BIOINFORMATICS

Vol. 18 no. 9 2002
Pages 1216–1226

Multi-class cancer classification via partial least squares with gene expression profiles

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

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

Received on July 17, 2001; revised on March 14, 2002; accepted on March 25, 2002

ABSTRACT

Motivation: Discrimination between two classes such as normal and cancer samples and between two types of cancers based on gene expression profiles is an important problem which has practical implications as well as the potential to further our understanding of gene expression of various cancer cells. Classification or discrimination of more than two groups or classes (multi-class) is also needed. The need for multi-class discrimination methodologies is apparent in many microarray experiments where various cancer types are considered simultaneously.

Results: Thus, in this paper we present the extension to the classification methodology proposed earlier (Nguyen and Rocke, 2002b, *Bioinformatics*, **18**, 39–50) to classify cancer samples from multiple classes. The methodologies proposed in this paper are applied to four gene expression data sets with multiple classes: (a) a hereditary breast cancer data set with (1) *BRCA1*-mutation, (2) *BRCA2*-mutation and (3) sporadic breast cancer samples, (b) an acute leukemia data set with (1) acute myeloid leukemia (AML), (2) T-cell acute lymphoblastic leukemia (T-ALL) and (3) B-cell acute lymphoblastic leukemia (B-ALL) samples, (c) a lymphoma data set with (1) diffuse large B-cell lymphoma (DLBCL), (2) B-cell chronic lymphocytic leukemia (BCLL) and (3) follicular lymphoma (FL) samples, and (d) the NCI60 data set with cell lines derived from cancers of various sites of origin. In addition, we evaluated the classification algorithms and examined the variability of the error rates using simulations based on randomization of the real data sets. We note that there are other methods for addressing multi-class prediction recently and our approach is along the line of (Nguyen and Rocke, 2002b, *Bioinformatics*, **18**, 39–50).

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

INTRODUCTION

Since the introduction of DNA microarray technology to quantitate thousands of gene expressions simultaneously (Schena *et al.*, 1995; Lockhart *et al.*, 1996), there have been increasing activities in the area of cancer classification. For example, Golub *et al.* (1999) used a weighted voting scheme for the molecular classification of acute leukemia. Alon *et al.* (1999) used a clustering technique based on the deterministic-annealing algorithm to classify cancer and normal colon tissues. Scherf *et al.* (2000) used average-linkage clustering for tumor tissues from various sites of origin. Support Vector Machines (SVM) were applied to the classification of tumor and normal ovarian tissues by Furey *et al.* (2000). The use of gene expression profiles to distinguish between negative and positive for *BRCA1* and *BRCA2* mutation in hereditary breast cancer was described by Hedenfalk *et al.* (2001). Nguyen and Rocke (2002a,b) proposed binary classification methods that combine the use of partial least squares (PLS) as a dimension reduction method together with Logistic Discrimination (LD) or Quadratic Discriminant Analysis (QDA).

The need for multi-class classification methods is apparent in the various microarray gene expression studies described above. For example, Hedenfalk *et al.* (2001) sought to classify primary breast cancers from three classes: *BRCA1* mutation, *BRCA2* mutation, and sporadic cancer cases based on the observed gene expression profiles. Hedenfalk and co-workers employed the ‘one versus all’ strategy of classification. Multi-class cancer prediction using gene expression data is an important problem and, recently, various methods have appeared in the literature.

In this paper we also address this problem of multi-class cancer classification using multivariate partial least squares (MPLS) dimension reduction together with PD or QDA. Microarray experiments are characterized by many measured variables (genes), p, on only a relatively few observations or samples N. Hence, the need for dimension reduction methods. We first describe the methodologies consisting of the dimension reduction method (MPLS) and classification methods (PD and QDA) in the Section **Dimension Reduction Methods** and the Section **Classification Methods**. A preliminary gene screening method based on pairwise comparison and the analysis of variance is also described. The classification methods were applied to four cancer gene expression data sets (see **Results**). We evaluated the classification algorithms and examined the variability of the error rates. Results were compared to other classifiers, namely Diagonal Quadratic Discriminant Analysis (DQDA), Diagonal Linear Discriminant Analysis (DLDA), and classification trees (Dudoit *et al.*, 2000; Zhang *et al.*, 2001). We also examined the effect of varying K, the PLS (and PCA) dimension, on classification performance. Most technical details are deferred to the Supplemental Appendix, which can be found at http://stat.tamu.edu/∼dnguyen/supplemental.html.

DIMENSION REDUCTION METHODS

Suppose that a qualitative response variable y takes on a finite number of unordered values, say $0, 1, \dots, G$ often referred to as classes (or groups). That is, y indicates the cancer class $0, 1, \dots,$ or G, for instance. The problem of multi-class cancer classification is to predict the cancer class based on a vector of gene expression values $\mathbf{x} = (x_1, x_2, \dots, x_p)'$. Most classification methods, such as classical discrimination analysis or polychotomous discrimination are based on the requirement that there are more observations (N) than there are explanatory variables or genes (p). One strategy to approach the problem of classification when $N < p$ is to reduce the dimension of the gene space from p to say K, where $K \ll N$. This is done by constructing K gene components and then classifying the cancers based on the constructed K gene components.

The dimension reduction process is illustrated in Figure 1 using the NCI60 data set consisting of cell lines derived from cancers of various origins. For illustration, we have reduced a gene expression matrix, $\mathbf{X}$ of size $N \times p = 35 \times 167$, to three gene components, t_1, t_2, t_3, using multivariate PLS (see next section). It can be seen from the 3-dimensional plot (Figure 1, bottom) that the three MPLS gene components separate the five cancer classes well (leukemia=*, colon=o, melanoma=+, renal=× and CNNS=$\diamond$).

