Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2007) 438–451
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Class prediction and gene selection for DNA microarrays using regularized sliced inverse regression

Luca Scrucca*

Dipartimento di Economia, Finanza e Statistica, Università degli Studi di Perugia, 06100 Perugia, Italy

Available online 17 February 2007

Abstract

The monitoring of the expression profiles of thousands of genes have proved to be particularly promising for biological classification. DNA microarray data have been recently used for the development of classification rules, particularly for cancer diagnosis. However, microarray data present major challenges due to the complex, multiclass nature and the overwhelming number of variables characterizing gene expression profiles. A regularized form of sliced inverse regression (REGSIR) approach is proposed. It allows the simultaneous development of classification rules and the selection of those genes that are most important in terms of classification accuracy. The method is illustrated on some publicly available microarray data sets. Furthermore, an extensive comparison with other classification methods is reported. The REGSIR performance is comparable with the best classification methods available, and when appropriate feature selection is made the performance can be considerably improved.

Keywords: Microarray; Classification; Dimension reduction; SIR; Regularization; Feature selection

1. Introduction

Gene expression data from DNA microarrays may be employed to define classification rules to predict the diagnostic category of a sample unit on the basis of its gene expression profile. Classification of microarray data is particularly problematic due to under-resolution, i.e. the presence of a large number of features (genes) from which to predict classes compared to the relatively small number of observations (samples). This characteristic makes most traditional classification methods inapplicable. In such cases, feature selection is routinely used to alleviate the under-resolution problem, but it also plays an important role for improving the prediction performance. Ideally, a classification rule should be based only on information-carrying genes. However, this is a difficult task. Since the inclusion of irrelevant or redundant genes negatively affect prediction accuracy, we usually require that a classifier should be based only on those genes which contribute most to classification accuracy.

Several classification methods based on gene expression profiles in microarray experiment have been proposed. These include classical methods (such as linear and quadratic discriminant analysis, *k*-nearest neighbor and logistic regressions), classification trees and aggregating classifiers (such as CART, bagging, boosting and random forest), machine learning algorithms (for example, neural networks and support vector machines), and other specialized algorithms (flexible discriminant analysis, nearest shrunken centroids, among others). Dudoit et al. (2002a), and more

* Tel.: +39 075 5855233; fax: +39 075 5855950.
E-mail address: luca@stat.unipg.it.

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.02.005

recently Lee et al. (2005), made a comparison of the above methods evaluating their performances on widely available gene expressions data sets.

Dimension reduction techniques have already been studied in the microarray context. A two-stage approach based on sliced inverse regression (SIR) was proposed by Chiaromonte and Martinelli (2002), with the first stage aimed at eliminating the dimension redundancy through an eigenanalysis which identified a set of linear combinations of the original genes which retained the main variability structure. Then, the SIR method was applied on such reduced set of eigengenes. A final gene selection stage was proposed based on a ranking of genes in accordance to the concordance with the final estimated directions. A related approach was also discussed by Bura and Pfeiffer (2003), in which they applied both SIR and SAVE, i.e. sliced average variance estimation, to microarray data. Their main focus was on the comparison of the two dimension reduction methods, but they did not address the under-resolution problem, nor any gene selection procedure. Antoniadis et al. (2002) proposed a dimension reduction method based on applying a nonparametric discriminant rule to directions estimated by MAVE (Xia et al., 2002). They used a filtering step to select a small subset (at most 200) of relevant genes, but this step was independent of the classifier and the final directions were linear combinations of all and only the starting genes selected.

In this paper we propose an approach based on a regularized form of sliced inverse regression for class prediction and gene selection from DNA microarray data. In Section 2 we briefly review the SIR methodology, then Section 3 illustrates the application of SIR to gene expression microarray data. The regularized SIR (REGSIR) approach is proposed to overcome the under-resolution problem. Class prediction based on REGSIR is discussed and a backward gene selection method is proposed to select those genes that are relevant to the prediction of the disease diagnosis. Section 4 shows the application of the methodology proposed to data on small round blue cell tumors of childhood (Khan et al., 2001), leukemia data (Golub et al., 1999) and colon data (Alon et al., 1999). The following section describes an extensive comparison of REGSIR with other classification methods based on several publicly available microarray data sets. The last section contains a summary of the proposed method and some concluding remarks.

2. Sliced inverse regression

SIR is a dimension reduction method introduced by Li (1991) which seek to find a few directions in the p-dimensional predictors space such that the regression of $Y|X$ can be fully studied on such dimension reduction subspace (drs) without loosing any relevant information contained in X about Y. SIR assumes that the relationship between a response variable and a set of predictors can be expressed through the model $Y = f(\beta_1^{\mathrm{T}}X, \ldots, \beta_d^{\mathrm{T}}X, \varepsilon)$, where ε is a random error term, $f()$ is an unknown function, and $d \leqslant p$. The directions $(\beta_1, \ldots, \beta_d) \equiv \boldsymbol{\beta}$ span the drs $\mathscr{S}(\boldsymbol{\beta})$ (Cook, 1998). Its dimension is $\dim(\mathscr{S}(\boldsymbol{\beta})) = d$, and provided that the assumed model holds, we can write $Y \perp\!\!\!\perp X|\boldsymbol{\beta}^{\mathrm{T}}X$. However, $\boldsymbol{\beta}$ is not unique, since we can always multiply $\boldsymbol{\beta}$ by any non-zero constant and still $Y \perp\!\!\!\perp X|\boldsymbol{\beta}^{\mathrm{T}}X$ holds. To avoid such non-uniqueness, the *central drs* $\mathscr{S}_{Y|X}$ has been defined as the intersection over all drs. A central drs exists under minor conditions, and when it exists it is the unique minimum drs (Cook, 1998, Chapter 6). Thus, the dependence of Y on X may be fully studied through $\boldsymbol{\beta}^{\mathrm{T}}X$, the coordinates of the projection of X onto the d-dimensional subspace spanned by the columns of $\boldsymbol{\beta}$. Often d is smaller than p; if this is the case, we have effectively reduced the dimensionality of the problem.

Li (1991, Theorem 3.1) showed that, under the linear design condition on the distribution of X, the population version of SIR is obtained through the eigendecomposition of $\boldsymbol{\Sigma}_{X|Y} = \mathrm{Cov}(\mathrm{E}(X|Y))$ with respect to $\boldsymbol{\Sigma}_X = \mathrm{Cov}(X)$, i.e.

$$\begin{aligned} &\boldsymbol{\Sigma}_{X|Y}\boldsymbol{v}_i = l_i\boldsymbol{\Sigma}_X\boldsymbol{v}_i \quad \text{with } l_1 \geqslant l_2 \geqslant \cdots \geqslant l_p, \\ &\qquad \text{and } \boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{\Sigma}_X\boldsymbol{v}_i = 1 \quad \text{for } i = 1, \ldots, p. \end{aligned} \tag{1}$$

The first d eigenvectors $\boldsymbol{V}_d = [\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d]$ provide a consistent estimate of a basis for the central subspace $\mathscr{S}_{Y|X}$. The matrix form of (1) is given by

$$\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{X|Y} = \boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}}, \tag{2}$$

and the spanning matrix of, at least part of, the central drs $\mathscr{S}_{Y|X}$ is obtained by setting $\boldsymbol{\beta} \equiv \boldsymbol{V}_d$. The sample version of SIR can be easily computed by defining a sliced version of Y with fixed number of slices, and replacing the above matrices with sample estimates (Li, 1991, 2000).

The linear design condition requires that the predictors are linearly related, i.e. $\mathrm{E}(\boldsymbol{b}^{\mathrm{T}}X|\boldsymbol{\beta}^{\mathrm{T}}X)$ is linear in $\boldsymbol{\beta}^{\mathrm{T}}X$ for all $\boldsymbol{b} \in \mathbb{R}^p$. Li emphasized that this condition is not a severe restriction, since most low-dimensional projections are close to being normal. Furthermore, the condition is required to hold only for the basis $\boldsymbol{\beta}$ of the drs. Since $\boldsymbol{\beta}$ is unknown, in practice it is required to hold for all possible $\boldsymbol{\beta}$, which is equivalent to elliptical symmetry distribution (such as multivariate normal) of X (Cook and Weisberg, 1991). In practice, transforming predictors such that they are approximately multivariate normal (Velilla, 1993) may help when gross nonlinearities are present.

