, 1994). An alternative is to use equal weights of $1/p$, which aims to spread the load among the X variables in making predictions. In this paper, we use only the variance weights.

After the components $T_1, T_2, \cdots, T_k$ are determined, model (1) is used for estimating the hazard function by the standard partial likelihood method. After an estimate of the regression model (1) has been determined, equations (3)-(7) can be used to write the risk score function $f(X)$ in model (1) in terms of original variables X, rather than the components $T_1, \cdots, T_k$, i.e.,

$$f(X) = \sum_{j=1}^{p} \beta_j^* V_{1j} = \sum_{j=1}^{p} \beta_j^* (X_j - \bar{x}_{.j}),$$

for some coefficients β_j^*. This can then be used for estimating the hazard function for future samples on the basis of their X values. Note that the sample mean of the score function $f(X)$ is zero. Finally, by examining the coefficients of X values in the final model, one can rank the gene effects by the absolute values of the coefficients.

The time dependent ROC curves and area under the curves

For classification or linear regression problem, one can use cross-validated misclassification rate or the sum of square residuals as a criteria to assess how well the model predicts the outcomes. However, due to censoring neither criteria can be used in censoring data regression setting. In order to assess how well the model predicts the survival outcome, we propose to employ the idea of time dependent ROC curve for censored data and area under the curve (AUC) as our criteria. These methods were recently developed by Heagerty *et al.* (2000) in the context of the medical diagnosis. For a given risk score function $f(X)$, we can define time dependent sensitivity and specificity functions as

$$\begin{aligned} \text{sensitivity}(c, t|f(X)) &= Pr\{f(X) > c|\delta(t) = 1\}, \\ \text{specificity}(c, t|f(X)) &= Pr\{f(X) \leq c|\delta(t) = 0\}, \end{aligned}$$

with c being the cutoff value and t being the time and define the corresponding ROC($t|f(X)$) curve for any time t as the plot of sensitivity$(c, t|f(X))$ vs $1 -$ specificity$(c, t|f(X))$ with cutoff point c varying, and the AUC as the area under the ROC($t|f(X)$) curve, denoted by AUC($t|f(X)$). Here $\delta(t)$ is the event indicator at time t. A nearest neighbor estimator for the bivariate distribution function is used for estimating these conditional probabilities accounting for possible censoring (Akritas, 1994). Note that larger AUC at time t based on a risk score function $f(X)$ indicates better predictability of time to event at time t as measured by sensitivity and specificity evaluated at time t. In our application presented in the next section, we study several different methods of constructing the risk score function $f(X)$ in the Cox model (1) and compare their predictive performance based on the AUCs.

Application to Real Data Set

To demonstrate the proposed PCR methods, we re-analyzed a recently published data set of DLBCL by Rosenwald *et al.* (2002). This data set includes a total of 240 patients with DLBCL, including 138 patient deaths during the followups with median death time of 2.8 years. The gene expression measurements of 7,399 genes are available for analysis. We applied a nearest neighbor technique to estimate those missing values. Specifically, for each gene, we first identified 8 genes which are the nearest neighbors according to Euclidean distance. We then filled the missing with the average of the nearest neighbors. Rosenwald *et al.* divided the 240 patients into a training set of 160 patients and a validation set or testing set of 80 patients and built a multivariate Cox model. The variables in the Cox model included the average gene expression levels of smaller sets of genes in four different gene expression signatures together with the gene expression level of BMP6. It should be noted that in order to select the gene expression signatures, they performed a hierarchical clustering analysis for genes across all the samples (including both testing and training samples). In another words, the information from the testing samples was indeed used in the clustering step in their analysis.

A comparison with principal components analysis We considered to construct the PCR components using all the 7399 genes and using only those genes which are significant in a univariate Cox regression analysis at the level of 0.05 (1836 genes) and 0.01 (506 genes). We built the PCR components based on the gene expression data of these genes using the 160 patients in the training data set defined in Rosenwald *et al.* (2002). Table 1 shows the p-values from univariate Cox regression analysis for the first ten PCR components constructed with different number of genes. We observed that the first seven PCR components are significantly associated with the risk of death at the 0.05 level, all other PCR components are not significant. As a comparison, we list in the same table the p-values for the first 10 PCs with the largest variances. Note that there are only two PCs among the top 10 PCs with the largest variances which are significantly associated with the risk of survival. This demonstrates that the PCR components are more significantly related to survival and fewer PCR components are needed in order to explain the variability in survival than the PCs.

