Statistical Applications in Genetics and Molecular Biology

Volume 5, *Issue* 1 2006 *Article* 6

Dimension Reduction for Classification with Gene Expression Microarray Data

Jian J. Dai, *University of California, Davis*
Linh Lieu, *University of California, Los Angeles*
David Rocke, *University of California, Davis*

Recommended Citation:
Dai, Jian J.; Lieu, Linh; and Rocke, David (2006) "Dimension Reduction for Classification with Gene Expression Microarray Data," *Statistical Applications in Genetics and Molecular Biology*: Vol. 5: Iss. 1, Article 6.
DOI: 10.2202/1544-6115.1147

Dimension Reduction for Classification with Gene Expression Microarray Data

Jian J. Dai, Linh Lieu, and David Rocke

Abstract

An important application of gene expression microarray data is classification of biological samples or prediction of clinical and other outcomes. One necessary part of multivariate statistical analysis in such applications is dimension reduction. This paper provides a comparison study of three dimension reduction techniques, namely partial least squares (PLS), sliced inverse regression (SIR) and principal component analysis (PCA), and evaluates the relative performance of classification procedures incorporating those methods. A five-step assessment procedure is designed for the purpose. Predictive accuracy and computational efficiency of the methods are examined. Two gene expression data sets for tumor classification are used in the study.

KEYWORDS: partial least squares, sliced inverse regression, feature extraction, gene expression, tumor classification

Author Notes: The research reported in this paper was supported by grants from the National Science Foundation (ACI 96-19020, and DMS 98-70172), the National Institute of Environmental Health Sciences (P43-ES04699 and P01-ES11269), the National Cancer Institute (P30-CA093373), and the UC Davis Health System. We thank the editor and two reviewers for helpful comments and suggestions.

1. Introduction

A characteristic of gene expression microarray data is that the number of variables (genes) p far exceeds the number of samples n, commonly known as the "large p, small n" problem (West et al., 2001; Dudoit et al., 2002). In addition, the gene expression measures can be highly correlated. These features present a challenge to modeling the relationship between phenotype and gene expression profiles and classifying (predicting) samples into known categories such as tumor types in cancer research (Speed, 2003). There are several ways to deal with the problem. One can reduce dimension of the data by selecting a subset of interesting genes (gene selection), or producing gene components or super genes – combinations of genes (dimension reduction), or using combination of the strategies. Gene selection is usually based on some univariate measure related to the classification (e.g. Hedenfalk et al., 2001; Dettling and Buhlmann, 2003). Gene components can be constructed using multivariate techniques with the premise that, although the microarray data contain numerous genes, there may be actually a small number of underlying variables that account for most of the variation in the data (West et al., 2001). For example, a few linear combinations of genes may explain most of the response variation. Each approach has its own advantages and limitations (Boulesteix, 2004). A combination of the strategies is often used in practice for classification with gene expression data. Such classification procedures often consist of the following steps: the first step is gene selection/dimension reduction, in which a few gene components are constructed from a large number of genes; the second step is classification, in which the samples are classified into categories by applying standard statistical models on the gene components (Nguyen and Rocke, 2002a, 2002b).

Dimension reduction is a subject of study in several research areas including high-dimensional data analysis, pattern recognition, and machine learning, where one seeks to explain observed high-dimensional data using an underlying low-dimensional representation. Dimension reduction has many applications in bioinformatics and computational biology. The purpose of this study is to evaluate some of those recently proposed for tumor classification with gene expression data. Specifically, we focus on three dimension reduction methods: partial least squares (PLS) (Nguyen and Rocke, 2002a, 2002b; Huang and Pan, 2003; Boulesteix, 2004), sliced inverse regression (SIR) (Chiaromonte and Martinelli, 2002; Antoniadis et al., 2003; Bura and Pfeiffer, 2003), and principal component analysis (PCA) (Ghosh, 2002). These methods have been shown highly useful for classification with gene expression data. However, there is lack of comparison studies on those methods. For example, the relative performance of PLS and SIR dimension reduction for classification is largely unknown.

In this paper, we evaluate the relative performance of several classification procedures incorporating those dimension reduction methods (PLS, SIR and PCA). We discuss the methodological presumptions of the methods, address issues involved in comparing the models, design a five-step assessment procedure, and present results of evaluations based on two gene expression data sets in cancer research: the leukemia data set of Golub et al. (1999) and the colon data set of Alon et al. (1999). The paper is organized as follows. In Section 2, we describe the methods of dimension reduction, classification, gene selection, model selection and validation, and design a procedure for assessing the relative performance of the models. In Section 3, we describe the microarray data sets and the experiments, and present the results of evaluations. Summaries and discussions are presented in Section 4.

