Statistical Applications in Genetics and Molecular Biology

Volume 6, *Issue* 1 2007 *Article* 16

Cox Survival Analysis of Microarray Gene Expression Data Using Correlation Principal Component Regression

Qiang Zhao* Jianguo Sun†

*Texas State University, qiang.zhao@txstate.edu

†University of Missouri-Columbia, sunj@missouri.edu

Cox Survival Analysis of Microarray Gene Expression Data Using Correlation Principal Component Regression*

Qiang Zhao and Jianguo Sun

Abstract

Statistical analysis of microarray gene expression data has recently attracted a great deal of attention. One problem of interest is to relate genes to survival outcomes of patients with the purpose of building regression models for the prediction of future patients' survival based on their gene expression data. For this, several authors have discussed the use of the proportional hazards or Cox model after reducing the dimension of the gene expression data. This paper presents a new approach to conduct the Cox survival analysis of microarray gene expression data with the focus on models' predictive ability. The method modifies the correlation principal component regression (Sun, 1995) to handle the censoring problem of survival data. The results based on simulated data and a set of publicly available data on diffuse large B-cell lymphoma show that the proposed method works well in terms of models' robustness and predictive ability in comparison with some existing partial least squares approaches. Also, the new approach is simpler and easy to implement.

KEYWORDS: survival analysis, Cox model, microarray gene expression data, correlation principal component regression

---

*The authors wish to thank two referees for their constructive comments, which greatly improved the paper.

1 INTRODUCTION

One of the major advantages of microarray technology is that it allows simultaneous monitoring or measurements of expression levels of thousands of genes. In consequence, this has produced a huge amount of high-dimension gene expression data and led to a great deal of research effort in statistics spent on the development of statistical approaches appropriate for the analysis of such data. In the analysis of microarray gene expression data, several questions have drawn a lot of attention. One is the classification or cluster analysis of gene expression data with purposes of, for example, classifying samples into categories such as types or grades of tumors. The developed methods for this have helped, for example, identify previously undetected subtypes of cancer. Another question is how to identify differentially expressed genes. This could be a single gene or a group of genes related to or determining a clinical outcome of interest. Also the genes to be identified could be unknown one or more specific groups of genes to the researchers. This paper discusses survival analysis of microarray gene expression data with purposes of relating gene expression information or patterns to clinical variables representing times to certain events such as survival times of cancer patients.

In the analysis of microarray gene expression data, one of the main challenges to statisticians is the number of variables (genes) needed to be dealt with, which of course could be thousands, being larger than or far exceeding the number of samples. It is well-known that this makes invalid or inappropriate most of existing statistical methods such as commonly used approaches for linear or nonlinear regression analysis and survival analysis. The same problem was encountered in chemometrics where one of the main purposes is to predict variables such as the quality of a product or human blood analytes using spectrum information that can be measured at hundreds or thousands of different wavelengthes simultaneously. For spectrum analysis, a number of regression techniques that allow a large number of variables have been developed and investigated. These include the standard principal component regression (SPCR), linear and nonlinear partial least squares (PLS) approaches and correlation principal component regression (CPCR). The CPCR has been successfully used in chemometrics to achieve a simpler and better prediction model than SPCR and PLS regression (Sun, 1995). Note that all of statistical methods developed in chemometrics are for the purpose of regression analysis, not survival analysis.

For the survival analysis of microarray gene expression data, in addition to facing a large number of variables, one has to deal with censoring, a common and unique feature of survival data. Due to this and the special interest in sur-

