56–65 *Nucleic Acids Research, 2005, Vol. 33, No. 1*
doi:10.1093/nar/gki144

Multi-class cancer classification by total principal component regression (TPCR) using microarray gene expression data

Yongxi Tan, Leming Shi[1], Weida Tong[1] and Charles Wang*

Department of Medicine, Cedars-Sinai Medical Center, David Geffen School of Medicine, UCLA, Los Angeles, CA 90048, USA and [1]Center for Toxicoinformatics, Division of Systems Toxicology, National Center for Toxicological Research, FDA, Jefferson, AR 72079, USA

Received August 13, 2004; Revised November 9, 2004; Accepted December 3, 2004

ABSTRACT

DNA microarray technology provides a promising approach to the diagnosis and prognosis of tumors on a genome-wide scale by monitoring the expression levels of thousands of genes simultaneously. One problem arising from the use of microarray data is the difficulty to analyze the high-dimensional gene expression data, typically with thousands of variables (genes) and much fewer observations (samples), in which severe collinearity is often observed. This makes it difficult to apply directly the classical statistical methods to investigate microarray data. In this paper, total principal component regression (TPCR) was proposed to classify human tumors by extracting the latent variable structure underlying microarray data from the augmented subspace of both independent variables and dependent variables. One of the salient features of our method is that it takes into account not only the latent variable structure but also the errors in the microarray gene expression profiles (independent variables). The prediction performance of TPCR was evaluated by both leave-one-out and leave-half-out cross-validation using four well-known microarray datasets. The stabilities and reliabilities of the classification models were further assessed by re-randomization and permutation studies. A fast kernel algorithm was applied to decrease the computation time dramatically. (MATLAB source code is available upon request.)

INTRODUCTION

Improvements in cancer classification have been of great importance in cancer treatment. It is difficult to distinguish tumors, which have similar histopathological appearance but different clinical course and response to therapy, by the traditional cancer diagnostic methods that are based primarily on morphological appearance of tumors (1). DNA microarray technology provides a powerful approach to the diagnosis and prognosis of various tumors on a genome-wide scale. By simultaneously monitoring the expression of thousands of genes in cells to obtain quantitative information about the complete transcription profile of cells, microarray technology makes tailored therapeutics to specific pathologies possible (1–4). Despite the usefulness of microarray technology, analyzing and understanding the obtained data has been a complex and challenging task. Microarray data analysis methods can be categorized roughly into unsupervised learning, including various clustering techniques such as self-organizing map (5) and hierarchical clustering (6), and supervised learning, including various classification and prediction techniques (7,8). Some recent applications of supervised learning techniques include molecular classification of acute leukemia (1), classification of human cancer cell lines (9), support vector machine classification of cancer tissue samples (10), classifying cancers using artificial neural networks (11), mapping of the physiological state of cells and tissues and identification of important genes using Fisher discriminant analysis (12), tumor classification by polychotomous discrimination and quadratic discriminant analysis after dimension reduction using principal component analysis (PCA) or partial least squares (PLS) (13), PCA disjoint models for cancer classification (14), multiclass tumor classification by discriminant PLS and assessment of classification models (15), classification by incorporating PLS within the interactively re-weighed least square steps for multinomial or binary logistic regression (16) and classification using PLS with penalized logistic regression (17).

DNA microarray gene expression data are usually characterized by thousands of variables (genes) with much fewer observations (samples), resulting in a high degree of multicollinearity. This makes it difficult or even impossible to apply

*To whom correspondence should be addressed. Tel: +1 310 4237363; Fax: +1 310 4237452; Email: charles.wang@cshs.org

directly classical statistical methods to the analysis of microarray data. To tackle this kind of collinearity problems, latent variable methods, such as PCA (18) and PLS (19), have been developed to reduce the dimensionality of gene expression data and mitigate the collinearity. These methods assume that the independent variables (gene expression profiles) are inherently located in a low-dimensional linear subspace, i.e. they have an intrinsic latent variable structure. PCA attempts to find a set of orthogonal principal components (linear combinations of original independent variables) to account for the maximum variations in independent variables (18). Since the information about the sample classification provided by dependent variable (class membership) is not taken into account in the extraction of the principal components in PCA, the performance of PCA in classification or prediction may not be satisfactory. And this is one of the reasons why PLS was developed several decades ago to provide a better performance in calibration and prediction than PCA (19). In addition, it is well-known that microarray experiments are influenced by many potential sources of variation/error and these kinds of variability can be roughly classified into three categories: biological variation (genetic or environmental factors, pooled or non-pool), technical variation (during extraction, labeling and hybridization) and measurement error (signal detection, etc.) (20–22). These kinds of variability have been analyzed using the ANOVA model to determine the sources and magnitudes of error/variation in gene expression profiles (22–25). However, in classification or prediction using microarray data, most of the statistical methods (e.g. PCA and PLS) relating the gene expression profiles, **X**-independent variables, to other information of interest, **Y**-dependent variables (e.g. tumor type or survival time), do not account for errors in the independent variables, which is one of the important characteristics of measured data. To tackle these problems mentioned above, a novel method, total principal component regression (TPCR), is proposed to not only incorporate the information of the dependent variables into the construction of latent structure but also take into account the errors in both independent variables and dependent variables. A salient feature of TPCR is that it extracts the latent structure from the augmented subspace of both independent and dependent variables, using a weighted least square fitting. This enables the proposed method to construct latent variables approximating optimally the actual latent structure and eliminate the collinearity to a certain degree.

METHODS

Total principal component regression

The basic goal of various projection or dimension-reduction approaches, for example PCA (18) and PLS (19), is to project the observations (samples) from the high-dimensional variables (genes) space to a low-dimensional subspace spanned by several linear combinations of the original variables, in order to satisfy a certain criterion. PCA attempts to find a set of orthogonal principal components to explain as much variance as possible in independent variables (**X**). The performance of PCA in classification may not be satisfactory from the predictive point of view, because there is no guarantee that the principal component representing the large variance in **X** should necessarily be the component strongly related to dependent variables (**Y**). To solve this problem, the information of dependent variables should be taken into account during the construction of orthogonal components. One way to do so is to maximize the sample covariance between the linear combination of dependent variables and the orthogonal component of independent variables, which is the essence of PLS (15,26,27), as shown by the objective criterion of PLS:

$$\{w, c\} = \underset{\substack{w^{\mathrm{T}}w=1 \\ c^{\mathrm{T}}c=1}}{\operatorname{argmax}} \operatorname{cov}^2(\mathbf{X}w, \mathbf{Y}c) \quad \mathbf{1}$$

where w and c denote the weight vectors of **X** and **Y**, respectively. It has been proved that w and c are related to the following eigenvalue problems (26,27):

$$\mathbf{X}^{\mathrm{T}}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}\mathbf{X}w = aw$$
$$\mathbf{Y}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{Y}c = ac \quad \mathbf{2}$$

where a is the maximum eigenvalue and the weight vectors w and c can thus be calculated as the first left and right singular vector of $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$, respectively.

Another simple way to make use of the information of dependent variable would be to construct the orthogonal components, rather than only from the subspace of **X**, from the augmented subspace of both **X** and **Y**, that is, finding a low-dimensional subspace to best fit the subspace spanned by both X and Y, which is one of the motivations of TPCR.

