## Title
Dimension Reduction Methods for Microarrays with Application to Censored Survival Data



## Authors
Li, Lexin
Li, Hongzhe

## Publication Date
2004-02-01

# Dimension Reduction Methods for Microarrays with Application to Censored Survival Data

*Lexin Li* and Hongzhe Li*

*Department of Biological Chemistry and Rowe Program in Human Genetics, School of Medicine, University of California, Davis, CA 95616, USA*

Running title: Micorarrays and Censored Survival Data

*To whom correspondence should be addressed:
Lexin Li, Ph.D.
Department of Biological Chemistry
School of Medicine
University of California
Davis, CA 95616-8500, USA
Tel: (530) 754-6911; Fax: (530) 754-7269
E-mail: lexli@ucdavis.edu

## ABSTRACT

**Motivation:** Recent research has shown that gene expression profiles can potentially be used for predicting phenotypes such as cancer types and survival time in biomedical research. Microarray technology which simultaneously measures expression values of thousands of genes provides a powerful tool as well as new challenges in relating gene expression profiles to phenotypes. Expression data are often very high-dimensional, which makes statistical modeling more difficult and complex, especially when the phenotypes such as time to death or cancer recurrence are subject to right censoring. We consider in this paper a model-free sufficient dimension reduction technique to reduce the dimension of microarray data in the context of analyzing censored survival data.

**Results:** We propose a dimension reduction technique which does not assume a particular model for survival time given gene expression values. After dimension reduction, the constructed gene expression components are used as covariates for predicting the survival probabilities in the framework of censored data regression analysis. In particular we use the popular Cox proportional hazards model to build a predictive model for survival. We demonstrate the use of the methodology by applying to a large diffuse large B-cell lymphoma gene expression data set, which consists of 240 patients and 7399 genes. The Cox proportional hazards model with the derived gene expression components is shown to provide a good predictive performance for patient's survival as demonstrated by the receiver operator characteristics analysis. The predictive model built using the training data set predicted highly significant survival difference in the testing data.

**Availability:** R programs are available on request from the authors.

**Contact:** lexli@ucdavis.edu; hli@ucdavis.edu

# INTRODUCTION

DNA microarray technology is a ground-breaking advance in biomedical and genomic research. It enables the simultaneous measurements of the expression levels of thousands of genes per sample. It has been shown in various studies that the gene expression profiles can be used successfully in molecular classification of tumor types (Golub *et al.* 1999), in therapeutic prediction of drug response (Scherf *et al.* 2000), and in genomic prediction of patients' survival (Rosenwald *et al.* 2002).

Among those applications, cancer class prediction using gene expression data has been studied extensively in recent years (See Dudoit *et al.*, 2000 for a review.) However, there has been less development in relating gene expression profiles to other phenotypes, such as survival time, due to a number of challenges. First, the microarray-based high-throughput technology generates a huge number of potential predictors, i.e., gene expression levels of thousands of genes, and the expression levels of many genes are often highly correlated. On the other hand, the sample size of patients or cell lines is usually very small compared to the number of genes in the study. Modeling such high-dimensional data is a complex and challenging problem. The problem becomes more difficult when the phenotypes such as time to death or time to cancer recurrence are subject to right-censoring. Additionally, microarray data often possess a great deal of noise.

There has been some recent development in relating gene expression profiles to censored survival endpoints. One popular approach is built on clustering analysis. For example, Rosenwald *et al.* (2002) first identified a small number of "signature" gene clusters using hierarchical clustering analysis, and then built a Cox proportional hazards model for predicting time to death in patients with large B-cell lymphoma based on the mean values of the expression levels of genes in those gene clusters. One disadvantage of clustering genes is that the sample phenotypes are not efficiently used. Bair and Tibshirani (2003) re-analyzed the lymphoma data set of Rosenwald *et al.* by applying the nearest shrunken centroid supervised clustering and the partial least squares techniques. Nguyen and Rocke (2002) proposed to construct a partial least squares proportional hazards model using residuals for the Cox model. However, the use of residuals in the estimation of parameters in the Cox model is not well-established in the survival analysis literature since there are many different ways of defining residuals (Barlow and Prentice, 1988). In addition, smaller sum of squares of residuals in the Cox regression model context does not always imply a better fit of the model. Li and Luan (2003) proposed

a penalized Cox proportional hazards model within the framework of kernel estimation, and they evaluated their method using a number of survival microarray data sets. But they did not study how different choices of kernel functions affect the predictive performance of their methods.

In this article, we introduce a dimension reduction strategy, in the context of survival prediction, to transform the high-dimensional predictors to a low-dimensional space. The proposed method capitalizes on the correlations of gene expressions among all the genes to identify a small number of linear combinations of the gene expression levels. Those linear combinations of genes may be regarded as "supergenes", and they are then used for building predictive survival model. The method differs from other approaches in that it does not impose any probabilistic model in the dimension reduction process, thus it allows the investigators to fit any model in the subsequent model building stage of analysis. We particularly consider sliced inverse regression (SIR) based on the theory of sufficient dimension reduction, which was first proposed by Li (1991) and Cook and Weisberg (1991) and was later formulated by Cook (1998). Similar sufficient dimension reduction techniques have been used in the microarray data analysis literature, including the studies by Chiaromonte and Martinelli (2002), Bura and Pfeiffer (2003), and Antoniadis *et al.* (2003). However, all those studies focus on tumor classification, in which the phenotype is binary or multi-class, rather than censored survival time.