3. Applying SIR to gene expression data

Suppose we have an expression array $\boldsymbol{X}$ of dimension $(n \times p)$ for n samples and p genes. The biologists view would consider $\boldsymbol{X}^{\mathrm{T}}$, in which each column represents the gene expression profile for a particular sample. We assume that gene expression measures are log transformed ratios to a baseline or a reference condition, and they have been normalized. Normalization is a key step in the pre-processing of DNA microarray data; it aims at removing the systematic variation in microarray experiments arising from several sources, such as sample preparation, hybridization, printing, or scanning artifacts. A categorical response variable Y with K levels representing biological outcomes, such as tumors type, is also recorded along with gene expression levels.

3.1. Regularized SIR

Applying SIR to gene expression data appears in principle straightforward. There is no need to slice the response variable since Y is categorical with a level for each biological class. But, since $p \gg n$, $\hat{\boldsymbol{\Sigma}}_X$ has rank at most n, and is hence singular and cannot be inverted (on this point see also Chiaromonte and Martinelli, 2002). Usually, although a preliminary screening of genes is performed (see Section 3.2), the number of genes showing a certain degree of variation across samples could still be larger than n. Therefore, a sort of regularization is required to overcome ill- and poorly-posed inverse problems (Friedman, 1989).

We first note that SIR directions solve the following optimization problem:

$$\arg\max_{\boldsymbol{\beta}} \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_{X|Y}\boldsymbol{\beta} \quad \text{subject to} \quad \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_X\boldsymbol{\beta} = \boldsymbol{I},$$

where $\boldsymbol{I}$ denotes the $(d \times d)$ identity matrix. This can be shown by direct constrained maximization. For the first direction, the function $\phi = \boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{\Sigma}_{X|Y}\boldsymbol{\beta}_1$ must be maximized with respect to the vector $\boldsymbol{\beta}_1$, under the constraint $\boldsymbol{\beta}_1^{\mathrm{T}}\boldsymbol{\Sigma}_X\boldsymbol{\beta}_1 = 1$. The solution of the problem satisfy $(\boldsymbol{\Sigma}_{X|Y} - l\boldsymbol{\Sigma}_X)\boldsymbol{\beta}_1 = 0$, where l is the value attained by ϕ at maximum. This equation may be rewritten as $(\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{X|Y} - l\boldsymbol{I})\boldsymbol{\beta}_1 = 0$, so l must be an eigenvalue, and $\boldsymbol{\beta}_1$ must be an eigenvector of $\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{X|Y}$. Since l is the maximum of ϕ, $\boldsymbol{\beta}_1$ must be the eigenvector corresponding to the largest eigenvalue. This defines the first direction where the ratio of between-group variability is the largest relative to the total variation. The same arguments extend directly to the following directions $\boldsymbol{\beta}_j$ $(j = 2, \ldots, d)$, under the constraints $\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{\Sigma}_X\boldsymbol{\beta}_j = 1$ and $\boldsymbol{\beta}_j^{\mathrm{T}}\boldsymbol{\Sigma}_X\boldsymbol{\beta}_i = 0$ (for any $i = 1, \ldots, j-1$), resulting in the matrix of eigenvectors $\boldsymbol{\beta}$.

Using results from Hastie et al. (1995) on penalized discriminant analysis, we may define a regularized SIR criterion as

$$\arg\max_{\boldsymbol{\beta}} \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}_{X|Y}\boldsymbol{\beta} \quad \text{subject to} \quad \boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{\Sigma}_X + \boldsymbol{\Omega})\boldsymbol{\beta} = \boldsymbol{I},$$

which is solved by the eigendecomposition of $\boldsymbol{\Sigma}_{X|Y}$ with respect to $(\boldsymbol{\Sigma}_X + \boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is a symmetric $(p \times p)$ penalizing matrix. This can be defined as $\boldsymbol{\Omega} = \lambda(\mathrm{tr}(\boldsymbol{\Sigma}_X)p^{-1}\boldsymbol{I} - \boldsymbol{\Sigma}_X)$, where the tuning parameter λ controls the amount of shrinkage applied $(0 \leqslant \lambda \leqslant 1)$. Thus, the regularized covariance matrix is given by the following convex combination:

$$\boldsymbol{\Sigma}_X(\lambda) = (1-\lambda)\boldsymbol{\Sigma}_X + \lambda\frac{\mathrm{tr}(\boldsymbol{\Sigma}_X)}{p}\boldsymbol{I}.$$

For $\lambda = 0$ we have $\boldsymbol{\Sigma}_X(\lambda) = \boldsymbol{\Sigma}_X$, the original covariance matrix. As $\lambda \to 1$, $\boldsymbol{\Sigma}_X(\lambda)$ approaches a diagonal matrix with diagonal elements given by $\mathrm{tr}(\boldsymbol{\Sigma}_X)/p$, i.e. the average eigenvalue of the covariance matrix. The adopted definition of the penalizing matrix ensures that the total variation, as measured by the trace (Seber, 1984), is constant for any λ. This can be easily seen by direct calculation: $\mathrm{tr}(\boldsymbol{\Sigma}_X(\lambda)) = (1-\lambda)\mathrm{tr}(\boldsymbol{\Sigma}_X) + \lambda\mathrm{tr}(\boldsymbol{\Sigma}_X)\mathrm{tr}(\boldsymbol{I})/p = \mathrm{tr}(\boldsymbol{\Sigma}_X)$.

Thus, the REGSIR eigendecomposition is defined as follows:

$$\boldsymbol{\Sigma}_X(\lambda)^{-1}\boldsymbol{\Sigma}_{X|Y} = \boldsymbol{V}\boldsymbol{L}\boldsymbol{V}^{\mathrm{T}}. \tag{3}$$

The regularized covariance matrix $\boldsymbol{\Sigma}_X(\lambda)$ can now be inverted if $\lambda > 0$ even for $p > n$, and the drs is spanned by the columns of $\boldsymbol{\beta} \equiv \boldsymbol{V}_d$, the eigenvectors corresponding to the first d largest eigenvalues. Again, sample estimates may be used to compute the eigendecomposition (3) in practical applications.

3.2. Filtering of genes

The large number of gene expression profiles usually collected in microarray experiments can be drastically reduced since many of them exhibit near constant expression levels across samples. In general, there are three basic approaches to feature selection: (i) filtering methods, which select a subset of genes independently of the classifier, are often used as pre-processing step to remove *irrelevant features*, i.e. those genes whose expression level is not correlated with the diagnosis; (ii) wrapper methods, which are an inherent part of the classification building process looking for the feature subset which provides the most accurate classification, and (iii) embedded methods, which simultaneously estimate a classifier and perform feature selection. Among the filtering criteria we mention the Fisher discriminant ratio (FDR) which ranks genes based on the ratio of between-groups to within-groups sum of squares (Dudoit et al., 2002a), while among wrapper methods a popular approach is the recursive feature elimination (RFE) proposed by Guyon et al. (2002). Embedded methods are used by classification trees, such as CART, and aggregation classifiers, like boosting and random forest. Here we discuss a preliminary filtering step, while in Section 3.4 we will introduce a wrapper gene selection method based on the estimated REGSIR directions.