2. Methods

The application context is prediction of response classes such as tumor types using gene expression microarray data. We view the problem as a multivariate regression problem where the number of variables far exceeds the number of observations (Stone and Brooks, 1990; Frank and Friedman, 1993; Krzanowski, 1995; Kiers, 1997). A classification procedure for the purpose may consist of two basic steps: the first step is dimension reduction, in which the data are reduced from the high p-dimensional gene space to a lower K-dimensional ($K<n$) gene component space; the second step is class prediction, in which response classes are predicted using a standard class prediction model on the gene components. A step of preliminary gene selection can be easily incorporated into the procedure. In this section, we first discuss three dimension reduction methods (PLS, SIR and PCA) and a standard classification model (logistic discrimination), and then describe the methods for gene selection, model selection and validation, and finally design and present a five-step procedure for model assessment.

2.1. Dimension Reduction: PCA, PLS and SIR

One way to achieve dimension reduction is to transform the large number of original variables (genes) to a new set of variables (gene components), which are uncorrelated and ordered so that the first few account for most of the variation in the data. The K new variables (gene components) can then replace the initial p variables (genes), thereby reducing the data from the high p-dimension to a lower K-dimension. PCA, PLS and SIR are three of such methods for dimension reduction. To describe them, let $\mathbf{X}$ be the $n \times p$ matrix of n tissue samples and p

DOI: 10.2202/1544-6115.1147

genes, **y** be the $n \times 1$ vector of response values, and $\mathbf{S}_x$ be the $p \times p$ covariance matrix of the gene expressions.

2.1.1. Principal Component Analysis

PCA is a well-known method of dimension reduction (Jolliffe, 1986). The basic idea of PCA is to reduce the dimensionality of a data set, while retaining as much as possible the variation present in the original predictor variables. This is achieved by transforming the p original variables $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p]$ to a new set of K predictor variables, $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_K]$, which are linear combinations of the original variables. In mathematical terms, PCA sequentially maximizes the variance of a linear combination of the original predictor variables,

$$\mathbf{u}_K = \arg\max_{u'u=1} Var(\mathbf{Xu}) \qquad (1)$$

subject to the constraint $\mathbf{u}_i' \mathbf{S}_X \mathbf{u}_j = 0$, for all $1 \le i < j$. The orthogonal constraint ensures that the linear combinations are uncorrelated, i.e. $Cov(\mathbf{Xu}_i, \mathbf{Xu}_j) = 0$, $i \ne j$. These linear combinations

$$\mathbf{t}_i = \mathbf{Xu}_i \qquad (2)$$

are known as the principal components (PCs) (Massey, 1965). Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system. The new axes represent the directions with maximum variability and are ordered in terms of the amount of variation of the original data they account for. The first PC accounts for as much of the variability as possible, and each succeeding component accounts for as much of the remaining variability as possible. Computation of the principal components reduces to the solution of an eigenvalue-eigenvector problem. The projection vectors (or called the weighting vectors) **u** can be obtained by eigenvalue decomposition on the covariance matrix $\mathbf{S}_X$,

$$\mathbf{S}_X \mathbf{u}_i = \lambda_i \mathbf{u}_i \qquad (3)$$

where λ_i is the i-th eigenvalue in the descending order for $i=1,\ldots,K$, and $\mathbf{u}_i$ is the corresponding eigenvector. The eigenvalue λ_i measures the variance of the i-th PC and the eigenvector $\mathbf{u}_i$ provides the weights (loadings) for the linear

transformation (projection). The maximum number of components K is determined by the number of nonzero eigenvalues, which is the rank of $\mathbf{S}_X$, and $K \leq \min(n,p)$. The computational cost of PCA, determined by the number of original predictor variables p and the number of samples n, is in the order of $min(np^2 + p^3,\ pn^2 + n^3)$. In other words, the cost is $O(pn^2 + n^3)$ when $p > n$.

2.1.2. Partial Least Squares

The objective of constructing components in PLS is to maximize the covariance between the response variable **y** and the original predictor variables **X**,

$$\mathbf{w}_K = \arg\max_{w'w=1} Cov(\mathbf{Xw}, \mathbf{y}) \tag{4}$$

subject to the constraint $\mathbf{w}_i'\mathbf{S}_X\mathbf{w}_j = 0$, for all $1 \leq i < j$. The central task of PLS is to obtain the vectors of optimal weights $\mathbf{w}_i$ (i=1,…,K) to form a small number of components that best predict the response variable **y**. Note that PLS is a "supervised" method because it uses information on both **X** and **y** in constructing the components, while PCA is an "unsupervised" method that utilizes the **X** data only.

