Classification using Partial Least Squares with Penalized Logistic Regression

Gersende Fort and Sophie Lambert-Lacroix
CNRS/LMC-IMAG, BP 53, 38041 Grenoble cedex 9, France

October 27, 2004

Abstract

Motivation: One important aspect of data-mining of microarray data is to discover the molecular variation among cancers. In microarray studies, the number n of samples is relatively small compared to the number p of genes per sample (usually in thousands). It is known that standard statistical methods in classification are efficient (*i.e.* in the present case, yield successful classifiers) particularly when n is (far) larger than p. This naturally calls for the use of a dimension reduction procedure together with the classification one.

Results: In this paper, the question of classification in such a high dimensional setting is addressed. We view the classification problem as a regression one with few observations and many predictor variables. We propose a new method combining Partial Least Squares (PLS) and Ridge penalized logistic regression. We review the existing methods based on PLS and / or penalized likelihood techniques, outline their interest in some cases and theoretically explain their sometimes poor behavior. Our procedure is compared with these other classifiers. The predictive performance of the resulting classification rule is illustrated on three data sets: Leukemia, Colon and Prostate.

Availability: Software that implements the procedures and data source on which this paper focuses are freely available at `http://www-lmc.imag.fr/SMS/membres/Gersende_Fort,Sophie_Lambert.html`

Contact: Gersende.Fort,Sophie.Lambert@imag.fr

Introduction

Microarray technology generates a vast amount of data by measuring, through the hybridization process, the levels of virtually all the genes expressed in a biological sample. One can expect that knowledge gleaned from microarray data will contribute significantly to advances in fundamental questions in biology as well as in clinical medicine.

One important goal of analyzing microarray data is to classify the samples. To cite a few, Golub *et al.* [13] have considered classification of acute leukemia, Alon *et al.* [2] have addressed the cluster analysis of tumor and normal

colon tissues. The approaches developed in these papers consist in discrimination methods and machine learning methods (see [6] for a comparative study).

In microarray studies, the number of samples, n, is relatively small compared to the number of genes, p usually in thousands. Unless a preliminary variable selection step is performed, standard statistical methods in classification perform poorly because there are far more variables than observations. One problem is multicollinearity : estimating equations become singular and have no unique and stable solution. For instance, the pooled within-class sample covariance matrix in Fisher's linear discriminant function is singular if $n < p + 2$. Even if all genes can be used as in support vector machines, it seems to be not sensible to use all the genes. Indeed, this use allows presence of the noise associated with genes of little or no discrimination power. That inhibits and degrades the performances of the classification rules in its application to unclassified tumor. In this situation, dimension reduction is needed to reduce the high p-dimensional gene space. In most previously mentioned works, the authors have used univariate methods for reducing the number of genes. Alternative approaches to handle the dimension reduction problem can also be used (see for instance [11, 22, 26, 3]).

Similar data structures have been encountered in the field of chemometrics. The method of Partial Least Squares (PLS, [27, 21, 15]) has been found to be a useful dimension reduction technique as well as Principal Component Regression (PCR, [20]) (see [9] for a statistical view of PLS and PCR). In the context of microarrays, the purpose of PCR is to produce orthogonal tumor descriptors that reduce the dimension to only few gene components (super-genes)[26]. But the dimension reduction is achieved without regard to the response variable and may be inefficient. chosen so that the sample covariance between the response and a linear combination of the p predictors (genes) is maximum.

Nguyen and Rocke [22] proposed using PLS for dimension reduction as a preliminary step to classification, based either on linear logistic discrimination, or linear or quadratic discriminant analysis. However, this seems to be intuitively unappealing because PLS is really designed to handle continuous responses and models that do not suffer from heteroscedasticity as it is the case for Bernoulli or multinomial data. Furthermore, in practice we have observed problems in the convergence of the Iteratively Reweighted Least Squares (IRLS) algorithm, which is the usual procedure for solving the maximum likelihood (ML) equation in the field of the generalized linear models (GLM). Indeed, for logistic regression, it is well known that convergence poses a long standing problem. Infinite parameter estimates can occur depending on the configuration of the sample points in the observation space ([1]).