In classical regression/prediction models, the independent variables are usually assumed to be non-stochastic, in other words, there is no error in the independent variables or at least the error is negligible. However, it is well-known that various variation/error may be introduced during a multi-step microarray experiment and these kinds of variation/error are usually not negligible. In order to account for errors in both independent variables and dependent variables, the error-in-variables (EIV) model (28–32) was used in this paper:

$$\mathbf{Y} = \widetilde{\mathbf{Y}} + \mathbf{E}_{\mathbf{Y}} \quad \mathbf{3}$$
$$\mathbf{X} = \widetilde{\mathbf{X}} + \mathbf{E}_{\mathbf{X}} \quad \mathbf{4}$$
$$\widetilde{\mathbf{Y}} = \widetilde{\mathbf{X}}\mathbf{B} \quad \mathbf{5}$$

where $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ denote the systematic or unobservable true values for independent variables $\mathbf{X}_{N\times P}$ and dependent variables $\mathbf{Y}_{N\times M}$, respectively; $\mathbf{E}_{\mathbf{X}}$, $\mathbf{E}_{\mathbf{Y}}$ represent the random error matrices whose rows are assumed to be independently, identically distributed (i.i.d) with common mean vector $\mathbf{0}$ and common covariance matrices $\sigma^2_{\mathbf{X}}\mathbf{I}_P$ and $\sigma^2_{\mathbf{Y}}\mathbf{I}_M$ ($\mathbf{I}_P$ and $\mathbf{I}_M$ denote appropriate identity matrices), respectively; **B** is a $P \times M$ matrix. Equation 5 implies an assumption of the EIV model, i.e. there exists a linear functional relationship between the systematic or true values of **X** and **Y** (29,30,32). In the case of microarray data analysis, in our opinion, the errors in the independent variables (gene expression profiles) may include the random fluctuations introduced by microarray technology itself, including measurement error and/or technical variation, while the biological variation (the true difference in gene expression) is embedded in the systematic part of independent variables.

Suppose the systematic or true values of the independent variables under observation is actually driven by a set of unobservable latent variables, i.e. they lie in a lower dimensional

linear subspace spanned by the latent variables, then we can define a column-wise orthonormal matrix $\mathbf{T}_{N\times K}$ ($K < P$ and $\mathbf{T}^{\mathrm{T}}\mathbf{T} = \mathbf{I}$, where superscript T denotes the transpose of a matrix), whose columns provide the basis for the subspace of both $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$:

$$\widetilde{\mathbf{X}} = \mathbf{TG} \quad 6$$

$$\widetilde{\mathbf{Y}} = \widetilde{\mathbf{X}}\mathbf{B} = \mathbf{TGB} = \mathbf{TF} \quad 7$$

where $\mathbf{G}_{K\times P}$ and $\mathbf{F}_{K\times M}$ are the corresponding loading matrices for $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$, respectively. $\mathbf{T}$ can be seen as the common latent structure for both $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$. On substitution of Equations 6 and 7 into the previous EIV model, we obtain the EIV latent variable model (33–35):

$$\mathbf{Y} = \mathbf{TF} + \mathbf{E}_{\mathbf{Y}} \quad 8$$

$$\mathbf{X} = \mathbf{TG} + \mathbf{E}_{\mathbf{X}} \quad 9$$

The assumption behind this model is that there is a linear functional or structural relationship between the systematic or true part of $\mathbf{X}$ and $\mathbf{Y}$ and this relationship can be linked by a set of unobservable underlying latent variables $\mathbf{T}$ (33–35). Note that just like the assumption of ordinary least squares is not strictly complied with in practical applications, although the i.i.d. assumption about the error structure in the TPCR model may not be rigorously valid in some cases, it should not degrade the performance of this model too much if the violation is not too severe. It is worthy to point out that a model taking into account the error information in the independent variable, even if incomplete, would provide more insight into the realistic characteristics and structure of data and better performance than those incorporating no such information. And the violation of this assumption may be corrected partially by some preprocessing techniques such as transformation and scaling.

A criterion for solving this EIV latent variable model is:

$$\min_{\substack{\mathbf{T},\mathbf{G},\mathbf{F}\\ \mathbf{T}^{\mathrm{T}}\mathbf{T}=\mathbf{I}}} \left(\frac{\|\mathbf{X} - \mathbf{TG}\|_{\mathrm{F}}^2}{\sigma_{\mathbf{X}}^2} + \frac{\|\mathbf{Y} - \mathbf{TF}\|_{\mathrm{F}}^2}{\sigma_{\mathbf{Y}}^2} \right) \quad 10$$

where $\|\mathbf{M}\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix, that is, $\|\mathbf{M}\|_{\mathrm{F}} = [\mathrm{tr}(\mathbf{MM}^{\mathrm{T}})]^{1/2}$.

To deal with the error in both independent variables and dependent variables, let meta parameter λ ($\geqslant 0$) be:

$$\lambda^2 = \frac{\sigma_{\mathbf{X}}^2}{\sigma_{\mathbf{Y}}^2} \quad 11$$

Then Equation 10 can be rewritten as

$$\min_{\substack{\mathbf{T},\mathbf{G},\mathbf{F}\\ \mathbf{T}^{\mathrm{T}}\mathbf{T}=\mathbf{I}}} \left(\|\mathbf{X} - \mathbf{TG}\|_{\mathrm{F}}^2 + \lambda^2 \|\mathbf{Y} - \mathbf{TF}\|_{\mathrm{F}}^2 \right) \quad 12$$

$$= \min_{\mathbf{T}} \left(\min_{\mathbf{G}} \|\mathbf{X} - \mathbf{TG}\|_{\mathrm{F}}^2 + \lambda^2 \min_{\mathbf{F}} \|\mathbf{Y} - \mathbf{TF}\|_{\mathrm{F}}^2 \right) \quad 13$$

Noting that $\mathbf{TT}^{\mathrm{T}}$ is the projection matrix, it is easy to see from least square analysis (36–38) that:

$$\|\mathbf{X} - \mathbf{TG}\|_{\mathrm{F}}^2 \geqslant \|\mathbf{X} - \mathbf{TT}^{\mathrm{T}}\mathbf{X}\|_{\mathrm{F}}^2 \quad 14$$

$$\|\mathbf{Y} - \mathbf{TF}\|_{\mathrm{F}}^2 \geqslant \|\mathbf{Y} - \mathbf{TT}^{\mathrm{T}}\mathbf{Y}\|_{\mathrm{F}}^2 \quad 15$$

Thus, Equation 13 is equivalent to

$$\min_{\mathbf{T}} \left(\|\mathbf{X} - \mathbf{TT}^{\mathrm{T}}\mathbf{X}\|_{\mathrm{F}}^2 + \lambda^2 \|\mathbf{Y} - \mathbf{TT}^{\mathrm{T}}\mathbf{Y}\|_{\mathrm{F}}^2 \right) \quad 16$$

$$= \min_{\mathbf{T}} \|\left(\mathbf{X} - \mathbf{TT}^{\mathrm{T}}\mathbf{X}, \lambda\mathbf{Y} - \lambda\mathbf{TT}^{\mathrm{T}}\mathbf{Y}\right)\|_{\mathrm{F}}^2 \quad 17$$

$$= \min_{\mathbf{T}} \|\mathbf{A} - \mathbf{TT}^{\mathrm{T}}\mathbf{A}\|_{\mathrm{F}}^2 \quad 18$$

$$= \min_{\mathbf{T}} \|\left(\mathbf{I} - \mathbf{TT}^{\mathrm{T}}\right)\mathbf{A}\|_{\mathrm{F}}^2 \quad 19$$

where $\mathbf{A}$ is the $N \times (P + M)$ augmented matrix of $\mathbf{X}$ and $\mathbf{Y}$, that is, $\mathbf{A} = (\mathbf{X}, \lambda\mathbf{Y})$.