vival analysis, models commonly used in survival analysis are quite different from those usually employed in regression analysis. Among them, the most widely used model is the proportional hazards or Cox model for right-censored data and several authors have investigated the use of its combination with the principal component analysis (PCA) or PLS for the survival analysis of gene expression data. For example, Nguyen and Rocke (2002) suggested to first apply the PLS approach to gene expression and survival data to reduce the dimension and then fit the Cox model to the resulting PLS components with the survival data. Park et al. (2002) proposed a method to transform the Cox survival analysis to a generalized linear regression problem and then apply the PLS technique to the transformed problem. Li and Gui (2004) also developed a clever analysis procedure that allows one to perform the PLS analysis under the Cox model. Note that the set of PLS components in these methods are determined using a complicated algorithm and the results of both gene expression and survival data. This could make the study of the properties of these methods more challenging. Of course, one could directly use the components resulting from PCA for the Cox survival analysis. However, this has been proved to tend to give less satisfactory results due to the fact that the components obtained may not be related to patients' survival (Bair and Tibshirani, 2004). Li and Li (2004) proposed to first select a subset of principal components and then apply a sufficient dimension reduction method to obtain fewer number covariates in the Cox model, which requires an additional assumption. Other approaches not using principle components or PLS components include penalized partial likelihood approaches (Gui and Li, 2005) and a boosting procedure using smoothing splines (Li and Luan, 2005).

In this paper, we present a new approach for the Cox survival analysis. Instead of PCA or PLS, the approach makes use of a variant of CPCR, which will still be referred as CPCR. Details of the methodology are presented in Section 2. The section also discusses the criteria for the selection and assessment of the prediction ability of models. In Section 3, for illustration, we apply the proposed method to a set of gene expression data with survival information on the diffuse large B-cell lymphoma (DLBCL) patients used by other authors to illustrate their methods. We also compare the proposed method with existing methods based on simulated data. The results suggest that the new method gives better prediction models and is more robust in situations considered. We conclude the paper with some remarks in Section 4.

DOI: 10.2202/1544-6115.1153

2 METHODS

2.1 Correlation Principal Component Regression

The SPCR has been used as a regression technique for a long time mainly for the situation where there exist strong multicollinearities among predictor variables and/or the number of predictor variables is close to or larger than the sample size such as in gene expression data. Multicollinearities could cause the variances of some of the estimated coefficients being very large and, thus, lead to unstable and potentially misleading estimates of the regression equations. CPCR is a variant of the SPCR and was developed particularly for the case where the prediction is of main interest. The idea behind SPCR is to regress a response variable on principal components ordered by their variability rather than on original predictor variables, while the idea behind CPCR is to regress a response variable on principal components ordered by their correlations with the response variable. In both cases, the number of principal components used in the final regression model is typically much less than the number of predictor variables.

Let X $(n \times p)$ denote the gene expression level matrix with n samples and p genes (predictor variables) and Y $(n \times 1)$ the sample of a response variable. In SPCR, the first step is to find the principal components of X given by $Z = (z_1, ... z_K) = UD$, where $K = min\{p, n\}$, the z_l's are ordered by their variances from the largest to the smallest and $X = U D V^t$ is the singular value decomposition of X, where $U(n \times p)$ and $V(p \times p)$ are matrices with $U^tU = V^tV = VV^t = I(p \times p)$, an identity matrix, and $D(p \times p)$ is a diagonal matrix with eigenvalues of X^tX on diagonal positions. For a given integer A, the least squares regression is applied by regressing Y on $Z_A = (z_1, ..., z_A)$, the first A principal components of X. For CPCR, the first step is the same as in SPCR, but in the second step, the least square regression is applied through regressing Y on $Z_A^* = (z_1^*, ..., z_A^*)$, the first A principal components with the highest correlations with Y.

Note that the idea behind SPCR is to retain as much as possible the variation present in the data. When prediction is mainly concerned, retaining the variability is no longer of interest. Instead one would be mostly interested in the variability of a predicted value that is usually evaluated by prediction errors discussed below. In other words, SPCR assumes that the main information of interest is contained in the directions of the predictor space with high variations. In some situations, however, the high variations may be generated by sources that are not related to the response variable under study and it's not uncommon that a component with lower variance is a good predictor in

a regression model (Jolliffe, 2002). In these cases, it is natural to use only the principal components that represent relevant directions in the regression as in CPCR. In general, the major information about a response variable is often included in the principal components with intermediate variances, especially in the case where the ratio of the sample size to the number of genes is small. In this case, there is not enough information available to estimate the regression parameters in all directions and only the parameters in important directions should be estimated. Also there always exists noise in the measurement, and thus the more principal components used in the model, the larger noise is incorporated. This is partly the motivation for CPCR, which employs only the relevant principal components (with large correlations) and ignores the irrelevant principal components.