Due to the high-dimensionality of microarray data, the SIR method cannot be applied directly to the microarray data. We propose to employ the principal components (PC) analysis in conjunction with SIR to achieve dimension reduction. Principal components analysis uses singular value decomposition (SVD) to recover the underlying structures and patterns of gene expression variation and has been applied widely in microarray data analysis (Holter *et al.*, 2000; Chiaromonte and Martinelli, 2002).

The rest of the paper is organized as follows: we first present the methods for sufficient dimension reduction for censored survival data. We propose an approach which combines both PC and SIR. We also present the idea of using the time dependent receiver-operator curve (ROC) and areas under the curves (AUCs) for evaluating the predictive performance of the proposed methods. Following the Methods section, we evaluate the proposed methods by analyzing the DLBCL data set of Rowenwald *et al.* (2002). Finally, we provide a summary of the paper and a brief discussion of the methods.

# METHODS

### Method of sufficient dimension reduction

The problem of classification, regression and survival time prediction can all be formulated as predicting a response outcome $Y$, which can be binary, multi-categorical, continuous, or censored, given a number of predictors $X$, with $X \in \mathbb{R}^p$. The goal of sufficient dimension reduction is to find a $p \times d$ matrix $\eta$, with $d \leq p$, such that

$$Y \perp\!\!\!\perp X \mid \eta^T X, \tag{1}$$

where $\perp\!\!\!\perp$ stands for the statistical independence. The statement (1) implies that the $p$-dimensional predictor vector $X$ can be replaced by $d$-dimensional $\eta^T X$ without loss of any information on regression of $Y$ given $X$, because given $\eta^T X$, $X$ contains no further information about $Y$. In practice, such $\eta$ exists with $d < p$, and in many applications $d$ is as small as 1, 2, or 3, therefore dimension reduction is achieved. Graphical data representation often becomes feasible as well, and it provides a powerful means to facilitate the subsequent model formulation.

It is easy to see that $\eta$ in (1) is not unique, because we can multiply $\eta$ by any non-zero constant and (1) still holds. Therefore, we seek the linear subspace $\text{Span}(\eta)$ which is spanned columns of $\eta$. Such a space is called a *dimension reduction subspace* (Li, 1991; Cook, 1996). The intersection of all the dimension reduction subspaces, which is also a dimension reduction subspace itself under minor conditions (Cook, 1994, 1996), provides the most parsimonious characterization of regression of $Y$ given $X$. It is named *central subspace*, denoted by $\mathcal{S}_{y|X}$, and is the main object of interest in our dimension reduction inquiry.

There are a number of methods to estimate $\mathcal{S}_{y|X}$ without making any model assumptions. Such methods include sliced inverse regression (Li, 1991) and sliced average variance estimation (SAVE) (Cook and Weisberg, 1991). SIR is employed in this article, but all ideas discussed here apply to SAVE and other sufficient dimension reduction methods as well. SIR first replaces $Y$ by a discrete version $\tilde{Y}$, constructed by partitioning its range onto $h$ intervals within which $\tilde{Y}$ is constant. It then shows that, under the linearity condition which will be discussed later, the inverse mean $\text{E}(X \mid \tilde{Y})$ belongs to $\mathcal{S}_{y|X}$, thus estimation of $\text{E}(X \mid \tilde{Y})$ provides useful information about $\mathcal{S}_{y|X}$. Operationally, SIR performs eigen-decomposition of matrix $\Sigma_{x|y} = \text{Cov}[\text{E}(X \mid \tilde{Y})]$, with respect to $\Sigma_x = \text{Cov}(X)$, i.e.,

$$\Sigma_{x|y}\, v_i = \lambda_i\, \Sigma_x\, v_i, \text{ with } \lambda_1 \geq \ldots \geq \lambda_p, \text{ and } v_i^T\, \Sigma_x\, v_i = 1. \tag{2}$$

With the linearity condition, the first $d$ eigenvectors of the above decomposition provide a consistent estimate of the basis of central subspace $\mathcal{S}_{y|X}$. There is also an asymptotic test available for finding the structural dimension $d = \dim(\mathcal{S}_{y|X})$ (Li, 1991). It involves a series of tests of hypotheses of $d = m$ versus $d > m$ for $m = 0, \ldots, p-1$. Estimate of $d$ is taken as the minimum $m$ that the null hypothesis $d = m$ is not rejected. Note that SIR does not impose any traditional assumption on the distribution of $Y \,|\, X$, therefore, any model can be applied in the subsequent analysis. On the other hand, SIR requires a condition on the marginal distribution of $X$, the linearity condition, which requires that $\mathrm{E}(X \,|\, \eta^T X)$ is a linear function of $\eta^T X$. When $X$ follows a normal distribution, the linearity condition is satisfied. Li (1991) and Hall and Li (1993) argued that this condition is not a severe restriction, because most low-dimensional projections of a high-dimensional data cloud are close to normal.