Marx [19] proposed an extension of PLS to categorical response variable and illustrates the developments from a spectroscopy example. His approach embeds the usual PLS steps within the IRLS. Unfortunately, we have observed that this algorithm does not converge in many cases of interest (such as in the applications considered in this paper). More recently, Ding and Gentleman [5] proposed an approach based on this procedure. They phrased the problem in a GLM setting and applied Firth's procedure to avoid (quasi)separation.

To deal with the high dimension problem, another approach consists in penalizing the likelihood. Eilers *et al.* [7] propose to use the Ridge penalized logistic regression in order to both stabilize the statistical problem and

remove numerical degeneracy due to multicollinearity. They have shown that this method appears to work well with microarray data. Note that this method is not a dimension reduction technique. Indeed all explanatory variables are allowed into the regression model. From the log-likelihood a so-called ridge penalty is subtracted. All the genes contribute, which can inhibit and degrade the performances of the classification rules. Note that we can find alternative approaches (see for example [16] and [12]) for which the classification problem is not viewed as a problem in a logistic regression.

In this paper, we extend the PLS method to binary response variable. To do that, we want to substitute the categorical response variable in the input of PLS by a continuous-valued pseudo-response variable whose expected value has a linear relationship with the covariates. The limiting pseudo-response variable in the IRLS algorithm seems to be a good candidate. Unfortunately, in the present situation "small n, large p", IRLS no longer works since the limiting pseudo-response variable is, in norm, infinite. The idea developed here is to penalize with a Ridge penalty the likelihood criterion in order to constrain the pseudo-response variable to be finite. That is, our procedure combines a Ridge penalty step and a PLS step and the dimension reduction step is incorporated in the classification step. Here we present classification rule for binary response variable indicating normal or colon tumor, for instance. Nevertheless, our approach remains valid for multi-categorical response variables. But the binary case is the simplest case which allows us to point out whether such a procedure works well or not and why.

This paper is organized as follows. The Methods section is the methodological part of this paper. It contains a description of the logistic regression and linear discrimination. We then recall the Ridge regression method and derive a weighted PLS algorithm in order to address the dimension reduction in heteroscedastic models. We then introduce an extension of PLS to GLM based on the Ridge penalty, and analyze the Nguyen and Rocke, Marx, Ding and Gentleman and Eilers *et al.* 's algorithms. Applications to disease classification through microarray are presented in the Results section.

Methods

Some basic ingredients

After introducing some notations, we recall the principle of linear logistic discrimination, some results on the existence of the maximum likelihood estimator and the classical algorithm used to compute it. Next, we present a regularization method, a penalized maximum likelihood method, and a dimension reduction technique, PLS.

Notations

Expression levels of the p genes for the n microarray samples are collected in a $n \times p$ *data matrix* $X = (x_{ij})$, $1 \leq i \leq n, 1 \leq j \leq p$. The entry x_{ij} is the expression level of the variable "gene" j in the microarray sample i, and the i-th row $X_{i,\cdot}$ is the vector of a gene expression profile for sample i. More generally, for a matrix A, $A_{i,j}$ denotes the entry (i,j), $A_{\cdot,j}$ (resp. $A_{i,\cdot}$) denotes the column vector collecting the column $\#j$ (resp. the row $\#i$). $A_{i_1:i_2,j_1:j_2}$

is the $(i_2 - i_1 + 1) \times (j_2 - j_1 + 1)$ matrix formed by picking out the rows i_1 to i_2 and columns j_1 to j_2 of A; $A_{\cdot j_1:j_2}$ is formed by picking out the columns j_1 to j_2 of A. The labels of the n microarray samples are collected in a $\{0, \ldots, (g-1)\}^n$-valued vector $\underline{\mathbf{y}}$. In supervised machine learning, each sample is thought to originate from a specific class $k \in \{0, \ldots, g-1\}$ where the number of possible classes g is known and fixed. A classifier can be regarded as a function $G : \mathbb{R}^p \to \{0, \ldots, g-1\}$ that predicts the unknown class label of a new tissue sample $x \in \mathbb{R}^p$ by $G(x)$. We assume that the data $(\underline{\mathbf{y}}, X)$ collect observations of n statistically independent and identically distributed random pairs $(Y, \mathbf{X})$. We choose a logit model for the data (see e.g. [8]), and the Logistic Discrimination (LD) method for the classification procedure (see e.g. [25]). In the terminology of the regression analysis, $(X_{\cdot j})_{1 \le j \le p}$ are the predictor variables and $(\underline{\mathbf{y}}_i)_{1 \le i \le n}$ the response variables. We include an intercept into the regression model, and denote by $Z = [\mathbb{1}_n \; X]$ the *design matrix* of size $n \times (p+1)$, where $\mathbb{1}_n = (1, \cdots, 1)'$ stands for the column vector of length n ($'$ denotes the transposition operator).