Noting that $(\mathbf{I}-\mathbf{TT}^{\mathrm{T}})$ stands for the $N \times N$ projection matrix, which projects on the orthogonal complement of the subspace spanned by $\mathbf{T}$, Equation 19 can be minimized when $\mathbf{T}$ is the first K largest principal component for the augmented matrix $\mathbf{A}$ (18). Let the singular value decomposition (SVD) of $\mathbf{A}$ be

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}} \quad 20$$

where left singular vectors $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_N) \in R^{N\times N}$ with $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_N$; right singular vectors $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{(P+M)}) \in R^{(P+M)\times(P+M)}$ with $\mathbf{V}^{\mathrm{T}}\mathbf{V} = \boldsymbol{I}_{(P+M)}$, Σ is a diagonal matrix containing singular values. Then let $\mathbf{T}$ be the first K columns of $\mathbf{U}$:

$$\mathbf{T} = (\mathbf{u}_1, \ldots, \mathbf{u}_K) \quad 21$$

Thus the estimates of $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$ can be obtained by:

$$\hat{\mathbf{X}} = \mathbf{TT}^{\mathrm{T}}\mathbf{X} \quad 22$$

$$\hat{\mathbf{Y}} = \mathbf{TT}^{\mathrm{T}}\mathbf{Y} \quad 23$$

The regression coefficients estimated using TPCR are given by

$$\mathbf{B}_{\mathrm{TPCR}} = \left(\mathbf{T}^{\mathrm{T}}\mathbf{X}\right)^{+}\mathbf{T}^{\mathrm{T}}\mathbf{Y} \quad 24$$

where superscript + denotes the generalized inverse of a matrix.

Fast kernel EVD algorithm for wide data

The speed and time of calculations have always been the practical and important problems in the implementation of algorithm or method in multivariate data analysis. Microarray dataset typically consists of thousands of variables and less than 100 samples ($P \gg N$). For such 'wide' data (39), the computation time needed for matrix decomposition using classical SVD algorithm (e.g. in MATLAB: [**U**, **S**, **V**] = **svd** (**A**), where **svd** is the build-in function to calculate the singular vectors, **U** and **V**, and singular values **S**) may be pretty long. The situation becomes even worse when cross-validation is applied to evaluate the methods, which is the typical case in tumor classification. Therefore, an efficient and fast algorithm is needed to calculate the singular vectors from microarray data.

Since the kernel matrix $\mathbf{AA}^{\mathrm{T}}$ contains the same information (e.g. eigenvalues) as the covariance matrix $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ while the size of $\mathbf{AA}^{\mathrm{T}}$ ($N \times N$) is much smaller than that of $\mathbf{A}_{N\times(P+M)}$ and $(\mathbf{A}^{\mathrm{T}}\mathbf{A})_{(P+M)\times(P+M)}$ ($P \gg N$), it would be much faster to calculate the left singular vectors or eigenvectors $\mathbf{U}$ from $(\mathbf{AA}^{\mathrm{T}})_{(N\times N)}$ by eigenvalue decomposition (EVD) than from $\mathbf{A}_{N\times(P+M)}$ by SVD (39).

Thus, the modified fast kernel EVD algorithm to calculate **T** is:

$$\left[\mathbf{U}, \mathbf{\Sigma}^2\right] = \mathbf{eig}\left(\mathbf{A}\mathbf{A}^{\mathrm{T}}\right) \quad 25$$

$$\mathbf{T} = (\mathbf{u}_1, \ldots, \mathbf{u}_K); \quad \text{the first } K \text{ columns of } \mathbf{U} \quad 26$$

where **eig** is the build-in function of MATLAB to calculate eigenvectors and eigenvalues of a matrix, **U** denotes the eigenvectors of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ or the singular vectors of **A** and $\mathbf{\Sigma}^2$ is a diagonal matrix containing eigenvalues of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ ($\mathbf{\Sigma}$ contains the singular values of **A**).

TPCR for discrimination

When TPCR is used for classification, the matrix of dependent variables (**Y**) contains the information about the class memberships, with element $y_{ik} = 0$ or 1 ($i = 1, \ldots, N$; $k = 1, \ldots, M$; where N and M is the number of samples and the number of tumor classes, respectively). If the i-th sample belongs to class k, then $y_{ik} = 1$, otherwise $y_{ik} = 0$.

The prediction of dependent variables on a new set of samples is made by:

$$\mathbf{Y}_{\mathrm{new}} = \mathbf{X}_{\mathrm{new}}\mathbf{B}_{\mathrm{TPCR}} \quad 27$$

where $\mathbf{X}_{\mathrm{new}}$ is the gene expression profiles for the new set of samples, and $\mathbf{Y}_{\mathrm{new}}$ is the predicted values for these samples. The identity of the class membership of each new sample (each row in $\mathbf{Y}_{\mathrm{new}}$) is assigned as the column index of the element with the largest predicted value in this row.

Meta parameter λ

As stated above, the relative magnitude of errors in the independent variables and dependent variables is given by meta parameter λ in TPCR model. It is important to choose the appropriate λ to obtain the best prediction performance. Considering two extreme cases for λ (≥0):

(i) If λ = 0, i.e. $\sigma_{\mathbf{x}}^2 = 0$, which means no error is considered in **X** matrix, then the augmented matrix $\mathbf{A} = (\mathbf{X}, \lambda\mathbf{Y}) = \mathbf{X}$; therefore, **T** will be the principal components of **X** itself. In other words, the TPCR model degenerates to the classical principal component regression (PCR) model in this extreme case. Note that no information about **Y** is taken into account in the construction of latent variable in this case.

(ii) If λ is very large, the variation of the columns of λ**Y** would be much larger than that of the columns of **X** in the augmented matrix **A**. In this case, the major principal components **T** will come largely from λ**Y**, since the PCA always projects to the directions showing the largest variation. Therefore, the prediction performance of **T** for the new sample $\mathbf{X}_{\mathrm{new}}$ will be poor.

The choice of λ depends on the experience about the data and the computation power available. In practice, the optimal meta parameter λ (≥0) is chosen from an appropriate set according to leave-one-out cross-validation (LOOCV) procedure.

Gene selection

Variable (gene) selection is important for the successful analysis of gene expression data since most of the genes do not provide useful information for classification, and including all of them in the modeling process will degrade the performance of the model. Therefore, non-informative genes should be removed before building a classification model. There exist different approaches for gene selection such as neighborhood analysis (1), significance analysis of microarrays (40), Wilks' lambda (12), *t*-score and critical score (13,41) and classifier feedback approach (14).

In this paper, the sum of squared correlation coefficients between gene expression and each of the dependent variables is used to select the genes for analysis (15). For example, the $g^* = 100$ genes are taken as the first 100 genes with the largest values of sum of squared correlation coefficients.

Assessment of prediction method by leave-one-out and leave-half-out CV

LOOCV has become a standard procedure to evaluate the performance of various classification methods in microarray data analysis. Note that when gene selection or dimension reduction is used together with LOOCV procedure, a common mistake made in tumor classification using microarray gene expression profiles, was to perform gene selection or dimension reduction before CV loop. However, such incomplete LOOCV procedure is well known to be substantially biased and prone to generating spuriously good results since the information about all the samples is used for gene selection or dimension reduction before the CV loop (15,42–44). In this paper, the complete LOOCV (the gene selection and dimension reduction within the CV loop) was applied.