**Modification of SIR to censored survival data**

SIR cannot be applied directly to censored survival data. We propose here a modification of SIR to accommodate censoring. Let $X$ be the vector of gene expression values of $p$ genes. We first introduce the following notation related to survival data:

$$\begin{aligned}
Y^0 &= \text{the true unobservable survival time,} \\
C &= \text{the censoring time,} \\
\delta &= \text{the censoring indicator; } \delta = 1 \text{ if } Y^0 \le C, \text{ and } \delta = 0 \text{ otherwise,} \\
Y &= \text{the observed survival time; } Y = Y^0 \text{ if } Y^0 \le C, \text{ and } Y = C \text{ otherwise.}
\end{aligned}$$

Letting $\mathcal{Y}^0 = (Y^0, C)^T$, and $\mathcal{Y} = (Y, \delta)^T$, the goal of sufficient dimension reduction for survival data is to find $\eta$ such that

$$\mathcal{Y}^0 \perp\!\!\!\perp X \mid \eta^T X.$$

Implementation of SIR in this context requires estimation of $\mathrm{E}(X \,|\, \mathcal{Y}^0)$. However $\mathcal{Y}^0$ is not observable, instead what can be observed is $\mathcal{Y}$. Using the conditional probability arguments, we have the following relationship between $\mathrm{E}(X \,|\, \mathcal{Y})$ and $\mathrm{E}(X \,|\, \mathcal{Y}^0)$,

$$\mathrm{E}(X \,|\, \mathcal{Y}) = \mathrm{E}[\mathrm{E}(X \,|\, \mathcal{Y}, \mathcal{Y}^0) \,|\, \mathcal{Y}] = \mathrm{E}[\mathrm{E}(X \,|\, \mathcal{Y}^0) \,|\, \mathcal{Y}], \tag{3}$$

where the second equality holds because $\mathcal{Y}$ is a function of $\mathcal{Y}^0$, therefore $X \perp\!\!\!\perp \mathcal{Y} \,|\, \mathcal{Y}^0$. With the linearity condition, $\mathrm{E}(X \,|\, \mathcal{Y}^0) \in \mathcal{S}_{\mathcal{Y}^0|X}$. Then equation (3) implies that $\mathrm{E}(X \,|\, \mathcal{Y})$, which is

a linear combination of $\mathrm{E}(X \mid \mathcal{Y}^0)$, also belongs to central subspace $\mathcal{S}_{\mathcal{Y}^0|X}$ of interest. Operationally, we slice $\mathcal{Y} = (Y, \delta)^T$ to obtain its discrete version $\tilde{\mathcal{Y}}$. Specifically, we first partition $\mathcal{Y}$ to $\mathcal{Y}_1$ for $\delta = 1$ and $\mathcal{Y}_0$ for $\delta = 0$. We then partition $\mathcal{Y}_1$ and $\mathcal{Y}_0$ to $h$ intervals respectively. This procedure is called double slicing in Li *et al.* (1999). Once $\tilde{\mathcal{Y}}$ is obtained, the same eigenvalue decomposition as in equation (2) can be performed. See also Setodji (2003) for discussion of SIR for survival data.

**Combination of SIR and PC analysis**

Implementation of SIR requires the covariance matrix $\Sigma_x$ of $X$ to be non-singular. This condition is satisfied in many applications. However, for microarray data, the number of genes $p$ is often much larger than the number of sample $n$, and in this situation, $\Sigma_x$ is singular. To address this issue, we propose to first obtain $q$ principal components from original $X$, with $q < n$ and then to apply sufficient dimension reduction methods with principal components as input. Singular value decomposition is used to find principal components. Holter *et al.* (2000) suggested that SVD is useful to recover the fundamental structures of gene expressions. SVD has also been used in many other microarray studies (e.g., Chiaromonte and Martinelli, 2002). One advantage of using principal components, compared to alternative reduction strategy such as identifying significant individual genes based on simple $t$ test, is that PC analysis takes into account correlations among genes. We illustrate this, as well as how to choose the number of principal components $q$, in the following section.

**Time dependent ROC curves and area under the curves**

To evaluate the predictive performance of the proposed methods, we propose to utilize the idea of time dependent ROC for censored data and AUC as our criteria. These methods were recently developed by Heagerty *et al* (2002) in the context of the medical diagnosis. For a given score function $f(x)$, we can define time dependent sensitivity and specificity functions as

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

and define the corresponding $\mathrm{ROC}(t|f(x))$ curve for any time $t$ as the plot of $\text{sensitivity}(c, t|f(x))$ vs $1 - \text{specificity}(c, t|f(x))$ with cutoff point $c$ varying, and the area under the curve 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 $t$ indicates better predictability of time to event at time $t$ as measured by sensitivity and specificity evaluated at time $t$.

# RESULTS

We present the results of application of the proposed dimension reduction techniques to the DLBCL data set of Rosenwald *et al.* (2002).

### Description of the data set and preprocessing of the data

The data set consists of measurements of 7399 genes from 240 patients. Of those 240 patients, 160 were used for training the model and 80 were reserved for model validation in Rosenwald *et al.* (2002). To facilitate comparisons with results in Rosenwald *et al* and other analyses, the same training and testing sets were used in our analysis. A survival time was recorded for each patient, which ranges between 0 and 21.8 years. Among them, 138 were dead (uncensored) during the study, and 102 were alive at the end of the study (censored). Detailed description of the data can be found in Rosenwald *et al.* (2002).

There were a large number of missing gene expression values in the data set. Among the 7399 genes, only 434 genes have no missing values. We first applied a nearest neighbor technique (Troyanskaya *et al.*, 2001) 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. Our method is slightly different from that of Troyanskaya *et al.* (2001) in that the nearest neighbors are not restricted to those 434 genes with no missing. We also tried the method of Troyanskaya *et al.* (2001) for filling the missing value, and the results of survival time prediction with two methods were very close.

Principal components were then identified based on the complete data of the training samples. We chose $q = 40$ PCs, which accounts for about 70% of total variation, to construct the predictive components in the following analysis. Choice of the number of principal components is further discussed in later section.

