Kernel Cox Regression Models for Linking Gene Expression Profiles to Censored Survival Data

H. Li, Y. Luan

Pacific Symposium on Biocomputing 8:65-76(2003)

KERNEL COX REGRESSION MODELS FOR LINKING GENE EXPRESSION PROFILES TO CENSORED SURVIVAL DATA

HONGZHE LI, YIHUI LUAN
Department of Statistics and Medicine, University of California, Davis, CA 95616, USA

In functional genomics, one important problem is to relate the microarray gene expression profiles to various clinical phenotypes from patients. The success has been demonstrated in molecular classification of cancer in which gene expression data serve as predictors and different types of cancer are the binary or multi-categorical outcome variable. However, there has been less research in linking gene expression profiles to other types of phenotypes, in particular, the censored survival data such as patients' overall survival or cancer relapse times. In the paper, we develop a kernel Cox regression model for relating gene expression profiles to censored phenotypes in the framework the penalization method in terms of function estimation in reproducing kernel Hilbert spaces. To circumvent the problem of censoring, we use the negative partial likelihood as a loss function in the estimation procedure. The functional combinations of the original gene expression data identified by the method are highly correlated with the patients' survival times and at the same time account for the variability in the gene expression levels. We apply our method to data sets from diffuse large B-cell lymphoma, lung adenocarcinoma and breast carcinoma studies to verify its effectiveness. The results from these analysis indicate that the proposed method works very well in identifying subgroups of patients with different risks of death or relapse and in predicting the risk of relapse or death based on the gene expression profiles measured from the tumor samples taken from the patients.

1 Introduction

The recent development of DNA microarrays, which permits the simultaneous measurements of the expression levels of thousands of genes, raised the possibility of a powerful, genome-wide approach to the genetic basis of different types of tumors in the area of molecular classification of cancers, different levels of drug responses in the area of pharmacogenomics, and different patients' clinical outcomes in the area of clinical phenotype prediction. The problem of cancer class prediction using the gene expression data has been studied extensively in recent years. [1,2,3,4,5,6] This problem can be formulated as predicting binary or multi-category outcomes using gene expression data. However, there has been less development in relating gene expression profiles to other phenotypes such as quantitative continuous phenotypes, or the censored survival phenotypes. Relationship between gene expression profiles with other phenotypes such as quantitative phenotype or survival phenotypes is also important in clinical applications. For example, correlating gene expression profiles with

to the drug responses of thousand of potential drug compounds in 60 human cell lines, researchers were able to identify sets of genes whose expressions are highly related to drug response of a set of compounds, which will eventually help development of new drugs. [7] Correlating gene expression profiles obtained from tumor samples prior to treatment with the time to cancer relapse or death due to cancer can be very important in clinical practice.

The goal of this paper is to develop new statistical methods for relating gene expression profiles to censored survival data such as time to cancer recurrence or death. From the statistical point of view, one challenge is that the time to cancer recurrence or death is often right censored because during the course of followups, some patients may still be cancer-free or alive. Another challenge is that the microarray gene expression data are often measured with great deal of noise, and that the sample size of tissues or cell lines is usually very small compared to the number of genes from expression arrays. The problem of censoring make the problem difficult compared to binary or continuous phenotypes. One popular approach is to first cluster tumor samples into several clusters based on their gene expression patterns across many genes, and then to use the Kaplan-Meier curves or the log-rank tests to test whether there is a difference in survival times among different tumor groups. One drawback of this approach is that the phenotype information is completely ignored in the clustering step and therefore may result in loss of efficiency. Another approach is to cluster genes first based on their expression across different samples, and use the sample averages of the gene expression levels in a Cox model [8] for survival outcome. Both methods of course depend on the methods of clustering used. Hastie *et al* [9] proposed a tree-harvesting approach in which a stepwise regression approach is used to select genes or clusters of genes that are related to the phenotypes using the Cox proportional hazards model. This approach still requires the intermediate results of the hierarchical clustering analysis. Nyuyen and Rock [10] proposed to generalize the idea of the partial least square in the framework of the Cox model by using the residuals. However, their method is limited to linear function of the gene expression levels. In addition, use of residuals in the estimation of the parameters in the Cox model is not well-established in the survival analysis literature since there are many different ways of defining residuals. [11]