LOOCV is nearly unbiased, but often with unacceptably high variability, especially for dataset with small number of samples (45–48), implying that it may give unreliable good prediction results due to the effect of high variance/variability, especially when the number of samples is small (just like ordinary least square may give unreliable result due to the large variance if there exists severe collinearity in the data, even if it is unbiased). External validation can provide some protection against over-fitting caused by LOOCV, but there may not be enough samples for external test, especially for microarray data analysis (due to the relatively expensive cost, etc.). In this case, a re-randomization study, leave-half-out cross-validation (LHOCV), provides an alternative approach to evaluate the prediction method more realistically (14,15,41,48). Briefly, the whole datasets, including original training and test dataset, are pooled together and split randomly (half/half) into a new training dataset and a new test dataset; the randomly generated training dataset is used to derive a classification model (select genes, reduce dimension and calculate regression coefficients) that is then applied to classify the corresponding new test dataset. This LHOCV procedure is repeated 100 times (splitting the pooled dataset randomly for 100 times) to avoid chance factor.

RESULTS AND DISCUSSION

Acute leukemia data

The well-known leukemia dataset was measured by Golub *et al*. (1) using Affymetrix high-density oligonucleotide microarray containing probes for 6817 human genes and has become a benchmark for the evaluation of various cancer classification

algorithms. The original training dataset consisted of 38 bone marrow samples from acute leukemia patients, including 19 B-cell acute lymphoblastic leukemia (B-ALL), 8 T-cell acute lymphoblastic leukemia (T-ALL) and 11 acute myeloid leukemia (AML). The original independent (test) dataset consisted of 24 bone marrow and 10 peripheral blood samples (19 B-All, 1 T-ALL and 14 AML). The gene expression data was log transformed and centered to have mean zeros across samples during the cross-validation process.

The original 38 training samples and 34 test samples were pooled together and then classified by TPCR using LOOCV with $g^* = 50$, 100, 200, 500 and 1000 genes selected (λ was taken from the values: 0, 0.001, 0.01, 0.1, 1–20 with interval of 1, 40–100 with interval of 20, 200, 1000, 10 000 and 100 000 in this paper) and the results are shown in Table 1. One ALL sample (#17; error rate 1.39%; $\lambda = 19$) was misclassified by TPCR. In our previous study (15), discriminant PLS (D-PLS) was used to classify the same dataset using the same gene selection method and resulted in at least 2 (2.78%) misclassifications. This dataset was also analyzed by Nguyen and Rocke (13) using polychotomous discrimination and quadratic discriminant analysis together with dimension reduction by PCA or PLS and at least 8 (11.1%) and 3 (4.2%) samples were misclassified using PCA and PLS for dimension reduction, respectively (from A2 procedure; note that both A0 and A1 procedures are incomplete LOOCV and should not have been used to compare with our results).

Hereditary breast cancer data

The gene expression profiles of primary breast tumor samples from seven carriers of the BRCA1 mutation, eight carriers of the BRCA2 mutation and seven patients with sporadic cases of breast cancer were monitored with a microarray of 6512 cDNA clones of 5361 genes and a total of 3226 genes were selected for analysis (49). The gene expression data were centered to have mean zeros across samples during cross-validation process.

Classification results based on TPCR using LOOCV are presented in Table 2 and at the best, four samples were misclassified (error rate 18.18%, $\lambda = 9$), which is better than the best result (22.7%, 5 misclassification) of D-PLS (15) and the result (5 misclassifications) obtained by Hedenfalk *et al*. (49). In the study of Nguyen and Rocke (13), at least six (27.3%) misclassifications were found using A2 procedure. The percentage of misclassified samples by TPCR and other methods was pretty high, implying the difficulty to accurately classify these 22 tumor samples based on gene expression profiles. This may be due to the inherent lack of discriminating power of the dataset (e.g. the small sample size, diagnosis error or the lack of differentiating power for the expression profiles of 3226 genes to separate these 22 samples) (15,49). Another possibility could be that the underlying structure of the data (e.g. inner non-linear relationship) was not characterized by the methods used (15). As pointed out before by Hedenfalk *et al*. (49), 'the use of microarray covering a larger proportion of the genome and the analysis of larger numbers of tumor may make possible a more precise molecular classification of breast cancer'.

Small, round blue cell tumor data

The small, round blue cell tumors (SRBCTs) of childhood, including neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS), are difficult to distinguish due to their similar appearances in routine histology. Khan *et al*. (11) monitored the gene expression profiles of 6567 genes for these four types of malignancies using cDNA microarrays and reduced the number of genes to 2308 by quality filtering for a minimal level of expression. The original 63 training samples included both tumor biopsy materials (13 EWS and 10 RMS) and cell lines (10 EWS, 10 RMS, 12 NB and 8 Burkitt Lymphomas (BL, a subset of NHL). The original test samples contained both tumors (5 EWS, 5 RMS and 4 NB) and cell lines (1 EWS, 2 NB and 3 BL). This dataset was centered to have mean zeros across samples for analysis.

The pooled-together 83 samples were classified by TPCR using LOOCV with $g^* = 50$, 100, 200, 500 and 1000 genes selected and the results are presented in Table 3. All the samples were correctly classified using 3–6 TPCR components with $g^* = 100$, 200 and 500 genes selected ($\lambda = 200$), indicating a good prediction performance of TPCR on this dataset. The same result (0 misclassification) was also obtained in other studies (15,50–53), indicating good class separability of this dataset. Khan *et al*. (11) also found 0 misclassification using artificial neural networks for 88 samples, including 5 non-SRBCT samples. However, the dimension reduction by

Table 2. Classification results for hereditary breast cancer dataset

g^*	Number of TPCR components (%, $\lambda = 9$, LOOCV)					
	1	2	3	4	5	6
50	54.55	45.45	40.91	40.91	45.45	40.91
100	59.09	50.00	40.91	31.82	36.36	22.73
200	54.55	22.73	**18.18**	27.27	31.82	31.82
500	77.27	31.82	27.27	36.36	40.91	40.91
1000	59.09	63.64	45.45	45.45	40.91	40.91

Given are the percentages of misclassification out of 22 samples using LOOCV.

Table 1. Classification results for leukemia dataset

g^*	Number of TPCR components (%, $\lambda = 19$, LOOCV)					
	1	2	3	4	5	6
50	13.89	2.78	4.17	5.56	8.33	6.94
100	15.28	4.17	4.17	5.56	5.56	5.56
200	13.89	2.78	2.78	**1.39**	4.17	4.17
500	13.89	2.78	2.78	4.17	4.17	4.17
1000	15.28	2.78	4.17	4.17	4.17	4.17

Given are the percentages of misclassification out of 72 samples using LOOCV.

Table 3. Classification results for SRBCT dataset

g^*	Number of TPCR components (%, $\lambda = 200$, LOOCV)					
	1	2	3	4	5	6
50	36.14	14.46	1.20	**0.00**	0.00	0.00
100	34.94	14.46	0.00	0.00	0.00	0.00
200	34.94	14.46	0.00	0.00	0.00	0.00
500	34.94	15.66	0.00	0.00	0.00	0.00
1000	34.94	18.07	1.20	0.00	0.00	0.00

Given are the percentages of misclassification out of 83 samples using LOOCV.

PCA was performed using all the 88 samples before the validation procedure, which may suffer from bias.

NCI60 data

Using cDNA microarrays containing 9703 cDNA clones representing ~8000 unique genes, Ross *et al*. (9) and Scherf *et al*. (54) studied gene expression in the 60 human cancer cell lines used in the NCI anticancer drug screening program. The 60 human cell lines were derived from tumors from a variety of tissues and organs, which, in contrast to clinical tumors, have been characterized pharmacologically by treatment with >70 000 different agents, one at a time and independently (54). As in the study of Nguyen and Rocke (13), five cancer types were used for multi-class classification: eight melanoma, eight renal, six leukemia, seven colon and six CNS. A subset of 1376 genes selectively filtered from initial 9703 genes and 40 molecular characteristics (targets) individually assessed by various laboratories was used to discriminate the different types of cancers (54). Since there are some missing gene expression values in this dataset, genes with <2 missing values were used for classification through replacing the missing values (1 or 2) with the median of the gene expression (13). This resulted in a gene expression dataset with 35 samples and 1299 genes. This dataset was centered to have mean zeros across samples in analysis. TPCR was then applied to this dataset with $g^* = 50$, 100, 200, 500 and 1000 genes selected.

