Improve Survival Prediction Using Principal Components of Gene Expression Data

Yi-Jing Shen[1] and Shu-Guang Huang[2]*

[1] *Department of Statistics, University of California, Los Angeles, CA 90095-1554, USA;* [2] *Statistics and Information Science, Lilly Corporate Center, Eli Lilly and Company, Indianapolis, IN 46285, USA.*

The purpose of many microarray studies is to find the association between gene expression and sample characteristics such as treatment type or sample phenotype. There has been a surge of efforts developing different methods for delineating the association. Aside from the high dimensionality of microarray data, one well recognized challenge is the fact that genes could be complicatedly inter-related, thus making many statistical methods inappropriate to use directly on the expression data. Multivariate methods such as principal component analysis (PCA) and clustering are often used as a part of the effort to capture the gene correlation, and the derived components or clusters are used to describe the association between gene expression and sample phenotype. We propose a method for patient population dichotomization using maximally selected test statistics in combination with the PCA method, which shows favorable results. The proposed method is compared with a currently well-recognized method.

Key words: microarray, principal component analysis, survival

Introduction

The goal of many pharmacogenomics studies is to find the association between gene expression and sample characteristics, which could be either the conditions that the samples are exposed to, or the phenotypes of the subject that the samples are extracted from. One type of sample characteristics is multi-categorical (qualitative) classes such as cancer types, different treatments, treatment over a time course, response to a treatment (yes or no), polymorphism of nucleotides, and so on. The research interest for such studies is generally to find the genes showing differential expressions across the sample categories. The approaches for data analysis in this situation are typically of the ANOVA type. The expression profile of the differentially expressed genes is expected to provide insights on the basic understanding of issues such as disease mechanism, drug action, therapeutic target, gene function, metabolic pathway, or other aspects of cell biology. Another type of sample characteristics is continuous (quantitative) in nature, such as subject tumor size, survival time, cholesterol level, and so on. The analysis of interest for such studies is also to identify the genes whose expressions are correlated with the sample phenotypes, but the analysis approaches for such a situation are generally of the regression type. The genes whose expressions are associated with sample characteristics could be of prognostic value, and thus could be used to assess the likelihood of the subject's response to a drug treatment, or to predict the subject's clinical outcome such as tumor growth or time of survival. The genes with such properties could be good candidate biomarkers for drug targets (for example, mutated genes) or for the development of patient-specific therapy (gene expression serves as surrogate biomarker for drug response).

Microarray is a high-throughput technology for pharmagogenomics that allows the monitoring of gene expression at the genome level. Aside from the high dimensionality of microarray data, one well recognized challenge is the fact that gene expression could be complicatedly correlated. Many genes interact and highly co-regulate with each other; they may share the same molecular function, involve in the same biological pathway or other more complicated genetic network. Investigating genes independently is not optimal since the univariate approach totally ignores the combinatorial effect of gene expression. It is thus important that genes should be considered in groups

***Corresponding author.**
E-mail: huang_shuguang@lilly.com

rather than individually. Specifically, this research focuses on the case where survival time (subject to censoring) is the measured phenotype, and discerning the association between survival time and gene expression is the analysis of interest.

In the past several years, there are quite a few publications proposing novel methodologies to statistically quantify the association of gene expression with patient survival. A frequently used approach is to first classify patients into several groups, each of which shares a distinct expression profile (based on the whole genome or a subset of genes), followed by a comparison of the survival profile (such as Kaplan-Meier curves) among the patient clusters. A good separation of the survival curves indicates that gene expression pattern distinguishes (associates with) patient survival. To deal with the difficulties of high dimensionality and correlation, principal component analysis (PCA) is another common approach to building survival models because correlated genes would project onto the same direction, that is, load onto the same principal component (PC). The identified PCs could then be treated as covariates replacing all the gene expressions that they are derived from. The drawback of both methods is that the phenotype information is not considered when the principal components or the sample clusters are derived, since the data dimension-reduction is based solely on gene expression. Recently, more complicated methods have been developed to simultaneously take account of gene expression correlation and the phenotype association. These methods showed improved performance in association modeling (*1*, *2*).

In cancer clinical studies, it is of frequent interest to simply classify patients into subgroups based on prognostic factors such as gene expression. For instance, it is desirable to categorize patients into high/low risk groups. Such a categorization can provide useful input for optimizing patient treatment assignment. Motivated by these realistic considerations, we propose a method for patient dichotomization using maximally selected test statistics (for example, Chi-square test statistics, Wilcoxon rank sum statistic, and so on) in combination with PCA on gene expression. Li and Gui (*1*) explored patient group dichotomization using the risk score function, but the choice of cutoff for the two groups is arbitrary. Our method identifies the optimal cutoff such that the difference between the two groups is maximized. To evaluate the performance of the proposed method, Li and Gui's method is chosen as the benchmark in this research.

