BIOINFORMATICS

Vol. 18 no. 1 2002
Pages 39–50

Tumor classification by partial least squares using microarray gene expression data

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

[1]Center for Image Processing and Integrated Computing and [2]Department of Applied Science, University of California, Davis, CA 95616, USA

Received on November 23, 2000; revised on March 22, 2001; accepted on June 6, 2001

ABSTRACT

Motivation: One important application of gene expression microarray data is classification of samples into categories, such as the type of tumor. The use of microarrays allows simultaneous monitoring of thousands of genes expressions per sample. This ability to measure gene expression en masse has resulted in data with the number of variables p (genes) far exceeding the number of samples N. Standard statistical methodologies in classification and prediction do not work well or even at all when $N < p$. Modification of existing statistical methodologies or development of new methodologies is needed for the analysis of microarray data.

Results: We propose a novel analysis procedure for classifying (predicting) human tumor samples based on microarray gene expressions. This procedure involves dimension reduction using Partial Least Squares (PLS) and classification using Logistic Discrimination (LD) and Quadratic Discriminant Analysis (QDA). We compare PLS to the well known dimension reduction method of Principal Components Analysis (PCA). Under many circumstances PLS proves superior; we illustrate a condition when PCA particularly fails to predict well relative to PLS. The proposed methods were applied to five different microarray data sets involving various human tumor samples: (1) normal versus ovarian tumor; (2) Acute Myeloid Leukemia (AML) versus Acute Lymphoblastic Leukemia (ALL); (3) Diffuse Large B-cell Lymphoma (DLBCLL) versus B-cell Chronic Lymphocytic Leukemia (BCLL); (4) normal versus colon tumor; and (5) Non-Small-Cell-Lung-Carcinoma (NSCLC) versus renal samples. Stability of classification results and methods were further assessed by re-randomization studies.

Availability: The methodology can be implemented using a combination of standard statistical methods, available, for example, in SAS. Illustrative SAS code is available from the first author.

Contact: nguyen@wald.ucdavis.edu;
dmrocke@ucdavis.edu

INTRODUCTION

With the wealth of gene expression data from microarrays (such as high density oligonucleotide arrays and cDNA arrays) prediction, classification, and clustering techniques are used for analysis and interpretation of the data. Some important recent applications are in molecular classification of acute leukemia (Golub *et al.*, 1999), cluster analysis of tumor and normal colon tissues (Alon *et al.*, 1999), clustering and classification of human cancer cell lines (Ross *et al.*, 2000), Diffuse Large B-cell Lymphoma (DLBCL; Alizadeh *et al.*, 2000), human mammary epithelial cells and breast cancer (Perou *et al.*, 1999, 2000) and skin cancer melanoma (Bittner *et al.*, 2000). These techniques have also helped to identify previously undetected subtypes of cancer (Golub *et al.*, 1999; Alizadeh *et al.*, 2000; Bittner *et al.*, 2000; Perou *et al.*, 2000). The problem of 'prediction' may come in various forms of applications as well; the prediction of patient survival duration with germinal center B-like DLBCL compared to those with activated B-like DLBCL using Kaplan–Meier survival curves (Ross *et al.*, 2000).

Gene expression data from DNA microarrays are characterized by many measured variables (genes) on only a few observations (experiments), although both the number of experiments and genes per experiment are growing rapidly. The number of genes on a single array is usually in the thousands, so the number of variables p easily exceeds the number of observations N. Although the number of measured genes is large there may only be a few underlying gene components that account for much of the data variation; for instance, only a few linear combinations of a subset of genes account for nearly all of the response variation. In this situation (i.e. when $N < p$), dimension reduction is needed to reduce the high p-dimensional gene space to a lower K-dimensional gene component space.

Under similar data structure in the field of chemometrics, the method of Partial Least Squares (PLS) has been found to be a useful dimension reduction technique. PLS has been useful as a predictive modeling regression method in the field of chemometrics. For example, in

[*]To whom correspondence should be addressed.

© Oxford University Press 2002

spectroscopy one may be predicting chemical composition of a compound based on observed signals for a particular wavelength, where the number of wavelengths (variables) is large. (Applications of PLS are abundant in the *Journal of Chemometrics* (Wiley) and *Chemometrics and Intelligent Laboratory Systems* (Elsevier), for example.) An introduction to PLS regression is given by Geladi and Kowalski (1986). The use of PLS in calibration can be found in Martens and Naes (1989). Some theoretical aspects and data-analytical properties of PLS have been studied by chemometricians and statisticians (de Jong, 1993; Frank and Friedman, 1993; Helland, 1988; Helland and Almøy, 1994; Höskuldsson, 1988; Lorber *et al.*, 1987; Phatak *et al.*, 1992; Stone and Brooks, 1990; Garthwaite, 1994).

In this paper we propose a novel analysis procedure for binary classification (prediction) of human tumor samples based on microarray gene expressions. Here, the response variable is a binary vector indicating normal or ovarian tumor samples, for example. This procedure involves dimension reduction using PLS and classification using Logistic Discrimination (LD) and Quadratic Discriminant Analysis (QDA). That is, the procedure involves two steps, a dimension reduction step and then a classification step. The proposed methods are applied to five different microarray data sets involving various human tumor samples: (1) normal versus ovarian tumor samples; (2) Acute Myeloid Leukemia (AML) versus Acute Lymphoblastic Leukemia (ALL); (3) Diffuse Large B-cell Lymphoma (DLBCLL) versus B-cell Chronic Lymphocytic Leukemia BCLL; (4) normal versus colon tumor samples; and (5) Non-Small-Cell-Lung-Carcinoma (NSCLC) versus renal. To assess the stability of the prediction results and methods we used re-randomization studies (as described in the Section **Methods**).

We compared PLS to the well known dimension reduction method of Principal Components Analysis (PCA; Massey, 1965; Jolliffe, 1986). PCA is used to reduce the high dimensional data to only a few gene components which explain as much of the observed total gene expression variation as possible. This is achieved without regard to the response variation. Gene components constructed this way are called Principal Components (PCs). In contrast to PCA, PLS components are chosen so that the sample covariance between the response and a linear combination of the p predictors (genes) is maximum. The latter criterion for PLS is more sensible since there is no *a priori* reason why constructed components having large predictor variation (gene expression variation) should be strongly related to the response variable. Certainly a component with small predictor variance could be a better predictor of the response classes. The ability of the dimension reduction method to summarize the covariation between gene expressions and response classes may yield better prediction results. Thus, for PCA to be competitive, relative to PLS, one can pre-select the genes which are predictive of the response classes and then apply PCA. Otherwise, one might expect PLS to give better predictions. Using the leukemia data set of (Golub *et al.*, 1999) we illustrate a condition when PCA fails to predict well relative to PLS.

