Bioinformatics Advance Access published July 15, 2004

Dimension Reduction Methods for Microarrays with Application to Censored Survival Data

Lexin Li[1*], *and Hongzhe Li*[2]
[1]*Department of Biochemistry and Molecular Medicine, and* [2]*Rowe Program in Human Genetics, School of Medicine, University of California, Davis, CA 95616, USA*

Running title: Microarrays and Censored Survival Data

*To whom correspondence should be addressed:
Lexin Li, Ph.D.
Department of Biochemistry and Molecular Medicine
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 various clinical phenotypes such as tumor class, drug response, and survival time. While there has been extensive studies on tumor classification, there has been less emphasis in other phenotypic features, in particular, patient survival time or time to cancer recurrence, which are subject to right censoring. We consider in this paper an analysis of censored survival time based on microarray gene expression profiles.

Results: We propose a dimension reduction strategy, which combines principal components analysis and sliced inverse regression, to identify linear combinations of genes, that both account for the variability in the gene expression levels and preserve the phenotypic information. The extracted gene combinations are then employed as covariates in a predictive survival model formulation. We apply the proposed method to a large diffuse large-B-cell lymphoma data set, which consists of 240 patients and 7399 genes, and build a Cox proportional hazards model based on the derived gene expression components. The proposed method is shown to provide a good predictive performance for patient survival, as demonstrated by both the significant survival difference between the predicted risk groups, and the receiver operator characteristics analysis.

Availability: R programs are available upon request from the authors.

Supplementary Information: http://dna.ucdavis.edu/∼hli/bioinfo-surv-supp.pdf.

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

INTRODUCTION

DNA microarray technology, which simultaneously measures the expression levels of thousands of genes, is a ground-breaking advance in biomedical and genomic research. 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).

In recent years, tumor class prediction using gene expression data has been studied extensively (see Dudoit *et al.*, 2002 for a review.) However, there has been less development in relating gene expression profiles to other phenotypes, e.g., survival time, due to a number of challenges. First, the microarray-based high-throughput technology generates a huge number of potential predictors, i.e., 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 complex. 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.

Among a few recent microarray studies of censored survival time, Rosenwald *et al.* (2002) employed hierarchical clustering to first identify a small number of "signature" gene clusters. Based on those gene clusters, they then built a Cox proportional hazards model for predicting time to death in patients with diffuse large-B-cell lymphoma (DLBCL). One disadvantage of clustering genes is that the sample phenotypes are not efficiently used. Nguyen and Rocke (2002) proposed to use partial least squares for survival data by employing 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. Park *et al.* (2002) circumvented the problem of censoring by reformulating the problem as a standard Poisson regression, and then employed partial least squares. But their method is limited to the linear Cox proportional hazards model, because the transformation is valid only for that particular model. More recently, 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. Bair and Tibshirani (2003) re-analyzed the lymphoma data set of Rosenwald *et al.* (2002) by applying the nearest shrunken centroid supervised clustering and partial least squares techniques.

In this article, we introduce a dimension reduction strategy to transform the high-dimensional gene expression data to a low-dimensional space. A predictive survival model is then built upon the reduced dimensional space. The proposed dimension reduction strategy consists of two steps. We first employ principal components (PC) analysis to identify a number of linear combinations of genes that capture the underlying variation structures of gene expressions. We then apply sliced inverse regression (SIR, Li, 1991), a technique of sufficient dimension reduction (SDR, Cook, 1998), to produce linear combinations of genes that preserve all information of phenotype given gene expression. The extracted linear

combinations of genes are then employed as covariates in the subsequent survival model formulation. Our method differs from others in that it both accounts for the variability in the predictor space, and it preserves all response information through the extracted gene components, so to assure the predictive power. Additionally, no probabilistic model is imposed in the dimension reduction process, thus it allows the investigators to fit any model in the subsequent model building stage of the analysis. Principal components analysis has been widely applied in microarray studies, see for instance, Alter, *et al.* (2000), Holter *et al.* (2000), and Chiaromonte and Martinelli, (2002). Similar sufficient dimension reduction techniques have been studied in the microarray context, for instance, Chiaromonte and Martinelli (2002), Bura and Pfeiffer (2003), Antoniadis *et al.* (2003), and Pérez-Enciso and Tenenhaus (2003). However, all those studies focus on tumor classification, in which the phenotype is binary or multi-class, rather than censored survival time.