**Identification of the predictive components based on the training data set**

Examining the marginal scatter plot of the 40 principal components revealed no strong violation of the linearity condition. Sliced inverse regression was then applied based on those principal components. The p-values of asymptotic test for $d = 0, 1, 2$ and 3 were 0.063, 0.372, 0.679, and 0.873 respectively. It suggested that $d = 1$, and only the first SIR linear combination, abbreviated as $\text{SIR}_1$, is needed for subsequent analysis. Figure 1 shows the patients' survival time versus the first two SIR linear combinations, i.e., SIR covariates $\text{SIR}_1$ and $\text{SIR}_2$, obtained from the training data. It is clear that $\text{SIR}_1$ was able to differentiate between dead and surviving patients, while $\text{SIR}_2$ did not provide useful information. This agrees with the result of asymptotic tests. In addition, by examining Figure 1, we noticed that the difference of survival time with respect to SIR1 included both a location difference and a scale difference, which in turn suggests that both the first and second order terms of $\text{SIR}_1$ may be needed in the model.

Since SIR imposes no model assumption in the stage of dimension reduction, we are free to fit any model based on the identified SIR covariates. To compare our method with others, we fitted a Cox proportional hazards model. The model suggested that both the linear and quadratic terms of $\text{SIR}_1$ was significant (p-value $= 4.3 \times 10^{-11}$ and 0.087 respectively), and $\text{SIR}_2$ was insignificant (p-value = 0.2). This agrees with what is observed in Figure 1. The final model was

$$\lambda_i(t \,|\, \text{SIR}_1) = \lambda_0(t)\ \exp(0.2418\ \text{SIR}_{1i} - 0.0046\ \text{SIR}_{1i}^2),$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, and $\lambda_i(t \,|\, \text{SIR}_1)$ is the hazard function for the $i$th patient. In this model, the gene expression profile measured over $p$ genes is related to the risk of death through the score function $f(\text{SIR}_1) = 0.2418\ \text{SIR}_1 - 0.0046\ \text{SIR}_1^2$.

Figure 2 shows the Kaplan-Meier estimate of survival curves for two groups of patients, the high-risk patients and the low-risk patients, defined by the scores $f(\text{SIR}_1) > 0$ or $f(\text{SIR}_1) < 0$. The cutoff value of 0 was chosen for convenience and it was close to the median of all scores. Figure 2(a) plots the survival curves for 160 training patients. The log-rank test of difference between two survival curves yielded a p-value of $1.89 \times 10^{-15}$, indicating a large difference in overall survival between the two groups. Figure 2(b) shows the survival curves for 80 testing patients, where the scores were computed based on the coefficients estimated using the training samples only. The difference between the two risk groups is still very significant, with p-value of the log-rank test of $2.17 \times 10^{-5}$. Both the plots and the tests suggest that the

Cox model built based on the dimension reduction technique can indeed be used to identify patients with different risk of death.

**Effects of the number of PCs used in the model and time-dependent AUCs.**

We further examine the issue of choosing the proper number of principal components $q$ for building the SIR components. We evaluated a series of values of $q$ ranging between 10 and 150. For each $q$, we performed dimension reduction and fitted a Cox model based on 160 training patients. We then evaluated the model for both the training and the testing patients using the area under ROC curve as a comparison criteria. Figure 3 shows the AUC for each value of $q$ for time ranging from 1 to 10 years. Note the ranges of the AUCs in the two plots are different and are not from 0 to 1. We first noted that the AUCs are essentially the same for $q$ between 30 and 130. Second, the plots confirm what one would expect: under-fitting for small $q$ and over-fitting for large $q$. As an illustration, three $q$ values, $10, 40$, and 150 represented by annotations 1, 4, and $f$ respectively, were emphasized in the plot by thick lines. When $q = 1$, the area under ROC in both the training and the testing data were low due to the lack of fitting. When $q = 150$, the area under ROC was high in the training samples but low in the testing samples, indicating over-fitting of the model. The number of PCs of $q = 40$ seems to provide a nice balance.

**Comparisons with PC analysis**

We now compare the method which combines SIR and PC analysis with the principal components Cox regression analysis. Although a Cox proportional hazards model can be fitted with 40 PCs as covariates based on the training set of 160 patients, the model depends on 40 predictors, which makes its interpretation more complicated. In addition, with 40 predictors, there is much less freedom to choose the form of the fitted model such as higher-order terms. One possible solution is to use cross-validation methods to identify significant principal components out of the 40 PCs and to build a model based on the selected PCs. We applied the cross-validated partial likelihood (Verwij *et al.*, 1993; Huang and Harrington, 2002) method on the training data and identified that the model with the first three PCs gives the best predictive performance.

Figure 4 compares the performance of the Cox proportional hazards models using the combination of PC and SIR, using all 40 PCs and using only the three PCs chosen by cross-

validation. It is clear that the model with combination of PC and SIR outperforms the other two methods. The p-values of log-rank test of difference between two risk groups in the testing data set are 0.000022, 0.0034 and 0.0333 respectively. For AUCs, SIR was the best for the training samples, and SIR and PC regression analysis with all 40 components were the best for testing samples. Overall, the Cox model built based on the combination of PC and SIR shows the best predictive performance.

**Comparisons with other analyses**

While it is out of the scope of this paper to compare the proposed methods with all available methods for relating gene expression profiles to censored outcomes, we compare our results to a few other analysis of the DLBCL data set. We focused on the method's performance in predicting the patients' survival time. It should however be noted that such comparison cannot be comprehensive since methods proposed by the other studies have their own desirable properties other than the survival prediction. In Bair and Tibshirani (2003), low-risk and high-risk patients were classified according to the fitted model employing supervised clustering and partial least squares. Comparing our Figure 2(b) to Figure 6 of Bair and Tibshirani, we observe that the results are comparable, while our method shows slightly higher significance in overall survival between the two risk groups in the testing data sets. The p-value of log-rank test for the difference of two survival curves is 0.0000217 for our method and 0.000827 for that of Bair and Tibshirani (2003).

