

Assessing Patient Survival Using Microarray Gene Expression Data Via Partial Least Squares Proportional Hazard Regression

Danh V. Nguyen and David M. Rocke

Center for Image Processing and Integrated Computing
and
Department of Applied Science

University of California, Davis
Davis, CA 95616

Abstract

High dimensional data sets from microarray experiments where the number of variables (genes) p far exceed the number of samples N render most traditional statistical tools of little direct use. However, some of these statistical tools when used in conjunction with an appropriate dimension reduction method can be effective. In this paper we introduce the use of the proportional hazard (PH) regression (Cox 1972) in conjunction with dimension reduction by partial least squares (PLS), since the number of covariates p exceeds the number of samples N . This setting is typical of gene expression data from DNA microarrays. Specifically, for a given vector of response values which are times to event (death or censored times) and p gene expressions (covariates) we address the issue of how to assess (estimate) the survival experience (curve) when $N \ll p$. The approach taken to cope with the high dimensionality is to reduce the dimension via some dimension reduction (component extraction) method in the first stage and then estimate the survival distribution using a PH regression model in the second stage. The primary methods of component extraction considered is PLS. PLS achieves dimension reduction by constructing components to maximize the covariance between the response (survival times) and the linear combination of the covariates (gene expressions) sequentially. This is analogous to principal components analysis (PCA) but the optimization criterion in PCA is variance rather than covariance in PLS. We demonstrate the use of the methodology to a diffuse large B-cell lymphoma (DLBCL) *complementary* DNA (cDNA) data set.

Key Words: Gene Expression; Lymphoma; Principal components; Proportional hazard regression.

1 Introduction

The introduction of DNA microarray technology is a technical advance in biomedical research. Specifically, the use of microarray technology, such as *complementary* DNA (cDNA) and oligonucleotide arrays, allows simultaneous monitoring of thousands of gene expressions per sample (Schena et al. 1995). Notable is the use of

microarrays for human cancer research (DeRisi et al. 1996; Golub et al. 1999; Alon et al. 1999; Ross et al. 2000; Alizadeh et al. 2000; Perou et al. 1999, 2000; Bittner et al. 2000; Bubendorf et al. 1999; Wang et al. 1999).

The ability to measure gene expression en masse has also resulted in data with the number of variables p (genes) far exceeding the number of samples N . Of particular interest, for example, is when survival times of cancer patients are tracked. In this setting, it is of interest to estimate the patient survival probabilities using p gene expressions ($N \ll p$) and controlling for other covariates such as levels of clinical risk. For example, through gene expression profiling, Alizadeh et al. (2000) identified two distinct molecular subtypes of diffuse large B-cell lymphoma (DLBCL). Estimate of patient survival probabilities for the two groups were then compared using Kaplan-Meier (1958) survival curves.

In this paper, we suggest that estimates of patient survival probabilities can be based on the proportional hazard (PH) regression model after extracting gene components by partial least squares (PLS). This approach is termed partial least squares proportional hazard (PLSPH) regression. We illustrate PLSPH regression on a diffuse large B-cell lymphoma cDNA data set. The results of this application are given in section 3, as well as a brief description of the cDNA experiments producing the lymphoma cDNA data set analyzed. The methodology is described next in section 2.

2 Methods

In this section we describe the methodology used to assess patient survival 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. The second step involves fitting Cox PH regression model with the gene components as covariates. We briefly review PH regression and then describe PLS.

2.1 PH Regression with Gene Expressions as Covariates

Let Y be the time to some event, such as the survival time until relapse of diseased patients. Associated with each patient are p covariates which could be p gene expressions from DNA microarray data, for example. A data set of N samples (microarrays) consists of the triple $(T_i, \delta_i, \mathbf{x}_i)$ ($i = 1, \dots, N$), where $\mathbf{x}_i' = (x_{i1}, \dots, x_{ip})$ is the covariate pattern of the i th patient and 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 profile of the i th patient. Thus, the variate of interest, the survival times, can not be observed and, instead, we are only able to observe $T_i = \min(Y_i, Z_i)$ 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) model to study the relationship between the time to event and a set of covariates in the presence of censoring. The PH model is

$$h(t; \mathbf{x}_i; \boldsymbol{\beta}) = h_0(t) \exp(\mathbf{x}_i' \boldsymbol{\beta}) \quad (1)$$

where h is the hazard function associated with covariate \mathbf{x}_i and h_0 is an unspecified baseline hazard function. The hazard function $h(x)$ is defined as

