BIOINFORMATICS

Vol. 18 no. 12 2002
Pages 1625–1632

Partial least squares proportional hazard regression for application to DNA microarray survival data

Danh V. Nguyen[1,*] *and David M. Rocke*[2]

[1]*Department of Statistics, Texas A&M University, College Station, TX 77843, USA and* [2]*Department of Applied Science, University of California, Davis, CA 95616, USA*

Received on October 14, 2001; revised on March 12, 2002; accepted on June 18, 2002

ABSTRACT

Motivation: Microarrays are increasingly used in cancer research. When gene transcription data from microarray experiments also contains patient survival information, it is often of interest to predict the survival times based on the gene expression. In this paper we consider the well-known proportional hazard (PH) regression model for survival analysis. Ordinarily, the PH model is used with a few covariates and many observations (subjects). We consider here the case that the number of covariates, p, exceeds the number of samples, N, a setting typical of gene expression data from DNA microarrays.

Results: For a given vector of response values which are survival times and p gene expressions (covariates) we examine the problem of how to predict the survival probabilities, when $N \ll p$. The approach taken to cope with the high dimensionality is to reduce the dimension using partial least squares with the response variable as the vector of survival times. After dimension reduction, the extracted PLS gene components are then used as covariates in a PH regression to predict the survival probabilities. We demonstrate the use of the methodology on two cDNA gene expression data sets, both containing survival data. The first data set contains 40 diffuse large B-cell lymphoma (DLBCL) tissue samples and the second data set contains 49 tissue samples from patients with locally advanced breast cancer in a prospective study.

Availability: The methodology can be implemented using a combination of standard statistical methods, available, for example, in SAS. Sample SAS macro codes to implement the methods will be available at http://stat.tamu.edu/∼dnguyen/supplemental.html.

Contact: dnguyen@stat.tamu.edu; dmrocke@ucdavis.edut

INTRODUCTION

The introduction of DNA microarray technology is a technical advance in biomedical research. Microarray technologies, including cDNA (Schena *et al.*, 1995) and oligonucleotide arrays (Lockhart *et al.*, 1996) allow simultaneous monitoring of thousands of gene expressions per sample. Despite the need for improvement of the current technologies, there have been many applications of microarrays in human cancer research, including DeRisi *et al.* (1996); Ross *et al.* (2000); Alizadeh *et al.* (2000), and Perou *et al.* (2000) among others.

The ability to measure gene expression en masse has also resulted in data with the number of variables (genes), p, far exceeding the number of samples, N. Of particular interest, for example, is when survival times of N cancer patients are available in conjunction with their mRNA expression data. In this setting, it is of interest to predict the patient survival probabilities using p gene expressions ($N \ll p$). For example, through gene expression profiling, Alizadeh *et al.* (2000) identified two distinct molecular subtypes of diffuse large B-cell lymphoma (DLBCL): germinal centre (GC) B-like and activated B-like. Estimate of patient survival probabilities for the two groups were then compared using Kaplan and Meier (1958) survival curves.

In this paper, we demonstrate how prediction of patient survival probabilities can be based on the proportional hazard (PH) regression model after extracting gene components by partial least squares (PLS). Details of the methodology, which involves dimension reduction and PH regression, are described in the **Methods** Section. We applied the methodology to a diffuse large B-cell lymphoma cDNA data set of Alizadeh *et al.* (2000) and a breast carcinomas data set published by Sørlie *et al.* (2001). The predicted survival probabilities for various molecular subgroups previously identified in the literature, are summarized in the **Results** Section. Other issues, including the PH assumption, goodness-of-fit, and the choice of the PLS dimension reduction parameter, are also discussed in the **Results** Section. We conclude with a discussion of the limitations of the proposed method and directions for future work. Computational details

*To whom correspondence should be addressed.

© Oxford University Press 2002

are given in the Supplemental Appendix available at http://stat.tamu.edu/∼dnguyen/supplemental.html.

METHODS

In this section we describe the methodology for estimating survival probabilities using the gene expressions as covariates. The method involves reduction of the high p-dimensional covariate space to a lower K-dimensional gene component space. The dimension reduction method utilized is partial least squares (PLS). Next, the K PLS gene components are used as covariates in a proportional hazard (PH) regression model.