Finally, it is interesting to compare our results with those presented in Rosenwald *et al.* (2002). Instead of dividing the patients into high- and low-risk groups, Rosenwald *et al.* (2002) divided these patients into fours risk groups based on the quartiles of the estimated scores (see Figure 2 of Rosenwald *et al.* (2002)). Figure 5 (a), (b) and (c) show the plots of the Kaplan-Meier estimates of overall survival among patients in the training group, the testing group, and all patients based on the quartiles of the respective estimated scores. Significant results are observed for all three plots. For panels (b) and (d), scores were evaluated for the same testing patients, but the difference lies in how the SIR covariate was obtained. For (b), the SIR covariate was obtained by applying principal component analysis and sliced inverse regression to 80 testing patients, while for (d), the covariate was obtained solely based on 160 training patients. Panel (b) corresponds to panel (B) of Figure 2 in Rosenwald *et al.* (2002), but panel (d) provides a better way of measuring the predictive performance for a future

patient's survival, because no testing patients information was used in building the prediction model. The p-value of log-rank test of difference were also given. Again, both the survival plot and the test indicate a good predictive performance of our proposed method.

## DISCUSSION

We have proposed the use of sufficient dimension reduction techniques to reduce the high-dimensional microarray data to a low-dimensional space for censored survival phenotypes. The proposed dimension reduction method is non-parametric without making any distributional assumptions on the data, therefore, it allows a very flexible model formulation in the subsequent model building step. Visualization is also available to facilitate the data analysis. For DLBCL data of Rosenwald *et al.* (2002), the dimension reduction method, particularly the sliced inverse regression, was combined with a Cox proportional hazards model, which together provided a good predictive performance for a future patient's survival.

Sufficient dimension reduction in the context of censored data was addressed, where the goal is to recover the most parsimonious space, the central subspace, of the true survival time $Y^0$ and censoring time $C$ given dependent predictor variables. Since $Y^0$ and often $C$ are unobservable, reduction was achieved through observed survival time $Y$ and status $\delta$. In some situations, only the central subspace of $Y^0$ given predictors is of interest. In this case, the proposed method works without modification if $C$ is a constant, or $C$ is independent of the true survival time as well as the dependent variables. Otherwise, slight modification is needed, which was discussed in Li *et al.* (1999).

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. One approach which is often employed in microarray analysis is to first select a number of individual genes based on univariate analysis. In survival data, such selection is usually based on the univariate Cox proportional hazards model. A disadvantage of this method is that the significance of genes is measured individually without accounting for correlations among genes and possible combinatorial effects of genes on the risk of event. For example, for the DLBCL data set, applying an univariate Cox model to 160 training patients identifies 473 genes which are significant at 0.01 level. For 80 testing patients, however, only 67 genes were significant, out of which only 4 genes were identified significant in both groups. Applying the proposed methods on these 473 genes resulted in very poor performance due to the possible

combinatorial effects of the gene expressions on the survival (details not shown). An alternative 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 until the resulted model gives significantly worse performance in prediction. We are currently investigating this possibility.

In summary, we have proposed a procedure which combines the principal components analysis and efficient dimension reduction technique for censored survival data. The procedure can be used for building a parsimonious predictive model for survival based on microarray gene expression profiles.

## ACKNOWLEDGMENTS

This research was supported by NIH grants ES11269 (L. Li) and ES09911 (H.Li).

## REFERENCES

Akritas, M.G. (1994). Nearest neighbor estimation of a bivariate distribution under random censoring. *Annals of Statistics*, **22**:1299-1327.

Alter, O., Brown, P.O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of National Academy of Sciences, USA*, **97**, 10101-10106.

Antoniadis, A., Lambert-Lacroix, S., and Leblanc, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data. *B*ioinformatics **1**9, 563-70.

Bair, E., and Tibshirani, R. (2003). Semi-supervised methods to predict patient survival from gene expression data. Technical report, Department of Statistics, Stanford University.

Barlow, W.E., and Prentice, R.L. (1988). Residuals for relative risk regression. *B*iometrika **7**5, 65-74.

Bura, E., and Pfeiffer, R.M. (2003). Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. *Bioinformatics*, **19**, 1252-1258.

Cook, R.D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, **89**, 177-190.

Cook, R.D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, **91**, 983-992.

Cook, R.D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*, New York: Wiley.

Cook, R.D., and Weisberg, S. (1991). Discussion of Li (1991). *Journal of American Statistical Association*, **86**, 328-332.

Chiaromonte, F., and Martinelli, J. (2002). Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences*, **176**, 123-144.

Hall, P., and Li, K.C. (1993). On almost linearity of low dimensional projections from high dimensional data. *Annals of Statistics*, **21**, 867-889.

Heagerty, P.J., Lumley, T., Pepe, M. (2002). Time dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, **56**:337-344.

Holter, N.S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J.R., and Fedoroff, N.V. (2003). Fundamental patterns underlying gene expression profiles: Simplicity from complexity, *Proc. Natl. Acad. Sci.* USA, **97**, 8409-8414.

Huang, J., 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 on Biocomputing*, **8**, 65-76.

Li, K.C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, **86**, 316-327.