$$h(x) = \lim_{\Delta x \rightarrow 0} P(x \leq X < x + \Delta x | X \geq x) / \Delta x \quad (2)$$

where $X \sim f(x)$. For X continuous, we have the relation $h(x) = f(x)/S(x)$, where $S(x) = P(X > x)$ is the survival function. Alternatively, since $S(x) = \exp[-H(x)]$ where $H(x) = \int_0^x h(u)du$ is the cumulative hazard function, under the PH model the survival function is

$$S(t; \mathbf{x}_i; \boldsymbol{\beta}) = [S_0(t)]^{\exp(\mathbf{x}_i' \boldsymbol{\beta})}. \quad (3)$$

$S_0(t)$ is the baseline survival function. Cox (1972) suggested maximizing the partial likelihood to estimate the vector of parameters $\boldsymbol{\beta}$. The partial likelihood does not involve the baseline hazard function. Subsequent studies of the the analysis of the PH model includes Kalbfleisch and Prentice (1973), Breslow (1974, 1975), Breslow and Crowley (1974), Cox (1975) and Efron (1977).

In the context of data generated from DNA microarrays, the number of co-variates or genes (p) is in the thousands, but the number of samples (N) is quite small (say $N = 50$). Standard statistical methodologies including the classical PH regression method described above (with $N \gg p$) do not work. Suppose that gene expression data are available from complementary DNA (cDNA) microarray experiments where each array contains $p = 10,000$ probes for human genes. Furthermore, suppose that we have obtained data for $N = 50$ samples of tissues from diseased patients under a treatment program. It may be of interest to study the group's survival experience in relation to the p gene expressions. For instance, does more positive survival experiences coincide with certain gene expression patterns (profiles)?

When there are more genes (variables) than there are samples and the PH regression is not defined, how do we estimate the survival experience? To cope with the high dimensionality of the covariate space (p -dimension) typical of gene expression data from microarrays, one can utilize some dimension reduction method. We describe this in the next section.

2.2 Dimension Reduction via PLS

The main method of dimension reduction considered in this paper is partial least squares (PLS), which originated in the field of econometric (Wold 1966, 1975) and has been applied with much success in the field of chemometrics. (See, for instance, Martens and Naes (1989).) High dimensional data with the structure that only a few underlying components explaining a large portion of total predictor variability is suitable for application of PLS, traditionally when the response variable is continuous. Gene expression data from DNA microarrays displays similar characteristics as those found in chemical applications. Nguyen and Rocke (2000, 2000b, 2001, 2001c) studied dimension reduction of gene expression data based on PLS for classification (when the response variable is binary or polychotomous). The reader is referred there for details. The use of PLS in the classical setting where the response is continuous has been investigated (Helland 1988, 1990; Höskuldsson 1988; Helland and Almøy 1994; Naes and Helland 1993; Frank and Friedman 1993; Stone and Brooks 1990; de Jong 1993; Phatek, Reilly and Penlidis 1992). In the present context the response variable is continuous but some observed time points are (right) censored. How does dimension reduction based on PLS perform under this setting? Particularly, after dimension can we still assess survival experience with some degree of "accuracy"? The use of PLSPH regression to estimate survival probabilities based on a simulation model for gene expression indicates that it is a useful method (Nguyen and Rocke, 2001b). We describe the methodology here and refer the reader to the reference above for details of the simulation studies.

We first describe principal component analysis (PCA) to highlight the similarities to PLS as well as to elucidate the differences. In PCA, orthogonal linear combinations are constructed to maximize the variance of the linear combination of the predictor variables sequentially,

$$\mathbf{v}_k = \operatorname{argmax}_{\mathbf{v}'\mathbf{v}=1} \operatorname{var}^2(\mathbf{X}\mathbf{v}) \quad (4)$$

subject to the orthogonality constraint

$$\mathbf{v}'\mathbf{S}\mathbf{v}_j = 0, \quad \text{for all } 1 \leq j < k \quad (5)$$

where $\mathbf{S} = \mathbf{X}'\mathbf{X}$. Often in applications of PCA, the predictors are standardized to have mean zero and standard deviation of one. This is referred to as PCA of the correlation matrix, $\mathbf{R}_{p \times p} = (1/(N-1))\mathbf{X}'\mathbf{X}$. The constructed principal components (PCs), satisfying the objective criterion (4) are obtained from the spectral decomposition of \mathbf{R} ,

$$\mathbf{R} = \mathbf{V}\mathbf{\Delta}\mathbf{V}', \quad \mathbf{\Delta} = \operatorname{diag}\{\lambda_1 \geq \dots \geq \lambda_{N-1}\}, \quad (6)$$

where $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_{N-1})$ are the corresponding eigenvectors. The i th PC is a linear combination of the original predictors, $\mathbf{X}\mathbf{v}_i$. Roughly, the constructed components summarize as much of the original p predictors' information (variation), irrespective of the response information (survival times).