In the paper, we develop methods for relating gene expression profiles to censored phenotypes in the framework of the support vector machines (SVMs) by using kernels. [12]. The SVMs technique is a relatively new but popular methods in machine learning, and was applied successfully in the problem of tumor classification. [13] The robustness of the methods with respective to the sparse and noisy data is making them the methods of choice for many prob-

lems in bioinformatics. As demonstrated by Wahba *et al*[14], the SVM can be reformulated as a penalization method in terms of function estimation in reproducing kernel Hilbert spaces, where the objective function can be written as "loss+penalty". In this paper, we formulate a kernel Cox regression model for censored survival data by using the negative partial likelihood in a generalized Cox model as the loss function.

The rest of the paper is organized as the following. We first present a general Cox model for relating gene expression profiles to the censored survival phenotypes. We then present methods for estimating the parameter of the model in the framework of the kernel Cox regression model and penalization methods. In Section 3, we present results of analysis of three different cancer survival data sets. We conclude in Section 4 with a discussion of the results presented in this paper and offer some directions for future work.

2 Statistical Methods

2.1 A general Cox model for censored survival data

Suppose that we have collected n patients with a particular cancer. For the ith patients, let (t_i, δ_i) be the observed phenotype, where t_i is the survival time (e.g., time to cancer relapse after treatment) when δ_i=1, and is the censoring time (e.g., time of last known being cancer-free) when $\delta_i = 0$. Let $x_i = (x_{i1}, \cdots, x_{ip})$ be the vector of the gene expression levels of p genes for the ith sample taken from the ith patient. We assume the following general Cox model, where the hazard function for the ith patient is modeled as

$$\lambda_i(t|x_i) = \lambda_0(t) \exp(f(x_i)), \tag{1}$$

where $\lambda_0(t)$ is the unspecified baseline hazard function, $f(x_i)$ is an arbitrary function of the gene expression data x_i. In this model, the gene expression profile measured over p genes is related to the risk of death or cancer relapse through the score function $f(x)$.

2.2 Estimation of the model

Since the dimension of x_i vector is usually much larger than the sample size n, standard methods such as the Cox partial likelihood [8] for estimating the unspecified function f is unfeasible. To overcome this problem, a regularized formulation of the Cox regression is considered as a variational problem in a

reproducing kernel Hilbert space [15] (RKHS) H_K generated by the kernel K,

$$minR_{reg}(f) = \frac{1}{n}\sum_{i=1}^{n} V(t_i, \delta_i, f(x_i)) + \xi||f||^2_{H_K} \tag{2}$$

where $V(t_i, \delta_i, f(x_i))$ is the loss function which is a functional of f depending on only the values of $f(x)$ at the data points, $\{f(x_i)\}_{i=1}^n$, $||f||^2_{H_K}$ is the norm defined in H_K, and $f = b + h$ with $h \in H_K, b \in R$, and $\xi > 0$ is a tuning parameter. For the general Cox model (1), we propose to use the negative log partial likelihood [8] as the loss function and reformulate the problem as finding function $f(x)$ such that

$$R_{reg}(f) = -\frac{1}{n}\sum_{i=1}^{n} \delta_i[f(x_i) - \log\{\sum_{j \in R_i} \exp(f(x_j)\}] + \xi||f||^2_{H_k} \tag{3}$$

is minimized, where $R_i = \{j : t_j \geq t_i, j = 1, \cdots, n\}$ is the set of individuals who were at risk at time t_i. The solution to this problem was given by Kimeldorf and Wahba [15], and is known as the representer theorem. By this theorem, the optimal $f(x)$ has the form:

$$f(x) = b + \sum_{i=1}^{n} a_i K(x, x_i), \tag{4}$$

where K is the reproducing kernel of H_K. Since b can be absorbed into the baseline hazard function in model (1), we omit b in the following discussion. For the simplest case of natural inner product kernel with $K(x_i, x_j) =< x_i, x_j >$, the function $f(x)$ can be expressed as