We also considered to perform the PCR analysis by performing principal components analysis first on the gene expression matrix. For the training set of 160 individuals, there are a total of 159 PCs after centering the data. Treating these 159 PCs as a set of new variables, we constructed the PCR components using our proposed methods. We call such components PC-PCR components. Table 1 presents the p-values for the first 10 PC-PCR components in the univariate Cox regression analysis. We observed that only the first 3 or 4 PCR components are strongly associated with the survival in the univariate Cox regression analysis, indicating that the PC-PCR components result in further dimension reduction comparing to the PC or PCR analysis.

Evaluation of the predictive performance of the models To examine how well the Cox model with PCR components, principal components and PC-PCR components predict the survival of future patients, we built several models using training data and predicted the survival for patients in the testing data set. All the components were constructed using all the 7399 genes. We used only the components which are significantly related to the risk of death at the p=0.05 level in the univariate Cox regression analysis to build the final predictive model. We used the mean of the risk scores of the patients in the training set, which is zero, as the cutoff point to divide the patients into high and low risk groups. Based on the estimated coefficients, we estimated the risk scores for patients in the testing data set and divided these patients into two risk groups. Figure 1 (a)-(c) shows the Kaplan-Meier survival curves for the two risk groups defined by three different models. Although all three models give significant difference in the risk

Table 1. Results (p-values) for univariate Cox regression analysis for the first 10 PCR components, first 10 PC-PCR components and the top 10 PCs with the largest variances built using 7399, 1836 and 506 genes.

	Number of genes used								
	7399			1836			506		
	PCR	PC	PC-PCR	PCR	PC	PC-PCR	PCR	PC	PC-PCR
1	8E-13	0.373	0	7E-13	2E-08	0	3E-13	2E-11	0
2	4E-09	0.505	3E-11	2E-06	2E-03	3E-13	1E-04	0.189	2E-12
3	9E-10	1E-05	2E-03	3E-11	0.565	2E-05	6E-09	0.481	2E-07
4	2E-05	0.907	0.233	4E-06	0.429	0.047	1E-05	0.014	4E-03
5	2E-03	0.942	0.223	8E-04	0.686	0.280	3E-04	0.846	0.086
6	4E-03	0.873	0.280	0.024	0.454	0.432	2E-03	0.784	0.234
7	0.031	0.784	0.829	0.013	0.358	0.551	0.027	0.486	0.190
8	0.078	0.553	0.824	0.121	0.064	0.716	0.056	0.951	0.546
9	0.298	0.421	0.652	0.499	0.376	0.839	0.241	0.124	0.516
10	0.251	0.029	0.899	0.356	0.365	0.518	0.143	0.112	0.946

of death between the high and low risk groups, the Cox model with PC-PCR components seems to give the best separation between the two risk groups. The median survival times are 10 years and 2 years respectively for the low and high-risk groups defined by the PC-PCR model (p=0.0033). Figure 1 (d) shows the time-dependent area under the ROC curves based on the estimated risk scores of the patients in the testing data set. We observe that the Cox model with PC-PCR components gave the best predictive performance on the testing patients. The AUCs in the first 10 years are close to 65%. On the other hand, the Cox model with PCs results in very low AUCs, less than 60%.

To validate the better performance of using PC-PCR components observed in analysis of the training/testing sets defined by Rosenwald *et al.* (2002), we conducted the following training/testing sets analysis. In the absence of genuine test sets, the prediction performance of difference models are compared based on random division of the dataset into a leaning set and a testing set. We chose to use 2:1 scheme by choosing 160 patients as training set and 80 patients as testing set. For each learning set (LS)/testing set (TS) run, we consider to use the PC and PC-PCR procedures build the predictive model using all the 7399 genes. For a chosen number of genes, predictive models are constructed using the LS. The predictive model is then applied to the TS to obtain the risk scores. We then divide the patients in the TS into high and low risk group based on whether the score is positive or negative and calculate the p value for testing the risk difference between the two groups. We considered to estimate the risk scores using PCs and PC-PCR components. Each LS/TS run yields a set of p-values for the TS and the results are summarized in Table 2 for each model over a total of 100 runs. For a p-value of 0.01, we observed that 65 out of 100 runs showed a significant difference in risk between the two risk groups for the testing data set based on the Cox model with PC-PCR components, as compared to only 37 times using the PCs only. For a p-value of 0.05, 92 out 100 runs resulted in significance difference in survival between the two risk groups define by the PC-PCR components, as compared to only 67 times based on the PCs alone. It is also interesting to note that for all the 100 runs, the model with PC-PCR components uses only 3 components in the model, as compared to using about 10 PCs in the Cox regression model. These results indicate that the Cox model with a small number of PC-PCR components can give better predictive performance than the Cox model with a large number of principal components.