The rest of the paper is organized as follows: we first present sufficient dimension reduction method for censored survival data, and propose a strategy that combines PC and SIR. We then 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 method (Heagerty, *et al.*, 2002). Following the Methods section, we apply our method to the DLBCL data set of Rosenwald *et al.* (2002). Finally, we conclude with a brief discussion.

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 η, with $d \leq p$, such that

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

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 $\eta^{\mathsf{T}} X$ without loss of any information on regression of Y given X, because given $\eta^{\mathsf{T}} X$, X contains no further information about Y. In practice, such η exists, and d is often far less than p, hence dimension reduction is achieved. In many applications d is as small as 1, 2, or 3, therefore, a fully informative data visualization becomes feasible.

It is easy to see that η in (1) is not unique, because we can multiply η by any non-zero constant and (1) still holds. Therefore, we seek the linear subspace Span(η) which is spanned by the columns of η. Such a space is called a dimension reduction subspace (Cook, 1998). 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, and is a unique population parameter. It is called the 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 model-free methods to estimate $\mathcal{S}_{y|X}$, for instance, sliced inverse regression (Li, 1991) and sliced average variance estimation (SAVE) (Cook and Weisberg,

1991). We consider SIR in this article. SIR first replaces Y by a discrete version $\tilde{Y}$ constructed by partitioning its range onto h intervals within which $\tilde{Y}$ is constant. h is a tuning parameter of SIR, but the choice of h usually does not affect the SIR estimate as long as $h > d$ (Li, 1991). It is then shown that, under a linearity condition which is to be discussed later, the inverse mean $\mathrm{E}(X \mid \tilde{Y})$ belongs to $\mathcal{S}_{y|X}$, thus estimation of $\mathrm{E}(X \mid \tilde{Y})$ provides useful information about $\mathcal{S}_{y|X}$. Operationally, SIR performs eigen-decomposition of the matrix $\Sigma_{x|y} = \mathrm{Cov}[\mathrm{E}(X \mid \tilde{Y})]$, with respect to $\Sigma_x = \mathrm{Cov}(X)$, i.e.,

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

The first d eigenvectors $\{v_1, \ldots, v_d\}$ in (2) provide a consistent estimate of a basis for the central subspace $\mathcal{S}_{y|X}$. There are asymptotic tests available for determining the dimension $d = \dim(\mathcal{S}_{y|X})$ of the central subspace. It consists of a sequence of tests of hypotheses $d = m$ 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 $Y \mid X$, henceforth, it allows a full flexibility in the subsequent model formulation. On the other hand, SIR requires a condition on the marginal distribution of X, the linearity condition, which assumes that $\mathrm{E}(X \mid \eta^{\mathsf{T}} X = u) = A_0 + A_1 u$, where $A_0 \in \mathbb{R}^p$ and A_1 is a $p \times d$ matrix. The condition is satisfied when X follows a normal distribution, and it is shown not to be a severe restriction, because most low-dimensional projections of a high-dimensional data cloud are close to normal (Hall and Li, 1993). SIR is available in both statistical software *R* (Ihaka and Gentleman, 1996) and *Arc* (Cook and Weisberg, 1999).

Modification of SIR to censored survival data

SIR can not be applied directly to the survival data because of censoring. 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:

Y^0	=	the true unobservable survival time,
C	=	the censoring time,
δ	=	the censoring indicator; $\delta = 1$ if $Y^0 \leq C$, and $\delta = 0$ otherwise,
Y	=	the observed survival time; $Y = Y^0$ if $Y^0 \leq C$, and $Y = C$ otherwise.