The filtering method adopted is similar to the one encountered in discriminant analysis, where it is customary to use a preliminary screening of the genes based on the ratio of between-groups sum of squares (*BSS*) to within-groups sum of squares (*WSS*). This statistic is clearly related to the decomposition used in computing discriminant variates, but for SIR a more natural statistic, albeit equivalent in terms of ordering, would be the ratio of between-groups to total sum of squares (*TSS*), i.e.

$$BT_j = \frac{BSS_j}{TSS_j} = \frac{[\hat{\boldsymbol{\Sigma}}_{X|Y}]_{jj}}{[\hat{\boldsymbol{\Sigma}}_X]_{jj}} \quad \text{for } j = 1, \ldots, p,$$

where $[\boldsymbol{A}]_{jj}$ indicates the jth element along the diagonal of matrix $\boldsymbol{A}$. Recalling the ANOVA decomposition, $TSS_j = BSS_j + WSS_j$, it is evident that the ordering provided by statistic (3.2) is the same as that based on BSS_j/WSS_j, for any gene $j = 1, \ldots, p$. For Gaussian data, under the null hypothesis of equality of within-group means it is known that $F = (BSS_j/(K-1))/(WSS_j/(n-K)) \sim F_{K-1,n-K}$. Such statistic can also be expressed in term of Eq. (3.2) as

$$F = \frac{BT_j/(K-1)}{(1-BT_j)/(n-K)} \quad \text{for } j = 1, \ldots, p.$$

This result provides a way to identify those genes whose expression levels differ between groups. However, instead of pursuing formal assessment of significance, which would involve accounting for multiple comparisons, we draw Quantile–Quantile (Q–Q) plots as a visual aid for identifying genes with unusual large F-statistic. As Dudoit et al. (2002b) pointed out, Q–Q plots informally correct for the large number of comparisons, and the genes which deviate markedly from an otherwise linear relationship are those who are retained for subsequent analysis. An example of such approach is shown in the left panel of Fig. 1, where we draw the Q–Q plot for the F-statistic in (3.2) computed on the whole set of genes versus the corresponding quantiles of the F distribution under the null hypothesis: points above the dotted line identify those genes which show a sufficient degree of variation across groups.

3.3. Class prediction based on REGSIR

Expression profiles for the active genes can be projected onto the estimated drs yielding the REGSIR variates $\hat{\boldsymbol{\beta}}_j^{\mathrm{T}}\boldsymbol{x}$ ($j = 1, \ldots, d$). A d-dimensional plot using Y as marking variable may then be used to visually allocate each sample point to the closest class. A more formal procedure consists in classifying each sample to the nearest centroid in the

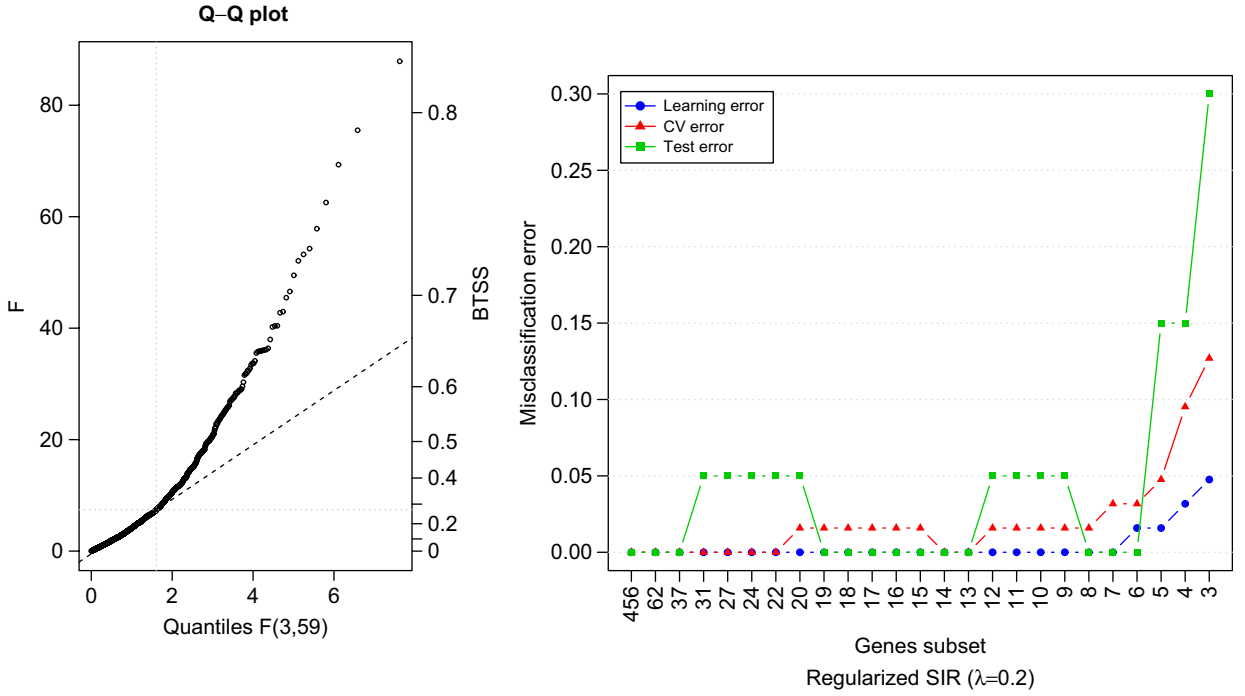

Fig. 1. Q–Q plot for pre-filtering of genes (left panel) and misclassification errors for gene subsets applied to the SRBCT data (right panel).

REGSIR subspace. Suppose we have a test sample with expression levels $\boldsymbol{x}^*$, then the discriminant score for class $Y = k$ is defined as

$$\delta_k(\boldsymbol{x}^*) = (\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}^* - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\bar{\boldsymbol{x}}_k)^{\mathrm{T}}\boldsymbol{W}^{-1}(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}^* - \hat{\boldsymbol{\beta}}^{\mathrm{T}}\bar{\boldsymbol{x}}_k) - 2\log(\pi_k). \tag{4}$$

The first term in the above expression is the Mahalanobis distance of the test sample $\boldsymbol{x}^*$ with respect to the centroid on the SIR subspace, using $\boldsymbol{W} = \sum_{k=1}^{K} f_k \boldsymbol{W}_k$ as the pooled within-class covariance matrix, where $\boldsymbol{W}_k = \sum_{i\in(Y=k)}(1/n_k)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)^{\mathrm{T}}$, and $f_k = n_k/n$ $(k = 1, \ldots, K)$ is the sample proportion of observations in each class. The second term is a correction, in analogy to Gaussian LDA, based on the class prior probability, with $\sum_{k=1}^{K}\pi_k = 1$; often, these are estimated by the sample class proportion f_k in the learning set. The classification rule is then

$$\mathscr{C}(\boldsymbol{x}^*) = \arg\min_k \delta_k(\boldsymbol{x}^*). \tag{5}$$

Discriminant scores can also be used, in analogy with classical linear discriminant analysis, to construct estimates of the class probabilities

$$\hat{p}_k(\boldsymbol{x}^*) = \frac{\exp\{-\frac{1}{2}\delta_k(\boldsymbol{x}^*)\}}{\sum_{j=1}^{K}\exp\{-\frac{1}{2}\delta_j(\boldsymbol{x}^*)\}}.$$

3.4. Backward gene selection

Gene selection is a relevant issue for several reasons. First of all, identifying a small set of genes that is able to accurately discriminate the samples allows to employ a less expensive diagnostic assay in practical applications. Furthermore, it may shed light into the mechanisms responsible for the disease. From a computational point of view, a classifier based on a small set of relevant genes is able to provide more accurate predictions. In the context of our approach, gene selection aims at identifying a subset of genes which is able to linearly explain the patterns variation in the estimated REGSIR subspace.

In the REGSIR approach we start by building a classifier using the relevant genes selected through the BSS/TSS ratio in Eq. (3.2).This filtering step removes the irrelevant genes. The REGSIR model estimated using such subset of relevant genes usually provides a perfect fit to the training data, hence 0 training error rate, but it tends to be a poorer classifier for future observations. However, many of the relevant genes are highly correlated to each other, thus it might be advisable to remove these *redundant features*. This task is addressed by applying a wrapper backward feature selection. For a set of REGSIR variates we would like to select the smallest subset of genes which maximizes the squared correlation coefficient computed taking into account the importance of each estimated REGSIR direction. This criterion amounts to find those features which best linearly explain the REGSIR variates.