As mentioned earlier, PLS has been used in the Cox survival analysis of gene expression data. Similar to CPCR, it assumes that the information of interest about a response variable is mainly contained in the directions of the predictor space which have both large variations and high correlations with the response variable. Note that in all of SPCR, CPCR and PLS, a sequence of models is generated. The selection of an appropriate model will be discussed below.

2.2 Cox Survival Analysis

By Cox survival analysis, we mean the analysis of survival data using the proportional hazards or Cox regression model, the most commonly used model in survival analysis. Suppose that one observes right-censored survival data given by $\{ (Y_i = \min\{T_i, C_i\}, \delta_i = I(Y_i = T_i), X_i); i = 1, ..., n \}$, where T_i denotes the true survival time, C_i the censoring time independent of the T_i, and X_i the vector of covariates or genes associated with subject i in the study. Let $\lambda(t|X_i)$ denote the hazard function given covariates X_i, the probability that a subject with X_i fails at time t given that the subject has survived up to time t. The Cox model assumes that

$$\lambda(t|X_i) = \lambda_0(t) \exp(\beta' X_i) , \tag{1}$$

where $\lambda_0(t)$ is an unknown baseline hazard function corresponding to subject with $X_i = 0$ and β is the vector of regression parameters. The objectives are to make inference about β and predict future survival.

To estimate β, the most frequently used procedure is to find β that maximizes the so-called partial likelihood function

$$L_p(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(\beta' X_i)}{\sum_{j=1}^{n} I(y_j \geq y_i) \exp(\beta' X_j)} \right\}^{\delta_i} .$$

DOI: 10.2202/1544-6115.1153

Note that L_p does not involve $\lambda_0(t)$, a main advantage of the partial likelihood approach. Of course all methods developed for the Cox model assumes that the sample size n is larger than the dimension of the X_i's. In the next subsection, we discuss how to perform the Cox survival analysis for microarray gene expression data.

2.3 A Proposed Survival Analysis Procedure

The proposed procedure is similar to that given in Nguyen and Rocke (2002). Instead of using PLS, we propose to use a variant of CPCR to first reduce the dimension of gene expression data by finding the principal components and order them based on the strength of their association with the survival time. Specifically, let $X = (X_1, ..., X_n)'$ $(n \times p)$ denote the gene expression data matrix, Y $(n \times 1)$ the response vector of interest representing survival times to a certain event such as death due to cancer or censoring times, and δ $(n \times 1)$ the vector of censoring indicators. We apply CPCR to X, Y, and δ to obtain the principal components $Z^* = (z_1^*, ..., z_K^*)$, ordered from the smallest to largest based on the $p-$values from testing $\beta = 0$ under model (1) when regressing Y on the principal components individually. Note that we propose to use $p-$values instead of correlation coefficients between the true survival time and individual components because the coefficients cannot be computed due to censoring. For X with centered columns, $K = n - 1$. For each A $(1 \leq A \leq K)$, replacing predictor variables in model (1) with $Z_A^* = (z_1^*, ..., z_A^*)$ gives

$$\lambda(t|Z_A^*) = \lambda_0(t) \exp(\beta_A^{*'} Z_A^*) = \lambda_0(t) \exp(f(X)), \tag{2}$$

where the risk score function $f(X)$ is a linear combination of the original data matrix X. Then one applies the partial likelihood approach to obtain an estimate $\hat{\beta}_A^*$ in model (2). The estimate, say $\hat{\beta}_A$, of the original β in model (1) can then be obtained by transforming Z_A^* back to X.

Note that if model (2) consists of all the PCs, it is equivalent to one including of all genes, so the problems caused by multicollinearities have not gone away. Thus, one key part of the above procedure is the selection of A, the number of the components used in model building. For this, note that it usually depends on the purpose of the modeling. In general, the number of components (A) in a model could be regarded as an indicator of the complexity of the model, and the larger A is, the bigger noise is incorporated and the more sensitive the model is to outliers. In other words, if the prediction is of main interest, one would prefer a model with fewer number of components and smaller prediction error.