Li, K.C., Wang, J.L., and Chen, C.H. (1999). Dimension reduction for censored regression data. *The Annals of Statistics*, **27**, 1-23.

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.

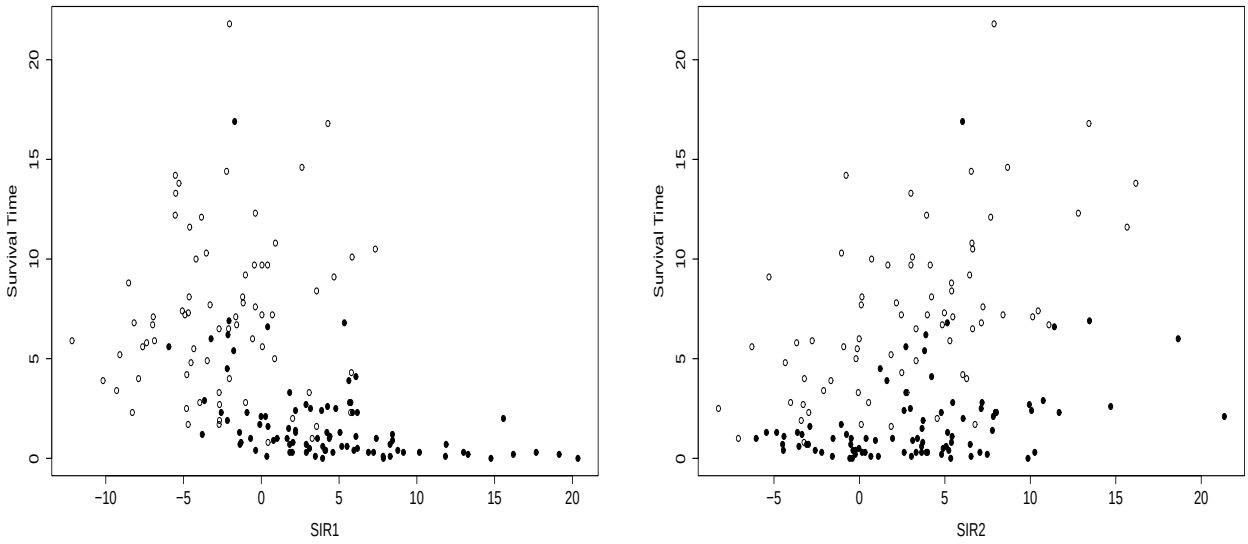
Rosenwald, A., Wright, G., Chan, W.C., Connors, J.M., Campo, E., Fisher, R.I., Gascoyne, R.D., Muller-Hermelink, H.K., Smeland, E.B., and Staudt, L.M. (2002). The use of

molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England Journal of Medicine*, **346**, 1937-1947.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R.B. (2001). Missing value estimation methods for DNA microarrays *B*ioinformatics, **17**, 520-525.

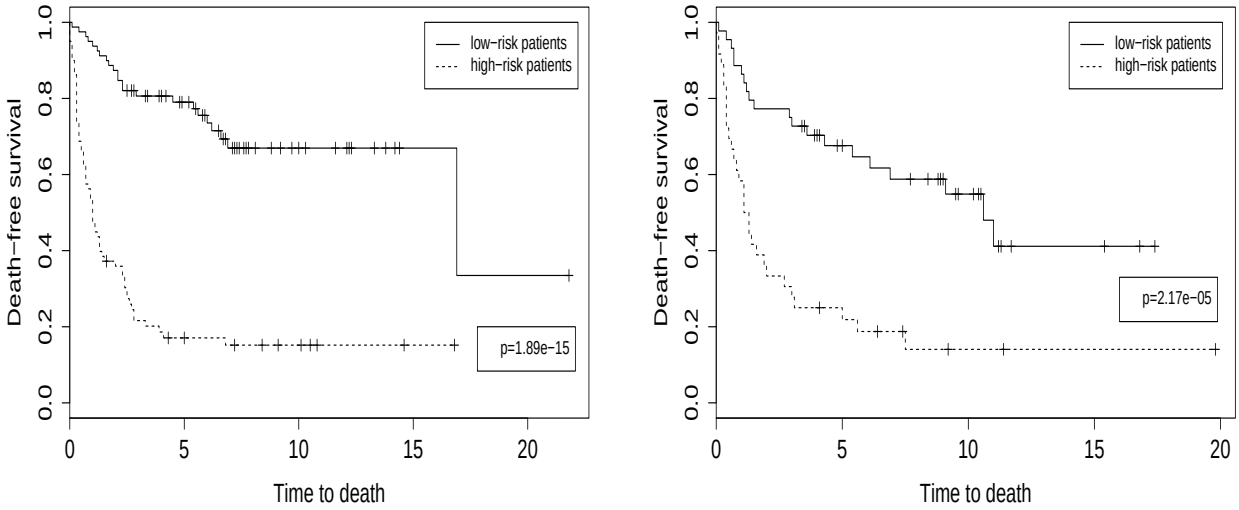
Setodji, M.C. (2003). Multivariate dimension reduction and graphics. Ph.D. Dissertation. School of Statistics, University of Minnesota.

Verwij, P.J.M., Van Houwelingen, J.C. (1993). Cross validation in survival analysis. *Statistics in Medicine*, **12**:2305-2314.