PH regression with gene expressions as covariates

Let Y be time to some event, such as the survival time of a diseased patient. Associated with each patient are p covariates which could be p gene expression measurements obtained from DNA microarray experiments, for example. A data set consists of N samples, each containing the triple $(T_i, \delta_i, \mathbf{x}_i)$ for $i = 1, \ldots, N$, where $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ip})$ is the covariate profile of the ith patient, T_i is the survival time if $\delta_i = 1$, and it is the right-censored time if $\delta_i = 0$. In the current context, $\mathbf{x}_i$ is the gene expression profile of the ith patient. The response values of interest, the survival times, are not observed for every patient. Instead, the clinical data recorded for the ith patient is only $T_i = \min(Y_i, Z_i)$, where Z_i is a censored value. It is assumed that the censoring mechanism or the censoring time distribution is independent of the survival time distribution.

Cox (1972) suggested the proportional hazard (PH) regression model to study the relationship between the time to event and a set of covariates in the presence of censoring. Model estimates are obtained by maximizing the partial likelihood. Details are given in the Supplemental Appendix B. Subsequent studies of the analysis of the PH model includes Kalbfleisch and Prentice (1973); Breslow (1975) and Cox (1975) among others.

In the context of data generated from DNA microarray experiments, the number of genes, p, is in the thousands, but the number of samples, N, is quite small. Standard statistical methodologies, including the classical PH regression method described above, do not work in this situation. For example, the breast carcinomas cDNA data set, considered in the next section, consists of $N = 49$ sample tissues from breast cancer patients. The number of genes (covariates) corresponding to each sample is $p = 3846$. Furthermore, survival information is recorded for each of the N patients. It may be of interest to study the survival experiences of the patients in relation to the p genes. For instance, does more positive survival experiences coincide with certain gene expression patterns?

When there are more genes (covariates) than there are samples and the PH regression is not defined, how do we predict the patient survival probabilities? One approach to cope with the high p-dimensional covariate space is to utilize some dimension reduction method to reduce the p-dimensional space. The survival probabilities can then be estimated using PH regression in the reduced space. We describe the PLS dimension reduction method next.

PLS dimension reduction

The method of dimension reduction we considered is partial least squares (PLS). Since its introduction the method has been applied with much success in the field of chemometrics. See, for instance, Martens and Naes (1989). Gene expression data from DNA microarray experiments display similar characteristics as those found in chemical applications. Dimension reduction of gene expression data based on PLS has been applied to binary tumor classification (Nguyen and Rocke, 2002a,b) as well as multi-class discrimination (Nguyen and Rocke, 2002c) problems. The reader is referred there for details. The uses of PLS in the classical setting where the response is continuous have been investigated by Helland (1988); Höskuldsson (1988) and Frank and Friedman (1993) among others. In the present context, the response variable is continuous but some observed values (times) are right censored.

In some aspects, PLS is similar to the well known method of principal component analysis (PCA). In a PCA orthogonal linear combinations are constructed to maximize the variance of the linear combination of the predictor variables (genes) sequentially. However, the optimization criterion may not be appropriate for some prediction problems associated with microarray data and this was pointed out earlier (Nguyen and Rocke, 2002a,b). Roughly, the constructed principal components summarize as much of the original p predictors' information (variation), irrespective of the response class information.

Maximizing the variance of the linear combination of the predictors, namely var($\mathbf{Xv}$), may not necessarily yield components predictive of the response variable, such as survival times. For this reason, a different objective criterion for dimension reduction may be more appropriate for prediction. The objective criterion for constructing components in PLS is to sequentially maximize the covariance between the response variable ($\mathbf{y}$), survival time, and a linear combination of the genes ($\mathbf{X}$). That is, in PLS, the components are constructed to maximize the objective criterion based on the sample covariance between $\mathbf{y}$ and a linear combination of $\mathbf{X}$. Thus, we find the weight vector $\mathbf{w}$ satisfying the following objective criterion,

$$\mathbf{w}_k = \underset{\mathbf{w}'\mathbf{w}=1}{\operatorname{argmax}} \operatorname{cov}^2(\mathbf{Xw}, \mathbf{y}) \tag{1}$$

subject to the orthogonality constraint

$$\mathbf{w}_k' \mathbf{S} \mathbf{w}_j = 0 \quad \text{for all} \quad 1 \leq j < k \tag{2}$$

where $\mathbf{S} = \mathbf{X}'\mathbf{X}$. The ith PLS component is also a linear combinations of the original genes, namely $\mathbf{t}_i = \mathbf{X}\mathbf{w}_i$, but the weights are non-linear functions of both $\mathbf{y}$ and $\mathbf{X}$.