To derive the components, [$\mathbf{t}_1$, $\mathbf{t}_2$, …, $\mathbf{t}_K$], PLS decomposes **X** and **y** to produce a bilinear representation of the data (Martens and Naes, 1989):

$$\mathbf{X} = \mathbf{t}_1\mathbf{w}'_1 + \mathbf{t}_2\mathbf{w}'_2 + ... + \mathbf{t}_K\mathbf{w}'_K + \mathbf{E} \tag{5}$$

and

$$\mathbf{y} = \mathbf{t}_1 q_1 + \mathbf{t}_2 q_2 + ... + \mathbf{t}_K q_K + \mathbf{F} \tag{6}$$

where **w**'s are vectors of weights for constructing the PLS components **t**=**Xw**, q's are scalars, and **E** and **F** are the residuals. The idea of PLS is to estimate **w** and q by regression. Specifically, PLS fits a sequence of bilinear models by least squares, thus given the name partial least squares (Wold, 1966, 1973, 1982).

At each step i (i=1,…,K), the vector $\mathbf{w}_i$ is estimated in such a way that the PLS component, $\mathbf{t}_i$, has maximal sample covariance with the response variable **y** subject to being uncorrelated with all previously constructed components. The first PLS component $\mathbf{t}_1$ is obtained based on the covariance between **X** and **y**. Each subsequent component $\mathbf{t}_i$ (i=2,…,K), is computed using the residuals of **X** and **y** from the previous step, which account for the variations left by the previous components. As a result, the PLS components are uncorrelated and ordered (Garthwaite, 1994; Helland, 1988, 1990).

DOI: 10.2202/1544-6115.1147

The maximum number of components, K, is less than or equal to the smaller dimension of $\mathbf{X}$, i.e. $K \leq \min(n,p)$. The first few PLS components account for most of the covariation between the original predictors and the response variable and thus are usually retained as the new predictors. The computation of PLS is simple and a number of algorithms are available (Martens and Naes, 1989). In this study, we used a standard PLS algorithm (Denham, 1995).

Like PCA, PLS reduces the complexity of microarray data analysis by constructing a small number of gene components, which can be used to replace the large number of original gene expression measures. Moreover, obtained by maximizing the covariance between the components and the response variable, the PLS components are generally more predictive of the response variable than the principal components.

The number of components, K, to be used in the class prediction model is considered to be a meta parameter and must be estimated in the application, which we will discuss later. PLS is computationally very efficient with cost only at $O(np)$, i.e. the number of calculations required by PLS is a linear function of n and p. Thus it is much faster than the other two methods (PCA and SIR).

2.1.3. Sliced Inverse Regression

SIR, one of the sufficient dimension reduction methods (Li, 1991, 2000; Duan and Li, 1991; Cook 1998), is a supervised approach, which utilizes response information in achieving dimension reduction. The idea of SIR is simple. Conventional regression models deal with the forward regression function, $\mathrm{E}(\mathbf{y}|\mathbf{X})$, which is a p-dimensional problem and difficult to estimate when p is large. SIR is based on the inverse regression function,

$$\boldsymbol{\eta}(\mathbf{y}) = \mathrm{E}(\mathbf{X} \mid \mathbf{y}) \tag{7}$$

which consists of p one-dimensional regressions and is easier to deal with. The SIR directions $\mathbf{v}$ can be obtained as the solution of the following optimization problem,

$$\mathbf{v}_K = \arg\max_{\mathbf{v}'\mathbf{v}=1} \frac{\mathbf{v}' Cov(E(\mathbf{X} \mid \mathbf{y}))\mathbf{v}}{\mathbf{v}' \mathbf{S}_X \mathbf{v}} \tag{8}$$

subject to the constraint $\mathbf{v}_i' \mathbf{S}_X \mathbf{v}_j = 0$, for all $1 \leq i < j$. Algebraically, the SIR components $\mathbf{t}_i = \mathbf{X}\mathbf{v}_i$ (i=1,…,K) are linear combinations of the p original predictor variables defined by the weighting vectors $\mathbf{v}_i$. Geometrically, SIR projects the data

from the high *p*-dimensional space to a much lower *K*-dimensional space spanned by the projection vectors **v**. The projection vectors **v** are derived in such a way that the first a few represent directions with maximum variability between the response variable and the SIR components. Computation of $\mathbf{v}_i$ is straightforward. Let $\mathbf{S}_\eta = Cov(\mathrm{E}(\mathbf{X} \mid \mathbf{y}))$ be the covariance matrix of the inverse regression function defined in (7) and recall that $\mathbf{S}_X$ is the variance-covariance matrix of **X**. The vectors $\mathbf{v}_i$ (i=1,…,K) can be obtained by spectral decomposition of $\mathbf{S}_\eta$ with respect to $\mathbf{S}_X$,

$$\mathbf{S}_\eta \mathbf{v}_i = \lambda_i \mathbf{S}_X \mathbf{v}_i \qquad (9)$$

where λ_i is the i-th eigenvalue in descending order for i=1,…,K; $\mathbf{v}_i$ is the corresponding eigenvector, and $\mathbf{v}_i' \mathbf{S}_X \mathbf{v}_j = 1$.