This paper is organized as follows. In the Section **Methods** we describe the dimension reduction methods of PCA and PLS, the classification methods of LD and QDA and a gene selection strategy based on the simple t-statistics. In the Methods section we also describe the re-randomization technique used to further assess the prediction methods and results. The results from applying the proposed methods to the five microarray data sets are given in the Results section. Also included in the Results section is the illustration of a condition when PCA fails to predict well relative to PLS. We then conclude and discuss generalizations and other potential applications of PLS to microarray gene expression data. An Appendix is included which contains the PLS algorithm and a brief discussion of the algorithm's computational feasibility.

METHODS

Traditional statistical methodology for classification (prediction) does not work when there are more variables than there are samples. Specifically, for gene expression data, the number of tissue samples is much smaller than the number of genes. Thus, methods able to cope with the high dimensionality of the data are needed. In this section we describe a novel combination of dimension reduction with traditional classification methods, such as logistic and QDA, for high dimensional gene expression data. PLS is the primary dimension reduction method utilized, although we also consider the related method of PCA for comparison. The approach taken here consists of two main steps. The first step is the dimension reduction step, which reduces the high dimension p down to a lower dimension $K(K < N)$. Since the reduced dimension is smaller than the number of samples, in the second step, we can apply readily available prediction tools, such as LD or QDA.

We introduce the method of PLS by first briefly describing the well known and related method of PCA. Classification methods, namely LD and QDA, are also briefly described. Prior to analysis, gene selection may be necessary. Hence, we also describe a simple gene selection strategy based on the t-statistics.

Dimension reduction: PCA and PLS

The goal of dimension reduction methods is to reduce the high p-dimensional predictor (gene) space to a lower K-dimensional (gene component) space. This is achieved by extracting or constructing K components in the predictor space to optimize a defined objective criterion. PCA and

PLS are two such methods. To describe these methods some notations are required. Let $\mathbf{X}$ be an $N \times p$ matrix of N samples and p genes. Also, let $\mathbf{y}$ denote the $N \times 1$ vector of response values, such as indicator of leukemia classes or normal versus tumor tissues.

In PCA, orthogonal linear combinations are constructed to maximize the variance of the linear combination of the predictor variables sequentially,

$$\mathbf{v}_k = \underset{\mathbf{v}'\mathbf{v}=1}{\operatorname{argmax}} \operatorname{var}^2(\mathbf{X}\mathbf{v}) \tag{1}$$

subject to the orthogonality constraint

$$\mathbf{v}'\mathbf{S}\mathbf{v}_j = 0, \quad \text{for all} \quad 1 \leqslant j < k \tag{2}$$

where $\mathbf{S} = \mathbf{X}'\mathbf{X}$. The maximum number of nonzero components is the rank of $\mathbf{X}$, which is the same as the rank of $\mathbf{X}'\mathbf{X}$. 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, $\mathbf{R}_{p\times p} = (1/(N-1))(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')$. The constructed PCs, satisfying the objective criterion (1) are obtained from the spectral decomposition of $\mathbf{R}$,

$$\mathbf{R} = \mathbf{V}\boldsymbol{\Delta}\mathbf{V}', \qquad \boldsymbol{\Delta} = \operatorname{diag}\{\lambda_1 \geqslant \cdots \geqslant \lambda_{N-1}\}, \tag{3}$$

where $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{N-1})$ are the corresponding eigenvectors. The ith 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), as possible irrespective of the response class information.

Note that maximizing the variance of the linear combination of the predictors (genes), namely var($\mathbf{X}\mathbf{v}$), may not necessarily yield components predictive of the response variable (such as leukemia classes). 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 and a linear combination of the predictors. That is, in PLS, the components are constructed to maximize the objective criterion based on the sample covariance between $\mathbf{y}$ and $\mathbf{Xc}$. Thus, we find the weight vector $\mathbf{w}$ satisfying the following objective criterion,

$$\mathbf{w}_k = \underset{\mathbf{w}'\mathbf{w}=1}{\operatorname{argmax}} \operatorname{cov}^2(\mathbf{X}\mathbf{w}, \mathbf{y}) \tag{4}$$

subject to the orthogonality constraint

$$\mathbf{w}'\mathbf{S}\mathbf{w}_j = 0 \quad \text{for all} \quad 1 \leqslant j < k \tag{5}$$

where $\mathbf{S} = \mathbf{X}'\mathbf{X}$. The maximum number of components, as before, is the rank of $\mathbf{X}$. The ith PLS components are also a linear combinations of the original predictors, $\mathbf{X}\mathbf{w}_i$. A basic algorithm to obtain $\mathbf{w}$ is given in the Appendix.

Based on the different objective criterion of PCA and PLS, namely var($\mathbf{X}\mathbf{v}$) and cov($\mathbf{X}\mathbf{w}, \mathbf{y}$), it is reasonable to suspect that if the original p predictors (genes) are already predictive of response classes then the constructed components from PCA would likely be good predictors of response classes. Therefore, prediction results should be similar to that based on PLS components. Otherwise, one might suspect that PLS should perform better than PCA in prediction. We give examples of this in the Section **Results**.

Classification: LD and QDA

After dimension reduction by PLS and PCA, the high dimension of p is reduced to a lower dimension of K gene components. Once the K components are constructed we considered prediction of the response classes. Since the reduced (gene) dimension is now low ($K < N$), conventional classification methods such as LD and QDA can be applied.

Let $\mathbf{x}$ be the column vector of p predictor values and y denotes the binary response value. For instance, $y = 0$ for a normal sample, $y = 1$ for a tumor sample and $\mathbf{x}$ is the corresponding expression values of p genes. In logistic regression, the conditional class probability, $\pi = P(y = 1|\mathbf{x}) = P(\text{tumor given gene profile } \mathbf{x})$, is modeled using the logistic functional form,

$$\pi = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta})}. \tag{6}$$

The predicted response probabilities are obtained by replacing the parameter $\boldsymbol{\beta}$ with its Maximum Likelihood Estimate (MLE) $\hat{\boldsymbol{\beta}}$. The predicted class of each sample (as a normal or tumor sample) is $\hat{y} = I(\hat{\pi} > 1 - \hat{\pi})$, where $I(\cdot)$ is the indicator function; $I(A) = 1$ if condition A is true and zero otherwise. That is, we classify a sample as a tumorous ($\hat{y} = 1$) if the estimated probability of observing a tumor sample given the gene expression profile, $\mathbf{x}$, is greater than the probability of observing a normal sample with the same gene profile. This classification procedure is called LD. As mentioned earlier, LD is not defined if $N < p$. Thus, in order to utilize the LD procedure, we need to replace the original gene profile, $\mathbf{x}$, by the corresponding gene component profile in the reduced space, obtained from PLS or PCA.