Multivariate PLS

In the well known method of Principal Components Analysis (PCA), the goal is to extract gene components sequentially which maximize the total predictor (gene) variability, irrespective of how well the constructed gene components predict cancer classes. In contrast to PCA, univariate PLS (orthogonal) components are constructed to maximize the sample covariance between the response values ($\mathbf{y}$) and the linear combination of the predictor or gene expression values ($\mathbf{X}$) (see Supplemental Appendix B or Nguyen and Rocke (2002b) for details on PCA and univariate PLS). When there are more than one response variable, the objective criterion for maximization (under orthogonality constraints) in multivariate PLS is

$$\mathrm{cov}^2(\mathbf{Xw}, \mathbf{Yc}) \tag{1}$$

where $\mathbf{w}$ and $\mathbf{c}$ are unit vectors. The MPLS components are denoted by $\mathbf{t}_k$ and are linear combinations of the gene expression values ($\mathbf{X}$) with coefficients given by $\mathbf{w}_k$ (satisfying the maximization criterion (1)). The PLS algorithm to obtain $\mathbf{w}$ (and $\mathbf{c}$) is simple and fast. The algorithm can be found in Höskuldsson (1988), Garthwaite (1994) and Helland (1988).

The response matrix $\mathbf{Y}$ in (1) consists of continuous response variables, which is the setting MPLS was designed for. However, in the current context, we have a qualitative response variable y consisting of classes $0, 1, \dots, G$, namely, cancer type 0 through cancer type G. We need to convert or recode the response information indicating cancer class, namely $\mathbf{y}$, into a response matrix $\mathbf{Y}$. To do this with the $G + 1$ cancer classes we created G 'design variables' representation (or 'reference cell coding') of $\mathbf{y}$. That is, we define the $N \times G$ response matrix $\mathbf{Y}$ with elements $y_{ik} = I(y_i = k)$ for $i = 1, \dots, N$ and $k = 1, \dots, G$. We have used $I(A)$ to denote the indicator function for event A, so that $I(A) = 1$ if A is true and it is 0 otherwise. Other strategies for constructing $\mathbf{Y}$ are possible.

Thus, $K \ll N$ multivariate PLS gene components, $\mathbf{t}_1, \dots, \mathbf{t}_K$, are extracted according to (1) using the original gene expression matrix, $\mathbf{X}$, and the response matrix, $\mathbf{Y}$, constructed from the vector of cancer class indicator $\mathbf{y}$.

CLASSIFICATION METHODS

In this section we describe two classification methods, which can be applied to make class prediction after dimension reduction. Polychotomous Discrimination (PD) is a generalization of logistic discrimination when there are more than two classes. QDA works for two or more classes. We also describe in this section a preliminary ranking and selection of the large number of genes used for the analyses.

Polychotomous discrimination

Assume that the qualitative response variable y can take on finite values, $y = k$, $k \in \{0, 1, \dots, G\} \equiv \mathcal{O}$. The distribution of y depends on predictors $x_1, \dots, x_p$. For example, the kth cancer type ($y = k$) depends on the p gene expression levels $x_1, \dots, x_p$ in a given experiment. The response variable y is a G-valued random variable and assume that $\pi(k \mid \mathbf{x}) = P(y = k \mid \mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X} \subset \mathcal{R}^{p+1}$ and $k \in \mathcal{O}$. For convenience we define the notation

$$g_k(\mathbf{x}) = \log\left(\frac{\pi(k \mid \mathbf{x})}{\pi(0 \mid \mathbf{x})}\right), \qquad \text{for } \mathbf{x} \in \mathcal{X} \text{ and } k \in \mathcal{O}. \tag{2}$$

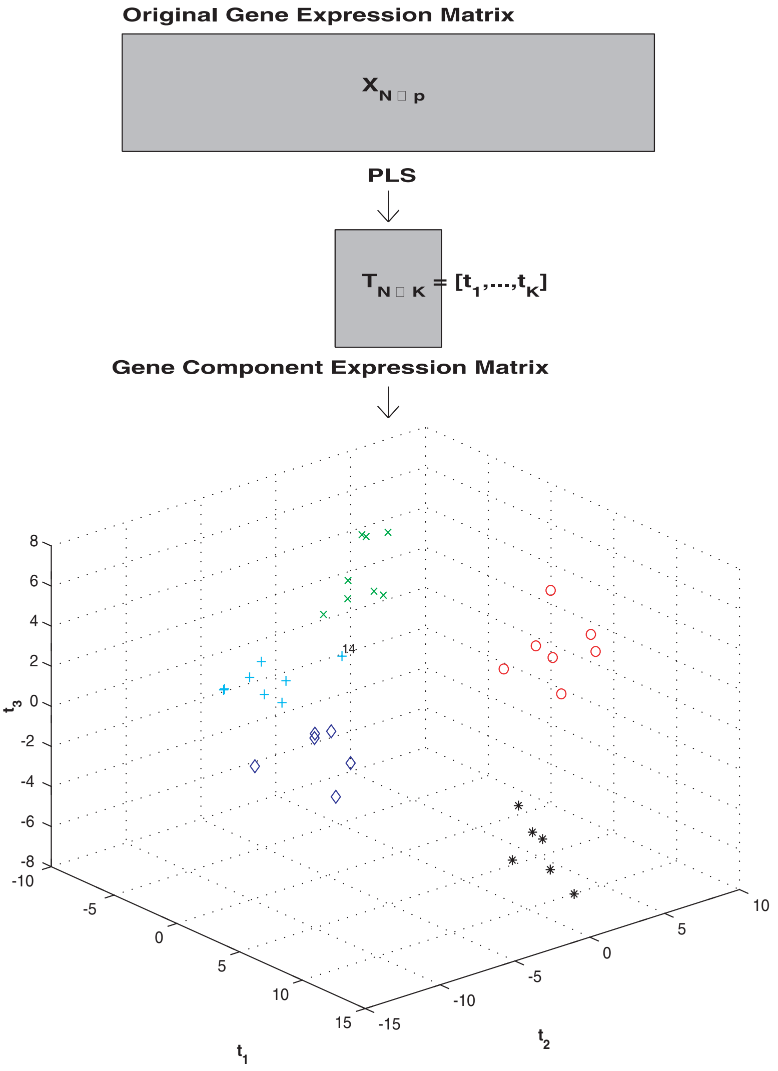

Fig. 1. Illustration of dimension reduction for NCI60 data. For the NCI60 data, the 'original' gene expression data set used here is $\mathbf{X}_{35\times 167}$ and $K = 3$ PLS gene components are constructed giving $T_{35\times 3} = [t_1, t_2, t_3]$. The 3-dimensional PLS gene components plot illustrate the separability of the cancer classes: leukemia=*, colon=o, melanoma=+, renal=× and CNNS=◇.