In order to assess how well a model predicts future survival outcomes, one can randomly set aside a set of observations in the dataset as validation data and use the remaining as training data for model fitting. Then the predictive performance of a fitted model can be evaluated using root mean squared error for prediction (RMSEP), area under a time dependent ROC curve (AUC), or correct prediction or classification rate if risk group information, such as stages in tumor development, of a subject is available.

To obtain the RMSEP, let R denote a random subset of $\{1, ..., n\}$ with m elements and $\hat{\beta}^*_{A,R}$ the partial likelihood estimator obtained based on training data $\{Y_R, Z^*_{A,R}\}$, where $\{Y_R, Z^*_{A,R}\}$ denotes $\{Y, Z^*_A\}$ with the information on subjects $i \epsilon R$ removed. That is, $\hat{\beta}^*_{A,R}$ is $\hat{\beta}^*_A$ with the information from the subjects $i \, \epsilon \, R$ removed. For each $i \, \epsilon \, R$ in the validation set, define the martingale residual as

$$M_{i,R} \, = \, \delta_i \, - \, \int \, I(Y_i \geq t) \, \exp(\hat{\beta}^{*'}_{A,R} \, z^*_{A,R,\,i}) \, d\hat{\Lambda}^*_0(t; \hat{\beta}^*_{A,R}) \, ,$$

where $z^*_{A,R,\,i}$ is the transformed X_i given by the same operation that gives Z^*_A from X and

$$\hat{\Lambda}^*_0(t; \hat{\beta}^*_{A,R}) \, = \, \int_0^t \, \left\{ \sum_i \, I(Y_i \geq s) \, \exp(\hat{\beta}^{*'}_{A,R} \, z^{**}_{A,\,i}) \right\}^{-1} \, \sum_i \, d \, I(Y_i \leq s, \delta_i = 1) \, ,$$

which is an estimate of the baseline cumulative hazard function. In the above, the summations $\sum_i$ and $\sum_l$ are over subjects not in R and $z^{**}_{A,\,l}$ is the lth row of $Z^*_{A,R}$. Then we propose to define the prediction error of a model as

$$RMSEP(A, m) \, = \, \left\{ \frac{1}{m} \sum_{i \epsilon R} \, M^2_{i,R} \right\}^{1/2} \, ,$$

the delete-m root mean square error of prediction. Alternatively, one can use

$$RMSEP(A) \, = \, \left(\frac{1}{n} \sum_{i=1}^{n} \, M^2_{i,\,i} \right)^{1/2} \, ,$$

the delete-one root mean square error of prediction, as a criterion to evaluate the prediction ability of a model. Note that in $RMSEP(A)$, the subset R contains only one element i.

Let A_0 denote A such that $RMSEP(A_0, m)$ for a given m or $RMSEP(A_0)$ minimizes $RMSEP(A, m)$ or $RMSEP(A)$, respectively. The optimal or final

DOI: 10.2202/1544-6115.1153

Cox survival model is then given by model (2) with β_A and $\lambda_0(t)$ replaced by $\hat{\beta}_{A_0}$ and $d\hat{\Lambda}_0(t;\hat{\beta}_{A_0})$, where

$$\hat{\Lambda}_0(t;\hat{\beta}_{A_0}) = \int_0^t \left\{ \sum_{i=1}^n I(Y_i \geq s) \exp(\hat{\beta}'_{A_0} X_i) \right\}^{-1} \sum_{i=1}^n d\, I(Y_i \leq s, \delta_i = 1)\,.$$

Note that in practice, one may want to choose a model or A_0^* such that $RMSEP(A_0^*, m)$ or $RMSEP(A_0^*)$ is close to the minimum, but A_0^* is less than A_0, which achieves the minimum RMSEP. This could give a simpler and more robust model with similar prediction ability. Since most statistical software packages such as SAS, S-PLUS, and R contain functions for the partial likelihood approach and SPCR, the proposed procedure can be easily implemented.