The algorithm used to obtain weights (1) as well as the relevant PLS quantities is detailed in the Supplemental Appendix A. We note that the PLS algorithm in the Appendix Section does not involve matrix inversion. For further details see Höskuldsson (1988) and Helland (1988).

As mentioned earlier, the constructed PLS gene components profiles are then used as covariates in a PH regression to predict the survival probabilities. However, interpretation of the fitted parameters of the PLS gene component profiles in terms of the original expression profiles, similar to the classical case where $N \gg p$, does not appear feasible directly. This is due to the fact that the PLS gene components are linear combinations of *all* the predictor genes.

RESULTS

In this section we describe the results of applying the PLSPH regression method to predict survival probabilities for two cDNA gene expression data sets with patient survival information. The first data set contains $N = 40$ tissue samples from patients with B-cell lymphoma (Alizadeh *et al.*, 2000) and the second data set consists of $N = 49$ breast carcinomas tissue samples (Sørlie *et al.*, 2001). We briefly described each data set below and the results from applying the proposed methodology follow. Specifically we fitted the PH regression model using PLS gene components as covariates. The model is then used to predict the patient survival probabilities. Predicted survival probabilities based on the fitted PH model are plotted for various cohorts defined by mRNA expression patterns.

Due to the small number of samples available we used leave-out-one cross-validation (LOOCV) to examine the stability of the estimated survival probabilities. Note that LOOCV is fitting the model using a training data set of $N - 1$ samples. Where possible, we assessed model fit and examined the proportional hazard (model) assumption. Also, as described in the **Methods** Section, PLS dimension reduction requires the selection of the number of PLS gene components, K. The selection of K is based on cross-validation and the proportion of response variation explained by the PLS dimension reduction model.

Diffused large B-Cell lymphoma data

The lymphoma cDNA data set consists of gene expression levels from cDNA experiments involving three prevalent adult lymphoid malignancies: diffuse large B-cell lymphoma (DLBCL), B-cell chronic lymphocytic leukemia (BCLL), and follicular lymphoma (FL). cDNA targets were prepared from experimental mRNA samples and were labelled with Cy5-dye during reverse transcription. A reference cDNA sample was prepared from a combination of nine different lymphoma cell lines and was labelled with Cy3-dye. Cy-labelled experimental and reference cDNAs were mixed and hybridized onto the microarray. See Alizadeh *et al.* (2000) for details. We analyzed the standardized log relative intensity ratios, namely the log(Cy5/Cy3) values.

For this gene expression data set survival times are also available for $N = 40$ DLBCL patients. There were 22 deaths. Associated with the survival times are mRNA expression ratios of $p = 5622$ genes. Using cluster analysis Alizadeh *et al.* (2000) identified two molecularly 'distinct' DLBCL subgroups based on the relative mRNA expression patterns: (1) germinal centre (GC) B-like (19 patients) and (2) activated B-like (21 patients). These two groups were identified by gene expression profiling using hierarchical clustering. Distinct expression patterns (profiles) for the two groups can be seen in Supplemental Figure 1, which displays the relative expression patterns of 50 genes differentially expressed across the two DLBCL subtypes. (The 50 genes were selected using the simple two sample t-statistics.) For details, including survival analysis using Kaplan–Meier survival curves for the two subgroups, see Alizadeh *et al.* (2000).

The patient survival probabilities were predicted using the PLS gene component profiles directly in a PH regression. We obtained $K = 2$ PLS gene components based on $p = 2000$ genes and then fitted the PH regression model using $K = 2$ PLS gene components as the covariates. Based on the fitted model, the survival probability estimates were obtained for an average PLS gene components profile vector for the GC B-like and the activated B-like group, namely $\bar{\mathbf{t}}_{GC}$ and $\bar{\mathbf{t}}_{Act.}$. The survival curves for the two groups are plotted in Figure 1 (top). Based on the fitted PLSPH model, the predicted survival probabilities for activated B-like group is distinctly lower than the GC B-like at the group average component expression profile (Figure 1, top).

Breast carcinomas data

There was little overlap in observed survival times (including censored times) in the DLBCL data set between the censored and non-censored group. With the exception of one patient, patients who died all died very early on in the study. This clustering of survival times could potentially have an effect on how well our method predicts survival. Hence, it is of interest to test the methods on survival data with different characteristics.