$$f(x) \quad = \quad \sum_{i=1}^{n} a_i K(x, x_i) = \sum_{j=1}^{p}(\sum_{i=1}^{n} a_i x_{ij})x^{(j)} = \sum_{j=1}^{p} \beta_j x^{(j)} \tag{5}$$

where $x = (x^{(1)}, \cdots, x^{(p)})$ the vector of the gene expression levels over p genes. Here the parameter β_j can be used as a measurement on how the gene expression level of gene j affects the risk of death or tumor recurrence. Because of this nice interpretation, this paper uses only this kernel in our analysis of real data sets. In the case when the data are not linearly separable, one can choose more general kernel such as the polynomial kernels with $K(x_i, x_j) = (< x_i, x_j > +1)^d$ or the Gaussian kernels with $K(x_i, x_j) = \exp(||x_i - x_j||^2/\sigma_d^2)$, where d and σ_d^2 are the kernel parameters.

From the representer formula (4), it can be shown that minimizing equation (3) is equivalent to minimizing over vector a the finite dimensional form:

$$R_a = -\delta^{'}(K_a a) + \delta^{'} \log\{\sum_{j \in R_i} \exp(K_a a)\}] + \xi a^{'} K_a a \tag{6}$$

where $a^{'} = (a_1, \cdots, a_n\}$, the regressor matrix $K_a = [K(x_i, x_j)]_{n \times n}$. Here the matrix K_a is called the kernel matrix.

For a fixed ξ, to find a, we set the derivative with respective to a in R_a to zero, and use the Newton-Raphson methods to iteratively solve the score equation. In case of singularity problem in the Newton-Raphson iterations, we used the downhill simplex algorithm to minimize the function (6). [16]

We propose to use the leave-one out method for selecting the tuning parameter ξ in equation (6). One approach is to choose ξ which minimizes the cross-validated sum of the square residuals. Another approach is to choose ξ which maximizes

$$\prod_{i=1,\delta_i=1}^{n} \frac{\exp(\hat{f}_{-i}(x_i;\xi))}{\sum_{j \in R_i} \exp(\hat{f}_i(x_j;\xi))}$$

where $f_{-i}(x;\xi)$ denote the kernel estimate of the function for a given ξ computed under model when δ_i is changed from one (uncensored) to zero (censored). This procedure essentially leaves only the uncensored individuals out in the cross validation procedure and has been demonstrated to work well for the problem of subset selection for the Cox regression model. [17]

3 Application to real data sets

We applied the proposed methods to three published data sets of three different cancers to illustrate the methods and to demonstrate that the proposed methods work well in separating patients into different risk groups and in predicting patient's risk of cancer relapse or overall survival. In all the analysis, we used the natural inner product kernel and chose the tuning parameter ξ by residual cross-validation.

3.1 Application to diffuse large B-cell lymphoma data set

Alizadeth *et al* [1] reported a genome-wide gene expression profiling analysis for diffuse large B-cell lymphoma (DLBCL), in which a total of 96 normal and malignant lymphocites samples were profiled over 17,856 cDNA clones. Details can be found in Alizadeth *et al*[1]. None of the patients included in the study has been treated before obtaining the biopsy samples. After biopsy, the patients

were treated at two medical centers using comparable standard chemotherapy regimens. Among 42 patients, 40 of them had followup information, including 22 death with death time ranging from 1.3 to 71.3 months (median 10.6) and 18 being still alive with the follow-up times ranging from 51.2 to 129.9 months (median 74.7).