The second criterion is to use AUC, originally developed by Heagerty *et. al.* (2000), computed from each subject in the validation dataset to assess the predictive ability of a model at different time points. For right-censored data, time-dependent event indicator $\delta(t)$ indicates whether a survival time is larger or smaller than time t. At each time t, risk score function $f(X)$ in model (2) can be used as a continuous diagnostic marker of a survival outcome for the binary variable $\delta(t)$. Following Li and Gui (2004), we can define

$$\textit{sensitivity}\,(c\,,t\,|\,f(X)\,) = P\{\,f(X)\,>\,c\,|\,\delta\,(\,t\,)\,=\,1\}, \text{ and}$$

$$\textit{specificity}\,(c\,,t\,|\,f(X)\,) = P\{\,f(X)\,\leq\,c\,|\,\delta\,(\,t\,)\,=\,0\}.$$

At time t, AUC is then defined as the area under a time-dependent receiver operator characteristic curve, ROC $(t\,|\,f\,(X)\,)$, which is a plot of *sensitivity* $(t\,|\,f\,(X)\,)$ vs. $1-$*specificity* $(t\,|\,f\,(X)\,)$ with c, the cutoff value for $f(X)$, varying. The larger the AUC is, the better predictive ability a model has at a certain time. Note that a nearest neighbor estimation of the bivariate distribution was used to estimate the sensitivity and specificity, which guarantees that both are monotonic in c (Heagerty *et al.* 2004 and Akritas, 1994). In practice, to construct a ROC $(t\,|\,f\,(\,X\,)$ curve and compute the corresponding AUC at time t, one may use a certain number, say r, of values for c between the smallest and the largest values of $f(X)$ based on validation data. One would choose a model that gives the largest AUC.

In addition, since $f(X)$, the risk score function in model (2), is a risk indicator of the survival event such as death in cancer studies, subjects can be classified into different risk groups such as different tumor development stages based on their values of the risk score function when comparing to a certain cutoff f_0, which may be determined with the help of a professional such as a

physician based on data in the past. When information about membership of a subject to a risk group is available, one can compute the correct classification rate, which is computed as percentage of times a subject is correctly classified into a group. This criterion is used in a simulation study in the next section to evaluation the proposed method.

3 NUMERICAL STUDIES AND RESULTS

3.1 DLBCL Data

To illustrate the procedure presented in the previous section and compare it to the two partial least squares methods proposed in Li and Gui (2004), referred as LG1 and LG2 thereafter, we reanalyzed the DLBCL gene expression data that were analyzed by Rosenwald *et. al.* (2002) and Li and Gui (2004) among others. The dataset includes gene expression measurements of 7,399 genes on 240 patients with DLBCL and their survival times. There exist some missing gene expression measurements and they are replaced by the average of gene expression measurements of the nearest 8 genes according to the Euclidean distance as described in Li and Gui (2004). Among the 240 patients, 138 (57.5%) deaths were observed during the study with the median death time of 2.8 years and, for others, right-censoring times were observed.

Considering the fact that most genes are irrelevant to patients' survival, we analyzed the reduced dataset given by 488 genes that are significantly related to the hazard rate of survival time at 0.01 significance level based on the univariate Cox regression analysis to reduce noise. To select a proper model, we then divided the 240 patients at randomly into two samples with $m = n/3 = 80$ patients in the validation sample and the remaining in the training sample, as in Li and Gui (2004) and Bair and Tibshirani (2004). We fit model (2) to the training data and calculated the root mean square error of prediction $RMSEP(A, m)$ based on the validation data. The above procedure is repeated for 50 times. Figure 1 (a) presents the average curves of the error of prediction obtained by connecting the values of $RMSEP(A, m)$ up to $A = 5$. Note that here we have three curves corresponding to three different Cox regression procedures. The curve corresponding to CPCR was obtained by using the procedure proposed in the previous section and the other two curves corresponds to the two procedures studied in Li and Gui (2004). Specifically, LG1 denotes the partial Cox regression procedure and LG2 means the application of LG1 to the principal components of the original data (Li and Gui, 2004). It can be seen from Figure 1 (a) that all three procedures require only one component, $A_0 = 1$, to achieve similar minimum values for