We considered a second gene expression data set with survival information, a breast carcinomas cDNA data set, published by Sørlie *et al.* (2001). Unlike the lymphoma data set, there is considerable overlap between

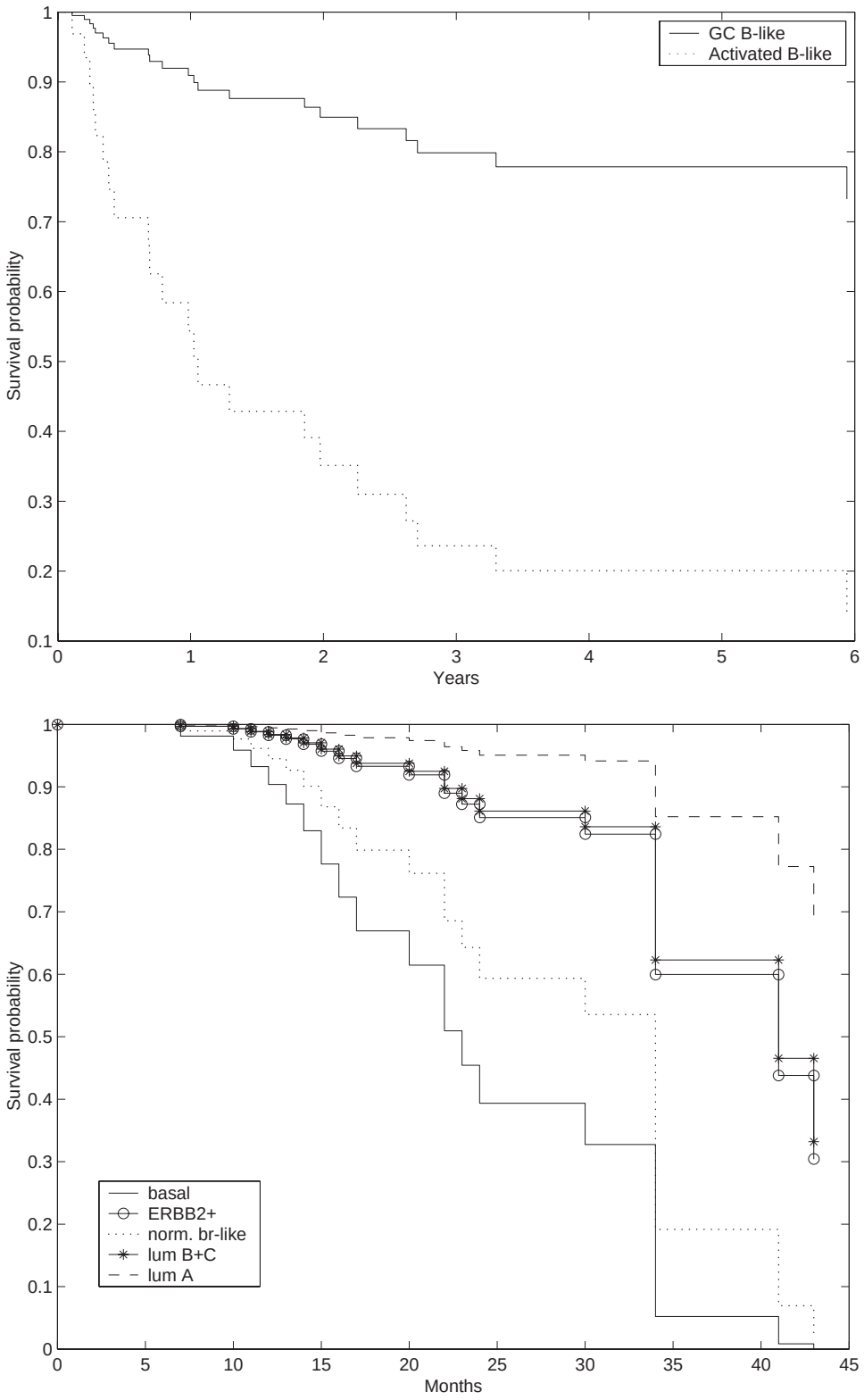

Fig. 1. Given are estimated survival curves from the PLSPH regression model fits to the (top) diffuse large B-cell lymphoma data and (bottom) and the breast carcinomas cDNA gene expression data. The curves are obtained for the group-average component profiles. See figure legend for group definitions.