Linear Logistic Discrimination

In logit models, the conditional class probability - or equivalently, the conditional expectation of Y given $\mathbf{X}$ - $\mathrm{P}(Y = 1 | \mathbf{X} = x; \gamma)$ is related to x and some parameter $\gamma \in \mathbb{R}^{p+1}$ through the relation $\mathrm{P}(Y = 1 | \mathbf{X} = x; \gamma) = h([1 \; x']\gamma)$ where $h(\eta) = 1/(1 + \exp(-\eta))$. The quantity $[1 \; x']\gamma$ is called the linear predictor. γ is an unknown parameter that has to be estimated from the data. In Logistic Discrimination (LD), it is usually estimated by $\hat{\gamma}^{\mathrm{ML}}$, the ML estimator. The log-likelihood of the observations for the value γ of the parameter, simply denoted by $l(\gamma)$, is given by

$$l(\gamma) = \sum_{i=1}^{n} \left\{ \underline{\mathbf{y}}_i \eta_i(\gamma) - \ln\left(1 + \exp(\eta_i(\gamma))\right) \right\}, \tag{1}$$

where for all $1 \le i \le n$, $\eta_i(\gamma) = (Z\gamma)_i$.

For a vector $z = [1 \; x']$, the predicted class $\hat{y}$ of each sample is 1 if $\hat{\pi} > 1 - \hat{\pi}$ and 0 otherwise, where $\hat{\pi} = h(z'\hat{\gamma}^{\mathrm{ML}})$. Nevertheless, as discussed below, in some cases, including in practice the case $n << p$, existence and unicity of $\hat{\gamma}^{\mathrm{ML}}$ for logit models is not guaranteed.

Maximum likelihood estimate and Iteratively Reweighted Least Squares (IRLS)

We say that the ML estimate exists if there exists $\gamma \in \mathbb{R}^{p+1}$ of finite norm which is a maximizer of the concave log-likelihood l. Hence, such an estimate is a solution to the normal equation $Z'(\underline{\mathbf{y}} - \pi(\gamma)) = 0$, where $\pi(\gamma)$ is the $\mathbb{R}^n$-valued mean vector with coordinates $\pi_i(\gamma) = h(\eta_i(\gamma))$.

If Z is full column-rank, the solution, when exists, is unique. Existence of a solution, when Z is full column-rank, depends on the configuration of the n samples points in the observation space $\mathbb{R}^p$ [1, 23]. There are three exclusive situations: separate, quasi-separate and overlap situations. In the first two cases, there exists $\hat{\gamma}$ such that $(Z\hat{\gamma})_i \ge 0$ for all i such that $\underline{\mathbf{y}}_i = 1$ and $(Z\hat{\gamma})_i \le 0$ for all i such that $\underline{\mathbf{y}}_i = 0$; roughly speaking, this means that there exists an hyperplane that exactly separates the two classes, except maybe some points that can belong to the hyperplane.

In such a case, l reaches its maximum as $\|\gamma\|$ tends to $+\infty$ and the ML estimate does not exist. In the third case, the estimate exists and is computed as the limit of a converging Newton-Raphson sequence; this algorithm is known as the Iteratively Reweighted Least Squares (IRLS) algorithm [14]. Let $W(\gamma)$ be the diagonal $n \times n$ matrix with diagonal entries $W_{i,i}(\gamma) = \pi_i(\gamma)(1 - \pi_i(\gamma))$. Each iteration divides into two steps,

$$z^{(t)} = Z\gamma^{(t)} + \left[W^{(t)}\right]^{-1} \left(\underline{\mathbf{y}} - \pi^{(t)}\right), \tag{2}$$

$$\gamma^{(t+1)} = \left(Z'W^{(t)}Z\right)^{-1} Z'W^{(t)}z^{(t)}, \tag{3}$$

where $W^{(t)}$ and $\pi^{(t)}$ are shorthand notations for $W(\gamma^{(t)})$ and $\pi(\gamma^{(t)})$. IRLS can thus be considered as iterative weighted least square regression of a $\mathbb{R}^n$-valued pseudo-variable $z^{(t)}$ onto the columns of Z.