SIR is implemented by appropriate discretization of the response. Let T(**y**) be a discretization of the range of **y**. SIR computes Cov(E(**X**|T(**y**))), the covariance matrix for the slice means of **X**, which can be thought of as the between covariance for the subpopulations of **X** defined by T(**y**). Usually, if the response is continuous, one divides its range into H slices. If the response is categorical, one simply considers its categories. In class prediction problems, the number of classes *G* is a natural choice for *H*, i.e. *H*=*G*. The maximum number of SIR components is *H* minus one, i.e. $K \leq \min(H\text{-}1, n, p)$. As discussed before, *K* is considered to be a meta-parameter and may be estimated by cross-validation. The cost of computing SIR directions using the standard algorithm is $O(np^2 + p^3)$, which is quite expensive comparing to the cost of PLS. We used a standard SIR algorithm (Härdle et al., 1995) in this study.

2.2. Class Prediction: Logistic Discrimination

After dimension reduction, standard statistical models can be used for class prediction based on a small number of new predictors. The class prediction model we use for this study is the logistic discrimination (LD). This model has been widely used for two-calss prediction problems and has been shown to perform well in previous studies (e.g. Nguyen and Rocke, 2002a). A number of statistical models can be used for the purpose (e.g. Dudoit et al., 2002).

To describe the model, let **Z** be the *n* by *K* matrix of predictor values (gene components) and **y** be the vector of binary responses (class labels), for example, y =1 for tumor type A, and y=0 for tumor type B. We want to predict the

DOI: 10.2202/1544-6115.1147

probability that the *i*-th tissue sample is of tumor type A given the gene expression profile $\mathbf{Z}_i$

$$\pi_i = P(\mathbf{y}_i = 1 / \mathbf{Z}_i) \tag{10}$$

and then use the probability to classify the sample. In LD, this probability is computed using the logistic function (Hosmer and Lemeshow, 2000)

$$\pi_i = \frac{\exp(\mathbf{Z}_i\boldsymbol{\beta})}{1+\exp(\mathbf{Z}_i\boldsymbol{\beta})} \tag{11}$$

where $\boldsymbol{\beta}$ is a vector of coefficients in the logistic regression, whose values can be estimated by the method of maximum likelihood estimation (MLE). The predicted probabilities $\hat{\pi}$ are computed by replacing $\boldsymbol{\beta}$ with the MLE estimates $\hat{\boldsymbol{\beta}}$. These probabilities are then used to classify each of the samples, *i* (for *i*=1,…,*n*), $\hat{y}_i = \mathbf{1}(\hat{\pi}_i > 1 - \hat{\pi}_i)$, where $\mathbf{1}(.)$ is the indicator function. $\mathbf{1}(A)$=1 if condition A is true and $\mathbf{1}(A)$=0 otherwise. The classification rule is simple: a tumor sample is classified as type A ($\hat{y}_i = 1$) if the predicted probability that the sample is of type A is greater than the probability that the sample is of type B; otherwise, the sample is classified as type B.

2.3. Methods of Assessment

With dimension reduction/logistic discrimination one can predict the response classes using gene expression data. The observed error rates can be used to compare the accuracy of the classifiers. Several issues need to be addressed in designing a procedure for the assessment. First, the procedure must provide protection against over-fitting the data. Cross-validation and re-randomization studies can be used for this purpose. Second, due to repeatedly fitting high dimensional data, the assessment studies can be very time-consuming. It would be useful to add a step of gene selection. Third, the number of gene components to be retained is a meta-parameter in the procedure and its value must be estimated. We now discuss these methods and design a five-step procedure for evaluating the relative performance of the classification procedures.

2.3.1. Cross-validation and re-randomization study

Cross-validation of prediction results can be achieved by leaving out part of the data, training a prediction rule on the remaining data (the training set), and predicting response values using the left-out data (the test set) (Stone, 1974). The prediction errors are used to evaluate the prediction accuracy of a model. Leave-one-out (LOO) cross-validation is often used when the number of samples in the data is relatively small. By this method, one of the samples is left out and a model is fitted based on all but the left-out sample. The fitted model is then used to predict the left-out sample. This is repeated for all samples. The error rate estimated through cross-validation is unbiased. It is important to treat dimension reduction as a step in building the prediction rule and therefore subject to cross-validation. We use cross-validation to choose the number of gene components (estimate the value of K).

Cross-validation provides some protection against overfitting the data, yet it may not be sufficient, because relatively small cross-validated errors can be achieved by capitalizing on chance properties. A further step to protect against overfitting is to do re-randomization studies. That is to re-randomize the entire data and then repeat the modeling and validation steps. Re-randomization studies help stabilize prediction errors.

2.3.2 Selection of gene subset

Although dimension reduction via PLS, SIR or PCA can handle a large number of genes, it is useful to include gene subset selection as part of the procedure. First, the assessment studies require fitting the data many times due to cross-validation and re-randomizations. A vary large p (number of genes) can be an impediment to the studies due to large computational time and other challenges. A usual approach to this is to select a subset of genes and use the subset for model comparisons. Second, it is often the case that only a subset of genes is of interest in practice. Thus, we include gene subset selection into the procedure and use subsets of genes in the assessment studies.