Another usual classification method is QDA based on the classical multivariate normal model for each class: $\mathbf{x}|y = k \sim \mathrm{N}_p(\boldsymbol{\Sigma}_k, \boldsymbol{\mu}_k), \mathbf{x} \in \mathcal{R}^p$ and $k = 0, 1, \ldots, G$. For binary classification, $G = 1$. The (optimal) classification regions are

$$R_k = \{\mathbf{x} \in \mathcal{R}^p : p_k f_k(\mathbf{x}) > p_j f_j(\mathbf{x}), \ j \neq k\} \tag{7}$$

where f_k is the pdf of $\mathbf{x}|y = k$ given above and $p_k = P(y = k)$. This is equivalent to classifying a given sample $\mathbf{x}$ into the class with $\max\{q_i(\mathbf{x}), i = 0, 1, \ldots, G\}$, where $q_i(\mathbf{x}) = \mathbf{x}'\mathbf{A}_i\mathbf{x} + \mathbf{c}'\mathbf{x} + c_i$ with $\mathbf{A}_i = -0.5\mathbf{\Sigma}_i^{-1}$, $\mathbf{c}_i = \mathbf{\Sigma}_i^{-1}\boldsymbol{\mu}_i$, and $c_i = \log p_i - 0.5 \log \mathbf{\Sigma}_i - 0.5\boldsymbol{\mu}_i'\mathbf{\Sigma}_i^{-1}\boldsymbol{\mu}_i$. The posterior probability of membership in class k is $\pi_k = P(y = k|\mathbf{x}) = \exp[q_k(\mathbf{x})]/\sum_{i=0}^{K} \exp[q_i(\mathbf{x})]$. Again, the full gene profile, $\mathbf{x}$, is replaced by the corresponding gene component profile in the reduced space obtained from PLS or PCA.

For details on QDA and other classical classification methods the reader is referred to Mardia *et al.* (1979); Johnson and Wichern (1992) and Flury (1997). Details on logistic regression can be found in Hosmer and Lemeshow (1989) and McCullagh and Nelder (1989).

Gene selection

Although the two-step procedure outlined above, namely dimension reduction via PLS followed by classification via LD or QDA, can handle a large number (thousands) of genes, only a subset of genes is of interest in practice. Even after gene selection, often, the number of genes retained is still larger than the number of available samples. Thus, dimension reduction is still needed. It is obvious that good prediction relies on good predictors, hence a method to select the genes for prediction is necessary. For two-class prediction, selection and ranking of the genes can be based on the simple t-statistics

$$t = \frac{\bar{x}_0 - \bar{x}_1}{\sqrt{s_0^2/N_0 + s_1^2/N_1}} \tag{8}$$

where N_k, $\bar{x}_k$ and s_k^2 is the size, mean and variance, respectively, of class k, $k = 1, 2$. For each gene, a t-value is computed. We retain a set of the top p^* genes, by taking $p^*/2$ genes with the largest positive t-values (corresponding to high expression for class 1) and $p^*/2$ genes with smallest negative t-values (corresponding to high expression for class 2).

We conducted this procedure for $p^* = 50$ genes for the ovarian, leukemia, lymphoma, colon, and NCI60 data. For the leukemia data set the individual gene discrimination is relatively strong, which is suggestive of the well-known separability of AML/ALL leukemia classes based on gene expression in this data set. However, this differentially expressed pattern is not as clear for normal and ovarian tumor tissue samples or normal and colon tissue samples. Some color figures illustrating this difference are available on the website handel.cipic.ucdavis.edu/∼dmrocke.

Assessing prediction methods and results

Following gene selection and dimension reduction, we predicted the response classes. The observed error rate can be used to give a rough assessment of a method relative to another. The strength or 'confidence' associated with any specific prediction (i.e. for each sample) can be assessed by examining the estimated conditional class probability $\hat{\pi}$ described above.

It is also important to get an idea of how the proposed method will perform in the light of new data. However, new data are usually not available, so a re-randomization study is an alternative. For re-randomization studies a relatively large sample size, N, is needed. If there are sufficient samples, we carry out a three steps procedure to assess the prediction methods.

(1) Randomly form a training data set consisting of N_1 of the N samples. These N_1 samples in the training data set will be used to fit the model. The remaining $N_2 = N - N_1$ samples are saved for model validation (testing).

(2) A model is fit to the training data and the fit to the training data is assessed by leave-out-one Cross-Validation (CV). That is, one of the N_1 samples is left out and a model is fitted based on all but the left out sample. The fitted model is then used to predict the left out sample. Leave-out-one CV is used for each of the N_1 samples in the training data set. This provides some protection against overfitting, but it is still possible accidentally to select a model that fits the training data especially well due to capitalizing on chance.

(3) The model fit to all N_1 training data items will be tested on the N_2 samples not used to fit the model. This provides additional protection against overfitting.

(4) Even with these precautions, a sufficiently intensive search over methods can produce good results for the test data by capitalizing on chance properties. A final protection against this problem is obtained by re-randomization. After determining the full method including any model selection steps, we automate it so that it can be run on an arbitrary data set. We then repeat steps 1–3 for several re-randomizations of the original data set. We take these results as the best indication of the success of the methodology.

We carried out this procedure for the leukemia and lymphoma data sets which contain enough samples. For the ovarian and NCI60 data sets, which contain few samples, we performed only leave-out-one CV prediction.

RESULTS

We demonstrate the usefulness of the proposed methodology described above to five microarray data sets with various human tumor samples: (1) ovarian (Furey *et al.*, 2000); (2) leukemia (Golub *et al.*, 1999); (3) lymphoma (Alizadeh *et al.*, 2000); (4) colon (Alon *et al.*, 1999); and (5) cancer cell lines from the NCI60 data set (Ross *et al.*, 2000). Data sets (2), (3) and (5) are published data sets and are publicly available at http://waldo.wi.mit.edu/MPR/, http://llmpp.nih.gov/lymphoma/ and http://genome-www.stanford.edu/ respectively. The ovarian data set is yet to be published but analyzed results were published by Furey *et al.* (2000).

Ovarian data

The microarray experiments consist of hybridizations of normal and ovarian tissues on arrays with probes for 97 802 DNA clones. The ovarian data set considered here consists of 16 normal tissue samples and 14 ovarian tumor tissue samples. The normal tissue samples consist of a spectrum of normal tissues: 2 cyst, 4 peripheral blood lymphocytes, 9 ovary and 1 liver normal tissue. All normal and tumorous samples are distinct, coming from different tissues (patients). We log transformed all the gene expression values due to the highly skewed data, typical of gene expression data. The expressions of all genes were also standardized to have mean zero and standard deviation of one across samples.