When Z is not full column-rank, the parameter is not identifiable and the ML estimate is not unique when exists; applying the above iterations (2-3) by replacing the inverse matrix (3) with the Moore-Penrose pseudo-inverse, yields the parameter estimate which is of minimal norm among all the solutions. In practice, in the present statistical framework $n << p$, $n = \text{rank}(Z)$ and the minimal norm solution verifies for all $1 \leq i \leq n$, $(Z\gamma)_i = \ln(\underline{\mathbf{y}}_i) - \ln(1 - \underline{\mathbf{y}}_i)$; it is thus of infinite norm and the ML estimate can not exist. This calls for regularization methods.

Ridge penalty and RIRLS

The ridge estimator [18] $\hat{\gamma}^R$ is defined as the (unique) maximizer of the penalized likelihood $l^*(\gamma) = l(\gamma) - 0.5\lambda\gamma'\Sigma^2\gamma$, where $\lambda > 0$ is the shrinkage parameter, and Σ^2 is a diagonal matrix with entries $\Sigma^2_{1,1} = 0$ and

$$\Sigma^2_{j,j} = \sum_{i=1}^{n} (Z_{i,j} - \mathbb{1}'_n Z_{\cdot,j}/n)^2, \quad j \in \{2, \cdots, p+1\}. \tag{4}$$

The weighted penalty term takes into account the non-scaling of the covariate matrix X, and does not apply to the location parameter γ_1. $\hat{\gamma}^R$ always exists, is unique and is computed as the limit of a Newton-Raphson sequence. We denote by RIRLS($\underline{\mathbf{y}}, X, \lambda$) (shorthand notation for Ridge-IRLS) this algorithm. It consists in replacing in IRLS, the weighted regression (3) by a weighted Ridge regression $\gamma^{(t+1)} = (Z'W^{(t)}Z + \lambda\Sigma^2)^{-1}Z'W^{(t)}z^{(t)}$, where $z^{(t)}$ is built as in (2).

λ controls the amount of shrinkage in the data and can be chosen as the minimum, over a given range, of the BIC criterion $-2l(\hat{\gamma}^R) + \log(n)\text{trace}[Z\left(Z'W(\hat{\gamma}^R)Z + \lambda\Sigma^2\right)^{-1} Z'W(\hat{\gamma}^R)]$ [17].

Weighted Partial Least Squares (WPLS)

Partial Least Squares (PLS) is both a tool for linear regression and a tool for dimension reduction [27, 21, 15]. Let $\underline{\mathbf{y}} \in \mathbb{R}^n$ be a response vector, X be a $n \times p$ data matrix and W be a positive definite $n \times n$ matrix. PLS (i) defines κ W-orthogonal *scores* $(t_k)_{1 \leq k \leq \kappa}$, linear combinations of the columns of Z and such that for all k, $\mathbb{1}'_n W t_k = 0$, and (ii) performs a W-weighted least squares regression of $\underline{\mathbf{y}}$ on $(\mathbb{1}_n, t_1, \cdots, t_\kappa)$. This yields the decomposition

$$\underline{\mathbf{y}} = q_0 \mathbb{1}_n + q_1 t_1 + \cdots + q_\kappa t_\kappa + f_{\kappa+1} = Z\hat{\gamma}^{\text{PLS},\kappa} + f_{\kappa+1}$$

where the residual term $f_{\kappa+1}$ is W-orthogonal to the vectors $(\mathbb{1}_n, t_1, \cdots, t_\kappa)$. Contrary to classical dimension reduction methods (such as Principal Component Regression), the scores depend on the response vector $\underline{\mathbf{y}}$; roughly speaking, given $(t_k)_{1\leq k\leq l}$, t_{l+1} is the linear combination of the columns of Z, *i.e.* is on the form $t_{l+1} = Zc$, which is the most informative on the residual response variable f_{l+1}, when information is defined in terms of the weighted covariance $|\mathrm{Cov}(\sqrt{W}Zc, \sqrt{W}f_{l+1})|$ ($\sqrt{W}$ denotes the square root matrix of W) [15]. While the maximal number of PLS scores $\kappa_{\max}$ can be lower than $\mathrm{rank}(X)$, in practice, it is often equal to $\mathrm{rank}(X)$. Helland [15] shows that the WPLS regression applied with $\kappa = \kappa_{\max}$ is nothing more than the Weighted Least Squares regression. In the literature, PLS is usually derived with $W = \mathrm{I}$, the identity matrix; we thus detail the algorithm in the weighted case. Let $\tilde{\Sigma}$ be the $p \times p$ positive-definite diagonal matrix with diagonal entries $\Sigma_{j,j}$, $j \geq 2$, given by (4).