DOI: 10.2202/1544-6115.1153

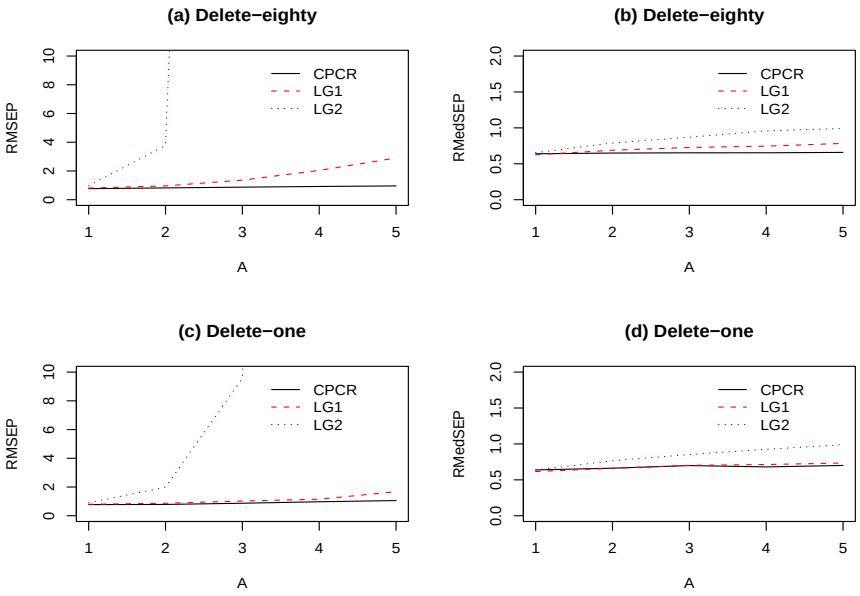

Figure 1: Prediction Error Comparison

RMSEP. However, CPCR yields more stable models because the prediction errors are much smaller as A increases.

Since it is known that the squared martingale residuals are skewed, to give a more robust estimate of the prediction error, we also calculated the root median of all squared residuals (RMedSEP) instead of $RMSEP(A, m)$ for the three procedures. Figure 1 (b) presents the three curves of the prediction errors given by connecting the mean $RMedSEP(A, m)$ values up to $A = 5$. It gives the similar conclusion as Figure 1 (a) and it is apparent that all procedures are more stable from the median point of view as expected.

In addition, we repeated the above analysis by using the delete-one root mean square error of prediction $RMSEP(A)$ and $RMedSEP(A)$ and the results are displayed in Figure 1 (c) and (d). As expected, it gives similar conclusions and again suggests that one component is good enough for the datasets discussed here.

We also applied the three methods to the full dataset with 7,399 genes and the results are very similar except for getting larger quantities for the prediction errors as expected.

3.2 Simulated Data

We compared the proposed method with LG1 and LG2 given in Li and Gui (2004) based on simulated gene expression and survival data by focusing on predictive ability. Results for correct classification rate and AUC will be reported later. Mimicking the DLBCL data and following Bair and Tibshirani (2004), we generated gene expression data X with $p = 5,000$ genes for $n = 240$ subjects for classification and 120 subjects for AUC comparison due to calculation time consideration. All gene expression values were generated from $N(0, 1)$ with a few exceptions. Genes 1-50 for subjects 1-$n/2$ had a mean of 1.0. For genes 51-100, 30% of all subjects were randomly chosen to have a mean of 2.0. Similarly, for genes 101-200, 50% of the subjects were randomly chosen to have a mean of 1.0. Finally for genes 201-300, 70% of the subjects were randomly chosen to have a mean of 0.5.

To generate survival data, we assume that subjects come from one of the two risk groups, high or low risk group. Survival times T_i are generated from exponential with a mean of $\mu = 10$ or 15 for subjects 1-$n/2$ (low risk) and with a mean of 8 for the rest (high risk). Independently, censoring time C_i is generated from exponential with a mean of 10 for each subject. Then we obtained survival data $\{Y_i = min(T_i, C_i), \delta_i = I(T_i \leq C_i), i = 1, ..., n\}$. Note that when $\mu = 10$, the two groups has a difference of 2 for mean survival time, which is smaller than that when $\mu = 15$. When $\mu = 10$, generated data has about 46.2% of right-censoring, which is close to that of the DLBCL data.