We considered $p^* = 50, 100, 500, 1000, 1500$ genes selected as described in the Section **Methods**. Since there are few samples, we made leave-out-one CV prediction. Classification of the 30 tissue samples based on $K = 3$ gene components constructed from p^* genes using PLS and PCA are given in Table 1. Overall, the classification results are good. All normal and ovarian tumor samples were correctly classified using LD with PLS and PCs. Results for QDA is the same, except, with PCs one normal (cyst) sample was misclassified ($p^* = 50$ and 500). Also, different analyses using $p^* = 1000$ and 1500 misclassified one normal ovarian sample. However, all classification methods using PLS gene components are 100% correct for the ovarian data. Furey *et al.* (2000) also used leave-out-one CV prediction for this data set as well, but using Support Vector Machines (SVM; Vapnik, 2000). Although it is not our intent to tune our analyses to theirs to make exact comparisons, a crude observation can be made. Furey *et al.* reported 3–5 normal samples and 2–4 ovarian tumor samples misclassified using SVM based on 25 to 100 genes. (See Table 1 of Furey *et al.*[†])

The strength or 'confidence' in the predictions made can be assessed by examining the estimated conditional class probability, namely $\hat{\pi}_i = \hat{P}(Y = k|\mathbf{x}_i^*)$, $k = 0, 1$, where $\mathbf{x}_i^*$ is the gene profile (pattern) in the reduced K-dimensional space. For $p^* = 50$ and 100 genes, the estimated conditional probability is essentially one for PLS and the lowest $\hat{\pi}$ is 0.973 for PCA. This holds for $p^* = 1000$ and 1500 genes as well. However, for $p^* = 500$ genes, two samples were correctly classified (PCA) with moderate estimated conditional class probability of 0.922 and 0.925. Sample 16 is a normal sample from a white blood cell line (HWBC3) and exhibits characteristics of both normal and tumor cells, which makes it a likely candidate for misclassification. SVM had problems classifying this sample (Furey *et al.*, 2000, p. 910) but PLS correctly classified this sample as normal tissue.

Table 1. Classification results for normal and ovarian tumor samples. Given are the number of correct classification out of 30 samples (16 normal and 14 ovarian tumor samples)

	LD		QDA		Sample
p^*	PLS	PC	PLS	PC	misclassified
50	30	30	30	29	#1
100	30	30	30	30	
500	30	30	30	29	#1
1000	30	30	30	29	#4
1500	30	30	30	29	#4

Leukemia data

The data set used here is the acute leukemia data set published by Golub *et al.* (1999). The original training data set consisted of 38 bone marrow samples with 27 ALL and 11 AML (from adult patients). The independent (test) data set consisted of 24 bone marrow samples as well as 10 peripheral blood specimens from adults and children (20 ALL and 14 AML). Four AML samples from the independent data set were from adult patients. The gene expression intensities were obtained from Affymetrix high-density oligonucleotide microarrays containing probes for 6817 genes. We log transformed the gene expressions to have mean zero and standard deviation one across samples. No further data preprocessing was applied.

We first applied the proposed methods to the original data structure of 38 training samples and 34 test samples for $p^* = 50$, 100, 500, 1000, and 1500 genes selected as described earlier. The results are given in Table 2. All methods predicted the ALL/AML class correctly 100% for the 38 training samples using leave-out-one CV. Prediction of the test samples using LD based on the training (PLS and PCA) components resulted in one misclassification: sample #66. This is based on $p^* = 50$ genes. This AML sample was also misclassified by Golub *et al.* (1999) using a weighted voting scheme[‡].

[†] Furey *et al.* included another sample tissue from the same patient. We only use samples from distinct patients since samples should be independent. However, inclusion of this one extra sample did not change the results reported here.

[‡] Participants of the Critical Assessment of Techniques for Microarray Data Mining (CAMDA'00, December 2000) Conference analyzing the leukemia data set all misclassified sample #66. Whether the sample was incorrectly labeled is not known.

Table 2. Classification results for the leukemia data set with 38 training samples (27 ALL, 11 AML) and 34 test samples (20 ALL, 14 AML). Given are the number of correct classification out of 38 and 34 for the training and test samples respectively

	Training data (leave-out-one CV)				Test data (out-of-sample)			
	LD		QDA		LD		QDA	
p^*	PLS	PC	PLS	PC	PLS	PC	PLS	PC
50	38	38	38	38	33	33	28	30
100	38	38	38	38	32	32	29	30
500	38	38	38	38	31	31	32	28
1000	38	38	38	38	31	31	31	28
1500	38	38	38	38	31	30	30	28

Table 3. Classification results for re-randomization study of the leukemia data set with 36/36 splitting. Each value in the table is the correct classification percentage averaged over 100 re-randomizations. Perfect classification is 36

	Training data (leave-out-one CV)				Test data (out-of-sample)			
	LD		QDA		LD		QDA	
p^*	PLS	PC	PLS	PC	PLS	PC	PLS	PC
50	36.00	34.08	35.99	34.92	34.72	33.66	34.63	34.63
100	35.88	33.29	35.89	34.89	34.30	32.92	34.80	34.58
500	36.00	34.32	36.00	35.09	34.73	34.08	34.53	34.60
1000	36.00	32.95	36.00	34.57	34.82	32.50	34.77	34.09
1500	36.00	32.51	36.00	33.79	34.71	32.11	34.67	33.66

To assess the stability of the results shown in Table 2 we carried out a re-randomization study as described in the Section **Methods**. We considered an equal random splitting of the $N = 72$ samples: $N_1 = 36$ training and $N_2 = 36$ test samples. The analysis above was repeated for 100 re-randomizations. Table 3 gives the average classification rates over the 100 re-randomizations. LD and QDA prediction based on PLS gene components resulted in near perfect classification (between 99 and 100% correct) for the training samples using leave-out-one CV. PCs fared slightly worse (between 90 and 97% correct). This is based on all p^* considered. For the test samples, PLS gene components in LD performed better than PCs. However, both PLS and PCs performed similarly in QDA.

We also classified the samples based on the 50 ‘predictive’ genes set reported by Golub *et al.* Leave-out-one CV predictions of the 38 training samples using QDA and LD with PLS gene components resulted in 100% correct and 36/38 for PCs. Based on only the training components, out-of-sample predictions of the 34 test samples were also made. LD with PLS gene components resulted in one misclassification (#66). Golub *et al.* associated with each prediction a ‘Prediction Strength’ (PS). (For details, see Golub *et al.*) Five test samples were predicted with low (PS < 0.30) to borderline PS: samples #54, 57, 60, 67 and 71 (PS $= 0.23$, 0.22, 0.06, 0.15 and 0.30) with one sample misclassified. These five samples were all correctly classified using LD with PLS gene components with moderate to high conditional class probabilities of 0.97, 1.00, 0.98, 0.89 and 1.00 respectively. Results for all 72 samples are given in Table 4 and re-randomization results, given in Table 5, showed the stability of the estimates.