The misclassification results using TPCR are given in Table 4. The best result was one misclassified sample (#15 ME:LOXIMVI, 2.86%, $\lambda = 20$), the same as the best result using D-PLS (15), while the best result obtained by Nguyen and Rocke (13) was 2 (5.7%) and 3 (8.6%) misclassifications using PCA and PLS for dimension reduction, respectively (A2 procedure). The classification performance of TPCR on this dataset is good, given the small sample-to-class ratio (35 samples and 5 classes). It would be worthy to note that ME:LOXIMVI, although supposedly a melanoma in origin, was reported to lack melanin and other useful marker for the identification of melanoma cells (55) and showed different characteristic pattern from the other seven melanoma lines (9,54). Furthermore, in an earlier study involving the clustering of the 60 cancer cell lines based purely on their sensitivity to tens of thousands of potential anticancer compounds, ME:LOXIMVI was also found to be different from other melanoma cell lines; instead, it was found to be more similar to a group of colon cancer cell lines (56).

Table 4. Classification results for NCI60 dataset

g^*	Number of TPCR components (%, $\lambda = 20$, LOOCV)					
	1	2	3	4	5	6
50	100.00	60.00	60.00	8.57	8.57	8.57
100	91.43	57.14	37.14	**2.86**	5.71	5.71
200	77.14	57.14	20.00	2.86	2.86	5.71
500	77.14	57.14	20.00	2.86	5.71	5.71
1000	77.14	48.57	20.00	2.86	2.86	2.86

Given are the percentages of misclassification out of 35 samples using LOOCV.

Evaluation of TPCR more realistically by LHOCV

As demonstrated in the above LOOCV studies with four well-publicized microarray datasets, TPCR showed better or at least comparable prediction performance compared with other published methods. However, LOOCV may provide unreliable

Table 5. Classification results using TPCR and PLS under LHOCV procedure for Leukemia, hereditary breast cancer, SRBCT and NCI60 datasets

g^*	Number of TPCR components						Number of PLS components					
	1	2	3	4	5	6	1	2	3	4	5	6
Leukemia dataset (%, LHOCV) ($\lambda = 19$)												
50	14.97	**4.19**	5.83	6.72	6.61	6.50	15.00	**4.53**	5.28	5.19	5.36	6.08
100	15.08	**4.08**	4.61	5.00	5.31	5.19	15.22	**4.17**	4.36	4.50	4.75	5.14
200	15.19	3.89	**3.75**	3.92	4.89	4.97	15.28	**4.03**	4.36	4.53	4.56	4.89
500	14.86	4.31	4.28	**4.06**	4.14	4.25	14.94	**4.36**	4.75	4.67	4.58	4.58
1000	14.72	5.25	4.83	4.44	**4.19**	4.19	14.72	4.72	4.89	**4.61**	4.61	4.72
Hereditary breast cancer dataset (%, LHOCV) ($\lambda = 9$)												
50	60.27	47.18	47.00	46.09	**45.55**	46.27	60.27	47.91	47.09	**46.27**	46.55	46.45
100	58.91	45.09	45.00	45.81	44.18	**43.82**	58.55	44.55	45.36	44.45	44.36	**44.18**
200	58.73	45.45	45.27	44.64	**44.00**	44.18	59.18	46.09	44.55	**44.00**	44.00	44.91
500	61.09	46.45	46.00	45.27	**44.18**	44.73	60.64	45.73	45.00	**44.45**	45.18	45.27
1000	64.64	54.73	49.00	48.27	47.55	**46.45**	63.55	50.91	48.91	47.73	**47.45**	47.45
SRBCT dataset (%, LHOCV) ($\lambda = 200$)												
50	40.85	18.63	0.83	**0.39**	0.56	0.59	45.07	20.98	0.73	**0.54**	0.56	0.68
100	41.29	18.80	0.66	**0.17**	0.29	0.22	46.54	21.66	0.61	**0.22**	0.39	0.37
200	42.59	19.22	1.07	0.32	**0.27**	0.32	49.66	21.54	0.90	**0.29**	0.39	0.39
500	43.83	20.71	2.20	0.78	0.56	**0.51**	52.71	23.63	1.88	0.66	**0.56**	0.63
1000	45.61	22.95	4.54	1.85	1.15	**0.80**	54.88	26.22	4.07	1.37	0.71	**0.61**
NCI60 dataset (%, LHOCV) ($\lambda = 20$)												
50	72.65	55.24	36.12	**13.18**	13.41	13.88	71.00	53.12	34.47	14.18	**13.88**	14.53
100	72.18	54.00	34.41	**11.12**	11.35	11.41	70.24	52.29	32.82	11.47	11.65	**11.41**
200	70.94	52.59	33.18	**9.47**	9.82	10.12	70.47	51.41	30.82	**9.65**	10.65	10.65
500	70.59	50.94	31.35	**8.82**	9.24	9.53	70.65	50.47	30.71	**9.06**	9.53	9.76
1000	70.76	49.29	30.35	9.53	**9.00**	9.41	71.35	49.82	30.88	**9.00**	9.82	9.71

Given are the percentages of misclassifications averaged over 100 re-randomizations. (Bold number denotes the minimum value in the same row; bold and underlined number means the minimum value in the whole 5×6 data matrix).

good prediction results due to the high variance (45,46). To assess the prediction performance of TPCR more realistically, TPCR was further compared using LHOCV procedure with the well-known PLS method, which has been widely used in microarray analysis (13,15–17,41,57–60), including tumor classification. This LHOCV procedure was repeated 100 times and the average misclassification error rates over 100 re-randomizations for the four microarray datasets are shown in Table 5.

It is obvious that the error rates using LHOCV were higher than those using LOOCV since the size of the dataset is small and only half of the total samples were used to construct the classification model under LHOCV procedure. On the other hand, it can be observed that the minimum LHOCV error rate obtained by TPCR is consistently lower than that obtained by PLS for each of the four microarray datasets. Actually, with the same number of genes selected, the minimum LHOCV error rate by TPCR (bold number in Table 5) is, in most cases, lower than that by PLS, indicating that the prediction performance of TPCR is better than or at least comparable to the well-known PLS method.

Assessment of the reliability of classification models by permutation analysis

Given the relatively small sample size of microarray datasets in cancer classification, especially for hereditary breast cancer dataset and NCI60 dataset, it is important to evaluate the stability and reliability of a classification model. There are various statistical methods to assess the reliability when there are not enough samples available to perform external validation (13–15,41). In this paper, so-called permutation or shuffle studies were performed to compare the misclassification error rates using TPCR with those expected at random. Initially, the class memberships of all the samples were permuted (the rows of **Y** matrix were shuffled) while keeping the gene expression profiles (**X** matrix) unchanged; then the newly generated random dataset with shuffled **Y** and unchanged **X** was analyzed