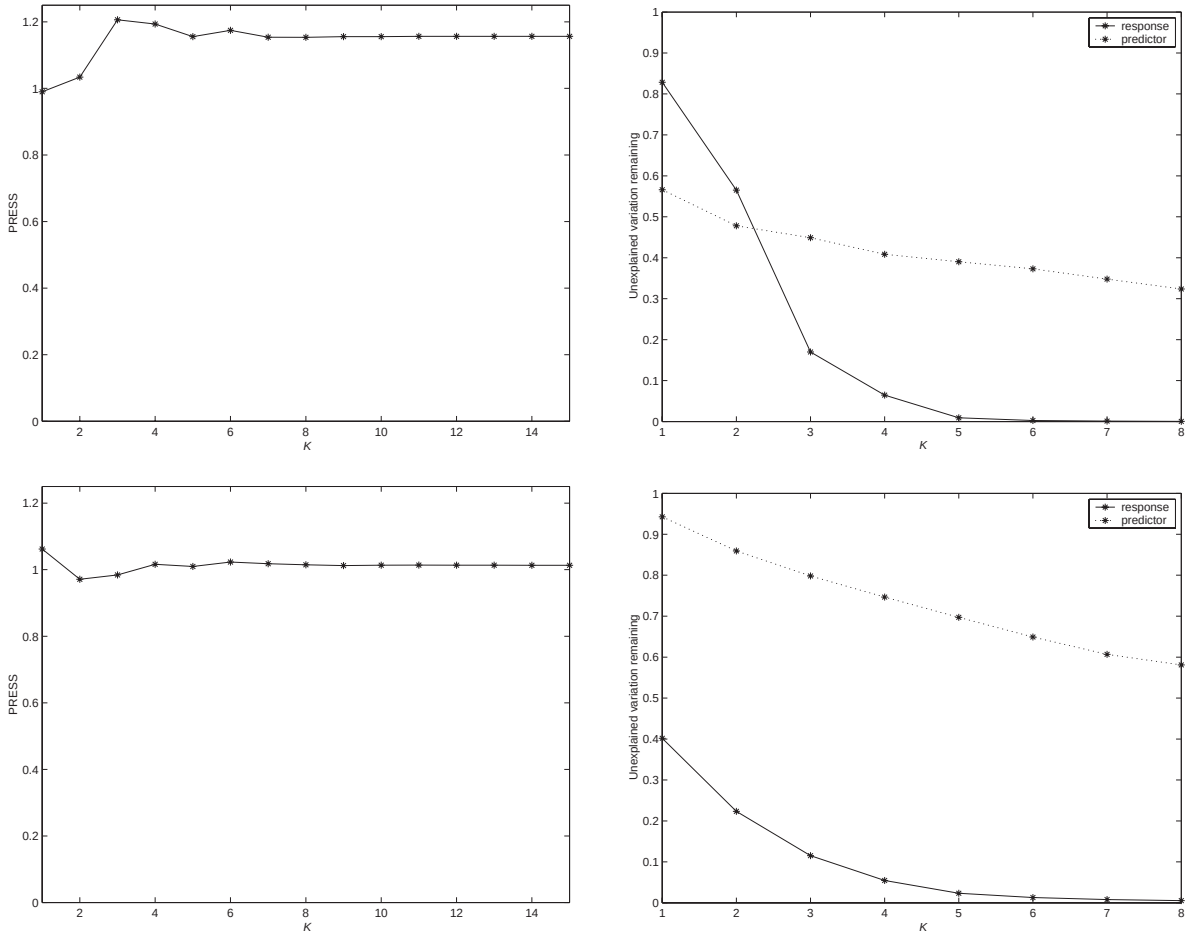

Fig. 2. Plotted are the cross-validated predicted residual sum of squares (PRESS) from the PLS fit for $K = 1, \ldots, 15$ (left column) and the proportion of response (y) and predictor (**X**) variation *unexplained* for dimension $K = 1, \ldots, 15$ (right column). Results are given for the diffuse large B-cell lymphoma data (top row) and the breast carcinomas data (bottom row).

the distribution of survival and censored times in the breast carcinomas data set. The data set analyzed consists of $N = 49$ samples and expression patterns for $p = 3846$ genes. Tissue samples were obtained from patients in a prospective study on locally advanced breast cancer with no distant metastases.

Sørlie *et al.* (2001) identified 6 clusters of gene expression profiles corresponding to basal-like, ERBB2+ (overexpression of ERBB2 oncoprotein), normal breast-like, luminal subtypes A, B, and C. Expression profiles for luminal subtypes B and C were similar and in some of their analyses, these were combined as one group (Luminal B+C). For details on the data set, including Kaplan–Meier analysis, see Sørlie *et al.* (2001).