Results

Simulations

In order to compare our method with Li and Gui's, data were simulated according to two scenarios. The first scenario assumes that there are two genes, X_1 and X_2, correlated with each other and with the survival time. The second scenario assumes five $(X_1, X_2, X_3, X_4, X_5)$ biomarkers, of which the first four are correlated to the survival time, with (X_1, X_2) independent of (X_3, X_4), and X_5 serving as an independent noise in the model. The first scenario demonstrates the simplest case, while the second scenario resembles a more realistic data setting. In each scenario, 150 samples were simulated. The correlation matrix and the variation levels for each variable are shown as follows for the two scenarios respectively:

Scenario 1:

$$Corr\begin{pmatrix} Y \\ X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 1 & 0.82 & 0.72 \\ 0.82 & 1 & 0.66 \\ 0.72 & 0.66 & 1 \end{pmatrix}, \quad Var\begin{pmatrix} Y \\ X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 0.7 \\ 0.3 \\ 0.2 \end{pmatrix}$$

Scenario 2:

$$Corr\begin{pmatrix} Y \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{pmatrix} = \begin{pmatrix} 1 & 0.43 & 0.35 & 0.5 & 0.35 & 0 \\ 0.43 & 1 & 0.93 & 0 & 0 & 0 \\ 0.35 & 0.93 & 1 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 1 & 0.93 & 0 \\ 0.35 & 0 & 0 & 0.93 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad Var\begin{pmatrix} Y \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{pmatrix} = \begin{pmatrix} 2 \\ 2.1 \\ 2.1 \\ 1.2 \\ 1.2 \\ 1 \end{pmatrix}$$

Evaluation of the above-mentioned two methods was based on comparing the best PC (having the smallest p-value) of the principal components for Cox regression (PC-CR) and that of the principal components for partial Cox regression (PC-PCR) (see Materials and Methods). In light of the simulation, only the first two PCs were considered. Data were randomly split into 75% training set and 25% testing set. To improve the reliability of the results, the process was repeated for 100 iterations. The two methods, PC-CR and PC-PCR, were applied simultaneously and compared in each random split. Table 1 summarizes the results across the 100 iterations. The first column corresponds to the smallest p-value of the 100 iterations; the second column gives the number of times that each method was "better" (smaller p-value) between the two; and the third column gives the average ranking for each method (2 for smaller p-value, 1 for bigger p-value, so bigger ranking corresponds to smaller p-values). In addition, the number of times that the p-values are below certain thresholds is given in Table 2. The Kaplan-Meier curves of the best separation based on the best component from the 100 iterations are plotted in Figures 1 and 2 for Scenarios 1 and 2, respectively.

According to the results, the PC-CR model seems to be comparable to the PC-PCR model for both datasets. Tables 1 and 2 show a similarity in the numbers of iterations that a method returns the smaller p-values from the two models.

Application to lung cancer data

The study on lung cancer was conducted by the Whitehead Institute and MIT Center for Genome Research, with the goal to identify subclasses of adenocarcinoma based on patient gene expression profile. The dataset consists of a total of 125 snap-frozen lung adenocarcinoma tumor samples and 17 normal lung specimens. For each of the 125 adenocarcinoma samples, there was available clinical and pathological information such as patient age, gender, survival time, smoking history, and cancer status. The total RNA extracted from samples was used to generate cRNA targets, and subsequently hybridized to human U95A oligonucleotide probe arrays according to standard protocols. Each array consists of 12,625 probe sets. See the website http://research.dfci.harvard.edu/meyersonlab/lungca/ for more details about the study (*3*). CEL files were downloaded from the website, and the gene expression data were subsequently extracted using Affymetrix MAS5.0 software (Affymetrix, Santa Clara, USA). The averaged expression was used for samples with replicates.

For the current research, the 12,625 probe sets were first tested for differential expressions between the cancer group and the normal group. Non-differentially expressed genes were excluded from further analysis. The significance was based on the following criteria: mean difference greater than 250, fold

Table 1 Comparison of the p-values Between PC-CR and PC-PCR (Simulation)

Model	Best p-value*[1]		No. of times that p-value is smaller*[2]		Average ranking*[3]	
	S1	S2	S1	S2	S1	S2
PC-CR	2.2E-9	6.0E-9	65	53	1.47	1.53
PC-PCR	2.2E-9	1.1E-6	70	47	1.52	1.47

*[1]The best p-value in the 100 iterations. *[2]The number of times that one method outperformed the other. Note: when a tie is presented, both numbers are incremented by one. *[3]The average ranking for each method (bigger ranking value means better performance).

Table 2 Number of Iterations that the p-values Are Below a Threshold (Simulation)*

Model	p-value<0.05		p-value<0.01		p-value<0.005		p-value<0.001	
	S1	S2	S1	S2	S1	S2	S1	S2
PC-CR	99	88	97	64	93	51	88	33
PC-PCR	100	83	97	61	95	50	83	27

*Varying p-value cutoffs are used to define the significance level. At each level, the number of iterations (out of 100) passing the significance level is reported.