Note that maximizing the variance of the linear combination of the predictors (genes), namely $\operatorname{var}(\mathbf{X}\mathbf{v})$, may not necessarily yields components predictive of the response variable (such as survival). For this reason, a different objective criterion for dimension reduction may be more appropriate for prediction. This issue was explored in some details using a leukemia data set by Nguyen and Rocke (2001).

The objective criterion for constructing components in PLS is to sequentially maximize the covariance between the response variable (survival) and a linear combination of the predictors (gene expressions). That is, in PLS, the components are constructed to maximize the objective criterion based on the sample covariance between \mathbf{y} and $\mathbf{X}\mathbf{c}$. Thus, we find the weight vector \mathbf{w} satisfying the following objective criterion,

$$\mathbf{w}_k = \operatorname{argmax}_{\mathbf{w}'\mathbf{w}=1} \operatorname{cov}^2(\mathbf{X}\mathbf{w}, \mathbf{y}) \quad (7)$$

subject to the orthogonality constraint

$$\mathbf{w}'\mathbf{S}\mathbf{w}_j = 0 \quad \text{for all } 1 \leq j < k \quad (8)$$

where $\mathbf{S} = \mathbf{X}'\mathbf{X}$. The i th PLS components are also a linear combinations of the original predictors, $\mathbf{X}\mathbf{w}_i$. (For algorithms, algebraic structures and relationships to linear regression see Höskuldsson (1988), Helland (1988) and Garthwaite (1994) respectively.)

PLS has been found to be useful in chemometrics where applications involve continuous response. Hence, after dimension reduction via PLS, prediction is by multiple linear regression. Here, due to censoring, prediction is by PH regression after dimension reduction via PLS. The use of PLS followed by logistic classification (binary response) has also been studied (Nguyen and Rocke, 2000, 2000b, 2001) as well. For instance, classification of normal and tumor tissues as well as different tumor types based on PLS gene components gave precise prediction results. Applications to (1) normal versus ovarian tumor, (2) acute leukemia, (3) diffuse large B-cell lymphoma, (4) normal versus colon tumor and (5) non-small-cell-lung carcinoma versus renal samples can be found in Nguyen and Rocke (2001).

3 Application

In this section we briefly describe a “typical cDNA experiment giving rise to DNA microarray gene expression data. We then describe the results of applying the method proposed here, PLSPH regression, to a cDNA lymphoma data set (Alizadeh et al., 2000). Specifically we fitted the PH regression model using PLS gene components as covariates to assess survival experience of the patients. Comparison of the survival curves from PH regression to those obtained (by Alizadeh et al.) using Kaplan-Meier survival curves are also given.

3.1 cDNA Experiments

Microarray technology utilizes the complementary base-pairing or hybridization property of DNA. On a single cDNA microarray (glass microscope slide) there are (thousands say) p printed DNA sequences (called *probes*). Two mRNA samples (called *targets*) are prepared; i.e. denatured (separate double strand into two single strands). By reverse transcription cDNA are obtained (from the mRNA) and are labeled with red-fluorescent dye Cy5 and green-fluorescent dye Cy3 (reference sample), for instance. The two cDNA samples are then combined and hybridized to the microarray. After hybridization the goal is to measure the abundance of the DNA sequences (on the microarray) present in the two fluorescent tagged cDNA samples. This is done by capturing the two (red and green) fluorescence signal intensities by some image processing analysis. With each fluorescent intensity measurement there is a background fluorescent. Often the *expression level* (abundance) of a gene is taken to be the measured fluorescent intensity (I) minus the background (B) fluorescent. Specifically, $z_r = I_r - B_r$ and $z_g = I_g - B_g$ denote the expression level of a gene in the red (r) and green (g) fluorescent tagged cDNA samples respectively. The *relative abundance* of a gene in the two cDNA samples is often taken to be the ratio of expression levels $x = z_r/z_g$. (For details see, for examples, Schena et al. (1995), DeRisi et al. (1996) and Schena, editor (1999).)