Define with $\hat{\boldsymbol{Z}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ the estimated REGSIR variates based on the set of current active p genes. Let $\mathscr{S}_g$ be the set of dimension g containing a subset of current p genes ($g < p$). Recalling the squared correlation coefficient for a multivariate linear regression model (Mardia et al., 1979, pp. 170–171), the statistic proposed can be defined as follows:

$$R^2(\mathscr{S}_g) = 1 - \mathrm{tr}\{(\hat{\boldsymbol{Z}}^{*\mathrm{T}}\hat{\boldsymbol{Z}}^{*})^{-1}\hat{\boldsymbol{L}}\hat{\boldsymbol{E}}^{\mathrm{T}}\hat{\boldsymbol{E}}\}, \tag{6}$$

where $\hat{\boldsymbol{Z}}^*$ is the column-centered version of $\hat{\boldsymbol{Z}}$, $\hat{\boldsymbol{E}} = \hat{\boldsymbol{Z}} - \boldsymbol{X}_g(\boldsymbol{X}_g^{\mathrm{T}}\boldsymbol{X}_g)^{-1}\boldsymbol{X}_g^{\mathrm{T}}\hat{\boldsymbol{Z}}$ is the matrix of residuals for the regression of $\hat{\boldsymbol{Z}}$ on $\boldsymbol{X}_g$, with the latter being the $(n \times (g+1))$ matrix containing the genes in the subset $\mathscr{S}_g$ plus a column of 1s. The diagonal matrix $\hat{\boldsymbol{L}}$ has elements equal to $\hat{l}_i / \sum_{j=1}^{d} \hat{l}_j$ for $i = 1, \ldots, d$, so it weights each estimated direction with the corresponding eigenvalue. For a two-class problem statistic (6) reduces to the determination coefficient for the regression of the first REGSIR variate on genes included in $\mathscr{S}_g$.

Gene selection is performed using an iterative scheme: at each step only those genes which contribute the most to the overall patterns, based on statistic (6), are retained, then the REGSIR model is re-estimated and the accuracy of the resulting classifier evaluated. The screening of redundant genes depends on the cut-off value used in the R^2 criterion: a large value, say 0.999, implies that one or few genes are removed at each step. On the contrary, the process can be accelerated for small values, say 0.9, since in this case a large number of redundant genes are eliminated at each iteration. This process is repeated until the final subset contains $K - 1$ active genes, where K is the number of classes for the problem at hand, or it can be stopped at a desired subset size. The classification accuracy of each gene subset may be assessed on the basis of its misclassification error on a test set, if available, or on a cross-validated set (see also Section 3.5). This criterion may guide in choosing the "best" gene subset or a set of candidates gene subsets.

The backward procedure discussed above is similar to RFE (Guyon et al., 2002), except that RFE ranks each feature based on its contribution to the overall accuracy, while our approach looks for the smallest set of features which is able to best approximate (up to a desired level) the estimated REGSIR subspace.

In our current proposal once a gene has been removed it is not considered in the following steps. In principle, this could be allowed, although an explosive number of potential gene subsets should be considered. This kind of stepwise feature selection is likely to reproduce, or even exacerbate given the large number of genes, the computational problems encountered in linear regression models. However, *ad hoc* solutions could be pursued, for example applying at each step the criterion in (6) to all genes and not only to those previously selected. Of course, this is computationally more expensive, and eventual benefits should be investigated. All these aspects deserve further studies.

3.5. Selecting the shrinkage parameter and the gene subset

The shrinkage parameter λ in (3) is a tuning parameter, and its choice can be based on an estimate of the misclassification error. This can be derived from a test set, if available, or from a cross-validated set. The same procedure can also be used to assess the classification accuracy provided by the gene subsets selected as described in Section 3.4.

The cross-validation (CV) procedure adopted to assess the misclassification error rate for a given value of λ is based on the following algorithm:

FOR $i = 1$ to *nfold* DO

(1) Pre-filtering: using $\boldsymbol{X}_{-\mathscr{F}_i}$, the gene expression matrix obtained removing from $\boldsymbol{X}$ the observations in the i-th fold $\mathscr{F}_i$, select a subset of $p^* \leqslant p$ genes as described in Section 3.2.
(2) Fit REGSIR: apply the eigendecomposition in (3) to the current set of active genes.
(3) Classification: use the estimated REGSIR model to predict the left out observations in $\mathscr{F}_i$.

(4) Gene selection: use statistic (6) to select a subset of genes which "best" approximate the estimated REGSIR directions, and with this subset form the new gene expression matrix.
(5) Repeat steps (2)–(4) until a stopping criterion for gene selection is met (see Section 3.4).

This algorithm for $nfold = n$ is the classical leave-one-out cross-validation (LOOCV). However, it is known that LOOCV is approximately unbiased for the true classification error, but has high variance since only one observation is removed at each step. Setting $nfold = 10$ overestimates the true classification error but has a smaller variance (see Hastie et al., 2001, Section 7.10). In practice, we use 10-fold cross-validation, dividing the set of samples at random into 10 approximately equal-sized parts. Each part is roughly balanced, ensuring that the classes are distributed proportionally within each fold.

For a given value of λ, we may draw a plot of the estimated error rate for each gene subset selected. For example, in the right panel of Fig. 1 we show the paths of the misclassification error rate for $\lambda = 0.2$, evaluated in the training set, in the test set, and by 10-fold CV. Similar graphs (not shown) have been obtained for values of λ in the grid 0–1(0.1). On such plots we may look for the smallest gene subset which provide the best accuracy. The main problem is that there can be many settings which give the same estimate of the classification error rate. In our experience, selecting a small value of λ is desirable since it adds a small amount of shrinkage, so retaining most of the correlation among genes. Furthermore, a stable path of misclassification error rates for different gene subsets seems preferable.

4. Data analysis examples and simulation studies

In this section we present the results from applying the proposed methodology to publicly available gene expression data sets that have been analyzed by several authors. Then, we investigate the behavior of REGSIR under two different settings using simulated data.

4.1. Small round blue cell tumors (SRBCT) of childhood

Khan et al. (2001) provided data on small round blue cell tumors of childhood. Expression measurements for 2308 genes were obtained from glass-slide cDNA microarrays. Tumors were classified as Burkitt lymphoma (BL), Ewing sarcoma (EWS), neuroblastoma (NB), and rhabdomyosarcoma (RMS). Sixty-three cases were used as learning samples and 25 as test samples, although five of the latter were not SRBCTs.

Based on Q–Q plot shown on the left panel of Fig. 1, we selected as relevant genes the top 456 genes with the largest BSS/TSS ratio based on equation (3.2). Then, we applied the REGSIR approach for several values of the shrinkage parameter; results for the selected $\lambda = 0.2$ are reported in the right panel of Fig. 1. This plot shows the misclassification error rates provided by REGSIR for subsets of decreasing size obtained removing the redundant genes (see Section 3.4). Ten-fold CV errors have been computed as discussed in Section 3.5. As expected the training error appears to be an optimistic estimate of the misclassification error when compared to the test set and the CV set. The error rates for both the CV set and the test set are equal to zero for the first three steps, with the smallest subset having $g = 37$ genes, and for the subsets with $g = 14$ and 13 genes. Thus, we may adopt the latter as the "best" subset since it is the smallest subset to achieve zero error rate on both the test set and the 10-fold cross-validated set.

Fig. 2 shows the learning samples projected onto the subspace spanned by the REGSIR directions estimated using the "best" 13 genes, along with decision boundaries.

Khan et al. (2001) achieved a test error of 0% using a neural network approach and selected 96 genes for classification. Tibshirani et al. (2002) using shrunken centroids selected 43 genes, still retaining a 0% error on the test set. The REGSIR approach also allows to achieve a 0% test error, but it uses fewer genes. In particular, if we consider the $g = 13$ subset, nine genes, those with the largest BSS/TSS ratio within the subset, are in common with those identified by Tibshirani et al. (2002). If we consider the subset with $g = 37$, the genes selected by both procedures rise to 14.

Fig. 3 displays the correlation matrix using all the available genes on panel (a) and the "best" genes selected by the REGSIR method on panel (b). Samples belonging to the same group are positively correlated on both graphs, but correlations from different tumor samples tend to be uncorrelated or negatively correlated when using the "best" genes subset, while spurious positive correlations are present if all genes are used.