There are different methods for subset selection and each has its own limitations (Parmigiani et al., 2003). The simplest and fastest one is to form random subsets, each consisting of p^* ($p^* < p$) genes from the set of all genes. This can be done by random partition of the whole gene set or simple random sampling. The use of random subsets works for the purpose of this study, i.e. assessing the relative performance of the models, which doesn't require a gene subset to be "optimal". In other words, regardless of whether a subset contains "good" or "bad" (more or less predictive) genes, all models will be applied to the same subset of genes, and thus the comparison of model performance is valid.

DOI: 10.2202/1544-6115.1147

A subset of genes can also be selected based on measures related to the classification. Use of the class information in gene selection can help select genes whose expressions are more correlated to the response and thus improve prediction accuracy of the models. Most gene selection methods use some univariate measures related to classification. For a two-class application, gene selection can be based on the simple t-statistic (Nguyen and Rocke, 2002a):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2 / n_1 + s_2^2 / n_2}}$$

where n_k, $\bar{x}_k$ and s_k^2 are the size, mean and variance of class k, k=1,2. Using this method, t-scores are computed for all genes and the top p* genes with the best scores are retained. We use both random subset selection and the t-score based gene selection in the assessment studies.

2.3.3. Selecting the number of components

The number of components (K) is a meta parameter in the procedure. It can be estimated by cross-validation (CV) on the learning set using leave-one-out (LOO) or leave percentage out procedures. The leave-one-out validation procedure is as follows: one of the samples in the learning set is left-out, and a subset of genes (p^* genes, $p^* < p$) is selected. The models are fitted to all but the left-out sample. The fitted models are then used to predict the left-out sample. This is repeated for all samples in the learning data set with K taking successively different values. The predicted residual sum of squares (PRESS) is computed for each value of K, and the one that minimizes PRESS is chosen and denoted as K*. In our studies, the K values are set to be 1,2,3,4,5, which seems to be a good balance between computation time and estimation accuracy for binary classification (Boulesteix, 2004; Nguyen and Rocke, 2002b).

2.3.4. Assessment procedure

The assessment procedure consists of the following steps:

1. Form a learning set L with n_L samples and a test set T with n_T samples (n_L+n_T=n). Denote X_L as the learning data matrix of size n_L by p, and X_T as the test data matrix of size n_T by p. Use the learning set to determine the number of gene components, K*, by cross-validation (See 2.3.3).

2. Select a subset of p* genes from the set of all genes using one of the gene selection methods, resulting in X^*_L (n_L by p^* matrix) and X^*_T (n_T by p^* matrix).

3. Perform dimension reduction using PLS, SIR, or PCA. Let W denote the p* by K* matrix containing the projection vectors. Compute the matrix Z_L of gene components for the learning data set: $Z_L = X^*_L \times W$, and the gene components for the test data set: $Z_T = X^*_T \times W$.

4. Fit the class prediction model (logistic regression) to the learning components Z_L. Predict the classes of samples in the test set using the fitted classifier and the test components, Z_T.

5. Repeat all above steps *R* times with re-randomizations of the whole data set. The total class prediction error (TCPE) for each method is computed by

$$TCPE = \sum_{r=1}^{R} \sum_{i=1}^{n_T} \mathbf{1}(y_i - \hat{y}_i)$$

where y is the observed response class, $\hat{y}$ is the predicted response class, ***1()*** is an indicator function, n_T is the number of test samples, and R is the number of re-randomization studies. The error rate (proportion of misclassification) is computed by $TCPE/(n_T \times R)$ based on the test data only.

3. Results

In this section, we present the results of evaluations of the relative performance of three classification procedures: PLSLD, SIRLD and PCALD, which combine PLS, SIR, PCA with logistic discrimination (LD). The assessment studies are based on two microarray data sets: the leukemia data set of Golub et al. (1999) and the colon cancer data set of Alon et al. (1999). Both data sets are from Affymetrix high density oligonucleotide microarrays and are publicly available. The leukemia data set contains 72 tissue samples on 7129 genes; the colon data set has 62 tissue samples on 2000 genes. We implemented the five-step assessment procedure in the R software environment (Ihaka and Gentleman,

DOI: 10.2202/1544-6115.1147

1996) and then applied it to each of the data sets. We describe the leukemia data set first.

The acute leukemia data contain 72 bone marrow samples on 7129 genes from patients with either acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML). The original data consist of a training set of 38 samples with 27 ALL and 11 AML and a test set of 34 samples with 20 ALL and 14 AML. We preprocessed the gene expression data using the standard procedure including background correction, transformation and normalization (Kerr, 2000; Rocke and Durbin, 2001). In data transformation, we applied the generalized logarithm (glog), $\ln(x+\sqrt{x^2+\lambda})$, where x is an intensity value (background corrected) and λ is a transformation parameter whose value can be estimated using the method of maximum likelihood (Durbin et al., 2002; Durbin and Rocke, 2003; Huber et al., 2002; Munson, 2001). The glog transformation is a generalization of and improvement over the log transformation as the latter can inflate the variance of the expression values near background.