3.2 Lymphoma cDNA Data

The lymphoma cDNA dataset was published by Alizadeh et al. (2000) and 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). Each cDNA target was prepared from an experimental mRNA sample and was labeled with Cy5 (red fluorescent dye). A reference cDNA sample was prepared from a combination of nine different lymphoma cell lines and was labeled with Cy3 (green fluorescent dye). As described above each Cy5 labelled target were combined with the Cy3 labeled reference target and hybridized onto the microarray. We analyzed the standardized log relative intensity ratios, namely $\log(Cy5/Cy3)$ values.

For this gene expression data set survival data are available for $N = 40$ DLBCL samples (patients). There were 22 deaths. Associated with the survival times we considered relative intensity expressions of $p = 5622$ genes (covariates). Using cluster analysis Alizadeh et. al. identified two clinically distinct DLBCL subgroups: (Germinal Centre) GC B-like (19 patients) and Activated B-like (21 patients). These two groups were identified by gene expression profiling. In fact, this can be seen from the quite distinct expression patterns (profiles) for the two groups. For example, Figure 1 displays the relative expression patterns of 50 genes differen-

tially expressed across the two DLBCL subtypes. (The 50 genes are selected using the simple two sample t-statistics.)

Also available with this data set is the International Prognostic Index (IPI) for 38 of the 40 DLBCL patients. IPI scores are used to define clinical risk categories: low clinical risk=IPI score of 0-2 and high clinical risk=IPI score of 3-4.

3.3 Results

Based on the the DLBCL subgroups obtained by gene expression profiling Alizadeh et al. assessed (compared) patient survival experience using Kaplan-Meier survival curves for (1) GC B-like (19 patients, 6 deaths) versus Activated B-like (21 patients, 16 deaths), (2) low clinical risk (24 patients, 9 deaths) versus high clinical risk (15 patients, 11 deaths) and (3) GC B-like (10 patients, 6 deaths) versus Activated B-like (14 patients, 3 deaths) among patients with low clinical risk. These three pairs of survival curves are displayed in red in Figures 2, 3 and 4 respectively. In all cases, the log-rank test indicates significant differences between the cohorts compared (with $P = 0.01$, 0.002 and 0.05 for (1), (2) and (3) respectively).

We considered assessing patient survival experience using the gene expressions (covariates) directly using PH regression. This allows for controlling for other covariates, such as clinical risk status, for instance. To address (1) we obtained 3 PLS gene components based on $p = 2000$ genes and fitted the PH regression model using the 3 PLS gene components as covariates. Based on the fitted model, survival probability estimates are obtained for the average PLS gene components profile vector for GC B-like and Activated B-like, namely $\bar{\mathbf{t}}_{GC}$ and $\bar{\mathbf{t}}_{Act}$. The survival curves (blue) are plotted in Figure 2 along with the Kaplan-Meier curves of groups (1) obtained by Alizadeh et al. The survival curves estimated from the PH regression with PLS gene components is consistent in direction with that obtained by Alizadeh et al. However, the prognosis for Activated B-like is much worse and the the prognosis for GC B-like is much better on average.

Survival experience between patients with low clinical risk and high clinical risk differs significantly as can be seen from the Kaplan-Meier curves (red) in Figure 3. Thus, we refitted the PH regression model with 3 PLS gene components together with the clinical risk indicator as covariates. From this model, the estimated survival curves evaluated at the the mean PLS gene components profile together with low (0) and high (1) clinical risk indicator were obtained. That is, we compared GC and Activated B-like among high risk patients ($(1, \bar{\mathbf{t}}'_{GC})$ versus $(1, \bar{\mathbf{t}}'_{Act})$) and similarly for low risk patients ($(0, \bar{\mathbf{t}}'_{GC})$ versus $(0, \bar{\mathbf{t}}'_{Act})$) These four survival curves are plotted in Figure 3 together with the the Kaplan-Meier curves for the low and high clinical risk groups ignoring gene expressions. The survival probabilities for GC and Activated B-like are highly differentiated in both low (blue) and high (black) clinical risk patients. Figure 4 gives the Kaplan-Meier survival curves (in red) for the GC and Activated B-like groups among the low clinical risk patients. The corresponding estimated curves from the PH regression model (in blue) suggest much worse prognostic for Activated B-like patients.