As in the DLBCL data analysis, we fitted the PLSPH model to the breast carcinoma data and the survival probability estimates for each subgroup are given in Figure 1 (bottom). (Color figures are available at http://stat.tamu.edu/∼dnguyen/supplemental.html.) Based on the fitted model, the predicted survival probabilities are quite similar for the ERBB2+ group and the luminal B+C group.

Cross-validation and choice of K

We have chosen $K = 2$ gene components from PLS to fit the PH regression. The number of gene components, K, to use for prediction can be chosen by cross-validation to minimize the predicted residual sum of squares (PRESS). However, this approach is not completely satisfactory because, often, the model which minimizes PRESS compared with a much simpler model (smaller K) will have only a slight difference in PRESS score. This situation was not encountered in the two data sets examined here, since a PLS fit with a larger K had a larger PRESS score. Figure 2 (left column) plots the PRESS score for $K = 1$ to 14 using leave-out-one-CV (LOOCV). (Five-fold CV

resulted in the same choice for K.) We find it useful to also augment the selection of K by examining the proportion of response (y) variation explained by K PLS gene components, for various values of K. This is given in Figure 2 (right column). For the breast carcinomas data a choice of $K = 2$ resulted in the smallest cross-validated PRESS score and the total response variation explained was 77.66%. The choice of $K = 1$ gives the smallest PRESS score for the B-cell lymphoma data, but with $K = 1$ the response variation explained was only 17.14%. As indicated by Figure 2 (top left) the difference in PRESS scores for $K = 1$ and 2 is small, however, the cumulative response variation explained increases to 43.48% for $K = 2$. Hence, we also selected $K = 2$ for the lymphoma data. The selection of K is based on CV and an estimate of the proportion of response variation explained (Wold, 1994) by the PLS dimension reduction method. Other methods developed specifically for estimating the proportion of variation explained in the Cox model may be of interest to investigators (Schemper, 1992).

Proportional hazard assumption and model fit

Models are approximations and often contain assumptions. The Cox PH regression model used to predict the survival probabilities makes the assumption of proportional hazards. The proportional hazard assumption may not be true or correct and needs checking. We checked the model assumption using standard methods based on testing for significant slope of the smooth curve through the scatter of the rescaled Schoenfeld residuals versus time (Grambsch and Therneau, 1994). The tests indicate that the proportional hazard assumption is reasonable in both data sets. (The residuals used were from models displayed in Figure 1.) We also checked the deviance residuals from the models. Patients with large deviance residuals are poorly predicted by the model. The residuals falls within reasonable limits.

Various tests for overall goodness-of-fit attempt to detect departures, such as over-fitting, from the null hypothesis that the model fits. See Schoenfeld (1982) and Arjas (1988) for examples in the traditional PH setting. These tests rely on large sample sizes, but only small sample sizes are available from micorarray data currently. For the two data sets examined, $N = 40$, with 45% censoring and $N = 49$ with 61.2% censoring for the DLBCL and breast carcinomas data sets respectively. This makes assessing model fit difficult in the traditional setting, not to mention now we also need to simultaneously assess both the PLS and the PH fits.

We examined the cross-validated prediction of the survival distribution graphically. More precisely, for the ith sample left out, $i \in \mathcal{F}$, we obtained the predicted survival probabilities using PLSPH regression. The set of failure times is denoted by $\mathcal{F}$. Figure 3 gives the LOOCV estimates for each $i \in \mathcal{F}$ for the DLBCL (top) and breast carcinomas data (bottom). The plots indicate that the predictions are relatively stable, particularly for the DLBCL data. Although changes in the predicted probabilities from LOOCV are more variable for the breast carcinomas data, the 'ordering' of the basal, normal breast-like, and luminal A groups is preserved. See Figures 1 and 3 (bottom plots). Predicted survival probabilities for the ERBB2+ and luminal B+C groups are similar.

DISCUSSION

DNA microarray technologies, such as high-density oligonucleotide arrays and cDNA arrays, produce high dimensional gene expression data. Scientists using array technologies seek useful statistical methodologies able to cope with the high dimension. The Cox PH regression method is one of the most widely used tools in scientific research, particularly in biological and medical science. We have demonstrated the use of PLS gene components as covariates to predict survival probabilities in a proportional hazards model.