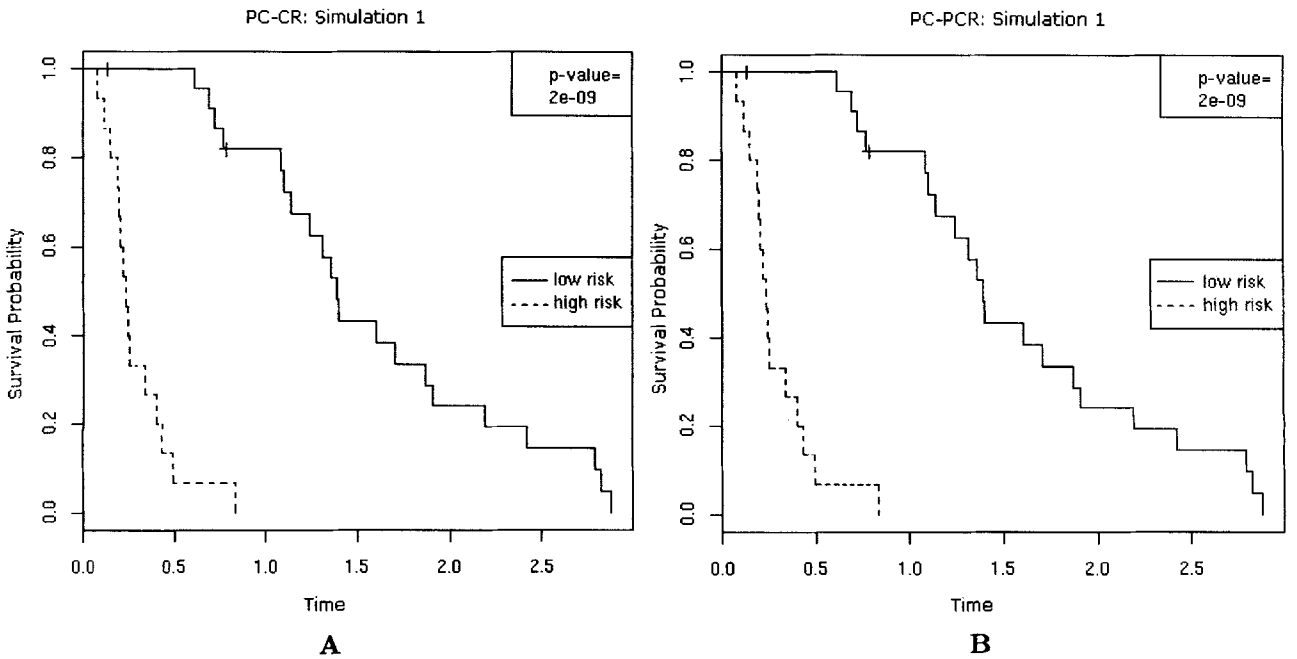


Fig. 1 Kaplan-Meier curves of the best separation from the 100 iterations on the simulated data for Scenario 1. **A.** The PC-CR model. **B.** The PC-PCR model. Each iteration randomly separates 75% data into the training set and 25% data into the testing set.