1. $X^s = X\tilde{\Sigma}^{-1}$, $t_0 = \mathbb{1}_n$, $E_0 = X^s$; $f_0 = \underline{\mathbf{y}}$.

2. For $k = 0, \cdots, \kappa$,

$$q_k = t_k' W f_k / (t_k' W t_k), \qquad f_{k+1} = f_k - q_k t_k,$$
$$E_{k+1} = E_k - t_k t_k' W E_k / (t_k' W t_k),$$
$$t_{k+1} = E_{k+1} E_{k+1}' W f_{k+1}.$$

Hereafter, this procedure is denoted by WPLS $(\underline{\mathbf{y}}, X, W, \kappa)$. If Z is full column-rank, this algorithm determines an unique estimate $\hat{\gamma}^{\mathrm{PLS},\kappa}$ satisfying $\underline{\mathbf{y}} - f_{k+1} = Z\hat{\gamma}^{\mathrm{PLS},\kappa}$; if Z is not full column-rank, the procedure above yields the minimal norm vector among all the vectors verifying $\underline{\mathbf{y}} - f_{k+1} = Z\gamma$.

Ridge Partial Least Squares (RPLS)

A direct application of PLS to GLM seems to be intuitively unappealing because PLS handles continuous responses. This is the reason why, in order to extend PLS to GLM, we want to replace the binary vector $\underline{\mathbf{y}}$ with a pseudo-response variable whose expected value has a linear relationship with the covariates. The pseudo-response variable z^∞ at convergence of RIRLS$(\underline{\mathbf{y}}, X, \lambda)$ verifies this condition and is thus our candidate : it is on the form $z^\infty = Z\hat{\gamma}^{\mathrm{R}} + \varepsilon$, where, conditionally to $\hat{\gamma}^{\mathrm{R}}$ being the true value of the parameter, ε is a centered vector of covariance matrix $(W^\infty)^{-1}$. The main advantage of choosing z^∞ instead of, for example, the pseudo-variable at convergence of IRLS - which has the linear structure too- is that this allows the combination of a regularization step and of a dimension reduction step. In addition, this extension is always well-defined : recall indeed that in some cases (including the case $n << p$), the ML estimate does not exist so that the pseudo-variable 'at convergence' of IRLS is of infinite norm.
As a consequence, we propose a new procedure which combines Ridge penalty - the regularization step - and PLS - the dimension reduction step - and so called Ridge-PLS (RPLS). Let λ be some positive real constant and κ be some positive integer. RPLS divides in two steps:

1. $(z^\infty, W^\infty) \longleftarrow$ RIRLS$(\underline{\mathbf{y}}, X, \lambda)$;

2. $\hat{\gamma}^{\mathrm{PLS},\kappa} \longleftarrow \mathrm{WPLS}(z^{\infty}, X, W^{\infty}, \kappa)$.

A detailed implementation is given in the Appendix. The first step builds a continuous response variable z^{∞} for the input of PLS, the "dispersion matrix" of which is $[W^{\infty}]^{-1}$. This explains the call, in the second step, to a weighted PLS procedure with weight W^{∞}. The use of X^s in WPLS and of Σ in the penalized ridge criterion, makes our procedure invariant to the scaling of the data matrix.

RPLS depends on two parameters, λ and κ. λ is determined at the end of Step 1, as minimizing the BIC criterion (see the Ridge penalty section), and thus independently of κ. In the linear regression setting, the optimal choice of κ when dimension reduction is achieved by PLS, is to our best knowledge, an open problem: the non linear dependence of $\hat{\gamma}^{\mathrm{PLS},\kappa}$ upon the response vector, makes an explicit control of the error term $f_{\kappa+1}$ impossible. Finally, observe that RPLS provides an estimate $\hat{\gamma}^{\mathrm{RPLS}}$ (which is unique, given $\underline{\mathbf{y}}$, X, λ and κ).