Table 2. Summary of significance results based on 100 training/testing sets. The number in the table is the number of times when the two risk groups in the testing data sets have significant difference in risk for the chosen p-value. PC: Cox model with PCs; PC-PCR: Cox model with PC-PCR components.

p-value	10^{-5}	10^{-4}	10^{-3}	10^{-2}	0.05	0.1
PC	0	1	14	37	67	76
PC-PCR	4	8	31	65	92	94

Comparison with the L_2 penalized estimation Finally, as a comparison, Figure 2 shows the results from the penalized procedure of Li and Luan (2003). While the estimated survival curves for the high and low risk groups in the testing data set are quite comparable to the those obtained by the PCR or PC-PCR procedures (see Figure 1 (a) and (c)), the AUC seems a bit lower than that estimated from the PC-PCR procedure. Of course, it is impossible to conclude which method performs better by analyzing only one data set. However, the proposed PCR procedure does have computational advantage over the L_2 penalized procedure since it does not require inversion of large matrix.

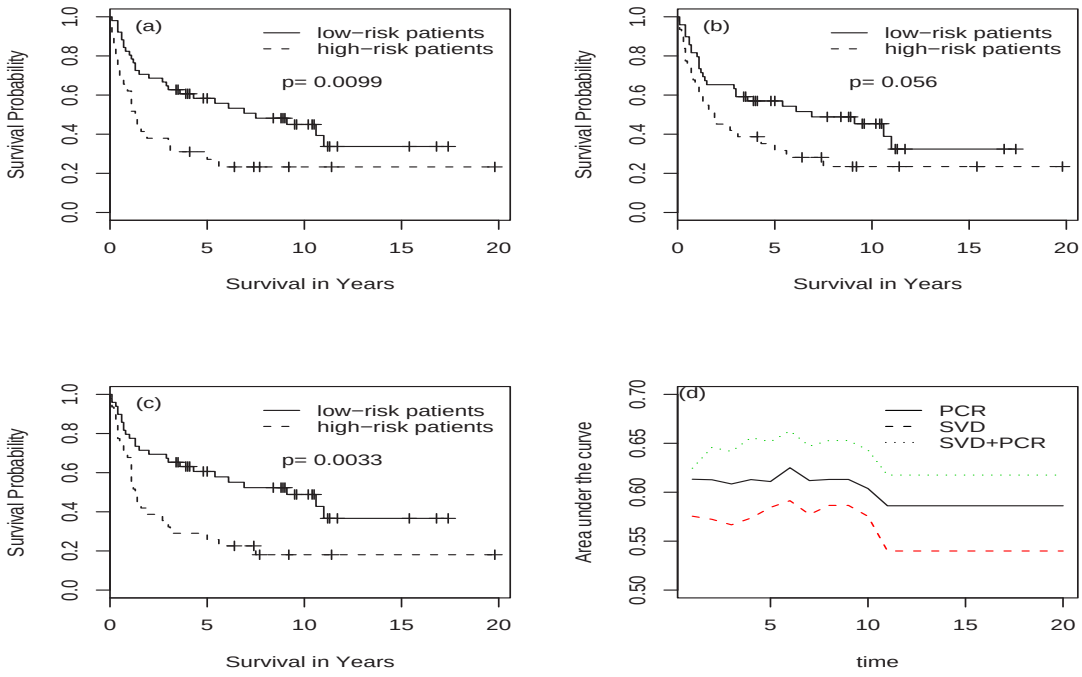

Fig. 1. Results from analysis of the DLBCL data set. (a)-(c) Survival curves for two groups of patients in the testing set defined by having positive (high risk) or negative (low risk) risk scores. The scores are estimated based on the model built from the training data set using all the 7399 genes. (a) Results based on the Cox model with 7 PCR components; (b) results based on the Cox model with 12 PCs; (c) results based on the Cox model with 3 PC-PCR components. (d) The area under the curves for the three models based on the estimated risk scores for patients in the testing data set.