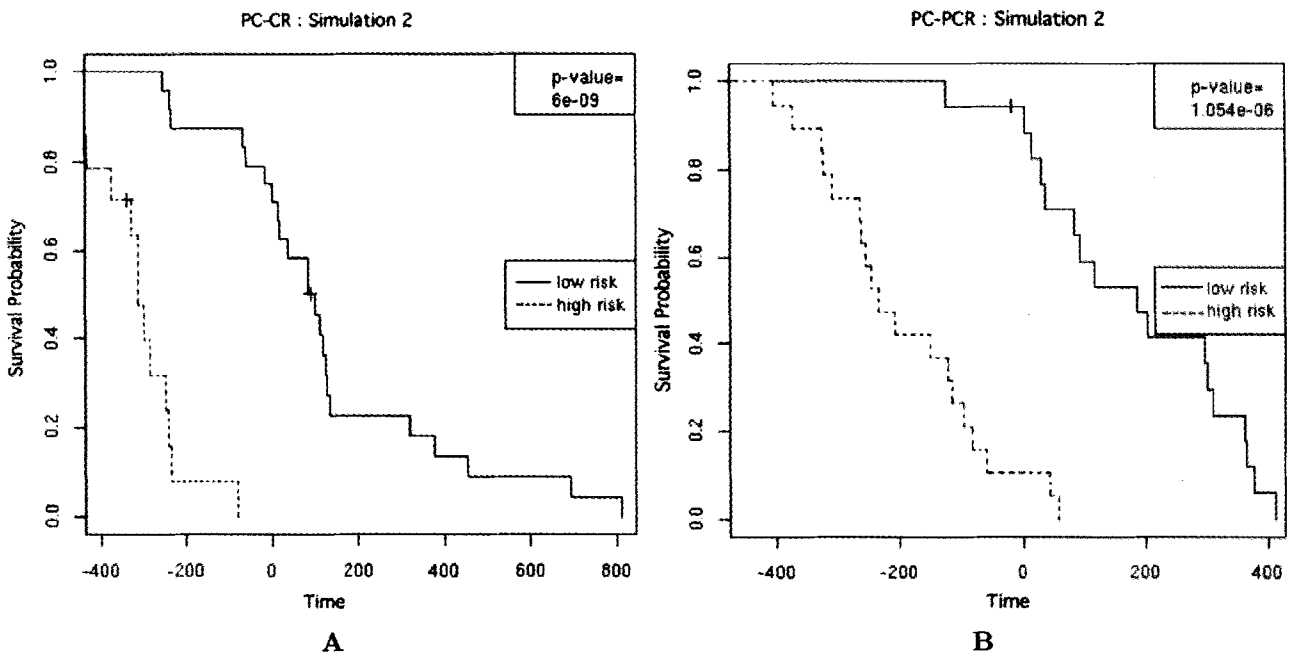


Fig. 2 Kaplan-Meier curves of the best separation from the 100 iterations on the simulated data for Scenario 2. **A.** The PC-CR model. **B.** The PC-PCR model. Each iteration randomly separates 75% data into the training set and 25% data into the testing set.

change greater than 2, and false discovery rate (*4*) smaller than 0.01. This resulted in 324 interesting probe sets that might be involved in the cancer mechanism (Figure 3).

In Figure 3, Three patient clusters (rows) and two gene clusters (columns) were requested. The last cluster of samples corresponds to the 17 samples from normal patients (with 0% classification error, which is not surprising because the probe sets were so selected and the whole data were used as the "training" set). Since the probe sets in each cluster show very similar expression pattern, a subset of 30 probe sets

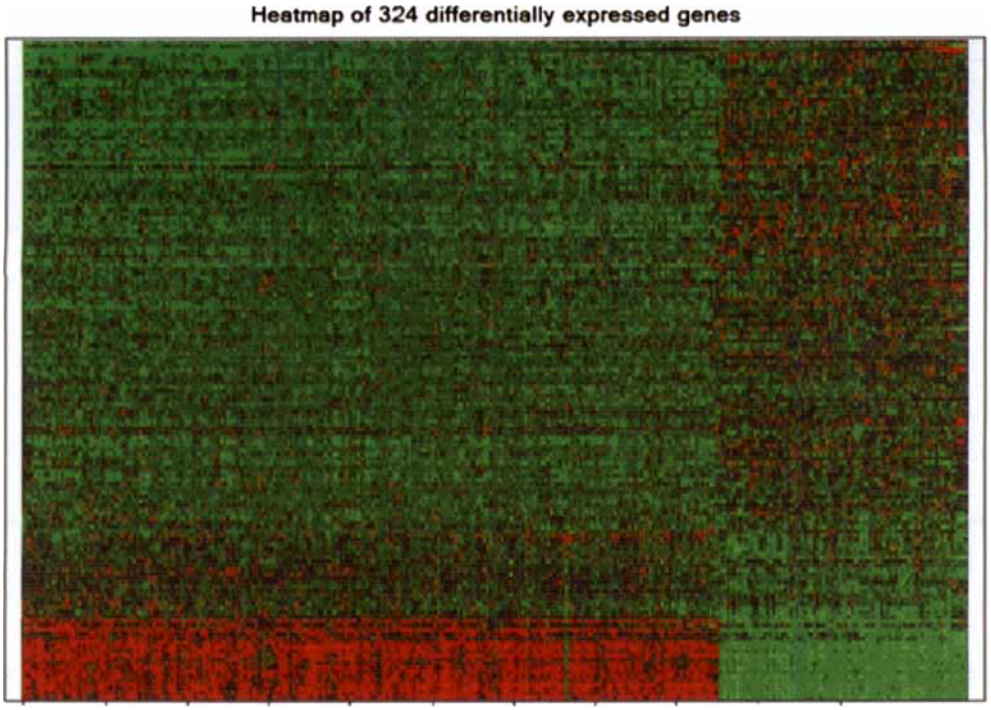


Fig. 3 Heat map of the cancer-related genes. Three patient clusters (rows) and two gene clusters (columns) were requested.

Table 3 Comparison of the p-values Between PC-CR and PC-PCR (Lung Cancer Data)

Model	Best p-value	No. of times that p-value is smaller	Average ranking
PC-CR	3.4E-4	85	1.85
PC-PCR	1.8E-2	15	1.15