This is the log of the ratio of the probability of a sample with gene expression profile $\mathbf{x}$ being of cancer type k relative to cancer type 0. Often this quantity $(g_k(\mathbf{x}))$ is modelled as a linear function of the p gene expressions, $\mathbf{x}$,

$$g_k(\mathbf{x}) = \log\left(\frac{\pi(k \mid \mathbf{x})}{\pi(0 \mid \mathbf{x})}\right) = \beta_{k0} + \beta_{k1}x_1 + \beta_{k2}x_2 + \cdots + \beta_{kp}x_p = \mathbf{x}'\boldsymbol{\beta}_k. \qquad (3)$$

Thus, the conditional class probabilities are

$$\pi(k \mid \mathbf{x}) = \frac{\exp(g_k(\mathbf{x}))}{1 + \sum_{k=0}^{K}\exp(g_k(\mathbf{x}))}, \qquad \mathbf{x} \in \mathcal{X} \text{ and } k \in \mathcal{O}. \qquad (4)$$

This is the probability that a sample with gene expression profile $\mathbf{x}$ is of cancer class k. We take (4) as the polychotomous regression model and note that $\pi(k \mid \mathbf{x}) \equiv \pi(k \mid \mathbf{x}; \boldsymbol{\beta})$ is a function of $\nu = G(p+1)$ parameters $\boldsymbol{\beta}' = (\boldsymbol{\beta}'_1, \ldots, \boldsymbol{\beta}'_K)$ $(\boldsymbol{\beta} \in \mathcal{R}^{G(p+1)})$, with $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \ldots, \beta_{kp})'$.

Estimate of $\boldsymbol{\beta}$ is obtained by maximum likelihood estimation (MLE) and it is described in the Supplemental Appendix A. The MLE of $\boldsymbol{\beta}$ is denoted $\hat{\boldsymbol{\beta}}$ and it can be obtained (if it exists) when there are more samples than there are parameters, i.e. when $N > \nu = G(p+1)$. Thus, after dimension reduction we can use PD by replacing the full gene profile $\mathbf{x}$ by the corresponding gene component profile in the reduced space obtained by MPLS or PCA.

From the estimated coefficient vector $\hat{\boldsymbol{\beta}}$, the estimated conditional class probabilities $\hat{\pi}(k \mid \mathbf{x})$ $k = 0, \ldots, K$ can be obtained by substituting $\hat{\boldsymbol{\beta}}$ into (4). A given sample with gene expression profile $\mathbf{x}$ is then predicted to be of cancer class k with maximum estimated conditional class probability $\hat{\pi}(k \mid \mathbf{x})$. That is, we classify a sample as a cancer of class k if the estimated probability of observing a cancer of this class given the gene expression profile, $\mathbf{x}$, is higher than the probability of observing any other class of cancer given the *same* gene expression profile. See Hosmer and Lemeshow (1989), Kooperberg *et al.* (1997) and Albert and Anderson (1984) for details on polychotomous regression.

Quadratic discriminant analysis

Another classification method that can be used after dimension reduction is quadratic discriminant analysis (QDA). QDA is based on the classical multivariate normal model for each class: $\mathbf{x} \mid y = k \sim \mathrm{N}_p(\boldsymbol{\Sigma}_k, \boldsymbol{\mu}_k)$, $\mathbf{x} \in \mathcal{R}^p$ and $k = 0, 1, \ldots, G$ (for binary classification, $G = 1$). The (optimal) classification regions are $R_k = \{\mathbf{x} \in \mathcal{R}^p : p_k f_k(\mathbf{x}) > p_j f_j(\mathbf{x}), j \neq k\}$, where f_k is the pdf of $\mathbf{x} \mid y = k$ given above and $p_k = P(y = k)$. The posterior probability of membership in class k is $\pi_k = P(y = k \mid \mathbf{x}) = \exp[q_k(\mathbf{x})] / \sum_{i=0}^{K} \exp[q_i(\mathbf{x})]$. As in PD, the full gene profile, $\mathbf{x}$, is replaced by the corresponding gene component profile in the reduced space obtained from MPLS 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).

Preliminary gene screening

For any given classification problem we may also select the genes which are 'good' predictors of the cancer classes. In the binary case, preliminary selection and ranking of the genes based on t-scores worked well. For more than two classes, we ranked and selected the genes for multi-class prediction as follows. We compared all $\binom{G+1}{2}$ pairwise (absolute) mean differences, $|\bar{x}_k - \bar{x}_{k'}|$ (for $k \neq k', k, k' \in \mathcal{O}$), to a critical score

$$t \sqrt{MS_E \left(\frac{1}{n_k} + \frac{1}{n_{k'}} \right)}. \tag{5}$$

MS_E (mean squared error) is the estimate of variability from the analysis of variance (ANOVA) model with one factor and $G+1$ (cancers) groups and t is the $t_{\alpha/2, N-(G+1)}$ value of the t-distribution. Each gene ($j = 1, \ldots, p$) is ranked according to the number of times the pairwise absolute mean difference exceeded the critical score.

RESULTS

We summarize the results of multi-class prediction using the proposed methodologies. The results are described separately for each of four gene expression data sets consisting of human cancer samples; (1) hereditary breast cancer; (2) NCI60 cell lines derived from cancers of various origins; (3) lymphoma; and (4) acute leukemia.