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

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

Implementation of SIR in this context requires estimation of $\mathrm{E}(X \mid \mathcal{Y}^0)$. 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 $\mathrm{E}(X \mid \mathcal{Y})$ and $\mathrm{E}(X \mid \mathcal{Y}^0)$,

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

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

the central subspace $\mathcal{S}_{\mathcal{Y}^0|X}$. Operationally, we slice $\mathcal{Y} = (Y, \delta)^\top$ to obtain its discrete version $\tilde{\mathcal{Y}}$. That is, 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 Σ_x of X to be non-singular, a condition that is often satisfied. However, for microarray data, the number of genes p is much larger than the number of samples n, in which case Σ_x is singular. To address this problem, we adopt the idea of Chiaromonte and Martinelli (2002) to combine SIR with principal components analysis. That is, we first obtain q principal components based on correlations among all genes with $q < n$. We then apply SIR with principal components as input. By doing so, the dimension reduction takes into account both the predictor variability and correlates the extracted linear combinations of genes with the response. Selection of the number of principal components q will be discussed in the Results section.

Time dependent ROC curves and area under the curves

To evaluate the predictive performance of the proposed method, we employ the idea of time dependent ROC for censored data and AUC as our criterion (Heagerty *et al.*, 2002). For a given score function $f(x)$, we 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 $\text{ROC}(t|f(x))$ curve for any time t as the plot of $\{\text{sensitivity}(c, t|f(x))\}$ versus $\{1 - \text{specificity}(c, t|f(x))\}$, with cutoff point c varying. The area under the curve, $\text{AUC}(t|f(x))$, is defined as the area under the $\text{ROC}(t|f(x))$ curve. 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

Data description and missing values

The DLBCL data set of Rosenwald *et al.* (2002) 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 their results, as well as other analyses of the same data in the literature, we use the same training and testing sets 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). More description of the data can be found in Rosenwald *et al.* (2002).

There are a large number of missing expression values in the data. Among the 7399 genes, only 434 genes have no missing values. We first apply a nearest neighbor technique (Troyanskaya *et al.*, 2001) to estimate those missing values. Specifically, for each gene, we first identify 8 genes which are the nearest neighbors according to Euclidean distance. We then fill the missing with the average of the nearest neighbors. Our method is slightly different from that of Troyanskaya *et al.* (2001) in that we do not restrict the nearest neighbors only to those 434 genes with no missing. We have tried both methods of filling missing values and the results on survival prediction are very close.

Identification of predictive components

Principal components are first identified based on the training samples. With the number of PCs q ranging between 10 and 120, the accounted percentage of variation ranges between 45% and 95%. We choose $q = 40$ PCs, which accounts for about 70% of total variation, for subsequent analysis. Choice of q will be further discussed later.

Examining the marginal scatter plot of the 40 principal components reveals no strong violation of the linearity condition. Sliced inverse regression is then applied with those PCs as input. The p-values of the asymptotic tests for $d = 0, 1, 2$ and 3 are 0.063, 0.372, 0.679, and 0.873 respectively. It suggests that the first SIR linear combination captures all response information. Let s denote this extracted linear combination of gene expression levels. Figure S1 (see web supplement) plots the patients survival time versus s with the censored status marked (circle denotes censored and dot denotes uncensored). It is clear that s is capable of differentiating between the dead and surviving patients. We also note that the difference of survival time of censored and uncensored patients with respect to s consists of both a location difference and a scale difference. Thus we may consider both the linear and quadratic terms of s in the subsequent modeling (Cook and Weisberg, 1999).