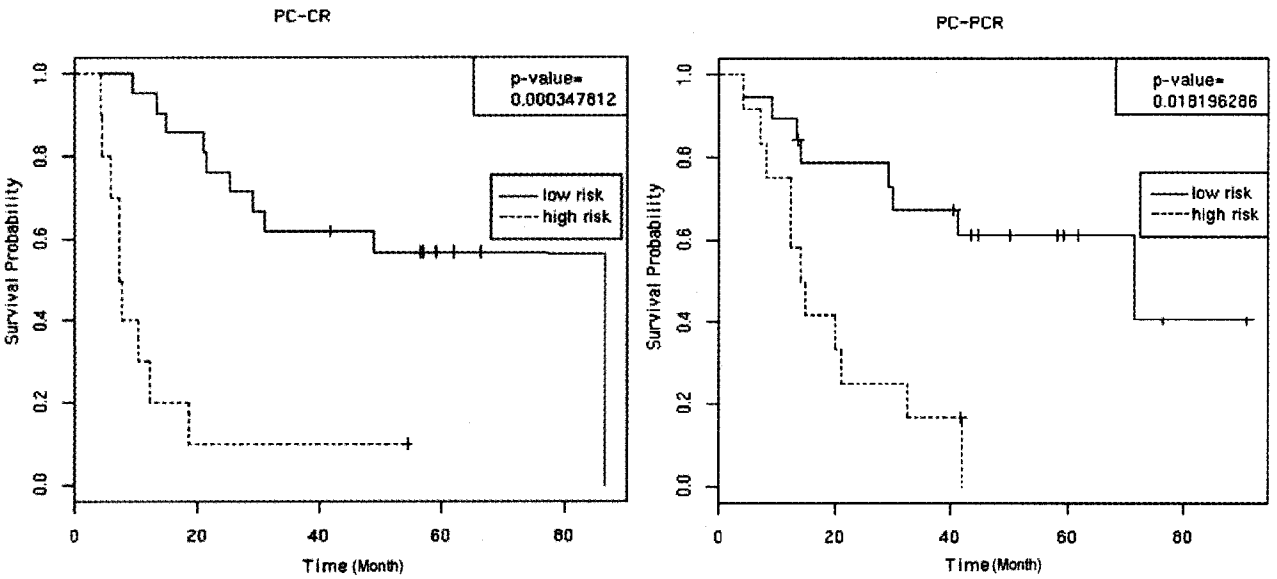
Table 4 Number of Iterations that the p-values Are Below a Threshold (Lung Cancer Data)

Model	p-value<0.05	p-value<0.01	p-value<0.005	p-value<0.001
PC-CR	54	19	13	3
PC-PCR	11	0	0	0

were "randomly" selected from the three clusters for the following analyses. The 17 patient samples were left out since there was no survival information available for them.

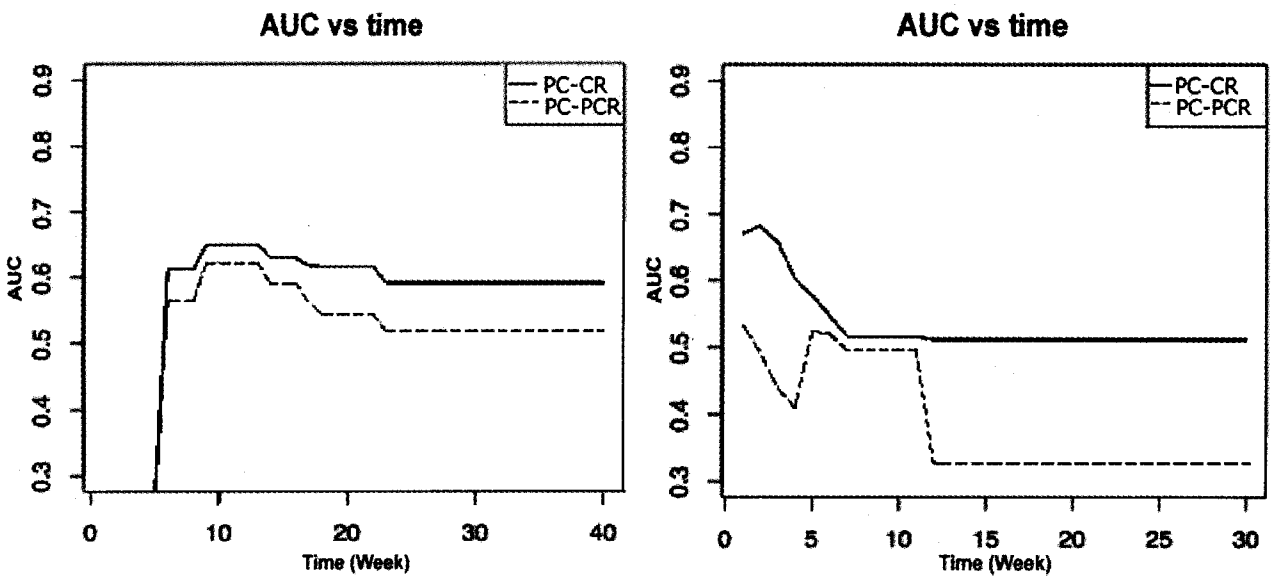
The two methods were again applied simultaneously and compared in each random split. Each split randomly separated 93 (out of the 125) patients into the training set, and the remaining 32 patients into the testing set. Tables 3 and 4 summarize the results from the 100 iterations. Based on these analysis results, it is obvious that the PC-CR method outperforms the PC-PCR method in classifying patient risk groups. The Kaplan-Meier curves of the best separation from the 100 iterations are plotted in Figure 4 for the two methods separately.