Hereditary breast cancer data

Hedenfalk *et al.* (2001) studied gene expression patterns in hereditary breast cancer (HBC). Many cases HBC are attributed to individuals with a mutant *BRCA1* or *BRCA2* gene. Breast cancers with *BRCA1* or *BRCA2* mutation have pathologically distinct features (e.g. high mitotic index, noninfiltrating smooth edges and lymphocytic infiltrate, grade level; see Hedenfalk *et al.* (2001, p. 539–540)). Furthermore, distinctive features of *BRCA1* and *BRCA2* cancers are used to distinguish them from *sporadic* cases of breast cancers. Previous experimental evidence indicates that generally cancers with *BRCA1* mutation lacks both estrogen and progesterone receptors but these hormones receptors are present in those with *BRCA2* mutations (Karp *et al.*, 1997; Johannsson *et al.*, 1997; Loman *et al.*, 1998; Verhoog *et al.*, 1998). Also, functional *BRCA1* and *BRCA2* proteins are involved in the repairing of damaged DNA, hence, cells with the mutant genes have decreased ability to participate in DNA repair.

Hedenfalk *et al.* (2001) monitored the global expression patterns of 7 cancers with *BRCA1* mutation, 8 with *BRCA2* mutation, and 7 sporadic cases of primary breast cancers using cDNA microarrays. There were 6512 cDNA used which represent 5361 unique genes. Selected for analysis were $p = 3226$ genes and these are available publicly. We considered multi-class classification methods to predict each sample as a breast cancer with *BRCA1* mutation, *BRCA2* mutation or as sporadic breast cancer based on the observed gene expression profiles.

Preliminary ranking and selection of the genes for analysis was carried out as described in the Section **Classification methods**. The number of genes with 0, 1, 2 or 3 pairwise absolute mean differences exceeding the critical score is 2269, 541, 405, or 11 respectively. Thus, of the 3226 genes 2269 showed no pairwise absolute mean difference and only 11 genes showed all 3 pairwise differences. The subset of genes selected for analysis is denoted by p^*. We considered two analyses based on $p^* = 11$ (genes with all 3 pairwise differences) and $p^* = 416$ (genes with at least 2 pairwise differences).

We applied multivariate PLS and PCA to reduce the dimension from $p^* = 11$ or $p^* = 416$ to $K = 3$ MPLS gene components and 3 PCs respectively. All analyses were based on standardized log expression ratios. Prediction of each of the $N = 22$ samples as

BRCA1, *BRCA2*, or as sporadic was carried out using PD and QDA based on the constructed gene components. Prediction results were based on leave-out-one cross-validation (LOOCV). The results are summarized in Table 1A0 (left). PD using MPLS gene components correctly classified all 22 samples using either $p^* = 11$ or $p^* = 416$ genes. For $p^* = 416$ MPLS components in QDA and PCs in PD also correctly predicted all samples into their cancer classes. For this data set MPLS gene components performed better than PCs in both PD and QDA.

An interesting sporadic sample misclassified by PCs using QDA ($p^* = 416$) is sample 20. When classifying all samples as either *BRCA1*-mutation-positive versus negative (binary classification) Hedenfalk *et al.* misclassified this sporadic sample as having a *BRCA1* mutation. We obtained similar results using the method reported in (Nguyen and Rocke, 2002b) for binary classification. (results not shown here). Studies have suggested that abnormal methylation of the promoter region is indicative of inactivation of the *BRCA1* gene (Catteau *et al.*, 1999; Esteller *et al.*, 2000); therefore, such samples show similar phenotypes as samples with *BRCA1* mutation. Thus, such samples are potential candidates for misclassification when using data at the molecular level. However, expression patterns (or lack thereof) of an inactivated gene is not identical to that of a mutated gene.

NCI60 data: cell lines derived from various cancer sites

This data set is from (Ross *et al.*, 2000) and Scherf *et al.* (2000). For illustration of the multi-class cancer classification methods we considered classification of 6 cancer types: leukemia ($n_1 = 6$), colon ($n_2 = 7$), melanoma ($n_3 = 8$), renal ($n_4 = 8$), and CNS ($n_5 = 6$). We used a subset of 1376 genes and 40 individually assessed targets ($p = 1416$) analyzed by Scherf *et al.* (2000) relative to drug activities of the same cell lines, which is publicly available. For this data set there are some missing gene expression values. Genes with 2 or fewer missing values (out of 35) were included for analysis by replacing the (1 or 2) missing values with the median of the gene's expression. This resulted in a subset of 1299 genes which we used for analysis.

Applying the preliminary gene ranking procedure resulted in the following ranking of the genes: 167 (0), 76 (1), 115 (2), 119 (3), 266 (4), 148 (5), 241 (6), 109 (7), 53 (8), 5 (9), 0 (10). That is 167 genes showed no pairwise absolute mean difference, 76 genes showed 1 pairwise difference, etc. We pooled all genes showing at least 8 pairwise differences ($p^* = 58$) and also all genes showing at least 7 pairwise differences ($p^* = 167$) for analysis. As before dimension reduction via MPLS and PCA and classification using PD and QDA were then used to predict the cancer class of each sample. The classification results based on LOOCV are displayed in Table 2A0 (left). With $p^* = 58$ genes 3 MPLS gene components and PCs correctly classified all cancer classes using PD. Three MPLS gene components constructed from $p^* = 167$ genes also correctly classified all cancer classes with PD. These components are plotted in Figure 1. QDA did not perform as well with one misclassification when using MPLS gene components (both $p^* = 58$ and 167). This commonly misclassified sample (#14), a melanoma sample, is marked in Supplemental Figure 1 (bottom) and it can be seen that this sample do not group with the other melanoma samples.

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: Diffuse Large B-Cell lymphoma (DLBCLL; $n_1 = 45$), B-Cell Lymphocytic Leukemia (BCLL; $n_2 = 29$) and Follicular Lymphoma (FL; $n_3 = 9$). We analyzed the standardized log relative intensity ratios, namely the log(Cy5/Cy3) values. We consider multi-class cancer classification of all 3 classes simultaneously here. We analyzed a subset of the data consisting of $p = 4151$ genes. Preliminary ranking resulted in 2168 genes with 0 pairwise absolute mean difference, 1003 with 1, 896 with 2, and 84 with all 3 pairwise absolute mean expression differences.