After data preprocessing, we applied the 5-step assessment procedure on the data set. We considered p* = 200, 500, 1000 genes and used 100 random subsets (R=100) for each p*. We randomly split each subset of genes into two data sets: a training set with 36 samples (n_L=36) and a test set with 36 samples (n_T=36). We used leave-one-out cross-validation on the training set to determine the number of gene components $\hat{K}$, and the test set for evaluating prediction errors. In total, 3600 class predictions (*36*100*=3600) were made using each of the three classifiers based on 100 random subsets. The prediction error rates of the classifiers were computed.

Table 1. Classification error rates of the three methods on the leukemia data set with 36/36 split of tissue samples averaged over 100 randomization studies.

p*	PLSLD	SIRLD	PCALD
200	0.086 (0.051) (2.1)	0.096 (0.051) (1.0)	0.120 (0.069) (3.4)
500	0.059 (0.045) (1.8)	0.059 (0.041) (1.0)	0.087 (0.065) (3.6)
1000	0.045 (0.033) (1.9)	0.046 (0.032) (1.0)	0.068 (0.055) (3.7)

The estimated error rates are presented in Table 1. The standard deviation of an error rate is shown in the first parentheses and the average value of the estimated meta parameter (K*) is in the second parentheses. It can be seen from the table that the error rates decrease with the increase of size of gene subset (p*). At any of the subset sizes, the error rates of PLS and SIR based procedures (PLSLD and SIRLD) are similar and they are lower than that of the PCA based procedure (PCALD). It suggests that PLS and SIR are about equally effective comparing to each other and are more effective than PCA in dimension reduction. To compare the relative computational cost, we computed the ratios of computational time of PLSLD and SIRLD to that of PLSLD. The ratios are 7.8 for PCALD/PLSLD and 30.8 for SIRLD/PLSLD with p*=500. It is clear that the PLS based procedure is much faster than those based on SIR or PCA.

The second data set used in this study is the colon data, which consist of gene expressions of 2000 human genes with 62 colon tissue samples (40 tumor and 22 normal). In pre-processing the data, we did background correction, glog transformation and normalization. For assessment, we considered p*=200, 500 and 1000 genes and generated 100 random subsets for each p*. Each subset was randomly partitioned into two parts: a training set with 36 samples (n_L=36) and a test set with 26 samples (n_T=26). The training set was used for dimension reduction and model selection by leave-one-out cross-validation, and the test set was used for prediction. In total, 2600 (*26*100*) class predictions were made using each of the three classifiers and misclassification rates were computed. The error rates and their standard deviations, based on 100 randomization studies, are reported in Table 2. It can be seen from the table that the classes in the colon data are less well separated than the leukemia data, as noted in the previous studies (Nguyen and Rocke, 2002a; Antoniadis et al., 2003).

Table 2. Classification error rates of the three methods on the colon data set with 36/26 split of tissue samples averaged over 100 randomization studies.

p*	PLSLD	SIRLD	PCALD
200	0.167 (0.055) (2.2)	0.193 (0.054) (1.0)	0.224 (0.096) (3.5)
500	0.152 (0.063) (2.1)	0.159 (0.046) (1.0)	0.201 (0.086) (3.9)
1000	0.151 (0.057) (2.4)	0.156 (0.040) (1.0)	0.187 (0.087) (3.7)

DOI: 10.2202/1544-6115.1147

The pattern of performance of the methods on the colon data is similar to that on the leukemia data. It is observed from Table 2 that the average error rates of PLSLD and SIRLD are consistently lower than that of PCALD. It is also noted that PLSLD and SIRLD have similar error rates although the misclassification rate of SIRLD seems to be a little higher than that of PLSLD. In terms of relative computational cost, the results are also comparable to those from the leukemia data. The ratios of computation time are 7.2 for PCALD/PLSLD and 27.2 for SIRLD/PLSLD as p*=500. Again, PLS dimension reduction is much faster than the other two.

The results presented so far for both the leukemia data and the colon data are based on the studies using random subsets of genes. Next we describe the results of evaluations using subsets of genes selected based on the t-scores, a univariate measure related to the classification. Use of this "supervised" gene selection method should improve the accuracy of the classification procedures.

As described in Section 2.3.2, we computed t-scores for all genes in the data sets, ranked the genes by the scores, and selected top p^* genes. For both the leukemia and the colon data sets, we used p*=1000 genes for illustration. We performed gene selection and dimension reduction within cross-validation using the training set only and estimated the error rates using the test set. The results on both data sets are presented in Table 3 below. Shown in the table are the estimated error rates, their standard deviations and the average values of estimated K* based on 100 re-randomizations studies.