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 fit a Cox proportional hazards model. It turns out that both the linear and quadratic terms of s are significant (p-value $= 4.3 \times 10^{-11}$ and 0.087 respectively), while the second SIR linear combination is insignificant (p-value = 0.2). This agrees with the results of asymptotic tests. The final model is

$$\lambda_i(t \mid s_i) = \lambda_0(t)\ \exp(0.2418\ s_i - 0.0046\ s_i^2),$$

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

Figure 1 shows the Kaplan-Meier estimate of survival curves for two groups of patients, the high-risk patients ($f(s) > 0$) and the low-risk patients ($f(s) < 0$). The cutoff value 0 is chosen for convenience. Figure 1(a) plots the survival curves for the 160 training patients. The log-rank test of difference between two survival curves yields a p-value of 1.89×10^{-15}, indicating a significant difference in overall survival between the two groups. Figure 1(b) shows the survival curves for the 80 testing patients, where the scores are computed based on the model that is estimated from the training samples only. The difference between the two risk groups is still significant, with p-value of the log-rank test equal to 2.17×10^{-5}.

Both the survival plots and the log-rank tests show that our proposed method works very well in predicting patients survival risks.

Number of PCs, and cross-validation

We next examine the problem of choosing a proper number of principal components q. For each q in a sequence of values ranging from 10 to 150, we perform SIR and fit a Cox model for the 160 training patients. We then evaluate the model for both the training and the testing patients using the area under ROC curves as a comparison criterion. Figure S2 (web supplement) shows the AUCs for each value of q with survival time ranging from 1 to 10 years. We observe that the AUCs are essentially the same for q between 30 and 130. Besides, the plots confirm possible under-fitting for a small value of q and over-fitting for a large q. As an illustration, three q values, $10, 40$, and 150 represented by annotations `1`, `4`, and `f` respectively, are highlighted in the plot by thick lines. When $q = 1$, the area under ROC in both the training and the testing data are low due to the lack of fitting. When $q = 150$, the area under ROC is 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.

We also verify the performance of our method using a 5-fold cross-validation. AUCs are again employed as a comparison criterion. Figure 2 shows the average AUCs plus and minus one standard error for training and testing data. The relative stable performance of the fitted models can be seen in the plot.

Comparisons with other analyses

While it is out of the scope of this paper to compare the proposed method with all available methods for relating gene expression profiles to censored outcomes, we compare our results to a few other analyses of the DLBCL data set. We focus on the method's performance in predicting the patients survival time. It should be noted, however, that such a comparison can not be comprehensive since methods proposed by the other studies may have their own desirable properties other than the survival prediction.

We first compare our model with the principal components Cox regression analysis, i.e., a Cox model based on PC alone. Although a Cox proportional hazards model can be fitted with 40 principal components as covariates, the model involving 40 predictors is difficult to interpret. In addition, with 40 predictors, there is much less freedom to choose the form of the fitted model such as including 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 apply the cross-validated partial likelihood (Verwij *et al.*, 1993; Huang and Harrington, 2002) method on the training data and identify that the model with the first three PCs (accounting for about 25% of total variation) gives the best relative predictive performance. Figure S3 (web supplement) 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 shown 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 2.17×10^{-5}, 3.40×10^{-3} and 3.33×10^{-2} for the three methods respectively. For

AUCs, SIR is the best for the training samples, and SIR and PC with all 40 components are the best for testing samples. PC with only three components performs poorly. Overall, the Cox model built based on the combination of PC and SIR shows the best performance in both prediction and interpretation aspects.

Bair and Tibshirani (2003) employed supervised clustering and partial least squares to classify patients to low-risk and high-risk groups. Comparing our Figure 1(b) to Figure 6 in their paper, we observe that the two 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 2.17×10^{-5} for our method and 8.27×10^{-4} for that of Bair and Tibshirani (2003).

We also compare our results with those presented in Rosenwald *et al.* (2002). Following Rosenwald *et al.* (2002), patients are divided into four risk groups based on the quartiles of the estimated scores (see Figure 2 of Rosenwald *et al.*, 2002). Figure S4 (web supplement) shows the plots of the Kaplan-Meier estimates of survival of the four groups. It is noteworthy to point out that a fair performance measure of prediction of a future patient survival should be based on scores that are estimated using training samples only (Figure S4-d). The testing samples information should not be used in model building. Again, both the survival plot and the test indicate a good predictive performance of our proposed method.