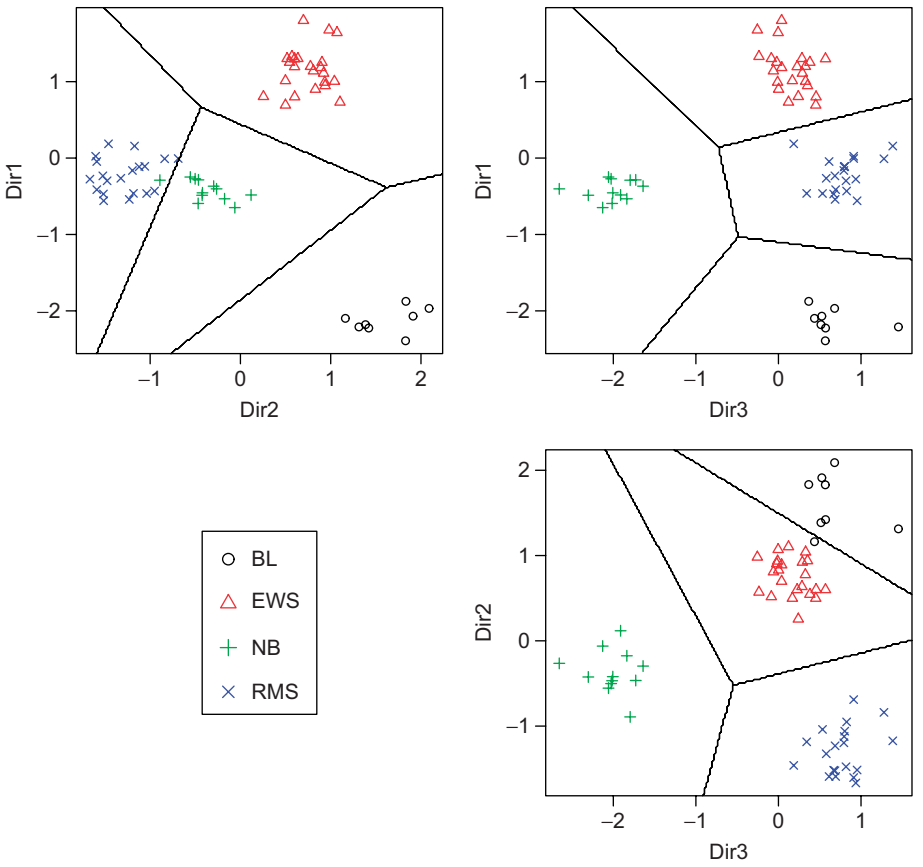

Fig. 2. Scatterplots of estimated REGSIR variates using the best subset of 13 genes for the SRBCT data with decision boundaries.

4.2. Leukemia data

This data set from high-density Affymetrix oligonucleotide arrays was originally studied by Golub et al. (1999). There were 3571 genes and 38 training samples: 27 in class ALL (acute lymphoblastic leukemia) and 11 in class AML (acute myeloid leukemia). The samples in class ALL could be further splitted into B and T cell types. A test set of 34 samples, 20 in ALL class and 14 in AML class, was also available.

Golub et al. (1999) report a test error rate of $\frac{4}{34}$ for their procedure using 50 genes. Tibshirani et al. (2002) provide a 10-fold CV error equal to $\frac{1}{38}$ and a test error of $\frac{2}{34}$ using 21 genes. The REGSIR procedure with $\lambda = 0.3$ and using 11 genes achieves a test error equal to $\frac{1}{34}$ (=0.0294), while the 10-fold CV error is zero (see the right panel of Fig. 4). This result was achieved after a preliminary filtering of genes which retained the top 250 genes based on the BSS/TSS ratio (see the left panel of Fig. 4).

The graph on the top-left panel of Fig. 5 reports the coefficients of the estimated REGSIR direction. Note that, since there are just two classes, only one direction can be estimated. The genes involved in such linear combination can be read off from the labels appearing on the y-axis of the graph. The projection along such direction is shown, separately for the learning set and the test set, on the box plots in the top-right panel of Fig. 5. Conditioning on the leukemia classes, the distributions for the learning set appear well separated; this separation is still evident for the samples in the test set, except for one ALL case which is misclassified to AML. Finally, it is interesting to look at the heat map of normalized expression levels for the selected gene subset (see bottom panel of Fig. 5). For this map we arranged genes by increasing value of REGSIR coefficients. Note that genes with negative REGSIR coefficient appear to be over-expressed in the ALL samples and under-expressed in ALL samples, while the opposite happen for those genes with positive REGSIR coefficient.

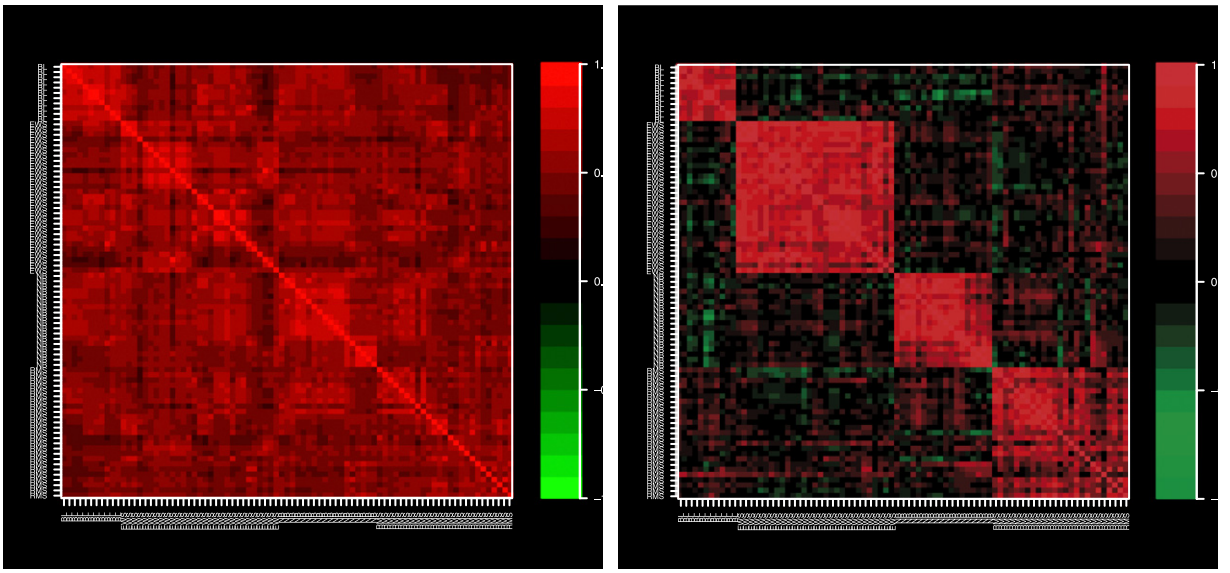

Fig. 3. Image plot of correlation matrix for the 63 samples from the SRBCT data. Correlations are computed based on expression profiles for all 2308 genes (left panel) and for the 13 "best" genes selected by REGSIR (right panel).