We have suggested the use of dimension reduction (PLS) in conjunction with proportional hazards regression (PH) as a possible way to analyze gene expression data with patient survival information. We hope that this will prompt further needed work on this method from other researchers. For example, as hinted in the previous section, the problem of goodness-of-fit for the PLS and PH model simultaneously needs to be further addressed. A potentially fruitful approach is to consider the problem in a Bayesian framework, where gene selection, prediction, and model fit may be addressed more fully. Also, the response variable is continuous but censored. Thus, PLS components constructed may contain some bias depending on the amount of censoring. We performed a simulation (data not shown) which suggests that there may be a small positive bias in predicting the true survival probabilities using PLSPH regression. This was observed with data generated under the condition that the first few principal components (e.g. 3) capture most of total gene expression variation ($\geqslant$ 70%). Although such microarray data structure is not typical in practice, it is an issue deserving of further investigation.

ACKOWLEDGEMENTS

We thank three referees for their careful reviews and helpful comments. The research reported in this paper was supported by grants from the National Science Foundation (ACI 96-19020, and DMS 98-70172), the National Institute of Environmental Health Sciences, National Institutes of Health (P43 ES04699), and the National Cancer Institute (CA90301).

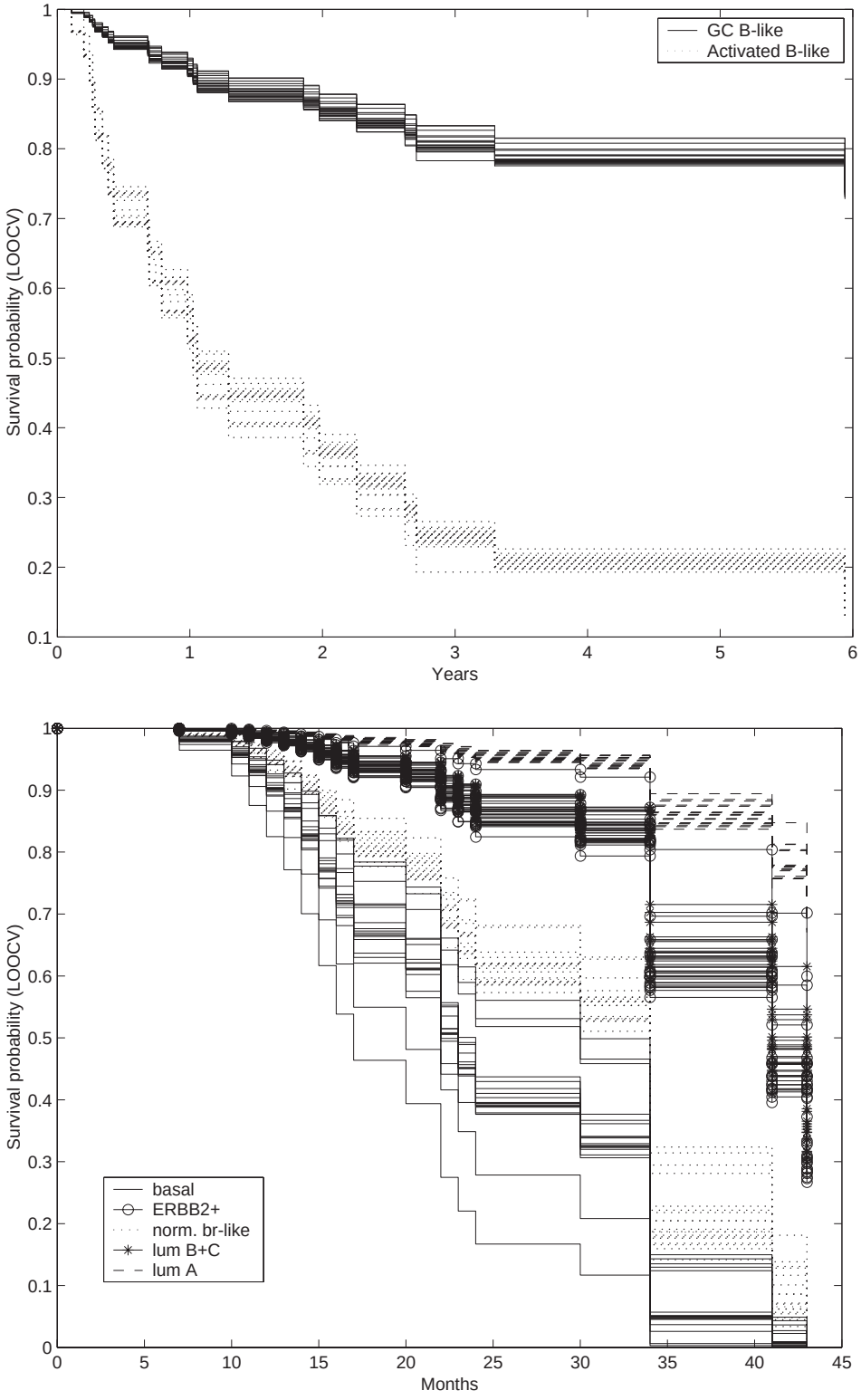

Fig. 3. Given are cross-validated survival curve estimates from the PLSPH regression model fits to the (top) diffuse large B-cell lymphoma data and (bottom) and the breast carcinomas data. See text for details. As in Figure 1 the curves are obtained for the group-average component profiles.

REFERENCES

Alizadeh,A.A., Eisen,M.B., Davis,R.E., Ma,C., Lossos,I.S., Rosenwald,A., Broldrick,J.C., Sabet,H., Tran,T., Yu,X. *et al.* (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Arjas,E. (1988) A graphical method for assessing goodness-of-fit in Cox's proportional hazards models. *J. Am. Stat. Assoc.*, **83**, 204–212.

Breslow,N. (1975) Analysis of survival data under the proportional hazard model. *Int. Statist. Rev.*, **43**, 45–57.

Cox,D.R. (1972) Regression models and life-tables (with discussion). *J. R. Stat. Soc.*, **B 34**, 187–220.

Cox,D.R. (1975) Partial likelihood. *Biometrika*, **62**, 269–276.