Using LOOCV, each sample was predicted to be DLBCL, BCLL, or FL based on 3 gene components constructed from $p^* = 84$ genes (with all 3 pairwise mean differences) and $p^* = 980$ genes (with at least 2 pairwise mean differences). The results are given in Table 3A0 (left) For PD MPLS gene components performed better than PCs with two misclassifications (97.6%). However, for this data set QDA performed best with only one misclassification (98.8%). A BCLL sample (#51) was misclassified by all (eight combinations) of the methods. MPLS gene components performed better than PCs for $p^* = 84$ and the results are equal for $p^* = 980$.

Acute 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 Acute Lymphoblastic Leukemia (ALL) and 11 Acute Myeloid Leukemia (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). It has been noted that global expression patterns of T-cell ALL (T-ALL) and B-cell ALL (B-ALL) are distinct and can be used to differentiate between the two sub-classes of ALL (Golub *et al.*, 1999). Thus, for multi-class cancer discrimination we pooled the two data sets to obtain $N =$

Table 1. Hereditary breast cancer data. $N = 22$, $n_1 = 7$ (*BRCA1*), $n_2 = 8$ (*BRCA2*), and $n_3 = 7$ (sporadic). Given (left side) are the proportion of misclassification out of $N = 22$ samples from classification studies **A0**, **A1**, and **A2**. Note that in study **A2** the genes are reselected N times, each time with one sample left out, so under the column p^* the numbers given are the min–mean–max of the number of genes over the N reselections (also, under column p^*, in parentheses, is the number of pairwise absolute mean differences). For comparison classification studies **A0**, **A1**, and **A2** were repeated (right side) using DQDA and DLDA from Dudoit et al. (2000). Note that DQDA and DLDA are special cases of QDA with $\mathbf{\Sigma}_g = \mathrm{diag}\{\sigma^2_{g1}, \ldots, \sigma^2_{gp}\}$ for $g = 0, 1, \ldots, G$ and $\mathbf{\Sigma}_g = \mathrm{diag}\{\sigma^2_1, \ldots, \sigma^2_p\}$ (not depending on cancer class g) respectively. The samples misclassified from study **A0**, with superscript 1, 2 and s indicating *BRCA1*, *BRCA2* and sporadic respectively, are given at the bottom of the table.

p^*	PD		QDA		DQDA		DLDA	
	MPLS	PCA	MPLS	PCA	MPLS	PCA	MPLS	PCA
A0								
11 (3)	0.000	0.046	0.091	0.136	0.046	0.000	0.046	0.000
416 (≥ 2)	0.000	0.000	0.000	0.091	0.000	0.000	0.000	0.000
A1								
11 (3)	0.136	0.000	0.227	0.591	0.091	0.591	0.046	0.182
416 (≥ 2)	0.000	0.000	0.091	0.555	0.000	0.591	0.000	0.000
A2								
8-11-15 (3)	0.409	0.318	0.364	0.591	0.364	0.500	0.409	0.273
343-391-438 (≥ 2)	0.318	0.364	0.318	0.500	0.409	0.727	0.273	0.272

Samples misclassified from **A0**.

$p^* = 11$		$p^* = 416$	
MPLS–PD		MPLS–PD	
PCA–PD	(#16^s)	PCA–PD	
MPLS–QDA	(#2^1, 15^2)	MPLS–QDA	
PCA–QDA	(#2^1, 13^2, 21^s)	PCA–QDA	(#16^s, 20^s)

Table 2. NCI60 data: 5 cancer classes. $N = 35$, $n_1 = 6$ (leukemia), $n_2 = 7$ (colon), $n_3 = 8$ (melanoma), $n_4 = 8$ (renal) and $n_5 = 6$ (CNS). For details see Figure 1 caption. The samples misclassified from studies **A0**, with superscript le, co, me, re, and cn indicating leukemia, colon, melanoma, renal, and CNS respectively are given at the bottom of the table.

p^*	PD		QDA		DQDA		DLDA	
	MPLS	PCA	MPLS	PCA	MPLS	PCA	MPLS	PCA
A0								
58 (≥ 8)	0.000	0.086	0.029	0.143	0.029	0.057	0.029	0.057
167 (≥ 7)	0.000	0.000	0.029	0.086	0.000	0.029	0.029	0.029
A1								
58 (≥ 8)	0.000	0.229	0.086	0.543	0.029	0.571	0.029	0.057
167 (≥ 7)	0.029	0.029	0.057	0.543	0.029	0.457	0.029	0.057
A2								
41-54-69 (≥ 8)	0.429	0.229	0.257	0.514	0.171	0.486	0.143	0.114
148-159-189 (≥ 7)	0.257	0.057	0.200	0.400	0.114	0.400	0.086	0.057

Samples misclassified from **A0**.

$p^* = 58$		$p^* = 167$	
MPLS–PD		MPLS–PD	
PCA–PD	(#29^{re}, 31^{cn}, 34^{cn})	PCA–PD	
MPLS–QDA	(#14^{me})	MPLS–QDA	(#14^{me})
PCA–QDA	(#14^{me}, 26^{re}, 29^{re}, 31^{cn}, 34^{cn})	PCA–QDA	(#1^{le}, 14^{me}, 30^{cn})

72 samples with three cancer classes: (1) AML ($n_1 = 25$), (2) B-ALL ($n_2 = 38$) and (3) T-ALL ($n_3 = 9$).

We log transformed the gene expressions to have mean zero and standard deviation one across samples. For

Table 3. Lymphoma data. $N = 83$, $n_1 = 45$ (DLBCL), $n_2 = 29$ (BCLL), and $n_3 = 9$ (FL). For details see Figure 1 caption. The samples misclassified from studies **A0**, with superscript D, B and F indicating DLBCL, BCLL and FL respectively, are given at the bottom of the table.

p^*	PD		QDA		DQDA		DLDA	
	MPLS	PCA	MPLS	PCA	MPLS	PCA	MPLS	PCA
A0								
84^+ (3)	0.024	0.060	0.036	0.072	0.048	0.060	0.048	0.048
980 (≥ 2)	0.048	0.048	0.012	0.012	0.012	0.048	0.036	0.024
A1								
84^+ (3)	0.036	0.060	0.036	0.470	0.048	0.386	0.048	0.193
980 (≥ 2)	0.048	0.145	0.012	0.554	0.012	0.265	0.036	0.145
A2								
70-84-112 (3)	0.024	0.084	0.048	0.458	0.060	0.422	0.072	0.205
878-971-1168 (≥ 2)	0.072	0.169	0.024	0.566	0.024	0.265	0.036	0.169

Samples misclassified from **A0**.