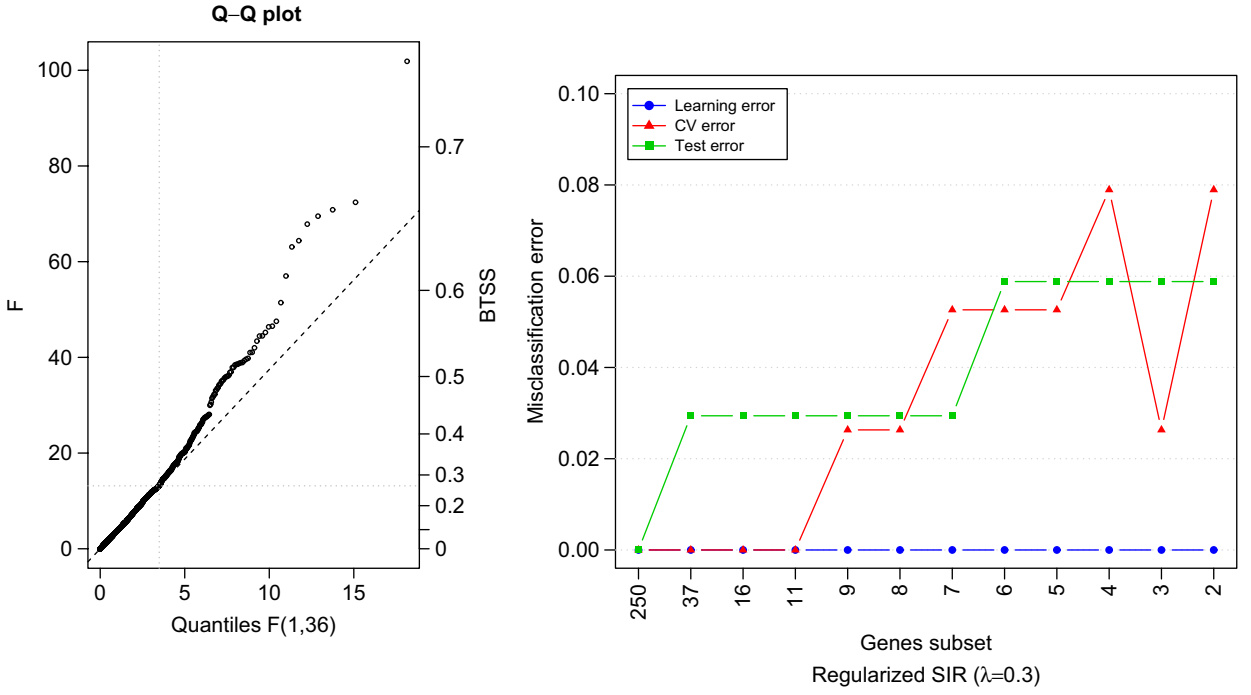

Fig. 4. Q–Q plot for pre-filtering of genes (left panel) and misclassification errors for gene subsets applied to the Leukemia data (right panel).

4.3. Simulation studies

4.3.1. Two-group independence structure

In the first setup we generate 100 independent samples for 1000 genes from two classes which differ only on a subset of genes. For the first class, 50 observations are generated from $p = 1000$ independent standard normal variables.

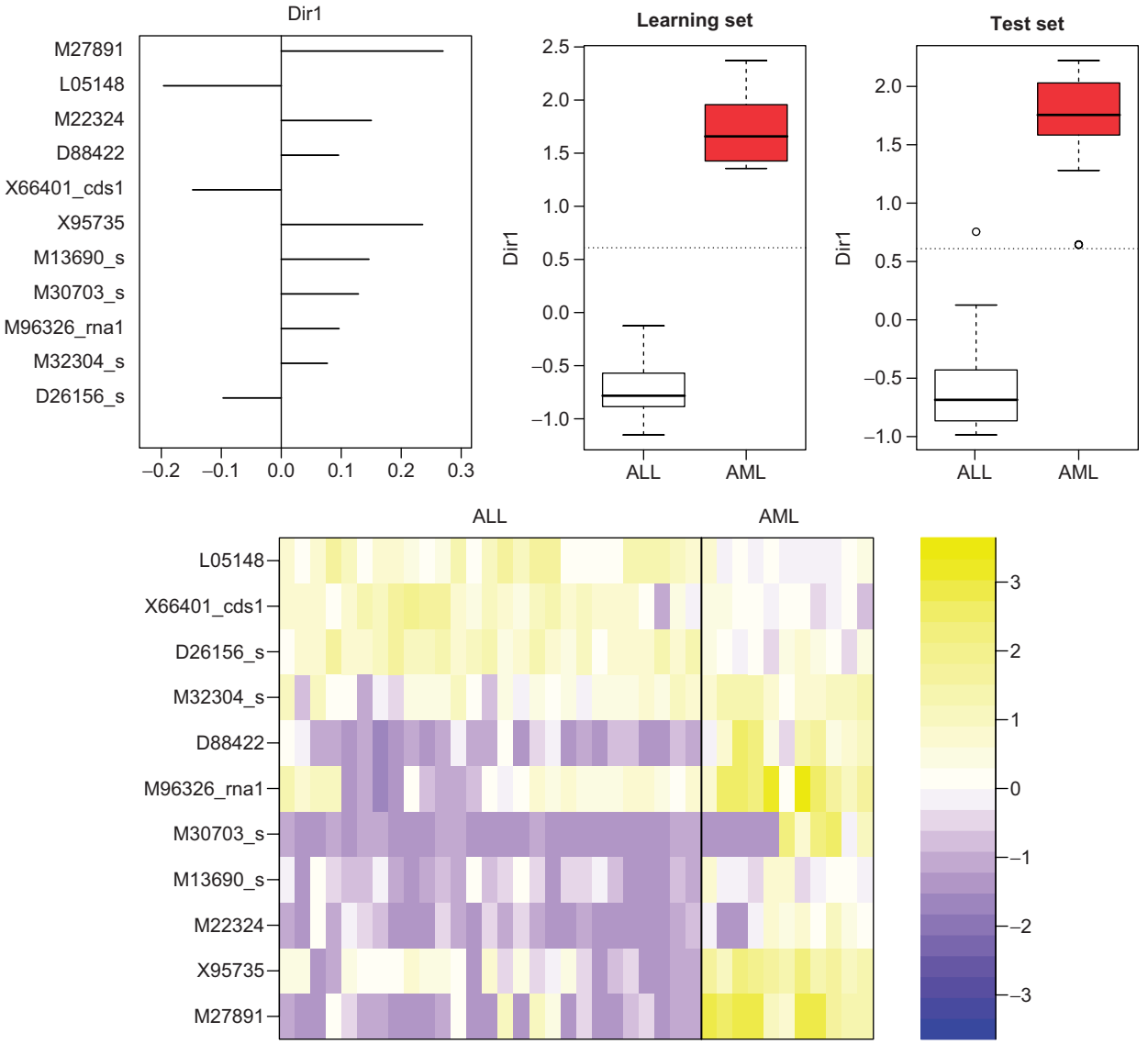

Fig. 5. Coefficients for the first REGSIR direction estimated on the Leukemia data (top-left panel) and box-plots for the corresponding REGSIR variate conditional on the leukemia tumor classes, separately for the learning set and the test set (top-right panel). Heat map of normalized expression levels for the selected gene subset, with gene ordered by increasing value of REGSIR coefficients (bottom panel).

For the second class, the first 10 variables are generated from N(1, 1), while the remaining 990 are generated from a standard normal, all independent. Following the same scheme we also simulate 5000 test samples from each class. Table 1 reports the results obtained from 100 random simulation: both the 10-fold CV error and the test error are the averages of the minimum error achieved by each simulation (with standard deviation within parenthesis). These are also shown graphically in the left panel of Fig. 6.

Two points are worthwhile to note: (i) on average, the 10-fold CV error provides an assessment of the misclassification rate quite close to the error obtained from the independent test sets; and (ii) given the identity covariance structure within groups, large values of λ, which shrink the covariance matrix toward a diagonal matrix as discussed in Section 3.1, tend to provide a better classification accuracy.