To answer the question on how well the model predicts the subjects who are likely to die and the subjects who are likely to survive, Figure 5 shows the area under the curve (AUC) versus time for the simulated and the lung cancer datasets, respectively. For both datasets, the PC-CR model demonstrates more accuracy since it has larger AUC for all time points. For the simulated data, the difference between the two



A B

Fig. 4 Kaplan-Meier curves of the best separation from the 100 iterations on the lung cancer data. **A**. The PC-CR model. **B**. The PC-PCR model. Each iteration randomly separates 93 (out of the 125) patients into the training set and the remaining 32 patients into the testing set.



A B

Fig. 5 Area under the curve (AUC) versus time (in weeks) for the simulated data based on the second scenario (**A**) and the lung cancer data (**B**). Estimations are based on the PC-CR model (solid line) and the PC-PCR model (dashed line).

AUC curves ranges between circa 0.05 and 0.1, for time between 7 to 23 weeks; while for the lung cancer data, it ranges between approximately 0.01 and 0.2, for time between 2 to 13 weeks, and the difference is larger in the first several weeks up to week 5.

Discussion

It has been realized that in clinical studies drugs show varying efficacy on different patients. The variation in reaction to a treatment can be due to many factors, an important one of which is the genetic difference in the patient population. Pharmacogenomics studies have provided a useful tool in developing strategies to identify differences in subject genotype or gene expression that could be included in prospective and randomized trials to develop a more individualized

treatment for patients. In other words, studies need to identify molecular biomarkers whose expression signature (or expression profile) can help predict response to the proposed therapy. For example, as discussed in this paper, a frequent interest for clinical studies is to simply divide the subject into high/low risk groups based on gene expression profile, so that the assignment of treatment could be "sub-population specific".

Multiple genes interacting with each other may affect response in complex ways. From our analysis, we conclude that the PC-CR method provides good ability to discriminate patient subgroups, comparable to the PC-PCR method proposed by Li and Gui. Appropriately utilizing gene-gene co-regulation can aid pharmacogenomics to hold the promise that drugs might one day be individualized based on genetic expression profile. Environment, sex, age, lifestyle, and many other factors all can influence a person's response to medicine, but understanding an individual's genetic profile is the key to creating personalized drugs more effectively and safely. Pharmaceutical companies desire to create better drugs based on the association between the expression of a specific group of genes and the outcome of a disease. It is highly desirable to be able to analyze a patient's genetic profile and prescribe the best therapy without guessing or taking any risky chances. Therefore, it is important to build accurate predictive models that enable better diagnose and prescription.

In this paper, we developed a potential multivariate methodology for identifying significant (groups of) genes associated with clinical outcomes or other sample characteristics. We were able to use the predictive genes to find a threshold that could satisfactorily classify patients into high/low risk groups. This method shall be further applied and assessed on different datasets in future research. Li and Gui's PC-PCR method has the advantage of finding the most informative PCs by incorporating the survival information when searching for the PCs. Compared with Li and Gui's method, our method has the advantage that the cutoff is optimally selected and it is easier to apply in practice. It is reasonable to expect that the PC-PCR method will do better if the cutoff point is chosen based on maximally selected test statistics, but that is outside the scope of this research. Even though our method is developed for survival data analysis, it could be similarly applied to more general phenotypes such as tumor size, treatment type, and so on.

Materials and Methods

Sample dichotomization using maximally selected test statistics

In terms of data analysis, the goal of risk assessment is to categorize patients into subgroups using gene expression as the surrogate biomarker, that is, to group samples in such a way that the association between expression values and phenotype measure is statistically significant. Suppose that the goal is to separate patients into two risk groups (high/low) based on the expression of one gene, the following gives the algorithm for identifying the best cutoff of the expression and subsequently calculating the significance level of the difference between the two risk groups. This method was originally developed by Miller and Siegmund (*5*).

Algorithm for optimal dichotomization of samples using gene expression:

1. Sort the dataset by increasing expression value of the gene (or the PC), and choose the searching range $[\varepsilon_1, \varepsilon_2]$ of the expression data (for example, ε_1 = 10%, ε_2= 90%. Cutoffs close to the ends give unstable results).

2. Start the threshold from quantile ε_1 of the expression values, separate the samples into two groups using this threshold as the cutoff.

3. Achieve the statistic (Chi-square test statistics, Wilcoxon rank sum statistic, log-rank, *etc.*) or p-value for the testing difference of the phenotype in the two groups.

4. Repeat Steps 2 and 3 for each distinct value in the interval $[\varepsilon_1, \varepsilon_2]$ of the expression intensities.

5. Find the optimal test statistic (largest test statistic or smallest p-value) and its corresponding threshold.

6. Calculate the significance of the difference between the two groups, which is defined by the optimal cutoff with the p-value adjusted by:

$$P_{adj} = \phi(z)(z - 1/z)\log\left(\frac{\varepsilon_2(1-\varepsilon_1)}{\varepsilon_1(1-\varepsilon_2)}\right) + 4\phi(z)/z \quad (1)$$

where z is an appropriate transformation of the test statistic (for example, square-root of the Chi-square statistics, or the untransformed Wilcoxon rank sum statistic) or a transformation of p-value given by $z = \Phi^{-1}(1 - P_{\min}/2)$, in which $P_{\min}$ is the observed minimal p-value; ϕ denotes the standard normal probability density function, and Φ is the standard normal distribution function. For details, please see Miller and Siegmund (*5*) and Halpern (*6*).