Based on molecular differences between the two lymphoma types accompanied by the differentiated clinical prognosis for GC and Activated B cell DLBCL groups indicated by Kaplan-Meier survival curves, Alizadeh et al. suggested that the two types should be regarded as distinct diseases. Our analyses of the differentiated clinical prognosis for GC and Activated B cell DLBCL groups using PH regression based on PLS gene components are consistent with the results of Alizadeh et al. However, our analyses suggest that the differentiated clinical prognosis may be

much worse for Activated B-like patients and better for GC B-like patients than previously suggested.

4 Conclusions and Discussions

DNA microarray technologies, such as high-density oligonucleotide arrays and complementary DNA arrays, produce high dimensional gene expression data sets. 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 tool in scientific research, particularly in the biological and medicinal science. We have demonstrated the use of PLS gene components in PH regression to assess overall survival of GC and Activated B-like lymphoma patients controlling for clinical risk indicator, for instance. The methodology proposed here can be used in conjunction with current techniques used to analyze microarray data. Simulation studies of the proposed methodology here are described in details in Nguyen and Rocke (2001b).

References

1. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Bredrick, J. C., Sabet, H. Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J. Jr., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000), "Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling," *Nature*, 403, 503-511.
2. Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999), "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proceedings of the National Academy of Sciences*, 96, 6745-6750.
3. Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., Sampas, N., Dougherty, E., Wang, E., Marincola, F., Gooden, C., Lueders, J., Glatfelter, A., Pollock, P., Carpten, J., Gillanders, E., Leja, D., Dietrich K., Beaudry, C., Berens, M., Alberts, D., Sondak, V., Hayward, N., and Trent, J. (2000), "Molecular Classification of Cutaneous Malignant Melanoma by Gene Expression Profiling," *Nature*, 406, 536-540.
4. Bubendorf, L., Kolmer, J. K., Koivisto, P., Mousses, S., Chen, Y., Mahlamäki, E., Schraml, P., Moch, H., Willi, N., Elkahloun, A. G., Pretlow, T. G., Gasser, T. C., Mihatsch, M. J., Sauter, G., and Kallioniemi, O. (1999), "Hormone Therapy Failure in Human Prostate Cancer: Analysis by Complementary DNA and Tissue Microarrays," *Journal of the National Cancer Institute*, 91, 1758-1764.
5. Breslow, N. (1974), "Covariance Analysis of Censored Survival Data," *Biometrics*, 30, 89-99.

6. Breslow, N. (1975), "Analysis of Survival Data Under the Proportional Hazard Model," *Int. Statist. Rev.*, 43, 45-57.
7. Breslow, N., and Crowley, J. (1974), "A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship," *Annals of Statistics*, 2 437-53.
8. Cox, D. R. (1972), "Regression Models and Life-Tables" (with discussion), *Journal of the Royal Statistical Society*, Series B 34, 187-220.
9. Cox, D. R. (1975), "Partial Likelihood," *Biometrika*, 62, 269-76.
10. de Jong, S. (1993), "SIMPLS: An Alternative Approach to Partial Least Squares Regression," *Chemometrics and Intelligent Laboratory Systems*, 18, 251-263.
11. 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 Genetics*, 14, 457-460.
12. Efron, B. (1977), "The Efficiency of Cox's Likelihood Function for Censored Data," *Journal of the American Statistical Association*, 72, 557-565.
13. Frank, I. E., and Friedman, J. H. (1993), "A Statistical View of Some Chemometrics Regression Tools" (with discussion), *Technometrics*, 35, 109-148.
14. Garthwaite, P. H. (1994), "An Interpretation of Partial Least Squares," *Journal of the American Statistical Association*, 89, 122-127.
15. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531-537.
16. Helland, I. S. (1988), "On the Structure of Partial Least Squares," *Communications in Statistics-Simulation and Computation*, 17, 581-607.
17. Helland, I. S. (1990), "Partial Least Squares Regression and Statistical Models," *Scandinavian Journal of Statistics*, 17, 97-114.
18. Helland, S., and Almøy, T. (1994), "Comparison of Prediction Methods When Only a Few Components are Relevant," *Journal of the American Statistical Association*, 89, 583-591.
19. Höskuldsson, A., (1988), "PLS Regression Methods," *Journal of Chemometrics*, 2, 211-228.
20. Kalbfleisch, J. D., and Prentice, R. L. (1973), "Marginal Likelihoods Based on Cox's Regression and Like Model," *Biometrika*, 60, 267-78.
21. Kaplan, E. L. and Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, 53, 457-481.
22. Martens, H. and Naes, T. (1989), *Multivariate Calibration*. John Wiley & Sons, New York.