Lymphoma data

The lymphoma data set was published by Alizadeh *et al.* (2000) and consists of gene expressions from cDNA experiments involving three prevalent adult lymphoid malignancies: DLBCLL, 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). Each Cy5 labeled target was combined with the Cy3 labeled reference target and hybridized onto the microarray. Separate measurements were taken from the red and green channels. We analyzed the standardized log relative intensity ratios, namely the log($Cy5/Cy3$) values. To test the binary classification procedures proposed in this paper, we analyze a subset of the data consisting of 45 DLBCL and 29 BCLL samples with $p = 4227$ genes.

Using leave-out-one CV, each sample was predicted to be DLBCL or BCLL based on 3 gene components constructed from $p^* = 50$, 100, 500 and 1000 genes. The results are given in Table 6. Of the 74 total samples, PLS gene components resulted in either one or two misclassifications at most. The two misclassified samples, #33 and 51, were consistently misclassified. PCs did not perform as well using LD, with at most four misclassifications. However, PCs used with QDA performed similar to PLS components.

As with the analysis of the leukemia data, we turned next to re-randomization studies to assess the stability of the classification results. Table 7 summarizes the results of 100 re-randomizations (with 37/37 random split). For this data set, PLS components in LD appear to perform best for leave-out-one CV (of the training data sets). Out-of-sample prediction results for PLS and PCs are similar. On average, classification of the training samples using leave-out-one CV is nearly 100% correct and about less than two misclassifications out of 37 for test samples.

Colon data

Alon *et al.* (1999) used Affymetrix oligonucleotide arrays to monitor expressions of over 6500 human genes with

Table 4. 50 Genes from Golub *et al.* Predicted (1 = ALL, 0 = AML) probabilities using leave-out-one CV for original 38 training samples and out-of-sample prediction for the 34 test samples using PLS and PC. PS is the prediction strength from Golub *et al.* For LD, $\hat{\pi}$ is an estimate of $\pi = P(Y = 1|\text{data})$, and for QDA it is the posterior probability or conditional class probability. Samples marked with an $*$ were misclassified

			Training data							Test data			
			LD		QDA					LD		QDA	
i	y_i	PS	$\hat{\pi}_{\mathrm{PLS}}$	$\hat{\pi}_{\mathrm{PC}}$	$\hat{\pi}_{\mathrm{PLS}}$	$\hat{\pi}_{\mathrm{PC}}$	i	y_i	PS	$\hat{\pi}_{\mathrm{PLS}}$	$\hat{\pi}_{\mathrm{PC}}$	$\hat{\pi}_{\mathrm{PLS}}$	$\hat{\pi}_{\mathrm{PC}}$
1	1	1.00	1.00	1.00	1.00	1.00	39	1	0.78	1.00	1.00	1.00	1.00
2	1	0.41	1.00	1.00	1.00	1.00	40	1	0.68	1.00	1.00	1.00	1.00
3	1	0.87	1.00	1.00	1.00	1.00	41	1	0.99	1.00	1.00	1.00	1.00
4	1	0.91	1.00	1.00	1.00	1.00	42	1	0.42	1.00	1.00	1.00	1.00
5	1	0.89	1.00	1.00	1.00	1.00	43	1	0.66	1.00	1.00	1.00	1.00
6	1	0.76	1.00	1.00	1.00	1.00	44	1	0.97	1.00	1.00	1.00	1.00
7	1	0.78	1.00	1.00	1.00	1.00	45	1	0.88	1.00	1.00	1.00	1.00
8	1	0.77	1.00	1.00	1.00	1.00	46	1	0.84	1.00	1.00	1.00	1.00
9	1	0.89	1.00	1.00	1.00	1.00	47	1	0.81	1.00	1.00	1.00	1.00
10	1	0.56	1.00	1.00	1.00	1.00	48	1	0.94	1.00	1.00	1.00	1.00
11	1	0.74	1.00	1.00	1.00	1.00	49	1	0.84	1.00	1.00	1.00	1.00
12	1	0.20*+	1.00	0.02*	1.00	0.00*	50	0	0.97	0.00	0.00	0.00	0.00
13	1	1.00	1.00	1.00	1.00	1.00	51	0	1.00	0.00	0.00	0.00	0.00
14	1	0.73	1.00	1.00	1.00	1.00	52	0	0.61	0.00	0.01	0.00	0.00
15	1	0.98	1.00	1.00	1.00	1.00	53	0	0.89	0.00	0.00	0.00	0.00
16	1	0.95	1.00	1.00	1.00	1.00	54	0	0.23^{+}	0.03	1.00*	0.00	0.15
17	1	0.49	1.00	1.00	1.00	1.00	55	1	0.73	1.00	1.00	1.00	1.00
18	1	0.59	1.00	1.00	1.00	1.00	56	1	0.84	1.00	1.00	1.00	1.00
19	1	0.80	1.00	1.00	1.00	1.00	57	0	0.22^{+}	0.00	1.00*	0.00	0.03
20	1	0.90	1.00	1.00	1.00	1.00	58	0	0.74	0.08	0.00	1.00*	0.01
21	1	0.76	1.00	1.00	1.00	1.00	59	1	0.68	1.00	1.00	1.00	1.00
22	1	0.37	1.00	1.00	1.00	1.00	60	0	0.06^{+}	0.02	1.00*	0.00	0.68*
23	1	0.77	1.00	1.00	1.00	1.00	61	0	0.40	0.35	1.00*	1.00*	0.02
24	1	0.82	1.00	1.00	1.00	1.00	62	0	0.58	0.00	0.63*	0.00	0.00
25	1	0.43	1.00	1.00	1.00	1.00	63	0	0.69	0.00	0.98*	0.00	0.00
26	1	0.89	1.00	1.00	1.00	1.00	64	0	0.52	0.00	0.27	0.00	0.01
27	1	0.82	1.00	1.00	1.00	1.00	65	0	0.60	0.00	0.21	0.00	0.00
28	0	0.44	0.00	0.00	0.00	0.00	66	0	0.27*+	0.93*	1.00*	1.00*	0.99*
29	0	0.74	0.00	0.02	0.00	0.00	67	1	0.15*+	0.89	1.00	1.00	0.20*
30	0	0.80	0.00	0.00	0.00	0.00	68	1	0.80	1.00	1.00	1.00	1.00
31	0	0.61	0.00	0.00	0.00	0.00	69	1	0.85	1.00	1.00	1.00	1.00
32	0	0.47	0.00	0.00	0.00	0.00	70	1	0.73	1.00	1.00	1.00	1.00
33	0	0.89	0.00	0.00	0.00	0.00	71	1	0.30^{+}	1.00	1.00	1.00	1.00
34	0	0.64	0.00	0.00	0.00	0.00	72	1	0.77	1.00	1.00	1.00	1.00
35	0	0.21^{+}	0.00	1.00*	0.00	1.00*							
36	0	0.94	0.00	0.00	0.00	0.00							
37	0	0.95	0.00	0.00	0.00	0.00							
38	0	0.73	0.00	0.00	0.00	0.00							
		# correct	38	36	38	36				33	27	31	31