Principal components for Cox regression (PC-CR)

PCA is a dimension reduction method widely applied to high-dimension microarray data, which is designed to capture the maximum variation of the dataset in terms of PCs. In other words, this method tries to reduce the dimension of the dataset, retaining the most important information while filtering out the noise. PCs are a set of variables that summarize the maximum amount of variation in the original dataset and at the same time are uncorrelated to each other. Specifically, the first PC is the combination of variables that explains the greatest amount of variation of the data. The second PC captures the next largest amount of variation and is independent to the first PC, and so on. PCA can also be viewed as finding a projection of the observations onto orthogonal axes contained in the space defined by the original variables. The criteria are: the first axis includes the maximum amount of variation, the second axis includes the maximum amount of variation orthogonal to the first, and the third axis contains the maximum amount of variation orthogonal to the first and second axis, *etc.*, until the last axis contains the least amount of variation (*7*).

The Cox regression model (*8*) is widely used in the analysis of survival time to delineate the effect of explanatory variables. Suppose there are n number of patients, and p number of genes. Let $X_{n\times p} = \{X_1, X_2, \ldots, X_p\}$ denote the gene expression data matrix that is used to predict the patient survival. The datum for the i^{th} subject is denoted by $(t_i, \delta_i, x_{i1}, x_{i2}, \ldots, x_{ip})$, where δ_i is the censoring indicator, and t_i is the time to event outcome such that t_i is the survival time if δ_i=0 or is the censoring time if δ_i=1. Meanwhile, i (i=1,...,n) represents the patient index, and j (j=1,...,p) represents the gene index.

For the PC-CR method, PCA is first performed on $X_{n\times p}$, and the resulted PCs are treated as new variables for the Cox regression model. Since the correlated genes would load onto the same PC, and that the PCs are orthogonal to each other, multicolinearity is not an issue. Due also to the orthogonality, each PC could be fitted into the Cox model separately and the significant ones could be independently identified, then the significant PCs can be put together into the final Cox model. Finally, the significant gene list can be obtained by observing which ones are highly correlated or heavily loaded onto the PCs in the final Cox model. The hazard function of the considered Cox model could be written as:

$$\begin{aligned}\lambda(t) &= \lambda_0(t)\exp(\beta_1' PC_1 + \beta_2' PC_2 + \cdots + \beta_k' PC_k) \\ &= \lambda_0(t)\exp\big(f(X)\big) \qquad (2)\end{aligned}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, X is a gene subset of $X_{n\times p}$, k is the number of PCs selected into the model, and $f(X)$ is the log hazard ratio, which is also called risk index score. Like the PCs, $f(X)$ is a linear combination of X.

The high/low risk patient subgroups is to be classified based on the values of $f(X)$. In particular, rather than choosing the cutoff point arbitrarily, we propose that the cutoff point is determined by the procedure developed by Miller and Siegmund (*5*) as discussed above. The significance of the difference between the two risk groups is correspondingly determined.

Principal components for partial Cox regression (PC-PCR)

Partial least square (PLS; ref. *9*) is a method for modeling linear regression equations by constructing new explanatory variables (often called latent components or projections) using linear combinations of the original variables. Unlike PCA, the PLS method makes use of the response variable in constructing the latent components, so it does not suffer the drawback that clinical outcome information is ignored in the projection step. Li and Gui (*1*) proposed the PC-PCR method as an extension of PLS to the Cox model. Their algorithm involves constructing predictive components by iterative fitting the Cox regression model and least square fitting to the residuals from the previous Cox model. The identified components are then used in the Cox model for building a predictive model for clinical outcomes. The goal of PC-PCR is to build the following model:

$$\begin{aligned}\lambda(t) &= \lambda_0(t)\exp(\beta_1 T_1 + \beta_2 T_2 + \cdots + \beta_k T_k) \\ &= \lambda_0(t)\exp\big(f(X)\big) \qquad (3)\end{aligned}$$

where each component of T_k and $f(X)$ is a linear combination of $X = \{X_1, X_2, \ldots, X_p\}$ (some coefficients may be zero). The first component is defined as $V_{1j} = X_j - \overline{x}_j$, where $\overline{x}_j$ is the sample means of the j^{th} column, and V_{1j} is a standard value of X_j. Then, for each column j, the components are fitted into the model $\lambda(t) = \lambda_0(t)\exp(\beta_{1j}V_{1j})$. Let $\hat{\beta}_{1j}$ be