The generated datasets are then treated in the same way as with the original DLBCL for analysis. Significance level 0.05 was used in selecting relevant genes before applying the three methods, and the resulting gene matrix X generally has more genes than the number of subjects. On average, less than half of the first 50 genes were selected. But more were selected when the difference between the two risk groups was larger ($\mu = 15$). Based on 250 replications, Figure 2 shows that the classification rate for correctly classifying a subject in the validation sample of size 80 into either high- or low-risk group based on whether the value of $f(X)$ is greater or less than 0, which is the mean risk score for all subjects in the training sample. It can be seen that CPCR outperforms both LG1 and LG2 for classification, especially when the mean difference in survival time is large between the two groups, which corresponds to $\mu = 15$ for the low-risk group. The correct classification rates for the three methods generally drop slightly as more components are involved in the Cox model. This may be caused noise introduced by excessive number of components in the model.

Figure 3 shows the comparisons of mean AUC curves across time using

DOI: 10.2202/1544-6115.1153

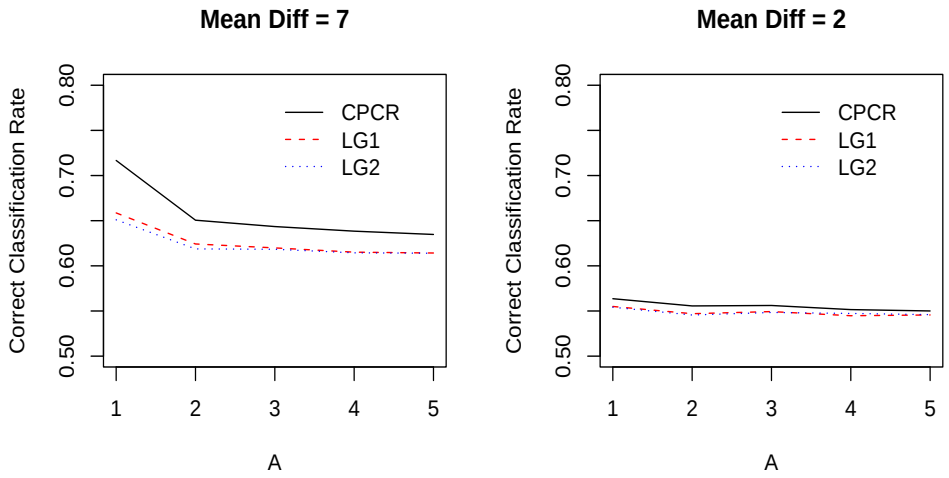

Figure 2: Classification Rate Comparison

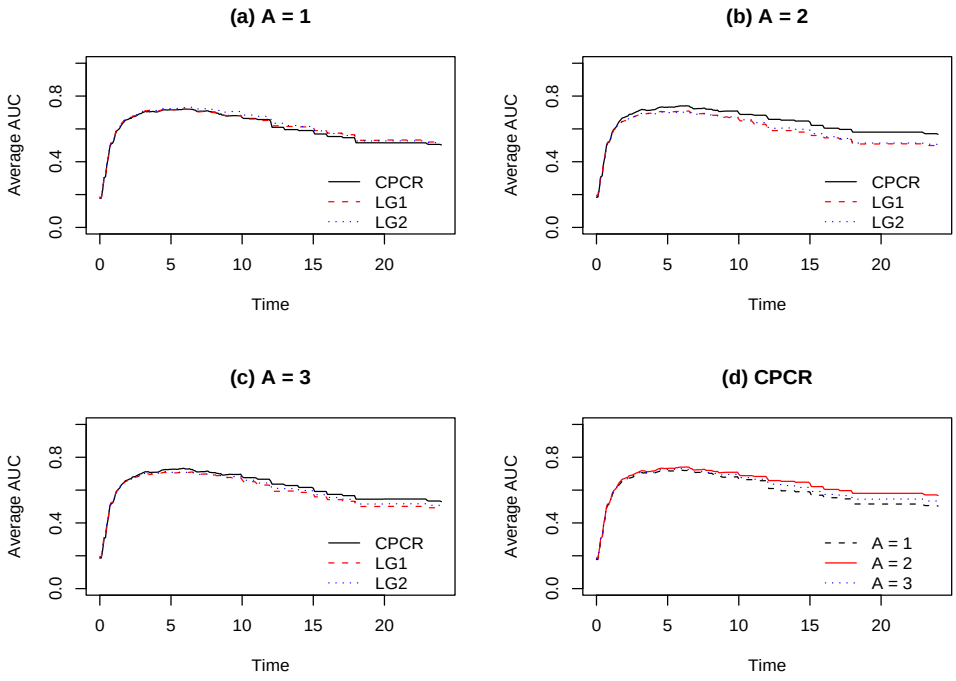

Figure 3: AUC Comparison

delete-forty cross validation based on 10 replications with $n = 120$ and $\mu = 15$. At each distinct time point from the validation sample, $r = 3$ cutoff values for $f(X)$ between the smallest and the largest values of $f(X)$ exclusively were used to construct a ROC curve and compute the AUC. Note that the computation could be time-consuming for large n and r. Since each replication may have a different set of distinct time points, interpolation and extrapolation were used in getting the average. Plots (a)-(c) in Figure 3 show that the proposed method outperforms LG1 and LG2 slightly when two or three components were used in the Cox model in terms of AUC. It has smaller but close AUC to LG1 and LG2 when only one component was used. Figure 3 (d) compares the AUC curves of CPCR with different values of A. It indicates that generally a two-component model yields the best predictive ability based on the 10 datasets generated.

4 CONCLUDING REMARKS

Microarray gene expression data have similar structures as spectrum data and one of the common features between them is that there exists strong multicollinearity in gene expression levels among different genes or spectra among different wavelengthes. No matter what are the purposes of the analysis, one has to deal with this multicollinearity first by transformation and/or dimension reduction.