samples of 40 tumor and 22 normal colon tissues. Using a clustering algorithm based on the deterministic-annealing algorithm, Alon *et al.* clustered the 62 samples into two clusters. One cluster consisted of 35 tumor and 3 normal samples (n8, n12, n34[†]). The second cluster contained 19 normal and 5 tumor tissues (T2, T30, T33, T36, T37). (See Figure 4 of Alon *et al.*) Furey *et al.* (2000) did leave-out-one CV prediction of the 62 samples using SVM and six tissues were misclassified, namely (T30, T33, T36) and (n8, n34, n36). As Furey *et al.* pointed out, the three misclassified tumors (T30, T33, T36) were among the five tumor samples which clustered into the normal group by Alon *et al.* Also, two normal samples (n8, n34) misclassified by Furey *et al.* were among the three normal samples clustered into the tumor group by Alon *et al.*

Classification of tumor and normal colon tissues using the method proposed here are displayed in Table 8. We carried out analyses for $p^* = 50, 100, 500$ and 1000. PLS gene components in LD for 50 and 100 genes performed

[†] The labeling for the 22 normal tissues in Alon *et al.* are not in consecutive order.

Table 5. Results from re-randomizations using the 50 genes obtained by Golub *et al.* Given are average classification rate from all re-randomizations (36 training/36 test samples splitting)

	LD		QDA	
	PLS	PC	PLS	PC
Training data	99.56	96.44	99.56	97.00
Test data	95.94	94.17	96.44	95.44

Table 6. Classification results for DLBCL and BCLL lymphoma samples. Given are the number of correct classification out of 74 samples (45 DLBCL and 29 BCLL samples). Samples misclassified are given in parentheses on the right side of the table

	LD		QDA		Sample(s) misclassified			
p^*	PLS	PC	PLS	PC	PLS	PC	PLS	PC
50	72	73	73	72	(33, 51)	(51)	(51)	(33, 51)
100	72	71	72	73	(33, 51)	(9, 33, 51)	(33, 51)	(51)
500	72	71	73	73	(33, 51)	(9, 45, 51)	(51)	(45)
1000	72	70	73	73	(33, 51)	(9, 32, 48, 51)	(51)	(51)

best with four misclassifications. For $p^* = 50$ genes T2, T11, T33, and n36 were misclassified. For $p^* = 100$ genes T11, T30, T33, and n11 were misclassified. Note that with the exception of T11 and n11 the samples misclassified here were also misclassifed using SVM and by clustering. Note from Table 8 that PCs is not competitive relative to PLS components for this data set. We also note that gene expression patterns for this data set are quite heterogeneous. Further, the samples that are most commonly misclassified by various methods of analysis have expression patterns that are quite different from their respective groups.

Table 7. Classification results for re-randomization study of the lymphoma data set with 37/37 splitting. Each value in the table is the correct classification percentage averaged over 100 re-randomizations. Perfect classification is 37

	Training data (leave-out-one CV)				Test data (out-of-sample)			
	LD		QDA		LD		QDA	
p^*	PLS	PC	PLS	PC	PLS	PC	PLS	PC
50	36.87	36.26	36.64	35.60	35.57	36.03	35.88	35.81
100	36.86	36.38	36.74	36.30	35.84	36.29	36.03	36.10
500	36.89	35.15	36.77	35.99	35.76	35.21	35.69	35.85
1000	36.83	34.94	36.90	35.68	35.58	33.87	35.32	35.12

Table 8. Classification results for normal and colon tissue samples. Given are the number of correct classification out of 62 samples (40 tumors and 22 normal samples)

	LD		QDA	
p^*	PLS	PC	PLS	PC
50	58	54	57	54
100	58	53	56	52
500	56	53	57	53
1000	57	52	56	54

NCI60 data

The NCI60 data set, published by Ross *et al.* (2000), consists of samples from human tumor cell lines. The data is from 60 cDNA arrays each containing 9703 spotted cDNA sequences. The cDNAs arrays contain approximately 8000 unique genes in 60 human cell lines obtained from various tumor sites: 7 breast, 5 Central Nervous System (CNS), 7 colon, 6 leukemia, 8 melanoma, 9 NSCLC, 6 ovarian, 2 prostate, 9 renal, and 1 unknown. The reference sample used in all hybridizations was prepared by combining an equal mixture of mRNA from 12 of the cell lines. As with the lymphoma (cDNA data) we analyzed the standardized log relative intensity ratios, namely the log(Cy5/Cy3) values. To illustrate our binary classification procedures to this cell lines gene expression data, we selected two of the largest groups: 9 NSCLC and 9 renal samples. Using a subset of 6814 genes we applied dimension reduction methods to the selected $p^* = 50, 100, 500$ and 1000 genes. The results are given in Table 9. PLS gene components predicted all NSCLC and renal cell lines samples correctly 100% in all instances. For each analysis, PCs only misclassified one sample, either sample 4 or 15. The expression patterns of these two misclassified samples are quite different from their respective groups.

A condition when PCs fail to predict well

The conditions under which PLS predicts well have not yet been fully characterized in the statistics or chemometrics literature. In this section we illustrate, by example, a condition when PCs fail to predict well, but PLS components continue to predict well. The example to be given is based on the leukemia data set of Golub *et al.* (1999).