Alizadeth *et al*[1] first identified 4026 genes which showed large variations across all the samples. We further selected 319 genes with the p-value less than 0.05 by using the Cox model for each of these 4026 genes. We used the inner product kernel to build a general Cox model for the time to death of the 40 patients using the gene expression levels of the 319 genes as predictors. Figure 1 (a) shows the estimated survival curves for patients in two different groups defined by the scores $f(x) > 0$ or $f(x) < 0$, indicating large difference in overall survival in the two groups. One group of 15 patients with $f(x) < 0$ were all alive during the followups, another group of 25 patients with $f(x) > 0$ includes 22 deaths. The group with $f(x) < 0$ includes 12 germinal centre B cell-like DLBCL, and the group with $f(x) > 0$ includes 18 activated B cell-like DLBCL, indicating much better overall survival in germinal centre B cell-like DLBCL patients. Figure 1 (b) shows the estimated coefficient for each of the 319 genes that were used in estimating the score $f(x)$. Higher expression levels of the genes with positive coefficients increase the risk of death due to lymphoma, and higher expression levels of the gene with negative coefficients decrease the risk of death due to lymphoma.

To further examine how the model predicts the survival time of future new patient, we performed a leave-one-out cross validation analysis. We left one patient out each time, and estimated the function $f(x)$ in model (1) with the rest of the data. We then estimated the score for the patient who was left out from model building process. Figure 1 (c) shows the plot of the cross-validated score versus the observed times to death or times at censoring, indicating the higher the score, the high risk of death from lymphoma. Figure 1 (d) shows the estimated survival curves based on these cross-validated scores being negative or positive. This plot suggests that the proposed method works quite well in predicting future patient's risk of death.

3.2 Application to lung cancer data set

Garber *et al.*[4] reported global gene expression profiles for 67 human lung tumors representing 56 patients whose clinical course was followed for up to 5 years. Among these tumors, there were 41 Adenocarcinomas (ACs), of those, 22 patients had survival information available, including 12 death ranging from 0 to 36 months (median 12.5 months) and 10 death-free during the 5 year

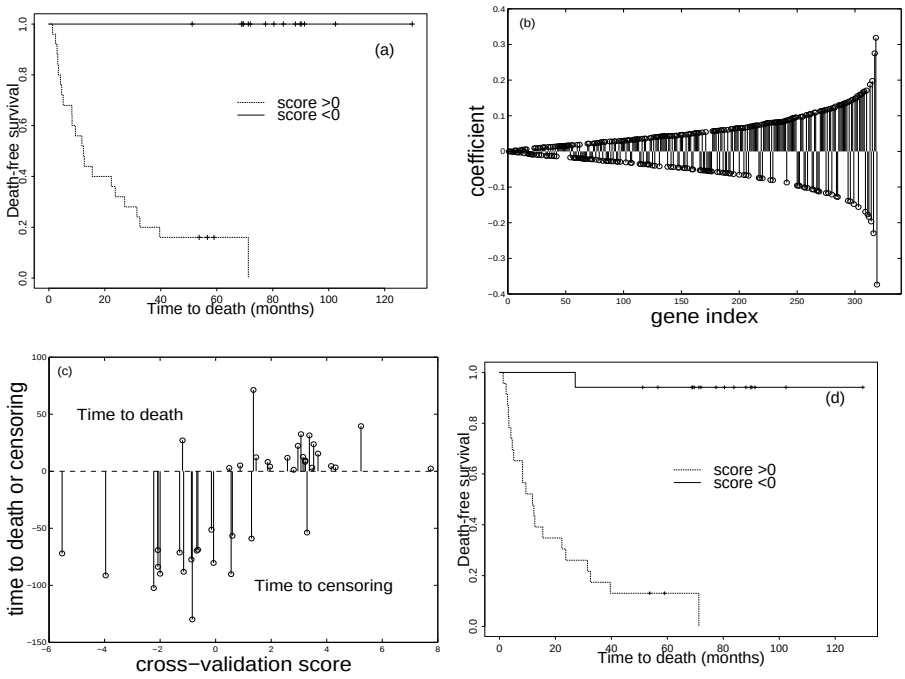

Figure 1: Analysis of lymphoma data set. (a) Survival curves for patients in two groups based on the estimated scores using gene expression profiles; (b) Coefficients of the gene effects sorted by the absolute values; (c) Plot of cross-validated scores versus observed data, where the death times are plotted as positive values on the y-axis, and the censored times are plotted as negative values on the y-axis; (d) Survival curves for patients in two groups based on cross-validated scores.

followups (followup times ranged from 18 to 54 months with median of 48 months). The original gene expression data include 23,100 cDNA clones representing 17,108 unique genes. Garber *et al* further reduced this list of gene to 918 cDNA clones representing 835 unique genes with large variabilities cross samples in different clusters.