$p^* = 84$		$p^* = 980$	
MPLS–PD	(#9^D, 51^B)	MPLS–PD	(#9^D, 32^D, 48^B, 51^B)
PCA–PD	(#9^D, 11^D, 18^D, 51^D, 55^D)	PCA–PD	(#9^D, 32^D, 48^B, 51^B)
MPLS–QDA	(#5^D, 11^D, 51^B)	MPLS–QDA	(#51^B)
PCA–QDA	(#9^D, 11^D, 18^D, 51^B, 55^B, 75^F)	PCA–QDA	(#51^B)

$^+$ Model without intercept.

the subsequent analyses we used a subset of $p = 3490$ genes. Preliminary ranking resulted in 1945 genes with 0 pairwise absolute mean difference, 732 with 1719 with 2, and 84 with all 3 pairwise absolute mean expression differences. As before, using LOOCV, each sample was predicted to be AML, B-ALL, or T-ALL based on 3 gene components constructed from $p^* = 94$ genes (with all 3 pairwise mean differences) and $p^* = 813$ genes (with at least 2 pairwise mean differences). The results are given in Table 4A0 (left) Classification methods compared similarly as for the lymphoma data set. Best classification results come from QDA with MPLS components constructed from $p^* = 813$ genes (all correct) and from $p^* = 94$ genes (1 incorrect). In all eight analyses combined there were 4 samples which were misclassified: two B-ALL (# 12, 17), one AML (#66), and one T-ALL (#67).

Assessment of classification algorithm

The classification results given in Tables 1A0–4A0 (left) is based on the following classification algorithm:

Algorithm A0

1. Select Genes: Select a set, $\mathcal{S}$, of p^* genes as described by (5) giving an expression matrix, $\mathbf{X}$, of size $N \times p^*$.
2. Dimension Reduction: Fit PLS (or PCA) to obtain PLS gene components matrix, $\mathbf{T}$, of size $N \times K$.
3. Classification/Prediction: Classification is based on LOOCV.

```
FOR $i = 1$ to $N$ DO
    Leave out sample (row) $i$ of $\mathbf{T}$. Fit classifier to the
    remaining $N - 1$ samples and use the fitted
    classifier to predict left out sample $i$.
END
```

Note that for a given expression matrix, $\mathbf{X}$, steps 1 (gene selection) and 2 (dimension reduction) are *fixed* with respect to LOOCV in algorithm A0. Thus, the effect of gene selection and dimension reduction on the classification can not be assessed. Based on a large simulation study of algorithm A0, using randomizations of the real data sets (BRCA, NCI60, Lymphoma, and Leukemia), the classification error rates were more optimistic than the expected error rates (results not shown). The potential sources of this problem may be with the gene selection (step 1) and/or the dimension reduction (step 2).

To assess the effects of gene selection and dimension reduction on classification we considered two modifications to algorithm A0. The first modification, given as algorithm A1 below, is to assess the affects of the dimension reduction step. In algorithm A1, the gene selection step is still fixed, but now the dimension reduction as well as the classifier is refitted N times, one for each sample left out.

Algorithm A1

1. Select Genes: Select a set, $\mathcal{S}$, of p^* genes as described by (5) giving an expression matrix, $\mathbf{X}$, of size $N \times p^*$.

Table 4. Acute leukemia data. $N = 72, n_1 = 25$ (AML), $n_2 = 38$ (B-ALL), and $n_3 = 9$ (T-ALL). For details see Figure 1 caption. Tthe samples misclassified in study **A0**, with superscript A, B and T indicating AML, B-ALL and T-ALL respectively, are given at the bottom of the table.

p^*	PD		QDA		DQDA		DLDA	
	MPLS	PCA	MPLS	PCA	MPLS	PCA	MPLS	PCA
A0								
94 (3)	0.056	0.056	0.014	0.042	0.028	0.042	0.028	0.042
813 (≥ 2)	0.042	0.056	0.000	0.028	0.014	0.042	0.014	0.042
A1								
94 (3)	0.056	0.111	0.028	0.222	0.028	0.181	0.042	0.042
813 (≥ 2)	0.056	0.153	0.028	0.417	0.028	0.319	0.042	0.083
A2								
69-90-100 (3)	0.056	0.111	0.056	0.236	0.056	0.264	0.056	0.111
710-804-850 (≥ 2)	0.056	0.167	0.042	0.431	0.042	0.306	0.056	0.111

Samples misclassified from **A0**.

$p^* = 94$		$p^* = 813$	
MPLS–PD	(#$12^B, 17^B, 66^A, 67^T$)	MPLS–PD	($17^B, 66^A, 67^T$)
PCA–PD	(#$12^B, 17^B, 66^A, 67^T$)	PCA–PD	(#$12^B, 17^B, 66^A, 67^T$)
MPLS–QDA	(#12^B)	MPLS–QDA	
PCA–QDA	(#$12^B, 66^A, 67^T$)	PCA–QDA	(#$12^B, 67^T$)

$^+$ Model without intercept.

```
FOR $i = 1$ to $N$ DO
    Leave out sample (row) $i$ of expression matrix $\mathbf{X}$,
    say $\mathbf{X}_{-i}$.
2.  Dimension Reduction: Fit PLS (or PCA) using $\mathbf{X}_{-i}$
    to obtain PLS gene components matrix, $\mathbf{T}_{-i}$.
3.  Classification/Prediction: Fit classifier to the
    remaining $N - 1$ samples, i.e. using $\mathbf{T}_{-i}$. Use the
    fitted classifier to predict left out sample $i$.
END
```

However, the choice of the gene set used for classification can have a large affect on classification. Thus, the second modification, given as algorithm A2 below, involves reselecting the gene set for classification each time a sample is left out.