In the analyses given above, although the results for PLS components were better than that for PCs, the results for PCs were competitive nonetheless. Examining the objective criterion of PLS and PCA, we noted earlier that it would be reasonable to expect predictions based on PCs to be similar to that from PLS if the predictors (genes) are highly predictive of response (leukemia) classes. This is the case of the analyses based on the 50 predictive genes reported by Golub *et al.*, for instance. However, to see when PCs fail to predict well, while PLS components

Table 9. Classification results for NSCLC and renal cell lines. Given are the number of correct classification out of 18 samples (9 NSCLC and 9 renal samples)

	LD		QDA		Sample
p^*	PLS	PC	PLS	PC	misclassified
50	18	17	18	18	#4
100	18	17	18	18	#4
500	18	18	18	17	#15
1000	18	18	18	17	#15

succeeded, we considered their prediction ability based only on expressed genes, but not exclusively expressed differentially for leukemia classes. This test condition is based on the simple fact that an expressed gene does not necessarily qualify as a good predictor of leukemia classes. For instance, consider a gene highly expressed across all samples, ALL and AML. In this case, the gene will not discriminate between ALL and AML well. We define five nested data sets consisting of all genes expressed on (A) at least one array ($p = 1554$), (B) 25% ($p = 1076$), (C) 50% ($p = 864$), (D) 75% ($p = 662$) and (E) 100% ($p = 246$) of the arrays. Note that these genes are expressed but not necessarily differentially expressed for ALL/AML. As before, we applied PLS and PCA to extract three gene components from these five data sets based on the 38 training samples. Predictions of the 38 training samples were based on leave-out-one CV and predictions of the 34 test samples were based on the training components only.

The results are given in Table 10. As can be seen, the decrease in performance of PCs relative to PLS is drastic compared to the result of the 50 predictive genes (Tables 4 and 5). To check the stability of the results in Table 10, we ran 50 re-randomizations. The results are given in Table 11. PCs did much worse relative to PLS gene components in the re-randomizations as well. The result here is not surprising since PCA aims to summarize only the variation of the p genes. However, only a subset of p expressed genes are predictive of leukemia classes. Why then do PLS components still perform well in this mixture of expressed genes, both predictive and non-predictive of leukemia classes? This is most likely attributed to the choice of objective criterion used, namely covariation between the leukemia classes and (the linear combination of) the p genes. Since PLS components are obtained from maximizing $\text{cov}(\mathbf{Xw}, \mathbf{y})$ it is more able to assign patterns of weights to the genes which are predictive of leukemia classes.

Table 10. LD and QDA of original (38 training/34 test samples splitting) based on class prediction using leave-out-one CV for training data set and out-of-sample prediction for test data set. The five data sets consist of expressed genes, but not all are differentially expressed for AML/ALL

	% correct, train		% correct, test	
Gene set	PLS	PC	PLS	PC
		LD		
Set A	100.00	84.21	91.18	73.53
Set B	100.00	81.58	91.18	73.53
Set C	100.00	84.21	91.18	73.53
Set D	100.00	81.58	91.18	73.53
Set E	100.00	76.32	79.41	64.71
		QDA		
Set A	100.00	84.21	91.18	82.35
Set B	100.00	84.21	94.12	82.35
Set C	100.00	81.58	91.18	82.35
Set D	100.00	81.58	91.18	88.24
Set E	100.00	57.89	71.05	50.00

Table 11. Average classification rates from 50 re-randomizations (36 training/36 test samples splitting) and prediction using leave-out-one CV for training data sets and out-of-sample prediction for test data sets

	% correct, train		% correct, test	
Gene set	PLS	PC	PLS	PC
		LD		
Set A	99.67	86.67	93.78	82.00
Set B	99.94	87.44	94.39	83.00
Set C	99.83	89.94	93.89	83.83
Set D	99.78	85.00	94.78	83.39
Set E	97.44	73.89	89.22	69.00
		QDA		
Set A	100.00	83.44	93.17	82.39
Set B	99.94	86.33	94.28	85.00
Set C	99.72	88.28	93.50	85.17
Set D	99.78	85.00	94.78	83.39
Set E	98.06	67.28	89.61	64.89

Further indication of this condition, where PCs fail to predict leukemia classes while PLS components succeeded, can be found in the table of response (leukemia classes) and predictor (genes) variation accounted for by the extracted gene components. For example, Table 12a summarizes variation explained by the constructed PLS components and PCs for gene set A ($p = 1554$). Note that three ($K = 3$) PLS components explained 93.8% of response variation and about 58.7% of predictor variability, compared to three PCs explaining 55.3 and 60.4% respectively. Thus, the total gene variability accounted by PCs and PLS components are similar, but PCs were unable to account for much of the leukemia class variation. Note also that the first PC accounted for 44.5% of total predictor (gene) variability but it accounted for

Table 12. Variability explained by PLS components and PCs. The number of components extracted is K

	K	Predictor		Response	
		Proportion	Cumulative proportion	Proportion	Cumulative proportion
(a) Gene set A					
PLS					
	1	26.4713	26.4713	50.0156	50.0156
	2	27.1942	53.6655	26.0319	76.0475
	3	5.0562	58.7217	17.7467	93.7942
PC					
	1	44.4644	44.4644	2.3520	2.3520
	2	10.5679	55.0323	38.2658	40.6177
	3	5.3219	60.3542	14.6836	55.3014
(b) 50 Predictive genes					
PLS					
	1	46.2635	46.2635	86.1931	86.1931
	2	14.7372	61.0006	3.4223	89.6154
	3	7.2307	68.2314	4.4394	94.0548
PC					
	1	46.3143	46.3143	84.9414	84.9414
	2	19.3407	65.6549	0.7407	85.6821
	3	5.3636	71.0185	0.1557	85.8377

only 2.4% of total response (leukemia class) variability. This is an indicator that it will poorly predict the leukemia classes, as it indeed did (Tables 10 and 11). Now consider the same analysis but with the 50 informative genes. This is given in Table 12b. This time, the first PC accounted for about 46.3% of predictor variability but also accounted for 84.9% of response (leukemia class) variation—this is a notable increase from 2.4 to 84.9%.

CONCLUSIONS AND DISCUSSIONS

We have introduced statistical analysis methods for the classification of tumors based on microarray gene expression data. The methodologies involve dimension reduction of the high p-dimensional gene expression space followed by logistic classification or QDA. We have also illustrated the methods' effectiveness in predicting normal and tumor samples as well as between two different tumor types. The samples varied from human tissue samples to cell lines generated from both one and two channels microarray systems, such as oligonucleotide and cDNA arrays. The methods were able to distinguish between normal and tumor samples as well as between two types of tumors from five different microarray data sets with high accuracy. Furthermore, these results hold under re-randomization studies. Finally, we have also illustrated a condition under which PLS components are superior to PCs in prediction.

The problem of distinguishing normal from tumor samples is an important one. Another problem of interest is in characterizing multiple types of tumors. A data set illustrating this multiple classification problem is the NCI60 data set, which contains nine types of tumors. The problem of multiple classification based on gene expression data is much more difficult than the problem of binary classification illustrated in this paper and is the topic of current research. The method of multivariate PLS (Höskuldsson, 1988; Garthwaite, 1994) could be of use for this problem.