Table 1
Average minimum misclassification error from 100 simulated data sets, evaluated on the basis of 10-fold CV and test sets for different values of the shrinkage parameter

λ	Two-group independence				Two-group block dependence			
	10-fold CV		Test		10-fold CV		Test	
0.1	0.1216	(0.0458)	0.12216	(0.0314)	0.0851	(0.0609)	0.08736	(0.0661)
0.2	0.1026	(0.0407)	0.11014	(0.0283)	0.0899	(0.0609)	0.08839	(0.0656)
0.3	0.0953	(0.0395)	0.09871	(0.0245)	0.0911	(0.0599)	0.09243	(0.0646)
0.4	0.0886	(0.0393)	0.09191	(0.0219)	0.0954	(0.0577)	0.09783	(0.0621)
0.5	0.0827	(0.0335)	0.08414	(0.0187)	0.1047	(0.0558)	0.10452	(0.0581)
0.6	0.0748	(0.0311)	0.07951	(0.0176)	0.1149	(0.0497)	0.11470	(0.0522)
0.7	0.0748	(0.0304)	0.07805	(0.0172)	0.1288	(0.0439)	0.12865	(0.0443)
0.8	0.0699	(0.0288)	0.07454	(0.0156)	0.1431	(0.0389)	0.14623	(0.0346)
0.9	0.0699	(0.0301)	0.07337	(0.0156)	0.1594	(0.0414)	0.16574	(0.0299)
1	0.0685	(0.0263)	0.07163	(0.0147)	0.1822	(0.0481)	0.19791	(0.0529)

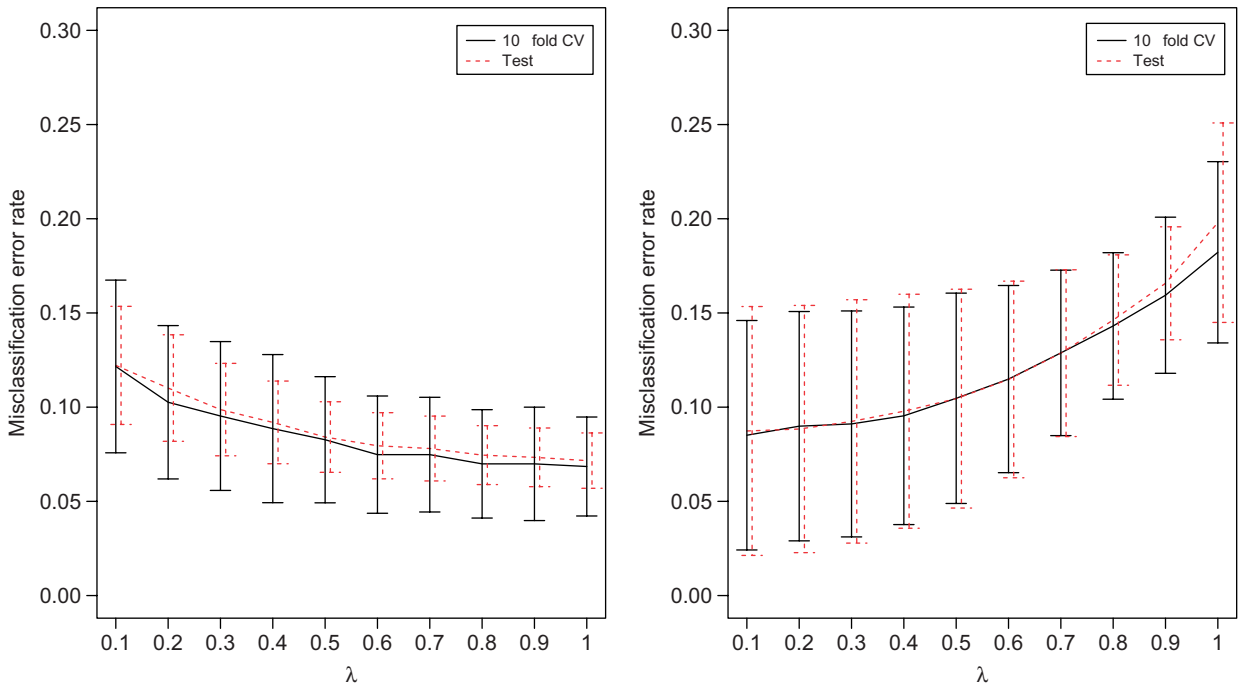

Fig. 6. Average 10-fold CV error and test set error for different values of the shrinkage parameter λ, with error bars at $\pm$ one standard deviation. The left panel refers to the two-group independence simulation scheme, while the right panel to the two-group block dependence structure.

4.3.2. Two-group block dependence structure

In this second setup, we generate a scenario which closer resemble the real microarray data. Again, we generate 100 training samples and 1000 test samples, with half observations from each of two groups, on $p = 1000$ genes. The simulation scheme has been designed to generate a two-group structure with block dependence. Within each block k ($k = 1, \ldots, 10$) there are 100 genes, with the mean for the first group equal to $[0, 0, \ldots, 0]^{\mathrm{T}}$, and equal to $[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, \ldots, 0]^{\mathrm{T}}$ for the second group. The covariance matrix is assumed constant for the two groups, with the value 1 along the diagonal and 0.8 elsewhere, so the genes within a block are highly correlated. On the contrary, genes from different blocks are drawn independently.

Results are shown in Table 1 and in the right panel of Fig. 6: as λ increases the average misclassification error increases too, so a small value of the shrinkage parameter seems to be preferable since it retains the correlation structure in the

data. As for the independence structure, the 10-fold CV error appears to be an accurate estimate of misclassification error obtained from the test sets.

5. A comparison with other classification methods

Lee et al. (2005) provided an extensive comparison of several classification methods applied to a set of publicly available microarray data sets. A brief description of these well-known data sets and of the several classifiers used in the comparison are given in that paper, so we invite the reader to refer to it for details. In this section we compare the classification accuracy of the REGSIR method with some of the "best" classification methods considered in the paper by Lee et al. (2005).

In Table 2 we report the average cross-validation error computed on 200 repetitions using the 2:1 scheme, i.e. using $\frac{2}{3}$ of the data chosen at random as training set and the remaining $\frac{1}{3}$ as test set. For each data set we report the classifier

Table 2
Average misclassification error estimated by 2:1 cross-validation

Data set	Method	*CV* error
Colon	kNN (rank-based)	0.13
(Alon et al., 1999)	LogitBoost and random forest (soft-thresholding)	0.14
$K = 2$	SVM-Rad (soft-thresholding)	0.12
$n = 62$ (40; 22)	PDA (soft-thresholding)	0.16
$p = 2000$	REGSIR ($\lambda = 0.5$, top *BSS*/*WSS* $g = 50$)	0.11
	REGSIR ($\lambda = 0.1$, best $g = 50$, 7/50)	0.07
Leukemia	kNN (soft-thresholding)	0.06
(Golub et al., 1999)	Random forest (*BSS*/*WSS*)	0.04
$K = 3$	SVM-Lin (*BSS*/*WSS*)	0.04
$n = 72$ (38; 9; 25)	PDA (*BSS*/*WSS*)	0.05
$p = 3571$	REGSIR ($\lambda = 0.7$, top *BSS*/*WSS* $g = 50$)	0.03
	REGSIR ($\lambda = 0.2$, best $g = 50$, 7/50)	0.01
Lymphoma	DQDA (rank-based)	0.08
(Alizadeh et al., 2000)	Random forest (*BSS*/*WSS* or soft-thresholding)	0.04
$K = 3$	SVM-Lin and SVM-Rad (*BSS*/*WSS*)	0.02
$n = 66$ (11; 9; 46)	PDA (*BSS*/*WSS* or soft-thresholding)	0.04
$p = 4026$	REGSIR ($\lambda = 0.4$, top *BSS*/*WSS* $g = 50$)	0.02
	REGSIR ($\lambda = 0.6$, best $g = 50$, 3/50)	0.00
SRBCT	DLDA and kNN (rank-based)	0.11
(Khan et al., 2001)	Random forest (*BSS*/*WSS* or soft-thresholding)	0.01
$K = 4$	SVM-Lin and SVM-Rad (*BSS*/*WSS* or soft-thresholding)	0.01
$n = 63$ (23; 20; 12; 8)	PDA (*BSS*/*WSS* or soft-thresholding) and	
$p = 2308$	PAM (soft-thresholding)	0.01
	REGSIR ($\lambda = 0.2$, top *BSS*/*WSS* $g = 50$)	0.00
	REGSIR ($\lambda = 0.2$, best $g = 50$, 15/50)	0.00
Lung	DLDA (*BSS*/*WSS*)	0.06
(Garber et al., 2001)	LogitBoost (*BSS*/*WSS*)	0.06
$K = 5$	SVM-Lin (*BSS*/*WSS*)	0.09
$n = 73$ (41; 6; 17; 5; 4)	PDA (*BSS*/*WSS*)	0.12
$p = 918$	REGSIR ($\lambda = 0.7$, top *BSS*/*WSS* $g = 50$)	0.05
	REGSIR ($\lambda = 0.2$, best $g = 50$, 18/50)	0.04
NCI60	DLDA (soft-thresholding)	0.23
(Scherf et al., 2000)	LogitBoost (soft-thresholding)	0.31
$K = 6$	SVM-Lin (*BSS*/*WSS*)	0.43
$n = 60$ (8; 14; 9; 11; 10; 8)	PDA (*BSS*/*WSS*)	0.31
$p = 1375$	REGSIR ($\lambda = 0.5$, top *BSS*/*WSS* $g = 50$)	0.09
	REGSIR ($\lambda = 0.1$, best $g = 50$, 11/50)	0.02

with the lowest error for any of the four groups of methods identified by Lee et al. (2005): classical, tree, machine learning and generalized discriminant analysis methods. Each classifier is followed by the gene selection filter applied to select the top 50 genes used to build the classifier; no other gene reduction is then pursued. The values referring to such methods are taken from Tables 1–4 of Lee et al. (2005).