We are now able to answer to the classification problem in a high dimensional setting : our classification procedure consists in applying LD with the estimate $\hat{\gamma}^{\mathrm{RPLS}}$.

Comparison with other approaches

We briefly review some regression procedures that use PLS as the dimension reduction tool to manage the high dimensional setting. We outline their interest and in some cases, explain their poor behavior.

Nguyen and Rocke's approach

Nguyen and Rocke [22] substitute the data matrix X by a $n \times \kappa$ matrix $\tilde{X}$, the columns of which are the first κ PLS-scores given by WPLS $(\underline{\mathbf{y}}, X, \mathrm{I})$. Then they estimate the parameter in the ML sense by running IRLS $(\underline{\mathbf{y}}, \tilde{X})$. This yields $\hat{\gamma}^{\mathrm{NR}}$. As mentioned above, applying PLS with a binary input $\underline{\mathbf{y}}$ is unappealing; in addition, the PLS-regression step does not take into account the heteroscedasticity of the response vector $\underline{\mathbf{y}}$; finally, in many applications, $\|\hat{\gamma}^{\mathrm{NR}}\| = \infty$ since the ML estimate does not exist.

In practice, IRLS is stopped after a maximal number of iterations $n_{\max}$ thus hiding the non-convergence of IRLS. Unfortunately, the estimate $\hat{\gamma}^{\mathrm{NR}}$ depends on $n_{\max}$ and this yields an unstable procedure for classification. We observed this phenomenon on the Leukemia data set. $\hat{\gamma}^{\mathrm{NR}}$ is estimated by using the data in the Golub's training set [13]; classification is performed on the samples from the test set. When $p = 150$ and $\kappa = 3$, there are 1 (resp. 2) samples incorrectly classified if $n_{\max} = 7$ (resp. $n_{\max} = 10$) .

Marx's approach

In Marx [19], the parameter γ is estimated in the ML sense and is obtained at convergence of IRLS$(\underline{\mathbf{y}}, \tilde{X})$, where $\tilde{X}$ is defined by IRPLS, an algorithm that extends PLS to GLM. More precisely, IRPLS can be understood as an IRLS algorithm in which the weighted least squares regression (3) is replaced with the weighted PLS regression, $\mathrm{WPLS}(z^{(t)}, X, W^{(t)}, \mathrm{rank}(E_1))$. $\tilde{X}$ collects the first κ components "at convergence" of IRPLS.

As recalled above, WPLS applied with the maximal number of PLS components is nothing else than Weighted Least Squares (note that Marx chooses $\kappa = \text{rank}(E_1)$ while in theory, $\kappa_{\max}$ may be strictly lower than $\text{rank}(E_1)$). Hence IRPLS and IRLS coincide, and, when X is full row-rank (which is most often the case when $n << p$), IRPLS never converges. In practice, IRPLS is stopped after a fixed number of iterations, thus hiding the non-convergence phenomenon. In addition, initializing IRPLS by choosing a linear predictor on the form $\eta^{(0)} = c_0\underline{\mathbf{y}} - c_0(\mathbb{1}_n - \underline{\mathbf{y}})$ (where for example $c_0 = \ln(3)$), as done in Marx, yields $\hat{\gamma}^{\text{M}} = \hat{\gamma}^{\text{NR}}$. A trivial induction shows that for all $t \geq 0$, $z^{(t)} = 2c_t\underline{\mathbf{y}} - c_t\,\mathbb{1}_n$ with $c_t = 1 + c_{t-1} + \exp(-c_{t-1})$, and $W^{(t)}$ is proportional to the identity matrix I_n. Since $\text{WPLS}(\underline{\mathbf{y}}, X, W, \kappa) = \text{WPLS}(\alpha\underline{\mathbf{y}} + \beta\mathbb{1}_n, X, W, \kappa)$ - in terms of the exhibited scores -, for all $\alpha, \beta \in \mathbb{R}$, WPLS $(z^{(t)}, X, W^{(t)}, \kappa)$ returns the same scores as WPLS $(\underline{\mathbf{y}}, X, \text{I}_n, \kappa)$, thus proving $\hat{\gamma}^{\text{M}} = \hat{\gamma}^{\text{NR}}$.