CONCLUSIONS AND DISCUSSION

It is clinically relevant and very important to predict patient's time to cancer relapse or time to death due to cancer after treatment using gene expression profiles of the cancerous cells prior to the treatment. Powerful statistical methods for such prediction allow microarray gene expression data to be used efficiently. In this paper, we have proposed to develop the partial Cox regression method for censored survival data in order to construct predictive components for survival using microarray gene expression data. The model searches for the genes whose expression levels are related to survival phenotypes and identifies the optimal combinations of the gene expression data in predicting the risk of cancer recurrence or death. Since the risk of cancer recurrence or death due to cancer may result from the interplay between many genes, methods which can utilize data of many genes, as in the case of our proposed model, are expected to show better performance in predicting risk. We have demonstrated the applicability of our methods by analyzing time to death of diffuse large B-cell lymphoma patients and obtained satisfactory results, as evaluated by both applying the model to the test data set and time dependent ROC curves. Our analysis of the DLBCL data set shows that by combining the principal components analysis with our proposed PCR analysis we obtain the best predictive results for the testing data sets.

Like the PC regression analysis, the major objective of PCR analysis is to replace the p-dimensional gene expression levels by a much smaller number k of PCR components. Because of the way of constructing the PCRs, the components constructed in earlier steps should be more predictive to the survival time than those constructed in later steps. However, determining the number of PCR components k used in the Cox regression model (1) is not obvious. In order to eliminate large variances due to multi-collinearity it is essential to delete all those components whose variances are very small, but at the same time, it is undesirable to delete components that have high correlation with the outcomes. In this paper, we propose to choose the first k PCR or PC-PCR components which are significant in a univariate Cox regression analysis to build the final predictive model since these k components are most predictive and have large variances. This simple method of components selection seems to work well for the DLBCL data set in term of predicting clinically relevant risk groups. We also examined other methods of components selection

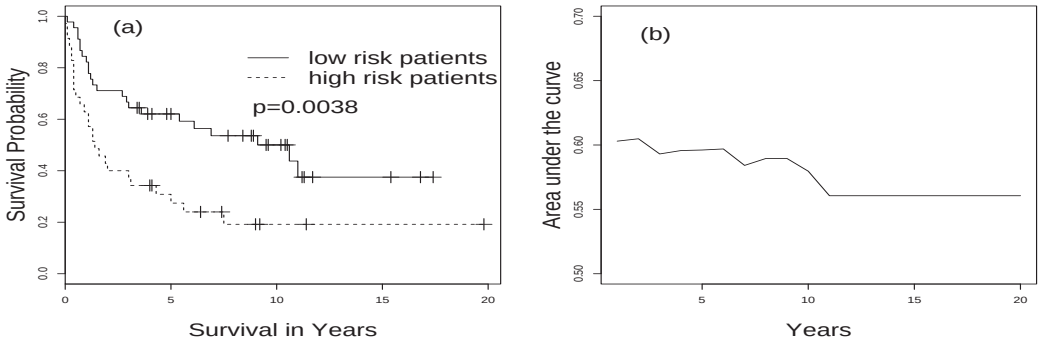

Fig. 2. Results from analysis of the DLBCL data set based on the penalized procedure of Li and Luan (2003). (a) Survival curves for two groups of patients in the testing set defined by having positive (high risk) or negative (low risk) risk scores. The scores are estimated based on the model built from the training data set using all the 7399 genes. (b) The area under the curves for the model based on the estimated risk scores for patients in the testing data set.

such as AICs and cross-validated partial likelihood (CVPL) (Huang and Harrington, 2002) methods and observed that the AICs tend to select fewer components and CVPL tends to select more components. Neither methods resulted in better predictive performance than the simple method used in this paper (details not shown).

Our proposed procedure is very different from some previous attempts in extending the PLS to the censored survival data (e.g., Nguyen and Rock, 2002; Park *et al.*, 2002). Instead of working directly with the Cox model, these previous extensions reduce the problem to a linear model problem and then use the PLS idea for the linear models to identify the predictive components. These components are then used in the Cox model as a set of covariates. Our approach constructs the components directly based on the Cox model and the components constructed in earlier steps are expected to be more predictive than those constructed in later steps. It would be interesting to compare the predictive performances of these different extensions. It is also important to point out that all these extensions assume a proportional hazard model, which is the most popular model for censored survival data. However, the proportional hazards assumption may not hold for gene expression data or for all diseases. As an alternative, we can consider the accelerated failure time models or more general semi-parametric transformation models (Wei, 1992; Cheng *et al.*, 1995). We are currently pursuing these alternative models.