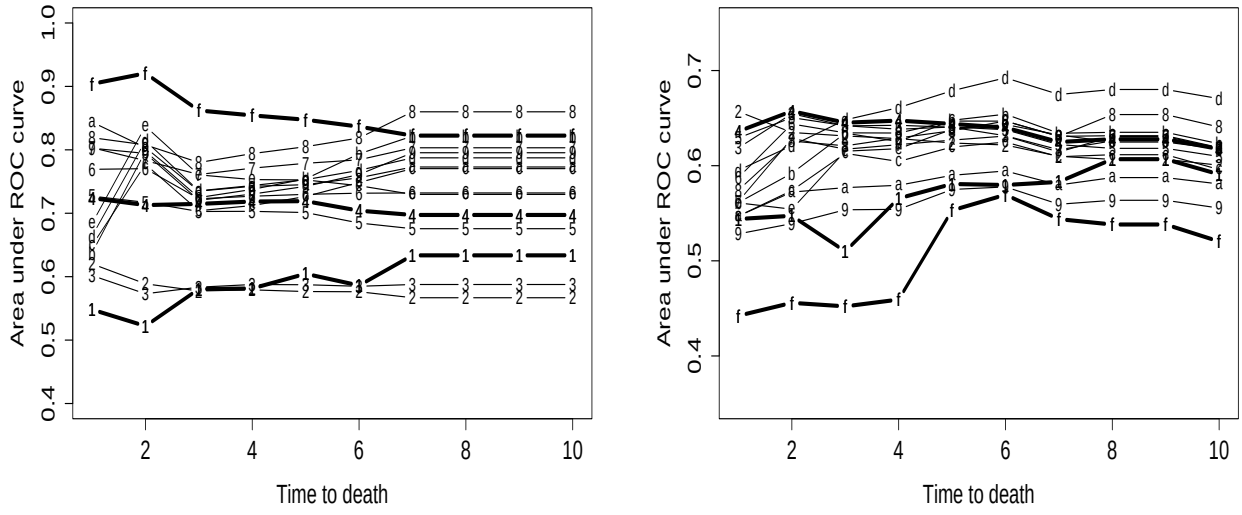

(a) SIR1 (b) SIR2

Figure 1: Survival time versus SIR covariates (a: SIR1, b: SIR2) for patients in the training data set; dot: dead, circle: alive.



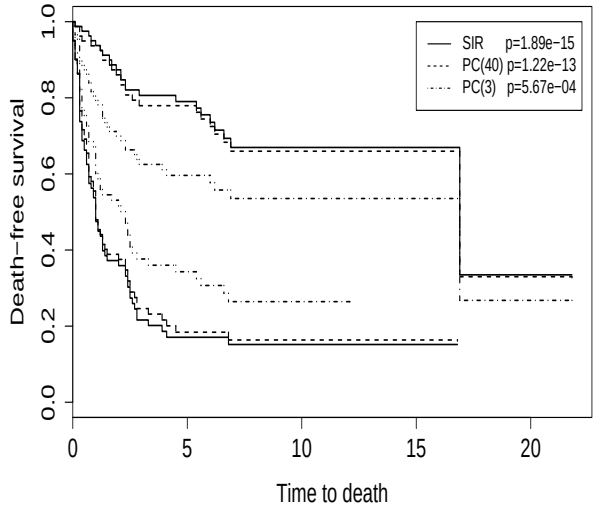

(a) Training data, p-value = 1.89e-15 (b) Testing data, p-value = 2.17e-05

Figure 2: Survival curves for patients in two groups of having positive and negative estimated scores using gene expression profiles. (a) 160 patients in the training set; (b) 80 patients in the testing set.

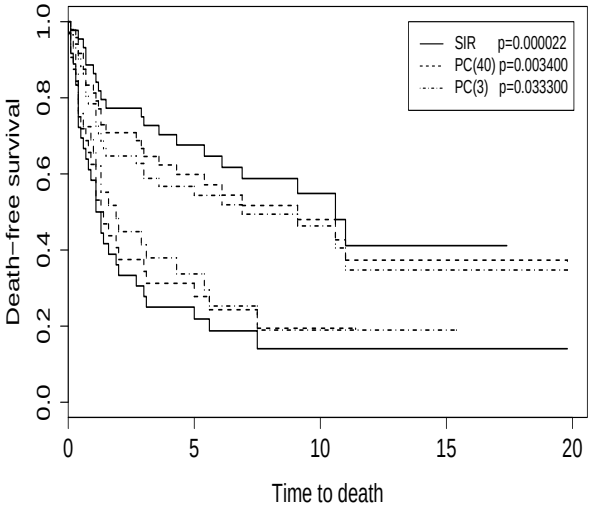



(a) Training data (b) Testing data

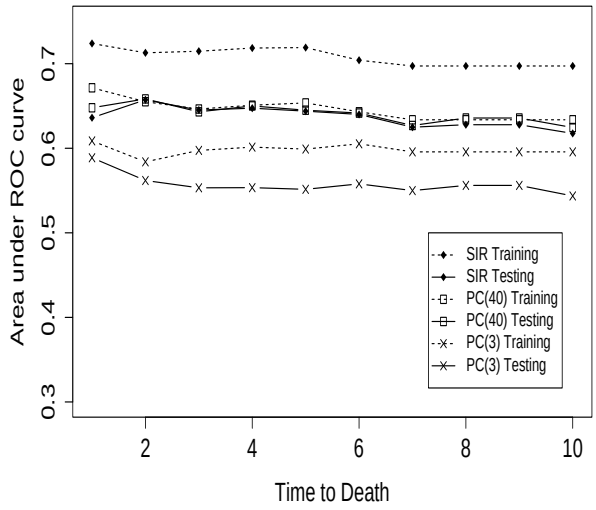
Figure 3: Area under ROC at time 1 year to 10 years for models based on different number of principal components ($q$) in estimating the SIR components. Numbers 1 through 9, plus characters $a$ through $f$ represent $q = 10, 20, \ldots, 150$ respectively.