This paper presents an approach to the survival analysis of gene expression data and it combines CPCR and partial likelihood approach together under the Cox model framework. Note that due to censoring, $p-$values are used in determining correlation instead of correlation coefficients calculated by treating censoring time as survival time or using imputed survival time. For the DLBCL data and the simulated data, using CPCR in conjunction with the Cox model can yield close or even better prediction and more robust models than using PLS approaches with the settings considered. Also the new approach can be easily implemented using common statistical software packages and is simpler in computation than the methods based on PLS (Li and Gui, 2004; Nguyen and Rocke, 2002). All computations were implemented using R.

Screening genes before applying dimension reduction techniques can lower the noise carried in the data and increase predictive ability. The number of genes screened out can be controlled by a significance level used for individual tests under the Cox model. The model selection criteria, RMSEP and AUC, were also discussed. Note that if estimation of β in the Cox model is the main objective instead of prediction, one may want to use a different model selection criterion.

DOI: 10.2202/1544-6115.1153

The proposed CPCR chooses PCs with strongest association with the survival time to be included in the Cox model. Another way to select PCs is to apply a standard model selection procedure such as backward or best subset selection, as pointed out by a referee. A direction for future investigation may focus on different ways of selecting PCs.

The simulation study focused on the performance of the proposed method in prediction. A more thorough examination of the method, such as on the ROC curves and components selected, may help gain more insight of the method. We will leave it for future investigation.

REFERENCES

Akritas, M. G. (1994) Nearest Neighbor Estimation of a Bivariate Distribution under Random Censoring. *Annals of Statistics*, **22**, 1299-1327.

Bair, E. and Tibshirani, R. (2004) Semi-Supervised Methods to Predict Patient Survival from Gene Expression Data. *PLoS Biology*, **2**, 0511-0522.

Gui, J. and Li, H. (2005) Threshold gradient descent method for censored data regression, with applications in pharmacogenomics. *Pacific Symposium on Biocomputing*, **10**, 272-283.

Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000) Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics*, **56**, 337-344.

Jolliffe, I. T. (2002) Principal Component Analysis, 2nd Ed. *Springer*.

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

Li, H. and Luan, Y. (2005) Boosting proportional hazards models using smoothing splines, with applications to high-dimensional microarray data. *Bioinformatics*, **21**, 2403-2409.

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

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

Park, P. J., Tian, L. and Kohane, I. S. (2002) Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, **18**, s120-s127.

Sun, J. (1995) Correlation principal component regression analysis of NIR data. *Journal of Chemometrics*, **9**, 21-29.

DOI: 10.2202/1544-6115.1153