There are 131 genes with the Cox score statistic larger than 3.84 (corresponding p-values less than 0.05). Using our proposed method with inner produce kernel, we obtained the estimate of $f(x)$ in the model (1) based on the 22 patients's survival information during the followups and the gene expression levels of 131 genes. Based on the estimated values of $f(x) > 0$ or $f(x) < 0$, we divided the 22 patients into two groups. Figure 2 (a) shows the

overall survival curves for these two groups, indicating that large difference in lung cancer patients' survival can be identified based on their gene expression profiles. Figure 2 (b) shows the estimates of the β coefficients in model (5), indicating the expression levels of different genes have different effects on patients's survival. It is important to note that it is the linear combination of these gene expression levels that is used for separating patients into different groups with different risks of death due to lung cancer.

We also performed similar cross-validation analysis as we did for the lymphoma example. Figure 2 (c) shows the cross-validated scores for each patient, plotted against the observed time to death or time to last followups. The plot indicates that the patients with larger cross-validated scores tend to have higher risk of being death. Figure 2 (d) show the overall survival curves for patients with the cross-validated scores greater or less than zero. This example further demonstrates that the proposed method can effectively predict the patient's risk of death from lung cancer.

3.3 Application to breast cancer data set

Sorlie *et al* [5] demonstrated the use of the gene expression profiles of breast carcinomas to distinguish tumor subclass with clinical implications. In the following analysis, only 49 of the patients from the prospective study of locally advanced disease and with no distant metastases were used. The gene expression levels of the 456 cDNA clones (427 unique genes) in the intrinsic gene list were obtained for tumor samples prior to chemotherapy. The followup information including the time to cancer relapse and overall survival was available for all the 49 patients. During the followups, 24 patients had cancer relapse at various time ranging from 0 month to 59 months after the treatment, and 25 patients had no relapse during the followup period ranging from 22 to 92 months.

Based on the univariate Cox regression, we first selected 68 genes with the scores greater than 3.85 (significant at 0.05 level). Based on the gene expression data of these genes and the patients' time to relapse data, we fitted the model (1) and obtained the estimated $f(x)$ function and the score for each patient. Figure 3 (a) shows the plot of the estimated scores versus the time to relapse or time to last followups. In general, we observed that the higher the score, the higher the risk of having relapse after treatment. Based on these scores, we can divide the patients into high (if $f(x) > 0.85$), medium (if $-1.45 < f(x) \leq 0.85$) and low risk (if $f(x) \leq -1.45$) groups with different risks of relapse. Figure 3 (b) give the relapse-free survival curves for these three groups. Therefore, based on the scores derived from the gene expression profiles, we are able to