23. Naes, T., and Helland, I. S. (1993), "Relevant Components in Regression," *Scandinavian Journal of Statistics*, 20, 239-250.
24. Nguyen, D. V. and Rocke, D. M. (2000), "Classification in High Dimension with Application to DNA Microarray Data," manuscript.
25. Nguyen, D. V. and Rocke, D. M. (2000b), "Classification of Acute Leukemia Based on DNA Microarray Gene Expressions Using Partial Least Squares," to appear in *Methods of Microarray Data Analysis*.
26. Nguyen, D. V. and Rocke, D. M. (2001), "Tumor Classification by Partial Least Squares Using Microarray Gene Expression Data," accepted *Bioinformatics*.
27. Nguyen, D. V. and Rocke, D. M. (2001b), "Partial Least Squares Proportional Hazard Regression for Application to DNA Microarray Data," manuscript.
28. Nguyen, D. V. and Rocke, D. M. (2001c), "Multi-Class Cancer Classification Via Partial Least Squares Using Gene Expression Profiles," manuscript.
29. 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., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A., Brown, P. O., and Botstein, D. (2000), "Molecular Portrait of Human Breast Tumors," *Nature*, 406, 747-752.
30. Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C. F., Lashkari, D., Shalon, D., Brown, P. O., and Botstein, D. (1999), "Distinctive Gene Expression Patterns in Human Mammary Epithelial Cells and Breast Cancer," *Proceedings of the National Academy of Sciences, USA*, 96, 9112-9217.
31. Phatak, A., Reilly, P. M., and Penlidis, A. (1992), "The Geometry of 2-Block Partial Least Squares," *Communications in Statistics-Theory and Methods*, 21, 1517-1553.
32. Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Rijn, M. V., Waltham, M., Pergamenschikov, A., Lee, J., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O. (2000), "Systematic Variation in Gene Expression Patterns in Human Cancer Cell Lines," *Nature Genetics*, 24, 227-235.
33. Schena, M., editor (1999), *DNA Microarrays: A Practical Approach*, Oxford University Press.
34. 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.
35. van de Voet. (1994), "Comparing the Predictive Accuracy of Models Using a Simple Randomization Test," *Chemometrics and Intelligent Laboratory Systems*, 25, 313-323.

36. Wang, K., Gan, L., Jeffery, E., Gayle, M., Gown, A. M., Skelly, M., Nelson, P. S., Ng, W. V., Schummer, M., Hood, L, and Mulligan J. (1999), "Monitoring Gene Expression Profile Changes in Ovarian Carcinomas using cDNA microarrays," *Gene*, 229, 101-108.
37. Wold, H. (1966), "Estimation of Principal Components and Related Models by Iterative Least Squares." In Krishnaiah, P. R., editor, *Multivariate Analysis*, 391-420, Academic Press, New York.
38. Wold, H. (1975), "Soft Modelling by Latent Variables: The Non-Linear Partial Least Squares (NIPALS) Approach." In Gani, J., editor, *Perspectives in Probability and Statistics, Papers in Honour of M. S. Barlett*, 117-142, Academic Press, London.