Algorithm A2

```
FOR $i = 1$ to $N$ DO
    Leave out sample (row) $i$ of original expression matrix
    $\mathbf{X}^O$ $(N \times p)$.
1.  Select Genes: Select a set, $\mathcal{S}_{-i}$, of $p^*$ genes as
    described by (5) giving an expression matrix,
    $\mathbf{X}_{-i}$, of size $N - 1 \times p^*$.
2.  Dimension Reduction: Fit PLS (or PCA) using $\mathbf{X}_{-i}$
    to obtain PLS gene components matrix, $\mathbf{T}_{-i}$.
3.  Classification/Prediction: Fit classifier to the
    remaining $N - 1$ samples, i.e. using $\mathbf{T}_{-i}$. Use the
    fitted classifier to predict left out sample $i$.
END
```

Classification was repeated on each of the four data sets using algorithm A1 and A2 to assess the effects of dimension reduction and gene selection. The results are given in Tables 1–4 under A1 and A2 (left). For the breast cancer (BRCA) and the NCI60 data (both small sample sizes) the effect of the set of genes selected (A2) for classification is large. The percentage of misclassification varies from 31.8–40.9% for the BRCA data and 5.7–42.9% for the NCI60 data. Refitting the dimension reduction (A1) had less effect, with classification error rates relatively low for all four data sets, with the exception of QDA using principal components (QDA–PCA) where the classification error rates are high whether the genes were reselected (A2) or not (A1). However, the effect the gene choice appeared to be diminished for the lymphoma and leukemia data (both with larger sample sizes) where classification error rates from A1 and A2 appear to be quite similar and relatively low. In fact the error rates from A1 and A2 are similar to those from A0 for the leukemia and lymphoma data, again, with the exception of QDA–PCA.

Evaluation of classification results: randomization studies

The reliability of classification results is an important issue and we further address this issue in relation to the small sample size associated with microarray data, especially in cancer microarray data. As can be seen from the previous section, for the BRCA and NCI60 data, both with small sample sizes, the variation of the observed classification

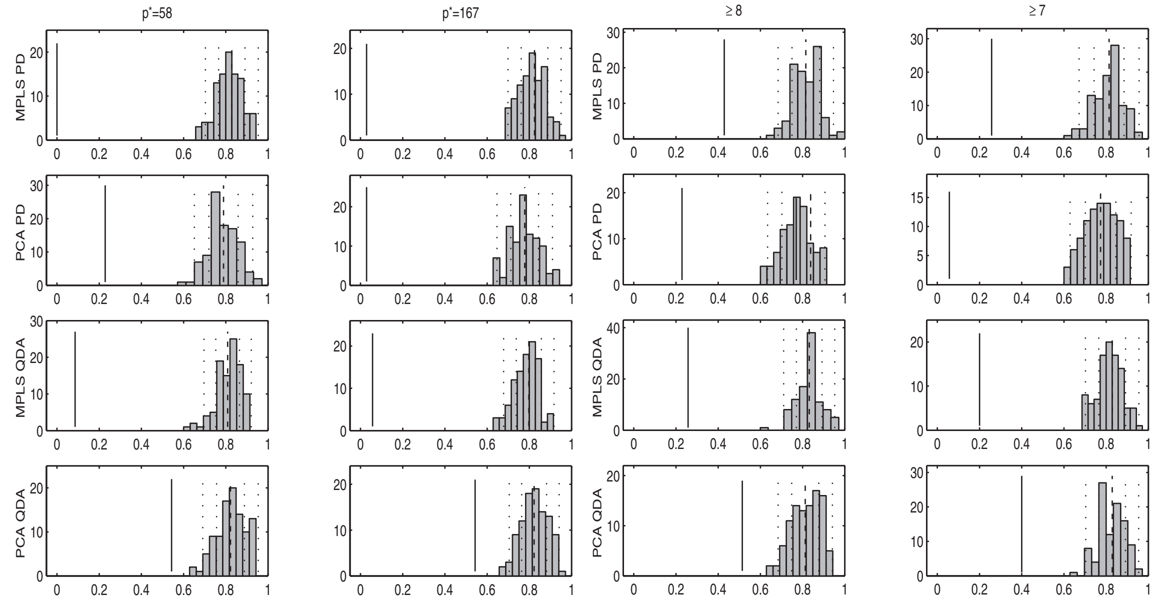

Fig. 2. Classification of NCI60 data under randomization—A1 & A2. Each histogram is of $B = 100$ classification error rates from B randomized data sets using algorithm A1 (columns 1, 2) and A2 (columns 3, 4). The observed gene expression profiles were randomly assigned cancer labels (leukemia, colon, melanoma, renal, or CNS). Class sizes (the $n_i s$) were the same as the original data set. The corresponding observed error rates given in Table 2.A1 and Table 2.A2 are indicated by the solid lines in the histograms.

error rates can be large.

As a minimum, we checked to see whether the observed classification error is lower that classification on 'random' data. That is, randomly assign the cancer labels (cancer group $0, 1, \ldots$ or G) to each gene expression profile (sample) to generate a 'new' permuted data set, say $\mathbf{X}^*$. We randomly generated $B = 100$ permuted data sets, $\mathbf{X}^*_{(1)}, \ldots, \mathbf{X}^*_{(B)}$ and obtained corresponding classification error rates $e_1, \ldots, e_B$ using both algorithm A1 and A2. The observed classification error rate from the original (real) data set, say e_{obs} (given in Tables 1A1–4A2 and 1A2–4A2), can be compared to the distribution of error rates obtained from randomization (the $e_i s$). This was carried out on all four data sets for every subset of genes and methods combination used.

For example, the distribution (histogram) of error rates from randomization corresponding to observed error rates in Table 1A2 (NCI60 data) is given in Figure 2. This distribution of error rates was obtained by fixing the gene set and refitting the dimension redution step as described by algorithm A1. The histogram is of $B = 100$ error rates. This was repeated by reselecting the genes and refitting the dimension at each step (algorithm A2; Figure 2). Similar disribution of error rates for the breast cancer, lymphoma, and acute leukemia data are given in Supplemental Figures 1–6.