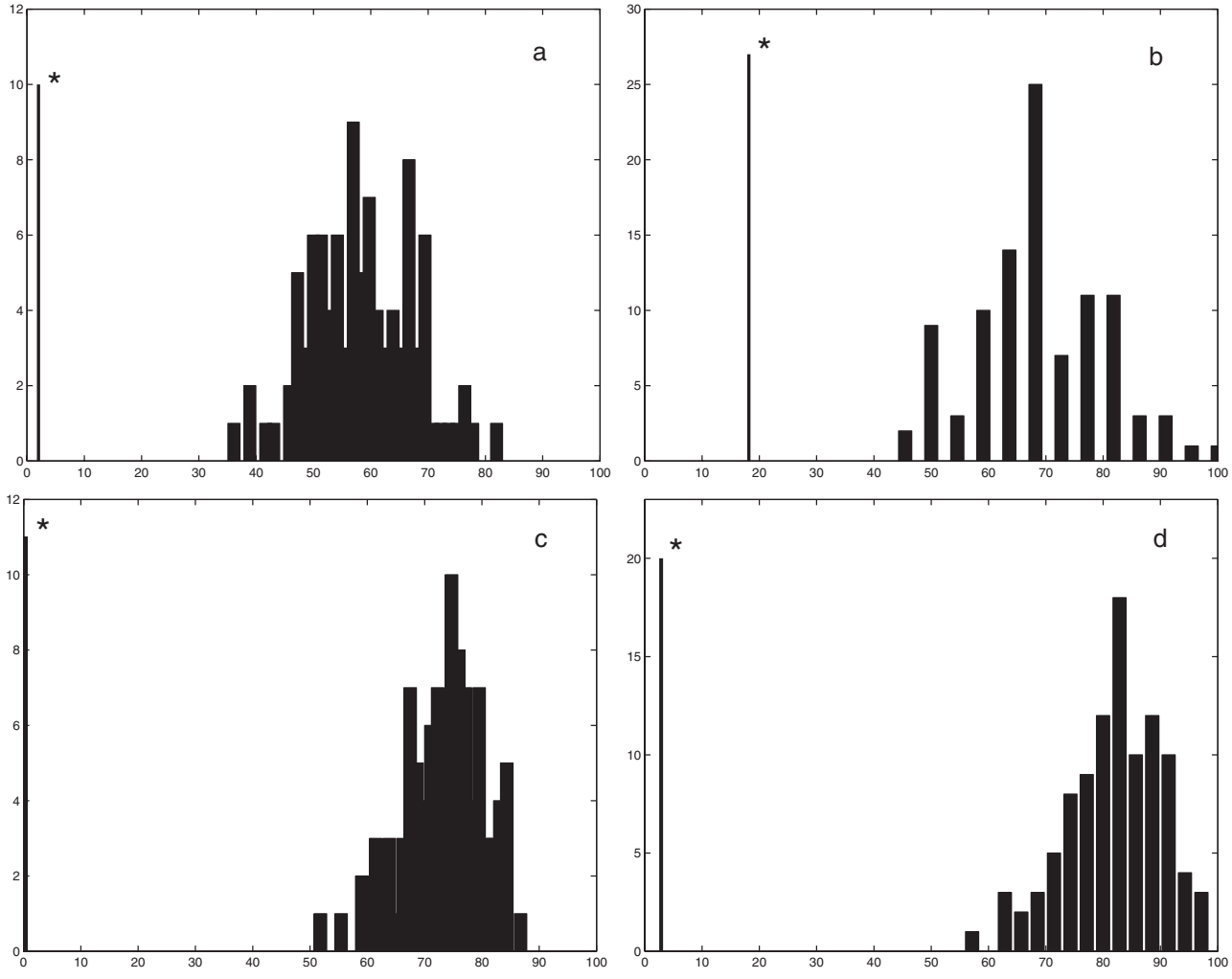

Figure 1. Distribution of error rate (percentage of misclassified samples) over 100 runs of permutation analysis (the original class memberships of all samples were randomly shuffled for 100 times and then used together with original gene expression profiles for classification by TPCR using the same LOOCV as applied before for original dataset). The solid line with asterisk labeled in each plot represents the minimum error rate using LOOCV for each original dataset: (**a**) Leukemia dataset (number of selected genes, TPCR component number, λ and corresponding LOOCV error rate for the original dataset: 200, 4, 19 and 1.39%); (**b**) Hereditary breast cancer dataset (200, 3, 9 and 18.18%); (**c**) SRBCT (50, 4, 200 and 0.00%); and (**d**) NCI60 (100, 4, 20 and 2.86%).

by TPCR using exactly the same LOOCV procedure as applied before to the original dataset (gene number, TPCR component number and meta parameter λ were the same as those chosen to obtain the minimum error rates for original datasets, as shown in Tables 1–4). This procedure was carried out 100 times and the distributions of the error rates over 100 permutations for the four datasets are plotted in Figure 1 and compared with the minimum misclassification error rates obtained from original datasets. It is obvious that, in all cases, the estimated error rate obtained by TPCR for original dataset is significantly lower than what would be expected at random. This kind of permutation analysis can be used to test whether a complete cross-validation is performed as well as whether there is some real structure or classification information inside the dataset. If an incomplete LOOCV is applied or a dataset with no classification information (e.g. random dataset) is analyzed, the estimated error rate obtained from original dataset will be close to that calculated from the shuffled dataset.

Another randomization analysis applied to further assess the stability and reliability of a classification model is to examine the distribution of the error rates over 100 re-randomizations obtained using LHOCV (14,15), especially when there is inadequate sets of data. If the classification model is unstable during the 100 re-randomization or perturbations introduced by this procedure, the estimate of the predictive ability is unlikely to be reliable (46). The distribution plots of misclassification error rates over 100 re-randomizations using LHOCV for the four microarray datasets are shown in Figure 2 (genes, the number of TPCR components and meta parameter λ were chosen according to the minimum averaged LHOCV error rate for each dataset). Substantial stability of the classification models can be observed on both Leukemia (Figure 2a) and SRBCT (Figure 2c) datasets with small averaged error rate and variance, while the classification model on hereditary breast cancer dataset (Figure 2b) is unstable with relatively large averaged error rate and variance, possibly due to the small