Although the proposed PCR analysis has no computational or methodological limitation in term of the number of genes that can be used in the prediction of patient's time to clinical event, since not all genes will be relevant to predicting censored survival phenotypes, we would expect better prediction results using only genes that are related to the phenotypes. However, we found in application to the lymphoma data set that selecting genes based on univariate analysis may result in poor predictive performance since such gene selection procedure totally ignores possible combinatorial effects of gene expressions on the risk of death. One interesting idea is to iteratively select genes based on the coefficients in the final Cox regression models, i.e., iteratively removing those genes with small coefficients and refitting the model. This deserves further investigation. Another possibility of selecting the relevant genes for building a predictive model is through the lasso type (Tibshirani, 1997) L_1 penalized maximum partial likelihood estimation subject to the sum of the absolute value of the coefficients being less than a constant. Our preliminary results from this lasso type estimation procedure in the high-dimensional settings are very encouraging and final results will be reported in another paper.

In summary, since the number of genes is usually much larger than the number of patients, it is crucial to develop methods for efficient dimension reduction. We have developed the partial Cox regression method for identifying uncorrelated predictive components for censored survival phenotypes. Our analysis of the diffuse large B-cell lymphoma data set shows that by combining the principal components analysis with our proposed PCR analysis we obtain the best predictive results for the testing data set. The proposed PC-PCR procedure can be very useful in building a parsimonious predictive model that can be used for classifying the future patients into clinically relevant high and low risk groups based on the gene expression profile and survival times of previous patients.

ACKNOWLEDGMENT

This work was supported by an NIH grant (ES09911).

REFERENCES

Akritas, M.G. (1994) Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics*, **22**,

1299-1327.

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, **96(12)**,6745- 6750.

Cheng, S.C., Wei, L.J., Ying, Z. (1995) Analysis of transformation models with censored data. *Biometrika*, **82**, 835-45.

Cox, D.R. (1972) Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B*, **34**, 187-220.

Garber, M.E., Troyanskaya, O.G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G.D., Perou, C.M., Whyte, R.I., Altman, R.B., Brown, P.O., Botstein, D. and Petersen, I. (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proceeding of National Academy of Science USA*, **98**, 13784-13789.

Garthwaite, P.H. (1994) An interpretation of partial least squares. *Journal of the American Statistical Association*, **89**, 122-127.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S. (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* , **286**,531-537.

Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A.A., Levy, R., Staudt, L., Chan, W.C., Botstein, D. and Brown, P. (2001) 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, **2**, research0003.1-003.21

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

Huang, J. and Harrington, D. (2002) Penalized Partial Likelihood Regression for Right-Censored Data with Bootstrap Selection of the Penalty Parameter. *Biometrics*, **58**,781-791.

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

Marx, B.D. (1996) Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics*, **38**, 374-381.

Nguyen, D., Rocke, D.M. (2002) Partial Least Squares Proportional Hazard Regression for Application to DNA Microarray Data. *Bioinformatics*, **18**, 1625-1632.

Park, P.J., Tian, L., Kohane, I.S. (2002) Linking Expression Data with Patient Survival Times Using Partial Least Squares. *Bioinformatics*, **18**, S120-127.

Rosenwald, A., Wright, G., Chan, W., Connors, J.M., Campo, E., Fisher, R., Gascoyne, R.D., Muller-Hermelink, K., Smeland, E.B. and Staut, L.M. (2002) The Use of Molecular Profling to Predict Survival After Themotheropy for Diffuse Large-B-Cell Lymphoma. *The New England Journal of Medicine*, **346**,1937-1947.

Segal, M.R., Dahlquist, K.D., Conklin, B.R. (2003) Regression approaches for microarray data analysis. *Journal of Computational Biology*, **10**: 961-980.

Tibshirani, R. (1997) The Lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385-395.

Tukey, J.W. (1993) Tighening the clinical trial. *Controled Clinical Trials*, **14**, 266-285.

Wei, L.J. (1992) The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. *Statistics in Medicine*, **11**, 1871-1879.

Wold, H. (1966) Estimation of principal components and related models by iterative least squares. In P.R. Krishaiaah (Ed.). *Multivariate Analysis*, New Academic Press, pp391-420.