The results displayed in Figure 2 and the Supplemental Figures 1–6 suggest that, in all cases, the observed classification error rate is significantly less than would be expected under randomization. Finally, we also note that the simulation studies given here only suggest that the observed classification rates reported in Tables 1–4 are much lower than would be expected at random, which is obvious. As with any other analytical method, further validation based on new real data will shed more light on its usefulness.

Comparisons to DQDA, DLDA and classification tree

Comparing classification performances from various methods is often of interest. Classification results from DQDA and DLDA using algorithms A0, A1, and A2 are summarized in the right side of Tables 1–4 (A0, A1, and A2). The 'simplest' classifier DLDA, particularly DLDA–MPLS, performed very well and other results are similar to those reported in the earlier section. We also compared to the recursive partitioning (classification tree) method of Zhang *et al.* (2001). The overall leave-one-out cross-validated classification errors for the leukemia, lymphoma, NCI, and BRCA data sets were 15.2%, 10.84%, 22.86% and 90.9%. For a more detailed description of these comparisons see the Supplemental Appendix D.

Selecting the number of gene components

We have chosen $K = 3$ gene components to fit classifiers. This is based on a study of K as a 'tuning' parameter. Increasing K above 5 does not drastically improve classification results. $K = 3$, 4, or 5 often provide similar results, so the simpler choice of $K = 3$ is more desirable. Details of the study of K is given in the Supplemental Appendix E and Supplemental Figure 7.

DISCUSSION

We have proposed multi-class cancer classification methods in this paper, which are extension of the methods proposed in an earlier paper for binary tumor classification. The utility of the methods will be further evaluated as more experimental data becomes available. An advantage of the methodologies proposed is that other classification methods can be utilized (other than PD and QDA) after dimension reduction via MPLS, for instance. As discussed in the Supplemental Appendix, numerical methods are needed to obtain the MLE in PD and the existence of the MLE depends on the data configuration. One disadvantage of using PD is when there is quasi-complete separation in the data. Detection of quasi-complete separation is numerically burdensome and classification is usually poor. (See Supplemental Appendix A for details.) Also, inversion problems can be encountered in the Newton–Raphson algorithm when searching for the MLE.

We have also provided a careful evaluation of the classification algorithm and demonstrated how the error rates are highly influenced by gene selection and LOOCV (often used with small sample sizes). Supplemental Appendix C contains a brief discussion of an alternative gene selection and normalization issues.

ACKNOWLEDGEMENTS

We thank three referees for helpful comments. The research reported in this paper was supported by grants from the National Cancer Institute (CA90301), the National Science Foundation (ACI 96-19020 and DMS 98-70172), and the National Institute of Environmental Health Sciences (P43 ES04699). We thank Neil Willits (Department of Statistics, University of California at Davis) for the many helpful discussions.

REFERENCES

Albert,A. and Anderson,J.A. (1984) On the existence of maximum likelihood estimates in logistic models. *Biometrika*, **71**, 1–10.

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.

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.

Catteau,A., Harris,W.H., Xu,C.F. and Soloman,E. (1999) Methylation of the BRCA1 promoter region in sporadic breast cancer: correlation with disease characteristics. *Oncogene*, **18**, 1957–1965.

Dudoit,S., Fridlyand,J. and Speed,T.P. (2000) Comparison of discrimination methods for the classification of tumors using gene expression data. *Technical Report # 576*, Department of Statistics, University of California, Berkeley.

Esteller,M., Silva,J.M., Domingquez,G., Bonilla,F., Matias-Guiu,X., Lerma,E., Bussaglia,E., Prat,J. *et al.* (2000) Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. *J. Natl Cancer Inst.*, **92**, 564–569.

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.

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

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. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Hedenfalk,I., Duggan,D., Chen,Y., Radmacher,M., Bittner,M., Simon,R., Meltzer,P., Gusterson,B., Esteller,M., Kallioniemi,O. *et al.* (2001) Gene-expression profiles in hereditary breast cancer. *The New England Journal of Medicine*, **344**, 539–548.

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

Höskuldsson,A. (1988) PLS regression methods. *Journal of Chemometrics*, **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.

Johannsson,O.T., Idvall,I., Anderson,C., Borg,A., Barkaròttir,V., Egilsson,V. and Olsson,S. (1997) Tumor biological features of BRCA1-induced breast and ovarian cancer. *Eur. J. Cancer*, **33**, 362–371.

Karp,S.E., Tonin,P.N. Begin,L.R. *et al.* (1997) Influence of BRCA1 mutations on nuclear grade and estrogen receptor status of breast carcinoma in Ashkenazi Jewish woman. *Cancer*, **80**, 435–441.

Kooperberg,C., Bose,S. and Stone,C.J. (1997) Polychotomous regression. *J. Amer. Stat. Assoc.*, **92**, 117–127.

Lachenbruch,P.A. and Mickey,M.R. (1968) Estimation of error rates in discriminant analysis. *Technometrics*, **10**, 1–11.

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

Loman,N., Johannsson,O., Bendahl,P.O. Borg,A. *et al.* (1998) Steroid receptors in hereditary breast carcinomas associated with BRCA1 and BRCA2 mutations or unknown suspectibility genes.

Cancer, **83**, 310–409.

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

Nguyen,D.V. and Rocke,D.M. (2002a) 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 microarray gene expression data. *Bioinformatics*, **18**, 39–50.

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. *et al.* (1999) Distinctive gene expression patterns in human mammary epitholial cells and breast cancer. *Proc. Natl Acad. Sci. USA*, **96**, 9112–9217.

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

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.

Scherf,U., Ross,D.T. Waltham,M. *et al.* (2000) A Gene Expression Database for the Molecular Pharmacology of Cancer. *Nature Genet.*, **24**, 236–244.

Verhoog,L.C., Brekelmans,C.T. Seynaeve,C. *et al.* (1998) Survival and tumor characteristics of breast-cancer patients with carmline mutations of BRCA1. *Lancet*, **351**, 316–321.

Zhang,H.P., Yu,C., Singer,B. and Xiong,M. (2001) Recursive partitioning for tumor classification with gene expression microarray data. *Proc. Natl Acad. Sci. USA*, **98**, 6730–6735.