The PLS method can be of use for gene expression analysis in other contexts as well. Predicting the expressions of a target gene based on the remaining mass of genes is one example. Here, PLS is used to reduce the dimension of the predictors and then multiple linear regression (or another prediction method for continuous response) is used to predict the expressions of the target gene. Quantifying the predicted gene expression values such that they are compatible with some clinical outcomes are of practical value.

A related problem which may benefit from PLS is the problem of assessing the relationship between cellular reaction to drug therapy and their gene expression pattern. For example, Scherf *et al.* (2000) assessed growth inhibition from tracking changes in total cellular protein (in cell lines) after drug treatment. Here, the response of cell lines to each drug treatment are the response variables. Associated with the cell lines are their gene expressions. Since the expression patterns are from those of untreated cell lines, Scherf *et al.* focused on the relationship between gene expression patterns of the cell lines and their sensitivity to drug therapy. This relationship can be studied via a direct application of univariate or multivariate PLS, which can handle the high dimensionality of the data.

A final example, in cancer research, is the prediction of patient survival times based on gene expressions. For example, Ross *et al.* (2000) compared patient survival duration with germinal center B-like DLBCL compared to those with activated B-like DLBCL using Kaplan–Meier survival curves (Kaplan and Meier, 1958). These groups were determined by gene expression analysis. A more general and useful approach is to model the observed survival (and censored) times as a function of the p gene expressions. A common tool widely used for this purpose is the proportional hazard regression proposed by Cox (1972). Again, straightforward application of this method is not possible since $N < p$. Hence, dimension reduction is needed, however, care is needed to address the observed censored times. Our preliminary studies indicate that PLS may be of use in this context as well.

ACKNOWLEDGEMENTS

We thank Michèl Schummer and Terrence Furey for providing the ovarian data set and for numerous helpful consultations during this project. We also thank Charles Drescher who provided the tumor samples to Michael Schummer. We thank Dan Notterman for providing the colon data set. We are grateful for helpful comments from two referees that greatly improved the paper. The research reported in this paper was supported by grants from the National Science Foundation (ACI 96-19020, and DMS 98-70172) and the National Institute of Environmental Health Sciences, National Institutes of Health (P43 ES04699).

APPENDIX

The following PLS algorithm is given in Höskuldsson (1988) and adopted in Garthwaite (1994). For details, see also Helland (1988) and Martens and Naes (1989).

```
1. FOR $k = 1$ to $d$ set $\mathbf{u}$ to first column of $\mathbf{Y}_{(k)}$ and DO:
2. $\mathbf{w} = \mathbf{X}'\mathbf{u}/(\mathbf{u}'\mathbf{u})$ and scale $\mathbf{w}$ to be of unit length.
3. $\mathbf{t} = \mathbf{Xw}$.
4. $\mathbf{c} = \mathbf{Y}'\mathbf{t}/(\mathbf{t}'\mathbf{t})$ and scale $\mathbf{c}$ to be of unit length.
5. $\mathbf{u} = \mathbf{Yc}$ and GO TO 6 IF convergence ELSE return to 2.
6. $\mathbf{p} = \mathbf{X}'\mathbf{t}/(\mathbf{t}'\mathbf{t})$.
7. $b = \mathbf{u}'\mathbf{t}/(\mathbf{t}'\mathbf{t})$.
8. Residual matrices: $\mathbf{X}_{(k+1)} = \mathbf{X}_{(k)} - \mathbf{tp}'$ and $\mathbf{Y}_{(k+1)} = \mathbf{Y}_{(k)} - b\mathbf{tc}'$ (with $\mathbf{X}_{(1)} = \mathbf{X}$, $\mathbf{Y}_{(1)} = \mathbf{Y}$).
9. END FOR
```

REFERENCES

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. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

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., 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.

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.

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

de Jong,S. (1993) SIMPLS: an alternative approach to partial least squares regression. *Chemometr. Intell. Lab. Syst.*, **18**, 251–263.

Flury,B. (1997) *A First Course in Multivariate Analysis*. Springer, New York.

Furey,T.S., Cristianini,N., Duffy,N., Bednarski,D.W., Schummer,M. and Haussler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

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

Garthwaite,P.H. (1994) An interpretation of partial least squares. *J. Am. Stat. Assoc.*, **89**, 122–127.

Geladi,P. and Kowalski,B.R. (1986) Partial least squares regression: a tutorial. *Anal. Chim. Acta*, **185**, 1–17.

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.

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

Helland,S. and Almøy,T. (1994) Comparison of prediction methods when only a few components are relevant. *J. Am. Stat. Assoc.*, **89**, 583–591.

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

Hosmer,D.W. and Lemeshow,S. (1989) *Applied Logistic Regression*. Wiley, New York.

Johnson,R.A. and Wichern,D.W. (1992) *Applied Multivariate Analysis*, 4th edn, Prentice-Hall, Englewood Cliffs, NJ.

Jolliffe,I.T. (1986) *Principal Component Analysis*. Springer, New York.

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

Lorber,A., Wangen,L.E. and Kowalski,B.R. (1987) A theoretical foundation for the PLS algorithm. *J. Chem.*, **1**, 19–31.

Mardia,K.V., Kent,J.T. and Bibby,J.M. (1979) *Multivariate Analysis*. Academic Press, London.

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

Massey,W.F. (1965) Principal components regression in exploratory statistical research. *J. Am. Stat. Assoc.*, **60**, 234–246.

McCullagh,P. and Nelder,J.A. (1989) *Generalized Linear Models*, 2nd edn, Chapman and Hall, London.

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. *Proc. Natl Acad. Sci. USA*, **96**, 9112–9217.

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.

Phatak,A., Reilly,P.M. and Penlidis,A. (1992) The geometry of 2-block partial least squares. *Commun. Stat. Theor. Meth.*, **21**, 1517–1553.

Press,S.J. (1982) *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*, 2nd edn, Krieger, Malabar, FL.

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., 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 Genet.*, **24**, 227–235.

Scherf,U., Ross,D.T., Waltham,M., Smith,L.H., Lee,J.K., Tanabe,L., Kohn,K.W., Reinhold,W.C., Myers,T.G., Andrews,D.T., Scudiero,D.A., Eisen,M.B., Sausville,E.A., Pommier,Y., Botstein,D., Brown,P.O. and Weinstein,J.N. (2000) A gene expression database for the molecular pharmacology of cancer. *Nature Genet.*, **24**, 236–244.

Stone,M. and Brooks,R.J. (1990) Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares, and principal components regression (with discussion). *J. R. Stat. Soc.,* B, **52**, 237–269.

Vapnik,V.N. (2000) *The Nature of Statistical Learning Theory*, 2nd edn, Springer, New York.