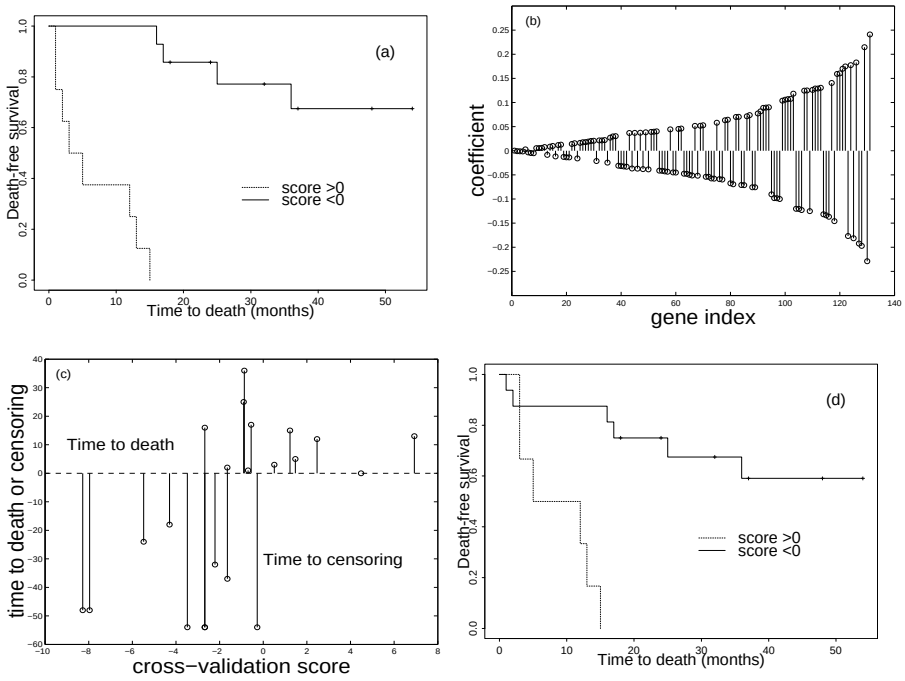

Figure 2: Analysis of lung cancer data set. (a) Survival curves for patients in two groups based on the estimated scores using gene expression profiles; (b) Coefficients of the gene effects sorted by the absolute values; (c) Plot of cross-validated scores versus observed data, where the death times are plotted as positive values on the y-axis, and the censored times are plotted as negative values on the y-axis; (d) Survival curves for patients in two groups based on cross-validated scores.

identify subgroups of the patients with different risks of breast cancer relapse after the chemotherapy.

In order to examine how well the model performs for predicting the risk of relapse for a new patient, we performed leave-one out cross-validation analysis, similar to the previous two examples, in which we left one patient out, and fitted the model with the rest of the patients, and estimated the score for the left-out patient. Figure 3 (c) shows the cross-validated scores versus the observed times to relapse/censoring. Comparing to the estimated scores in Figure 3 (a), it is less clear that there are three different risk groups for relapse. However, clear differences in risk of relapse can still been seen for those with very high or low scores. Figure 3 (d) shows the relapse-free survival curves for

the three groups defined by the cut-off values of -1.45 and 0.85. The cross-validated results still identified one group of patients with large scores who has higher risk of cancer relapse. However, the difference of the other two groups was not significant, which suggests that for predicting cancer relapse risk, there might only be two distinct groups with different risks.

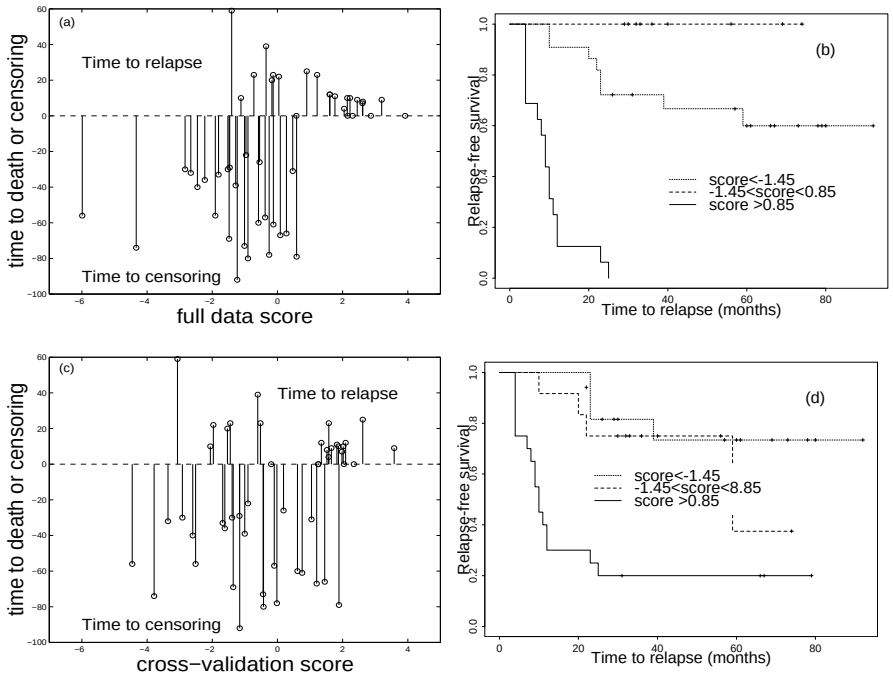

Figure 3: Analysis of breast cancer data set. (a) Estimated scores versus the observed times to cancer relapse or times at censoring, where the relapse times are plotted as positive values on the y-axis, and the censored times are plotted as negative values on the y-axis; (b) Survival curves for patients in three groups based on the estimated scores using gene expression profiles; (c) Cross-validated scores versus the observed times to cancer relapse or times to censoring; (d) Survival curves for patients in three groups based on cross-validated scores.