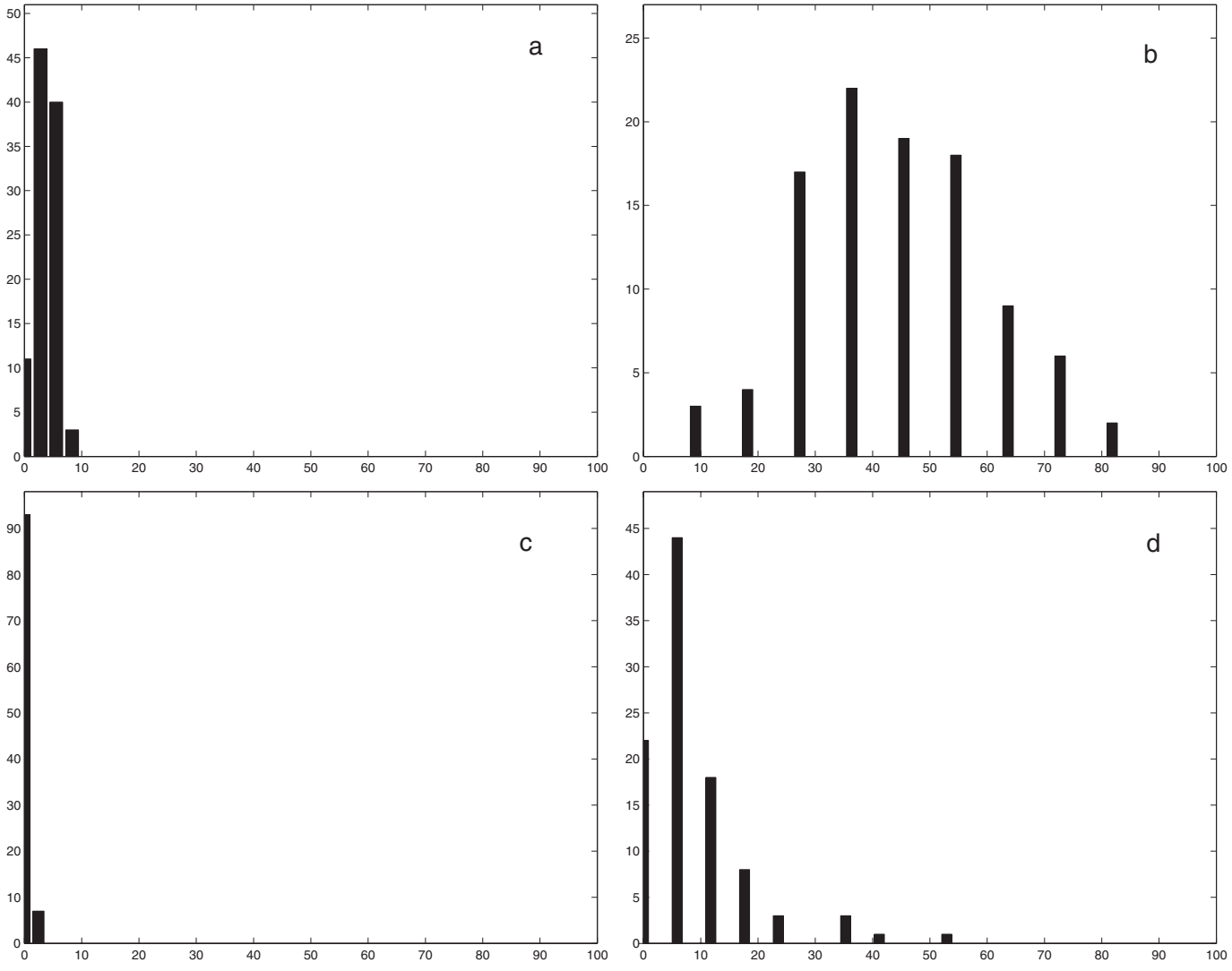

Figure 2. Distribution of error rate (percentage of misclassified test samples) over 100 runs of LHOCV or re-randomization analysis using TPCR (the original dataset was randomly split, half/half, into training and test samples for 100 times, then the new generated training dataset was used to predict the test dataset using TPCR): (**a**) Leukemia dataset (number of selected genes, number of TPCR components, λ and corresponding averaged LHOCV error rate: 200, 3, 19 and 3.75%); (**b**) Hereditary breast cancer dataset (100, 6, 9 and 43.82%); (**c**) SRBCT (100, 4, 200 and 0.17%); and (**d**) NCI60 (500, 4, 20 and 8.82%).

Table 6. Comparison of the speed of fast kernel EVD algorithm and classic SVD algorithm

Algorithm	Hereditary breast cancer (size: 22 × 3226)	Leukemia (72 × 7129)	SRBCT (83 × 2308)	NCI60 (35 × 1299)
SVD	9.7188	141.3125	16.3751	2.2969
Kernel EVD	0.0156	0.2031	0.0625	0.0156
Speed gain (fold)	623	696	262	147

Given are the time (second) used to calculate the singular vectors **U** from microarray gene expression profiles (**X** matrix).

sample size or inherent difficulty to discriminate the tumors. Taken together, the reliability of the four classification models is in the decreasing order: SRBCT > Leukemia > NCI60 > Hereditary breast cancer, the same order as obtained previously using PLS (15).

Comparison of the speed of classic SVD algorithm and fast kernel EVD algorithm

The classic SVD algorithm ([**U**, **S**, **V**] = **svd** (**X**), build-in function in MATLAB) and a fast kernel EVD algorithm originated from the work of Wu *et al*. (39), [**U**, Σ^2] = **eig** ($\mathbf{XX}^T$), were compared in terms of the time used to calculate the left singular vectors **U** from the four microarray gene expression profiles (**X** matrix), and the results are shown in Table 6 (Intel Pentium 4 1.70 GHz CPU; 512 MB RAM; Window 2000 Pro; MATLAB R14). It is clear that the fast kernel EVD algorithm is much faster than the SVD algorithm, with a speed increase ranging from 147 (NCI60) to 696 times (Leukemia). Generally speaking, the improvement of the speed increases as the size of the dataset increases. Considering that the time-consuming cross-validation procedure was applied to select optimal meta parameter λ and to evaluate the method, the improvement of speed is very significant, especially for LHOCV in which 100 runs of re-randomizations were performed.

CONCLUSIONS

In this paper, based on the EIV model, we proposed a novel method called TPCR to perform multi-class classification of tumor samples by taking into account the errors in microarray gene expression profiles. In addition, the latent variable structure underlying the microarray data was also taken into account in our method to mitigate the collinearity that resulted from the high-dimensional microarray data, which shows a salient advantage of our method over the classical EIV model. Four well-known microarray datasets were used to demonstrate that the performance of TPCR is better than or at least comparable to other published methods in the classification of tumors. A major advantage of TPCR over other methods is that it takes into account not only the errors in both independent variables and dependent variables but also the latent structure from the augmented subspace of independent and dependent variables. A fast kernel EVD algorithm was applied to decrease dramatically the time needed to extract the latent variables. The reported error rates of classification model using TPCR from the original datasets were shown to be significantly lower than what would be expected by chance from shuffling or permuting the memberships of samples. By half-half randomly splitting the original dataset into training and test samples for 100 times, the stability and reliability of the classification models were assessed by the distributions of the error rates over 100 re-randomizations and were shown in the decreasing order: SRBCT > Leukemia > NCI60 > Hereditary breast cancer.

ACKNOWLEDGEMENTS

We are grateful to Dr Robert R. Delongchamp of the FDA's National Center for Toxicological Research for carefully reviewing the manuscript. Y.T. and C.W. were partially supported by General Clinical Research Center grant M01-RR00425 from the National Center for Research Resources. Funding to pay the Open Access publication charges for this article was provided by Burns and Allen Research Institute Microarray core, Cedars-Sinai Medical Center.

REFERENCES

1. Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.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.
2. Brown,P.O. and Botstein,D. (1999) Exploring the new world of the genome with DNA microarrays. *Nature Genet*., **21**, 33–37.
3. Shi,L.M. (2001) Arrays, molecular diagnostics, personalized therapy and informatics. *Expert Rev. Mol. Diagn*., **1**, 363–365.
4. Young,R.A. (2000) Biomedical discovery with DNA arrays. *Cell*, **102**, 9–15.
5. Toronen,P., Kolehmainen,M., Wong,G. and Castren,E. (1999) Analysis of gene expression data using self-organizing maps. *FEBS Lett*., **451**, 142–146.
6. Eisen,M.B., Spellman,P., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expressed patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
7. Brazma,A. and Vilo,J. (2000) Gene expression data analysis. *FEBS Lett*., **480**, 17–24.
8. Brown,M.P., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M.,Jr and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.
9. Ross,D.T., Scherf,U., Eisen,M.B., Perou,C.M., Rees,C., Spellman,P., Iyer,V., Jeffrey,S.S., Rijn,M.V., Waltham,M. *et al*. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet*., **24**, 227–235.
10. 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.
11. Khan,J., Wei,J.S., Ringner,M., Saal,L.H., Ladanyi,M., Westermann,F., Berthold,F., Schwab,M., Antonescu,C.R., Peterson,C. and Meltzer,P.S. (2001) Classification and diagnostic prediction of cancers using expression profiling and artificial neural networks. *Nature Med*., **7**, 673–679.
12. Stephanopoulos,G., Hwang,D., Schmitt,W.A., Misra,J. and Stephanopoulos,G. (2002) Mapping physiological states from microarray expression measurements. *Bioinformatics*, **18**, 1054–1063.
13. Nguyen,D.V. and Rocke,D.M. (2002) Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, **18**, 1216–1226.
14. Bicciato,S., Luchini,A. and Bello,C.D. (2003) PCA disjoint models for multiclass cancer analysis using gene expression data. *Bioinformatics*, **19**, 571–578.
15. Tan,Y.X., Shi,L., Tong,W., Hwang,J.T.G. and Wang,C. (2004) Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models. *Comput. Biol. Chem*., **28**, 235–244.