DeRisi,J., Penland,L., Brown,P.O., Bittner,M.L., Meltzer,P.S., Ray,M., Chen,Y., Su,Y.A. and Trent,J.M. (1996) Use of cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet.*, **14**, 457–460.

Frank,I.E. and Friedman,J.H. (1993) A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, **35**, 109–148.

Grambsch,P. and Therneau,T.M. (1994) Proportional harzards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515–526.

Helland,I.S. (1988) On the structure of partial least squares. *Commun. Stat.-Simul. Comput.*, **17**, 581–607.

Höskuldsson,A. (1988) PLS regression methods. *J. Chem.*, **2**, 211–228.

Kalbfleisch,J.D. and Prentice,R.L. (1973) Marginal likelihoods based on Cox's regression and like model. *Biometrika*, **60**, 267–278.

Kaplan,E.L. and Meier,P. (1958) Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, **53**, 457–481.

Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittermann,M., Wang,C., Kobayashi,M., Horton,H. and Brown,E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.

Martens,H. and Naes,T. (1989) *Multivariate Calibration*. Wiley, New York.

Nguyen,D.V. and Rocke,D.M. (2002) Classification of acute leukemia based on DNA microarray gene expressions using partial least squares. In Lin,S.M and Johnson,K.F. (eds), *Methods of Microarray Data Analysis*. Kluwer, Dordrecht, pp. 109–124.

Nguyen,D.V. and Rocke,D.M. (2002b) Tumor classification by partial least squares using gene expression data. *Bioinformatics*, **18**, 39–50.

Nguyen,D.V. and Rocke,D.M. (2002c) Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, **18**, 1216–1226.

Perou,C.M., Sorlie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S., Rees,C.A., Pollack,J.R., Ross,D.T., Johnsen,H., Akslen,L.A., Fluge,O. *et al.* (2000) Molecular portrait of human breast tumors. *Nature*, **406**, 747–752.

Ross,D.T., Scherf,U., Eisen,M.B., Perou,C.M., Rees,C., Spellman,P., Iyer,V., Jeffrey,S.S., Rijin,M.V., Waltham,M. *et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, **24**, 227–235.

Schemper,M. (1992) Further results on the explained variation in proportional hazards regression. *Biometrika*, **79**, 202–204.

Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Schoenfeld,D. (1982) Chi-square goodness-of-fit tests for proportional hazards regression. *Biometrika*, **67**, 145–153.

Sørlie,T., Perou,C.M., Tibshirani,R., Aas,T., Geisler,S., Johnsen,H., Hastie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S. *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA*, **98**, 10869–10874.

Wold,S. (1994) PLS for Multivariate Linear Modeling. In van de Vaterbeemd,H. (ed.), *Chemometric Methods in Molecular Design*. Verlag-Chemie, Weinheim, pp. 195–218.