DISCUSSION

In this article we propose the use of principal components and sliced inverse regression to reduce the high-dimensional microarray data to a low-dimensional space while accommodating censored survival phenotypes. The proposed method is applied to the DLBCL data of Rosenwald *et al.*, (2002). In conjunction with a Cox proportional hazards model, the method is shown to provide a good predictive performance for patient survival.

Sufficient dimension reduction in the context of censored data is 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 is achieved through observed survival time Y and status δ. 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, as was discussed in Li *et al.* (1999). Additionally, sliced inverse regression is employed in this paper as an illustration to accommodate censoring. The same idea can be well applied to many other sufficient dimension reduction methods, because, with $\mathcal{Y}$ being a function of $\mathcal{Y}^0$, $\mathcal{S}_{\mathcal{Y}|X} \subseteq \mathcal{S}_{\mathcal{Y}^0|X}$. For instance, SAVE (Cook and Weisberg) is a more comprehensive method than SIR in estimating the central subspace; MAVE (Xia *et al.*, 2003) is a local SDR method which relaxes the linearity condition of SIR; DAME (Gather *et al.*, 2002) provides a robust version of SIR that is less prone to the influence of outliers. A similar double slicing procedure can be applied to all those methods for the survival data.

Since not all genes will be relevant to predict 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 the 160 training patients identifies 473 genes which are significant at 0.01 level. For the 80 testing patients, however, only 67 genes are significant, out of which only 4 genes are identified significant in both groups. Applying the proposed methods on those 473 genes results in a poor performance due to the possible combinatorial effects of the gene expressions on the survival (details not shown). An alternative idea is to select genes based on the coefficients in the final Cox regression models. For instance, in our analysis of the DLBCL data, we trace back the coefficient of each gene in the linear combination. The absolute magnitude of these coefficients may provide a useful measure of individual gene contribution. Such a gene selection may also be carried out in an iterative fashion, i.e., iteratively removing those genes with small coefficients and refitting the model until the resulting model gives significantly worse performance in prediction. We are currently investigating these ideas.

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.

Cook, R.D., and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. Wiley, New York.

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

Dudoit, S., Fridlyand, J., Speed, T.P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of American Statistical Association*, **97**, 77-87.

Gather, U., Hilker, T., and Becker, C. (2002). A note on outlier sensitivity of sliced inverse regression. *Statistics*, **13**, 271-281.

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.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299-314.

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.

Park, P.J., Tian, L, and Kohane, I.S. (2002). Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, **18**, 120-127.

Pérez-Enciso, M. and Tenenhaus, M. (2003). Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis approach. *Human Genetics*, **112**, 581-592.