16. Ding,B.Y. and Gentleman,R. (2004) Classification using generalized partial least squares. *J. Comput. Graph. Stat.*, in press.
17. Fort,G. and Lambert-Lacroix,S. (2004) Classification using partial least squares with penalized logistic regression. *Bioinformatics*, in press.
18. Jolliffe,I.T. (1986) *Principal Component Analysis*. Springer-Verlag, New York.
19. Wold,S., Ruhe,A., Wold,H. and Dunn,W.J.,III (1984) The collinearity problem in linear regression, the partial least squares approach to generalized inverse. *SIAM J. Sci. Stat. Comput.*, **5**, 735–743.
20. Churchill,G.A. (2002) Fundamentals of experimental design for cDNA microarrays. *Nature Genet.*, **32** (Suppl.), 490–495.
21. Novak,J.P., Sladek,R. and Hudson,T.J. (2002) Characterization of variability in large-scale gene expression data: implication for study design. *Genomics*, **79**, 104–113.
22. Chen,J.J., Delongchamp,R.R., Tsai,C.A., Hsueh,H.M., Sistare,F., Thompson,K.L., Desai,V.G. and Fuscoe,J.C. (2004) Analysis of variance components in gene expression data. *Bioinformatics*, **20**, 1436–1446.
23. Kerr,M.K., Martin,M. and Churchill,G.A. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.
24. Kerr,M.K., Afshari,C.A., Bennett,L., Bushel,P., Martinez,J., Walker,N.J. and Churchill,G.A. (2002) Statistical analysis of a gene expression microarray experiment with replications. *Statist. Sin.*, **12**, 203–217.
25. Oleksiak,M.F., Churchilll,G.A. and Crawford,D.L. (2002) Variation in gene expression within and among natural populations. *Nature Genet.*, **32**, 261–266.
26. Hökuldsson,A. (1988) PLS regression methods. *J. Chemom.*, **2**, 211–228.
27. Manne,R. (1987) Analysis of two partial-least-squares algorithms for multivariate calibration. *Chemo. Intell. Lab. Sys.*, **2**, 187–197.
28. Gleser,L.J. and Watson,G.S. (1973) Estimation of a linear transformation. *Biometrika*, **60**, 525–534.
29. Gleser,L.J. (1981) Estimation in a multivariate 'error-in-variables' regression model: large sample results. *Ann. Stat.*, **9**, 24–44.
30. Gleser,L.J. (1991) Measurement error models. *Chemo. Intell. Lab. Sys.*, **10**, 45–57.
31. Anderson,T.W. (1976) Estimation of linear functional relationships: approximate distributions and connetions with simultaneous equations in econometrics. *J. R. Stat. Soc. Ser. B*, **38**, 1–36.
32. Anderson,T.W. (1984) The 1982 Wald memorial lectures: estimating linear statistical relationships. *Ann. Stat.*, **12**, 1–45.
33. Martens,H. and Næs,T. (1989) *Multivariate Calibration.* Wiley, New York.
34. Burnham,A.J., MacGregor,J.F. and Viveros,R. (1999) Latent variable multivariate regression modeling. *Chemo. Intell. Lab. Sys.*, **48**, 167–180.
35. Burnham,A.J., MacGregor,J.F. and Viveros,R. (1999) A statistical framework for multivariate latent variable regression methods based on maximum likelihood. *J. Chemom.*, **13**, 49–65.
36. Rao,C.R. and Toutenburg,H. (1995) *Linear Models: Least Squares and Alternatives.* Springer, New York.
37. de Jong,S. and Kiers,H.A.A. (1992) Principal covariates regression. *Chemo. Intell. Lab. Syst.*, **14**, 155–164.
38. Burnham,A.J., MacGregor,J.F. and Viveros,R. (2001) Interpretation of regression coefficients under a latent variable regression model. *J. Chemom.*, **15**, 265–284.
39. Wu,W., Massart,D.L. and de Jong,S. (1997) The kernel PCA algorithms for wide data. Part I: theory and algorithms. *Chemo. Intell. Lab. Syst.*, **36**, 165–172.
40. Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
41. Nguyen,D.V. and Rocke,D.M. (2002) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39–50.
42. Ambroise,C. and McLachlan,G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA*, **99**, 6562–6566.
43. Simon,R., Radmacher,M.D., Dobbin,K. and McShane,L.M. (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl Cancer Inst.*, **95**, 14–18.
44. Simon,R. (2003) Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. *Br. J. Cancer*, **89**, 1599–1604.
45. Efron,B. (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Stat. Assoc.*, **78**, 316–330.
46. Kohavi,R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI, Montreal, Canada.
47. Efron,B. and Tibshirani,R. (1997) Improvements on cross-validation: the .632+ bootstrap method. *J. Am. Stat. Assoc.*, **92**, 548–560.
48. Weiss,S.M. (1991) Small sample error rate estimation for *k*-nearest neighbor classifiers. *IEEE Trans. Patt. Anal. Mach. Intell.*, **13**, 285–289.
49. Hedenfalk,I., Duggan,D., Chen,Y.D., Radmacher,M., Bittner,M., Simon,R., Meltzer,P., Gusterson,B., Esteller,M., Kallioniemi,O.P., Wilfond,B., Borg,A. and Trent,J. (2001) Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.*, **344**, 539–548.
50. Cho,J.H., Lee,D., Park,J.H. and Lee,I.B. (2003) New gene selection method for classification of cancer subtypes considering within-class variation. *FEBS Lett.*, **551**, 3–7.
51. Fu,L.M. and Fu-Liu,C.S. (2004) Multi-class cancer subtype classification based on gene expression signatures with reliability analysis. *FEBS Lett.*, **561**, 186–190.
52. Lee,Y.K. and Lee,C.K. (2003) Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*, **19**, 1132–1139.
53. Tibshirani,R., Hastie,T., Narasimhan,B. and Chu,G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.
54. 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. *et al.* (2000) A gene expression database for the molecular pharmacology of cancer. *Nature Genet.*, **24**, 236–244.
55. Stinson,S.F., Alley,M.C., Kopp,W.C., Fiebig,H.H., Mullendore,L.A., Pittman,A.F., Kenney,S., Keller,J. and Boyd,M.R. (1992) Morphological and immunocytochemical characteristics of human tumor cell lines for used in a disease-oriented anticancer drug screen. *Anticancer Res.*, **12**, 1035–1053.
56. Shi,L.M., Fan,Y., Lee,J.K., Waltham,M., Andrews,D.T., Scherf,U., Paull,K.D. and Weinstein,J.N. (2000) Mining and visualizing large anticancer drug discovery databases. *J. Chem. Inf. Comput. Sci.*, **40**, 367–379.
57. Huang,X.H. and Pan,W. (2003) Linear regression and two-class classification with gene expression data. *Bioinformatics*, **19**, 2072–2078.
58. Park,P.J., Tian,L. and Kohane,I.S. (2002) Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, **18**, S120–S127.
59. Pérez-Enciso,M. and Tenenhaus,M. (2003) Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. *Hum. Genet.*, **112**, 581–592.
60. Fort,G. and Lambert-Lacroix,S. (2004) Ridge-partial least squares for generalized linear models with binary response. In *Proceedings of the 16th Symposium of IASC on Computational Statistics (COMPSTAT'04)*, Prague, Czech Republic, August 23–27, pp. 1019–1026.