4 Discussion

We have introduced the kernel Cox regression models for relating gene expression profiles to censored survival data. The method generalizes the idea of the

support vector machine for binary or multi-categorical data to censored survival data. The model automatically searches for the genes whose expression levels are related to survival phenotypes and identifies the optimal combination of the gene expression data in predicting the risk of cancer recurrence or death. Since the risk of cancer recurrence of death due to cancer might be due to interplay of many genes in certain way, methods such as our proposed one are expected to show better performance in predicting the risks. We demonstrated the applicability of our methods by analyzing time to death or tumor recurrence of large cell lymphoma, lung carcinoma, and breast carcinoma. In all the analysis, we used the simple inner cross product kernels and the linear combination of the gene expression levels as our scores, and obtained satisfactory results. However, it is important to emphasize that our proposed method is not limited to obtaining only linear scores. One advantage of the method is that it can handle nonlinear scores easily by using alternative kernels without introducing any computational difficulty.

Another advantage of the proposed method is that there is no computational or methodological limitation in term of the number of genes that can be used in the prediction of patient's overall survival or time to cancer recurrence. One important future research is to examine how different number of genes used in model affects the results of prediction. Since not all genes will be important in predicting censored survival phenotypes, we would expect better prediction results using only genes that are related to the phenotypes. For all the three cancer data sets we analyzed, we selected these genes by using the univariate Cox score statistic. An alternative would be to iteratively select important genes and build predictive models. One approach for this problem is to iteratively delete those genes with β coefficient smaller than a preset cutoff values and build predictive model until no such genes can be deleted. We plan to study this approach in details in the future.

In summary, we formulated the problem of linking gene expression profiles to the censored survival data in the framework of the kernel Cox regression model. As demonstrated by applications to several real cancer data sets, the methods can potentially be useful for identifying important genes that are related to clinical outcomes and for predicting time to cancer relapse or death to cancer based on the tumor molecular gene expression profiles.

Acknowledgments

This work was supported by an NIH grant (ES09911) and an UC Davis Health System Research Award grant.

References

1. A. Alizadeh *et al., Nature* **403**, 503 (2000).
2. T. R. Golub *et al., Science* **286**, 531 (1999).
3. U. Alon *et al.*, *Proc Natl Acad Sci USA*, **96(12)**, 6745 (1999).
4. M. E. Garber *et al., Proc Natl Acad Sci USA* **98**, 13784 (2001).
5. T. Sorlie *et al., Proc Natl Acad Sci USA* **98**, 10869 (2001).
6. J. E. Staunton *et al., Proc Natl Acad Sci USA* **98**, 10787 (2001).
7. U. Scherf *et al., Nat Genet* **24**, 236 (2000).
8. D. Cox, *J. Roy Stat. Sco B* **74**, 187 (1972).
9. H. Trevor *et al., Genome Biology* **2**, research0003.1 (2001).
10. D. Nyuyen, D. M. Rocke, *Technical report*, UC Davis (2001).
11. W. E. Barlow, R. L. Prentice, *Biometrika* **75**, 65 (1988).
12. G. Wahba, *Technical report*, UW Madison (1998).
13. T. S. Furey *et al., Bioinformatics* **16**, 906 (2000).
14. G. Wahba, (SIAM, Philadelphia, 1990).
15. G. Kimeldorf, G. Wahba, *J. Math. Anal. Applic.*, **33**, 82 (1971).
16. J. A. Nelder, R. Mead, *Computer Journal* **7**, 308 (1965).
17. A. Y. C. Kuk, *Biometrika*, **71**, 587 (1984).