(a) Training data

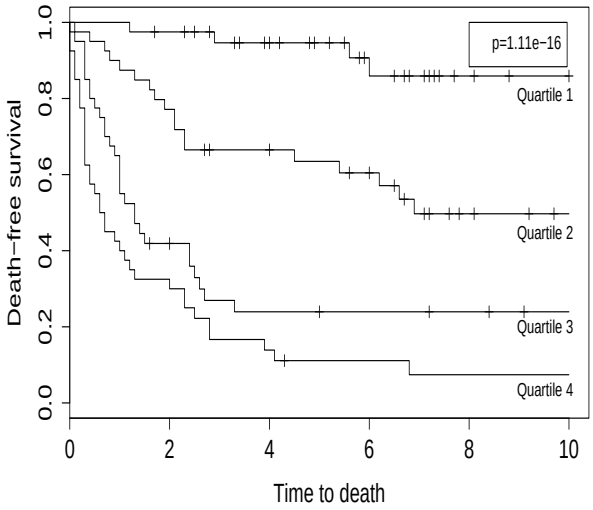



(b) Testing data

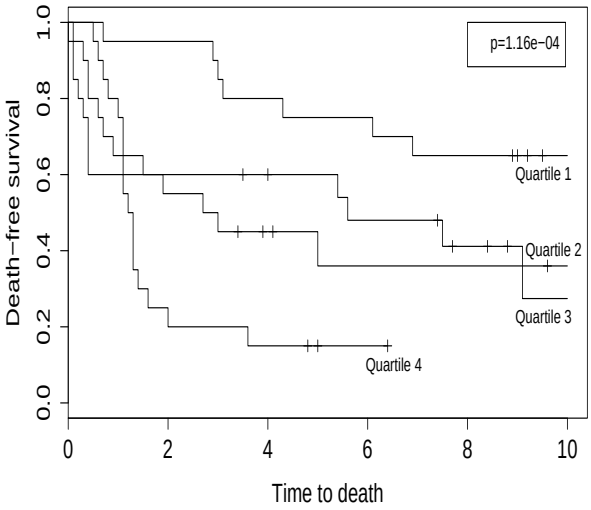



(c) Area under ROC

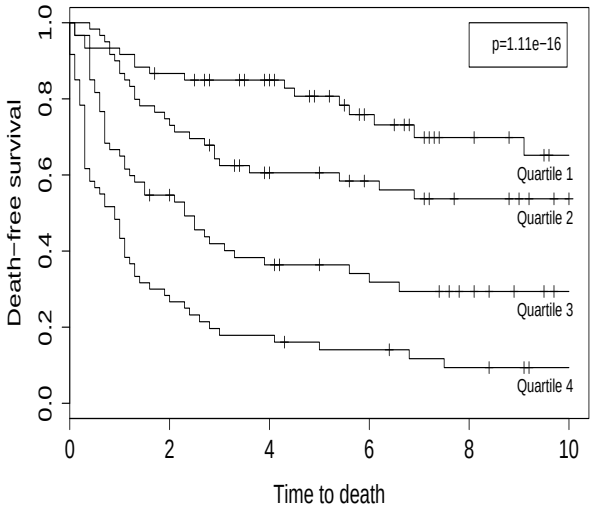
Figure 4: Comparisons between the principal components Cox models and the SIR Cox models. (a) Estimated survival curves for two groups of patients in the training data set. (b) Estimated survival curves for two groups of patients in the testing data set. (c) Time-dependent AUCs comparison of different models. The three models are: SIR=Cox model with components constructed by PC analysis and SIR; PC(40)=Cox model with 40 PCs; PC(3)=Cox model with 3 PCs selected by cross-validation.



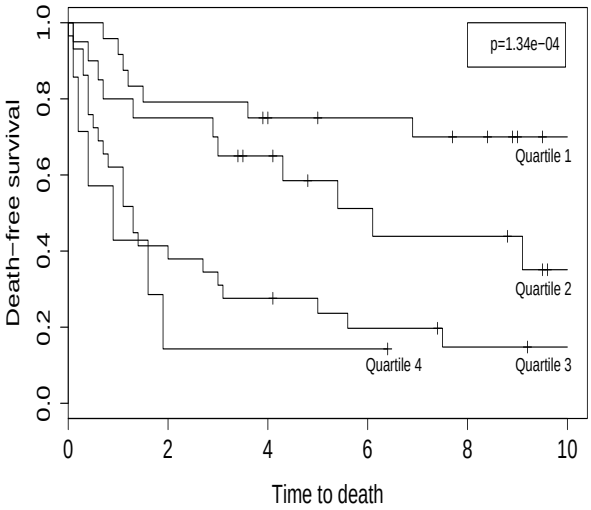

(a) Training data, p-value = 1.11e-16




(b) Testing data, p-value = 1.16e-04




(c) All data, p-value = 1.11e-16




(d) Testing data, p-value = 1.34e-04

Figure 5: Survival curves for patients based on the their estimated scores using gene expression profiles. (a) Four groups of patients in the training set defined by the quartile of their estimated scores; (b) four groups of patients in the testing set defined by the quartile of their estimated scores; (c) four groups of patients for all the patients defined by quartile of their estimated scores; (d) four groups of patients in the testing set based on the quartiles of the scores estimated by using the model derived from training data set.