Table 3. Classification error rates of the three methods on the leukemia data set with 36/36 split and the colon data set with 36/26 split, averaged over 100 re-randomization studies. p*=1000 top genes selected using the t-statistic.

Data Set	PLSLD	SIRLD	PCALD
Leukemia	0.025 (0.021) (1.7)	0.026 (0.020) (1.0)	0.042 (0.025) (2.1)
Colon	0.136 (0.045) (2.2)	0.141 (0.038) (1.0)	0.162 (0.069) (2.8)

Comparing the results in Table 3 with those in Table 1 and Table 2, one can see that the prediction accuracy of all three methods (PLSLD, SIRLD and PCALD) has been improved: the average prediction errors of all three methods are reduced and the mean squared errors of prediction are also decreased. The relative performance of the methods, however, remains basically the same: PLSLD and SIRLD have similar misclassification rates and both have done better than PCALD.

4. Discussion

An important application of microarray data is to classify biological samples or predict clinical or other outcomes. In this paper, we viewed the class prediction problem as a multivariate regression problem where the number of variables far exceeds the number of samples, and evaluated several classification procedures for dealing with the problem. Specifically, we compared three dimension reduction methods (PLS, SIR, PCA), examined the relative performance of classification procedures incorporating those methods, and designed a five-step procedure for assessment studies. The empirical analyses were based on two published gene expression data sets.

We found that PLS and SIR were both effective in dimension reduction and they were more effective than PCA. The PLS and SIR based classification procedures performed consistently better than the PCA based procedure in prediction accuracy. The empirical results are consistent with the analysis of the techniques. PLS and SIR construct new predictors using information on the response variable while PCA does not; thus PLS and SIR components are more likely to be good predictors than those from PCA. For similar reason, the use of "supervised" gene selection methods would be likely to improve the classification accuracy. We showed that a simple t-score based gene selection method worked well for two-class problems. In the study, we also evaluated the computational efficiency of the three dimension reduction methods and found that PLS had significant advantage over the other two. Considering both predictive accuracy and computational efficiency, we conclude that the PLS based procedure has provided the best performance among the three classification procedures.

Dimension reduction is a necessary part of multivariate analysis of high-throughput assay data such as gene expression data. Dimension reduction methods are frequently used but their relative performance has not been well studied. It would be difficult to compare the performance of dimension reduction methods based on results of published studies due to differences among the studies in data sets, data preprocessing, and methods of gene selection, model selection and validation. This study provides a systematic comparison of the three

DOI: 10.2202/1544-6115.1147

dimension reduction methods. Moreover, the assessment procedure developed in this study can be easily extended to include more methods into evaluation. The scope of the study is however quite limited. Many methods of gene selection/dimension reduction are available. In further studies, we intend to continue the investigation and include more methods into our evaluations.

References

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745-6750.

Antoniadis, A., Lambert-Lacroix, S. and Leblanc, F. (2003) Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, **19**, 563-570.

Boulesteix, A. (2004) PLS Dimension reduction for classification with microarray data. *Statistical applications in genetics and molecular biology*, **3**, 1-33.

Bura, E. and Pfeiffer, R.M. (2003) Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. *Bioinformatics*, **19**, 1252-1258.

Chiaromonte, F. and Martinelli, J. (2002) Dimension reduction strategies for analyzing global gene expression data with a response. *Mathematical Biosciences*, **176**, 123-144.

Cook, R.D. (1998) *Regression Graphics*. John Wiley & Sons, New York.

Denham, M.C. (1995) Implementing partial least squares. *Statistics and Computing*, **5**, 191-202.

Dettling, M. and Buhlmann, P. (2003) Boosting for tumor classification with gene expression data. *Bioinformatics*, **19**, 1061-1069.

Duan, N. and Li, K.C. (1991) Slicing regression: a link-free regression method. *The Annals of Statistics*, **19**, 505-530.

Dudoit, S., Fridlyard, J. and Speed, T.P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**, 77-87.

Dudoit, S., Shaffer, J.P. and Boldrick, J.C. (2003) Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18**, 71-87.

Durbin, B., Hardin, J., Hawkins, D.M., and Rocke, D.M. (2002) A variance-stabilizng transformation for gene-expression microarray data. *Bioinformatics*, **18**, 105S-110S.

Durbin, B. and Rocke, D.M. (2003) Estimation of transformation parameters for microarray data. *Bioinformatics*, **19**, 1360-1367.

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

Garthwaite, P.H. (1994) An interpretation of partial least squares. *Journal of American Statistical Association*, **89**, 122-127.

Ghosh, D. (2002) Singular value decomposition regression modeling for classification of tumors from microarray experiments. *Proceedings of the Pacific Symposium on Biocomputing*, 11462-11467.