the estimate of β_{1j} derived from the maximum partial likelihood estimate, in order to combine V_{1j} into one component, the first component is represented by $T_1 = \sum_{j=1}^{p} \varpi_{1j} \hat{\beta}_{1j} V_{1j}$, where ϖ_{1j} is a weight that is proportional to the variance of V_{1j}. The information in X_j that is not explained by T_1 would remain in the residuals when regressing T_1 against X_j, or equivalently against V_{1j} since V_{1j} is just another form of X_j. Iteratively, $V_{(i+1)j} = V_{ij} - \frac{V_{ij}' T_i}{T_i' T_i} T_i$ and the following Cox regression will be achieved:

$$\lambda(t) = \lambda_0(t) \exp(\beta_1 T_1 + \beta_2 T_2 + \cdots + \beta_i T_i + \beta_{(i+1)j} V_{(i+1)j}) \tag{4}$$

and each T_{i+1} will be constructed as

$$T_{i+1} = \sum_{j=1}^{p} \varpi_{(i+1)j} \hat{\beta}_{(i+1)j} V_{(i+1)j}$$

The risk score function $f(X)$ can be written in terms of X:

$$f(X) = \sum_{j=1}^{p} \beta_j^* V_{1j} = \sum_{j=1}^{p} \beta_j^* (X_j - \bar{x}_j) \tag{5}$$

Li and Gui 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.

Area under the time-dependent ROC curve

It is a common practice in statistics to use cross-validation to evaluate the performance of a predictive model. However, due to censoring, the general cross-validated misclassification cannot be easily applied to survival data. In the current literature, time-dependent receiver operating characteristics (ROC) curves (*10*) is one of the popular methods. The method is based on a generalization of the area under the ROC curve incorporating clinical outcomes, which was used by Li and Gui in evaluating their proposed PC-PCR (*1*). We also use this method to evaluate Cox regression on PCs (PC-CR). Briefly, we present some background on the use of time-dependent ROC curves.

Sensitivity and specificity for cut point value c at time t are defined as following:

$$\begin{aligned} \textit{Sensitivity}\left(c, t \,|\, f(X)\right) &= \Pr\{f(X) > c \,|\, \Delta(t) = 1\} \\ \textit{Specificity}\left(c, t \,|\, f(X)\right) &= \Pr\{f(X) \leq c \,|\, \Delta(t) = 0\} \end{aligned} \tag{6}$$

In the above equations, $f(X)$ represents the risk index score, and $\Delta(t)$ is the event indicator at time t, with $\Delta(t)$=1 representing an event (death) and $\Delta(t)$=0 a censoring. Therefore, sensitivity is the proportion of patients classified into the high risk group based on cut point c among the patients who died (that is, an event occured) at a given time point t; it is equivalent to the proportion of true positives. Correspondingly, specificity is the proportion of patients classified into the low risk group by cut point c among the patients who survived to time t or longer. In both proportions, greater value means better classification. The ROC curve, denoted by $ROC\left(t \,|\, f(X)\right)$, is the plot of $\textit{Sensitivity}\left(c, t \,|\, f(X)\right)$ versus $[1 - \textit{Specificity}\left(c, t \,|\, f(X)\right)]$ with the cutoff point c varying. It is worth mentioning that the sensitivity value is generally greater than the value of $(1 - \textit{Specificity})$ because there are greater proportions of true negative than false positive in practice. A greater ratio of $\textit{Sensitivity}/(1 - \textit{Specificity})$ indicates better prediction at the cut point. Graphically, if sensitivity is plotted on the y-axis and $(1 - \textit{Specificity})$ on the x-axis, the ROC curve would be above the 45 degree line. Consequently, greater AUC indicates better prediction. In the following, $AUC\left(t \,|\, f(X)\right)$ is used to denote the area under the curve of $ROC\left(t \,|\, f(X)\right)$.

Acknowledgements

We thank Chao-Feng Liu and Gary Sullivan for reviewing the manuscript and providing valuable suggestions.

Authors' contributions

SGH did most of the writing of the manuscript and was involved in the discussion and implementation of the methods. YJS conducted most of the computational work and method implementation. Both authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Li, H. and Gui, J. 2004. Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics* 20: I208-215.
2. Nguyen, D.V. and Rocke, D.M. 2002. Partial least squares proportional hazard regression for application to DNA microarray data. *Bioinformatics* 18: 1625-1632.
3. Bhattacharjee, A., *et al.* 2001. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. USA* 98: 13790-13795.
4. Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B* 57: 289-300.
5. Miller, R. and Siegmund, D. 1982. Maximally selected Chi square statistics. *Biometrics* 38: 1011-1016.
6. Halpern, J. 1982. Maximally selected Chi square statistics for small samples. *Biometrics* 38: 1017-1023.
7. Rencher, A.C. 1998. *Multivariate Statistical Inference and Applications.* John Wiley & Sons, New York, USA.
8. Cox, D.R. 1972. Regression models and life tables (with discussion). *J. Roy. Statist. Soc. B* 34: 187-220.
9. Wold, H. 1966. Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis*, pp.391-420. Academic Press, New York, USA.
10. Heagerty, P.J., *et al.* 2000. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 56: 337-344.