For the REGSIR method we computed the average CV error based on the 2:1 scheme. Our results are shown in the last two rows of each data set. In particular, the first row reports the CV error obtained using the "top" 50 genes based on the BSS/WSS criterion, which provides the same ordering of the filtering statistic in Eq. (3.2). This filtering criterion may help to make a direct comparison with the other classification methods. However, it is not necessarily the optimal way to perform feature selection for REGSIR. Furthermore, Lee et al. (2005) used three different filtering criteria, which are in some cases related to one or more classifiers: for instance, BSS/WSS is related to discriminant analysis classifiers, while soft-thresholding is strictly related to PAM (Tibshirani et al., 2002). For these reasons, we added a last row to each data set which provides the CV error for the "best" 50 genes selected using the backward procedure discussed in Section 3.4. In either cases, results are shown for the shrinkage parameter λ which provided on average the smallest CV error; since, in some cases, different values of λ gave equivalent error rates, we report only the smaller value of the shrinkage parameter. Finally, the second row also reports the fraction of the selected "best" genes which are among the "top" 50 genes based on the BSS/WSS ratio.

The results of REGSIR estimated using the "top" 50 genes with the largest BSS/WSS are remarkably good (see Table 2). Except for one case (Lymphoma), the average CV error provided by REGSIR is always smaller than the errors provided by the best methods for each group of classifiers considered by Lee et al. (2005). This improvement is usually in the order of few decimal points, but in the case of the NCI60 data set is quite large.

If we focus our attention to the performance of REGSIR using the "best" 50 genes, the results are even better: in all cases the average CV error is smaller than that obtained using the top 50 genes, except for the Khan data (SRBCT) where both methods achieve a zero error rate. We also note that only part of the genes included in the "best" subset are among the top 50 with the largest BSS/WSS ratio, ranging from 18 genes for the Lung data to only three genes for the Lymphoma data. Notwithstanding, such behavior highlights the importance of feature selection in building an accurate classifier, and that a "tailored" gene selection method can improve the performance of a classifier. This because each classification method uses different aspects of the data for training, so no single feature selection method is uniformly better than the others, but it must be chosen in conjunction with the characteristics of the classifier used.

In summary, REGSIR performance is comparable to the best classifiers for the data sets considered here, and when appropriate feature selection is made the performance of the method can be considerably improved.

6. Discussion

In this paper we proposed an approach based on SIR for the classification of diagnostic outcomes based on DNA expression profiles. This type of data is characterized by the overwhelming number of genes with respect to the limited number of samples available. SIR is a dimension reduction method which requires non-singular covariance structure, so a form of regularization is required. The proposed methodology, called REGSIR, allows to estimate SIR directions even in the presence of such under-resolution. The estimated REGSIR subspace can then be used to classify each sample.

As pointed out by Cook and Yin (2001) in the classification setting, SIR gains information for estimation from the variation in the class means, and it is able to recover at least a part of the drs. Since gene expression profiles within classes mainly differ on their mean structure, SIR often produces useful discriminant directions. Furthermore, the linear design condition discussed in Section 2 is likely to be satisfied in most gene expression data sets.

However, many included genes provide redundant information on classification, so a gene selection procedure would allow to improve the prediction performance of the classifier, and provide a better understanding of the underlying biological system that generates data. For this reason, a backward feature selection procedure is introduced with the aim of selecting at each step the genes which best approximate the estimated REGSIR subspace.

The proposed methodology has been applied to several publicly available microarray data sets, and the classification accuracy evaluated. Overall the REGSIR approach appeared to perform quite well, often achieving the smallest misclassification error provided by different classifiers. In particular, the backward feature selection approach adopted in conjunction with REGSIR seemed to be an effective way for screening the relevant features and removing the redundant genes.

Software written in the R language (R Development Core Team, 2005) implementing the method is available on request from the author.

Acknowledgments

I would like to thank you both the editors of this special issue of CSDA, and two anonymous referees, for their helpful suggestions and comments that have led to improvements of the paper.

References

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T.Jr.J.H., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Bostein, D., Brown, P.O., Staudt, L.M., 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, 503–511.

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J., 1999. Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA 96 (12), 6745–6750.

Antoniadis, A., Lambert-Lacroix, S., Leblanc, F., 2002. Effective dimension reduction methods for tumor classification using gene expression data. Bioinformatics 19 (5), 563–570.

Bura, E., Pfeiffer, R.M., 2003. Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. Bioinformatics 19 (10), 1252–1258.

Chiaromonte, F., Martinelli, J., 2002. Dimension reduction strategies for analyzing global gene expression data with a response. Math. Biosci. 176, 123–144.

Cook, R.D., 1998. Regression Graphics: Ideas for Studying Regressions through Graphics. Wiley, New York.

Cook, R.D., Weisberg, S., 1991. Comment on "Sliced inverse regression for dimension reduction" by K.C. Li. J. Amer. Statist. Assoc. 86, 328–332.

Cook, R.D., Yin, X., 2001. Dimension-reduction and visualization in discriminant analysis. Aust. N. Z. J. Statist. 43, 147–200.

Dudoit, S., Fridlyand, J., Speed, T.P., 2002a. Comparison of discrimination methods for the classification of tumors using gene expression data. J. Amer. Statist. Assoc. 97, 77–87.

Dudoit, S., Yang, Y.H., Callow, M.J., Speed, T.P., 2002b. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statist. Sinica 12, 111–139.

Friedman, J.H., 1989. Regularized discriminant analysis. J. Amer. Statist. Assoc. 84, 165–175.

Garber, M.E., Troyanskaya, O.G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., de Rijn, M.V., Rosen, G.D., Perou, C.M., Whyte, R.I., Altman, R.B., Brown, P.O., Botstein, D., Petersen, I., 2001. Diversity of gene expression in adenocarcinoma of the lung. Proc. Natl. Acad. Sci. USA 98 (24), 13784–13789.

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., Lander, E.S., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–537.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach. Learn. 46, 389–422.

Hastie, T., Buja, A., Tibshirani, R., 1995. Penalized discriminant analysis. Ann. Statist. 23, 73–102.

Hastie, T., Tibshirani, R., Friedman, J.H., 2001. The Elements of Statistical Learning; Data Mining, Inference and Prediction. Springer, Wiley, Berlin, New York.

Khan, J., Wei, J.S., Ringner, M., Saal, L.H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C.R., Peterson, C., Meltzer, P.S., 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. 7, 673–679.

Lee, J.W., Lee, J.B., Park, M., Song, S.H., 2005. An extensive comparison of recent classification tools applied to microarray data. Comput. Statist. Data Anal. 48, 869–885.

Li, K.C., 1991. Sliced inverse regression for dimension reduction. J. Amer. Statist. Assoc. 86, 316–342.

Li, K.C., 2000. High dimensional data analysis via the SIR/PHD approach. Unpublished manuscript.

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

R Development Core Team, 2005. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL ⟨http://www.R-project.org⟩.

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

Seber, G.A.F., 1984. Multivariate Observations. Wiley, New York.

Tibshirani, R., Hastie, T., Narashiman, B., Chu, G., 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS 99, 6567–6572.

Velilla, S., 1993. A note on the multivariate Box–Cox transformations to normality. Statist. Probab. Lett. 17, 315–322.

Xia, Y., Tong, H., Li, W.K., Zhu, L.X., 2002. An adaptive estimation of dimension reduction space. J. Roy. Statist. Soc. B. 64, 363–410.