Golub, T.R., Slonim, D.K., Tamayo, P., Huard,C., Gaasenbeek, M., Mesirov, P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531-537.

Gruber, M.H.J. (1998) *Improving Efficiency by Shrinkage*. Statistics: textbooks and monographs, volume 156. Marcel Dekker, Inc, New York.

Härdle, W., Klinke, S. and Turlach, B.A. (1995). *XploRe: an Interactive Statistical Computing Environment*, Springer-Verlag, New York.

Hawkins, D.M. and Yin, X. (2002) A faster algorithm for ridge regression of reduced rank data. *Computational Statistics & Data Analysis*, **40**, 253-262.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S.,

Gruvberger, D., Loman, N., Johannsson, O., Olsson, H., Wilfond, B, Sauter, G., Kallioniemi, O., Borg, A. and Trent, J. (2001) Gene expression profiles in hereditary breast cancer. *New England Journal of Medicine*, **244**, 539-548.

Helland, I.S. (1988) On the structure of partial least squares. *Communications in Statistics: Simulation and Computation*, **17**, 581-607.

Helland, I.S. (1990) Partial least squares regression and statistical models. *Scandinavian Journal of Statistics*, **17**, 97-114.

Hoerl, A.E. and Kennard, R.W. (1970) Ridge regression: biased estimation for non-orthogonal problems. *Technometrics*, **8**, 27-51.

Hosmer, D.W. and Lemeshow, S. (2000) *Applied Logistic Regression*. Wiley, New York.

Huang, X. and Pan, W. (2003) Linear regression and two-class classification with gene expression data. *Bioinformatics*, **19**, 2072-2978.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18**, 96S-104S.

Ihaka, R. and Gentleman, R. (1996) R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299-314.

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

Kerr, K., Martin, M., and Churchill, G. (2000) Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, **7**, 819-837.

Kiers, H.A.L. (1997) Discrimination by means of components that are orthogonal in the data space. *Journal of Chemometrics*, **11**, 533-545.

Krzanowski, W.J. (1995) Orthogonal canonical variates for discrimination and classification. *Journal of Chemometrics*, **9**, 509-520.

Li, K.C. (1991) Sliced inverse regression for dimension reduction. *Journal of American Statistical Association*, **86**, 316-342.

Li, K.C. (2000) High dimensional data analysis via the SIR/PHS approach. Unpublished manuscript dated April 6, 2000 obtained at the Internet site http://www.stat.ucla.edu/kcli/sir-PHD.pdf.

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

Massey, W.F. (1965) Principal components regression in exploratory statistical research. *Journal of American Statistical Association*, **60**, 234-246.

Munson, P. (2001) A "consistency" test for determining the significance of gene expression changes on replicate samples and two convenient variance-stabilizing transformations. *GeneLogic Workshop of Low Level Analysis of Affymetrix GeneChip Data*.

Neter, J., Kutner, M.H., Nachtsheim, C.J. and Wasserman, W. (1996) *Applied linear statistical models*, 4th edition, McGraw-Hill, New York.

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

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

Parmigiani, G., Garrett, E., Irizarry, R. and Zeger, S. (2003) *The Analysis of Gene Expression Data: Methods and Software*, Springer, New York.

Rocke, D.M. and Durbin, B.P. (2001) A model for measurement errors for gene expression arrays. *Journal of Computational Biology*, **8**, 557-569.

Speed, T. (2003) (eds.) *Statistical Analysis of Gene Expression Microarray Data*, Chapman & Hall/CRC, New York.

Stone, M. (1974). Cross-validation choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, Series B, **36**, 111-147.

Stone, M. and Brooks, R.J. (1990) Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of Royal Statistical Society, Series B*, **52**, 237-269.

DOI: 10.2202/1544-6115.1147

Xia, Y., Tong, H., Li, W.K. and Xi, Z.L. (2002) An adaptive estimation of dimension reduction space. *J. R. Statist. Soc*. *B*., **64**, 363-410.

West, M., Blanchette, C., Fressman, H., Huang, E., Ishida, S., Spang, R., Zuan, H., Marks, J.R. and Nevins, J.R. (2001). Predicting the clinical status of human breast cancer using gene expression profiles. *Proceedings of the National Academy of Sciences of the United States*, **98**, 11462-11467.

Wold, H. (1966) Nonlinear estimation by iterative least squares procedures. In *Research Papers in Statistics*, ed. F.N. David, pp. 411-444. Wiley, New York.

Wold, H. (1973) Nonlinear iterative partial least squares (NIPALS) modeling: some recent developments. In *Multivariate Analysis III*, ed. P. Krishnaiah, pp. 383-407, Academic Press, New York.

Wold, H. (1982) Soft modeling: the basic design and some extensions. In *Systems under Indirect Observation: Causality-Structure-Prediction*, ed. K. G. Joreskog and H. Wold, Vol. II, Ch. 1, pp. 1-54, North-Holland, Amsterdam.