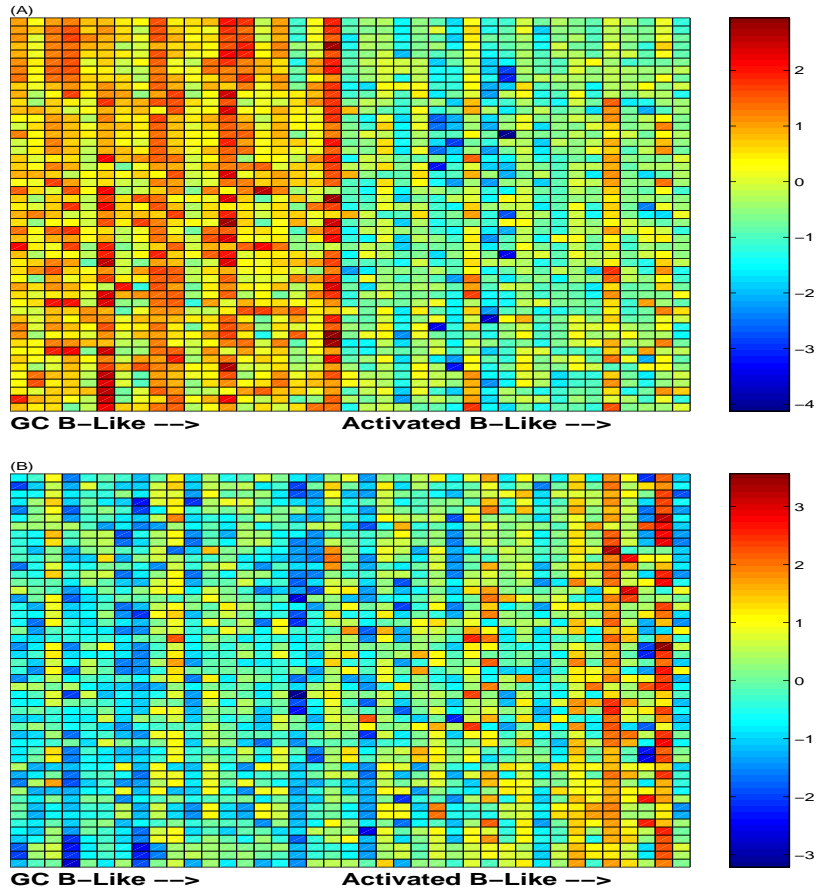


Figure 1: **DLBCL subtypes**. Displayed are the relative expression intensities of genes distinguishing DLBCL subgroups: GC B-like ($N = 19$) and Activated B-like samples ($N = 19$) and Activated B-like samples ($N = 21$) and Activated B-like samples ($N = 21$) samples. (A) Top 50 genes highly expressed in GC B-like relative to Activated B-like ($N = 21$) samples. (B) Top 50 genes highly expressed in Activated B-like relative to GC B-like samples. Relative expression values here are $\log(\text{Cy5}/\text{Cy3})$ values and Cy5 and Cy3 values are signal – background measurements. Expressions are standardized so that the mean is 0 and the standard deviation is 1.

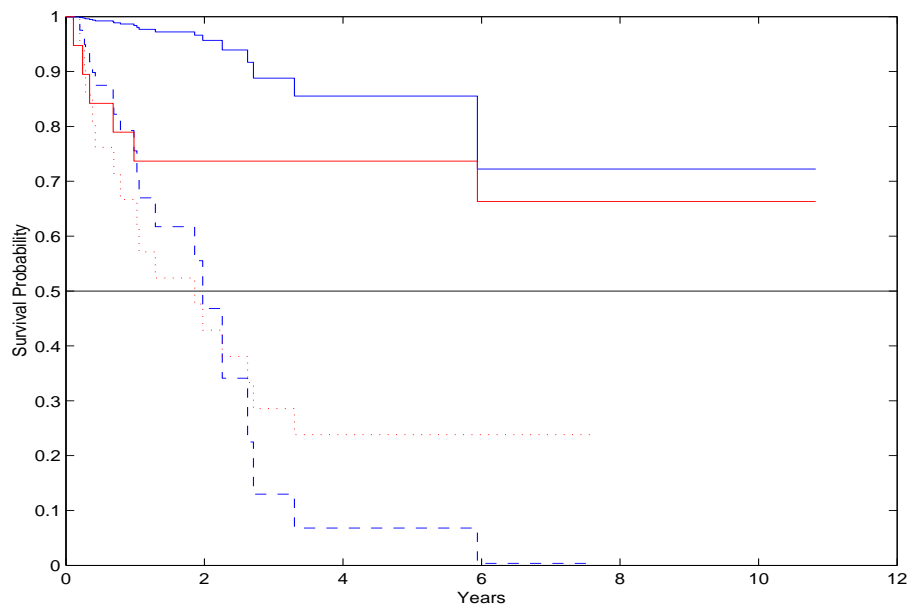


Figure 2: Given are Kaplan-Meier survival curves (in red) for 19 GC B-like patients (—) and 21 Activated B-like patients (---). The blue curves are obtained from the PH regression model with 3 PLS gene components (covariates) constructed from $p = 2000$ genes. The two (blue) survival curves from the PH regression were evaluated at the average covariate gene components vector for the GC and Activated B-like groups respectively.