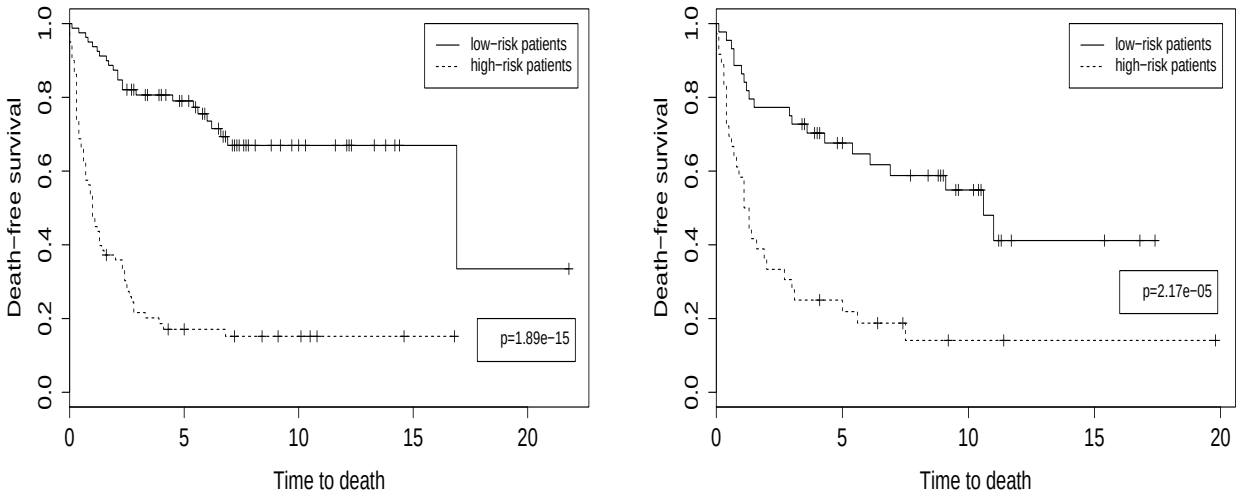
Rosenwald, A., Wright, G., Chan, W.C., Connors, J.M., Campo, E., Fisher, R.I., Gascoyne, R.D., Muller-Hermelink, H.K., Smeland, E.B., and Staudt, L.M. (2002). The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *The New England Journal of Medicine*,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

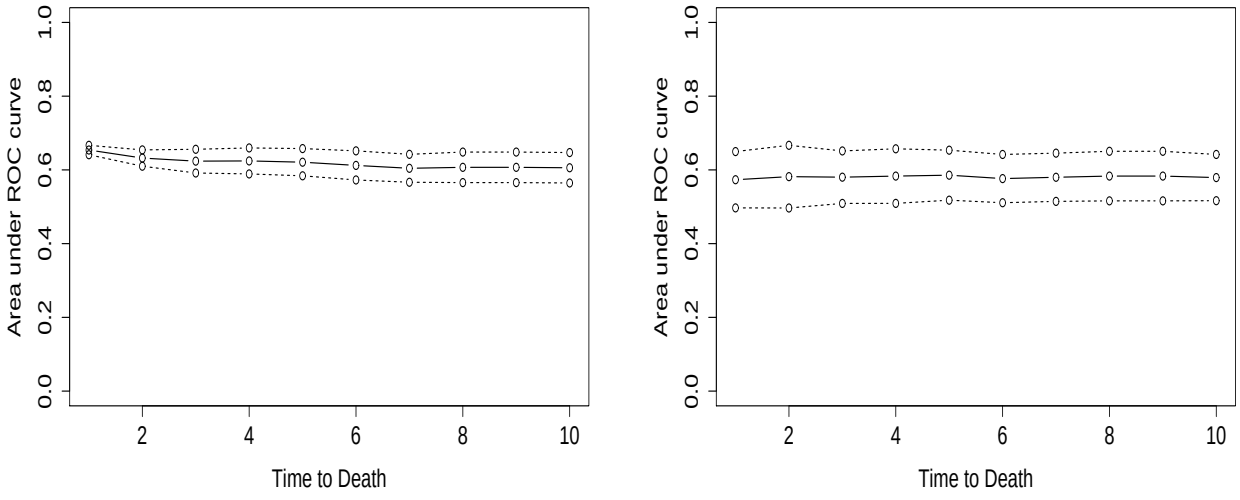
Verwij, P.J.M., Van Houwelingen, J.C. (1993). Cross validation in survival analysis. *Statistics in Medicine*, **12**:2305-2314.

Xia, Y., Tong, H., Li, W.K., and Zhu, L.X. (2002). An adaptive estimation of dimension reduction space, *Journal of Royal Statistical Society*, Series B, **64**, 363 - 410.



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

Figure 1: 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 2: Area under ROC at time 1 year to 10 years for 5-fold cross-validation: solid line the average of AUCs, dotted line the plus and minus of one standard error of AUCs. (a) patients in the training set of cross-validation; (b) patients in the testing set of cross-validation.