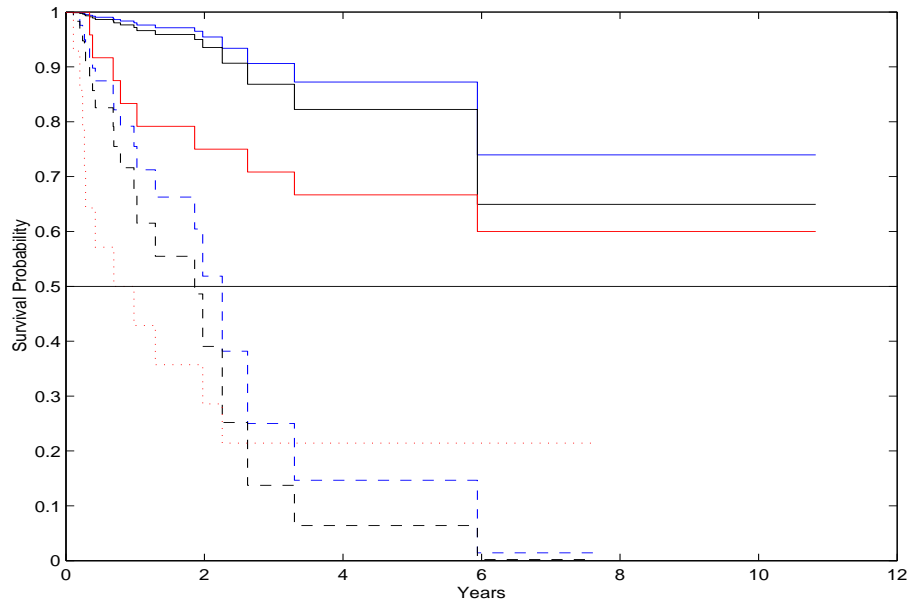


Figure 3: Given are Kaplan-Meier survival curves (in red) for 24 low clinical risk patients (\cdots) and 14 high clinical risk patients (---). The blue curves are the the low risk cohorts for GC B-like (---) and Activated B-like (---) group obtained from the PH regression model with 3 PLS gene components together with the clinical risk indicator as covariates. Similarly, the black curves are for the high risk cohorts (GC B-like --- , Activated B-like ---). The survival curves obtained from the PH regression (blue and black) were evaluated at the average covariate gene components for the GC and Activated B-like groups and corresponding risk indicator respectively. The PLS gene components (covariates) were constructed from $p = 2000$ genes.

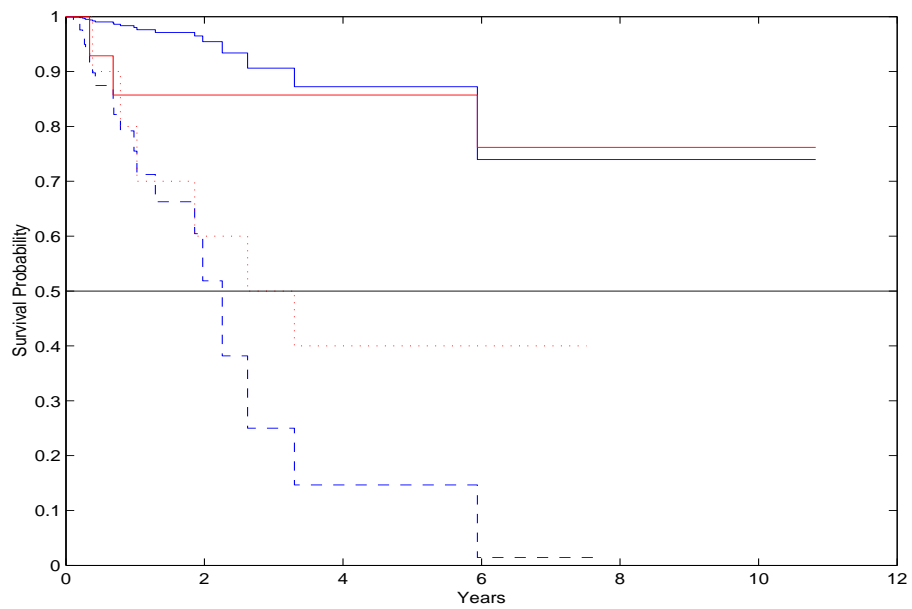


Figure 4: Given are Kaplan-Meier survival curves (in red) for 14 GC B-like patients (—) and 10 Activated B-like patients (---) with low clinical risk classification. The blue curves are the the low risk cohorts for GC B-like (—) and Activated B-like (---) group obtained from the PH regression model with 3 PLS gene components together with the clinical risk indicator as covariates. These are the blue curves given in Figure 3.