Ding and Gentleman's approach

The originality of their work [5] is that it simultaneously answers to the regularization question and to the dimension reduction one. They run an approximation of a Newton-Raphson (NR) algorithm for solving a Firth's penalized ML criterion. As in IRLS, any iteration of the NR algorithm is a Weighted Least Squares regression and Ding and Gentleman replace this Least-Square regression by a Weighted PLS one. We call this algorithm FPLS.
We run their method on the data sets described in the next section. On the Colon data set and on the prostate data set, the algorithm does not always converge: we observe a cyclic behavior : after a burn-in period the path is periodic; the estimate $\hat{\gamma}^{\text{DG}}$ and consequently the classification rule, may depend on the maximal number of iterations.
This approach is greatly promising since it addresses both the regularization and the dimension reduction problems. Comparisons of our results with their approach are of interest and will be explored in future research.

Eilers *et al.* 's approach

Their method [7] does not use PLS. We nevertheless mention their work since their estimate, $\hat{\gamma}^{\text{E}}$ is the Ridge-penalized ML estimate (with an un-weighted penalty term i.e. $\Sigma^2 = \text{I}$). The Eilers's *et al.* method does not reduce the dimension and only deals with the regularization question. In particular, all the explanatory variables are allowed and included into the regression model, which can deteriorate the performances of the classifier. In the next section, we will outline the high interest of combining a reduction step with the Ridge regularization.

Results

We illustrate the interest of RPLS by considering applications to classification of microarrays data. We compare the classification results from our procedure with those of other classifiers including RIRLS, FPLS, the effective dimension reduction (MAVE,[3]), diagonal linear discriminant analysis (DLDA), diagonal quadratic discriminant analysis (DQDA) and k-nearest neighbors (KNN) based on the Euclidean distance (see [4] for an overview of these last three methods).

DLDA, DQDA and KNN are thus introduced in the present paper as "classical statistical method". As commented in the abstract, our goal is to show that these methods poorly behave when applied to high-dimensional data sets. This is exactly what happens, thus stressing the interest of methods based on the regularization and dimension reduction.

In order to illustrate the interest of PLS over PCR for regression, we compare our algorithm RPLS to 'RPCR' (for Ridge-PCR). By nature, PCR handles continuous responses; this calls for an extension of PCR to GLM, in order to use it as a dimension reduction in GLM. The extension we derived for PLS remains valid for PCR: we exhibit the continuous-valued pseudo-response variable at convergence of the RIRLS algorithm and use this variable as the input variable for PCR. This yields RPCR.

Data, pre-processing and Gene selection

We will consider in turn the Leukemia, Colon and Prostate data sets. [1] The Leukemia data set, contains 72 tissue samples with $p_{\text{init}} = 7129$ genes: 47 cases of acute lymphoblastic leukemia (ALL), coded 0, and 25 cases of acute myeloid leukemia (AML), coded 1 [13]. The Colon data set contains 62 tissue samples with $p_{\text{init}} = 2000$ genes: 40 tumors tissues, coded 1, and 22 normal tissues, coded 0 [2]. The Prostate data set contains 102 tissue samples with $p_{\text{init}} = 12600$ genes: 52 tumors tissues, coded 1, and 50 normal tissues, coded 0 [24].

For Leukemia and Colon data (resp. Prostate), the pre-processing steps of [6] (resp. [24]) are applied: thresholding (floor of 100 (resp. 10) and ceiling of 16000)/ filtering (exclusion of genes with max/min≤ 5 and (max-min)≤ 500 (resp. 50)/ $\log_{10}$-transformation / standardization. Notice that the filtering step is applied using only the Learning set. This yields a resulting number of covariates $p_{\max}$ depending on the subdivision Learning and Testing set, lower than p_{init} but still far larger than the number of observations.

Although the procedures can handle a large number (thousands) of genes, the number of genes may be still too large for practical use. Furthermore, a considerable percentage of the genes do not show differential expression across groups and only a subset of genes is of interest. We perform the preliminary selection of gene based on the BSS/WSS criterion used in [6]. When training the rule for the selection of gene, we select p genes by the previous criterion with $p \in \mathcal{P}_l = \{50, 300, 500, 1000\}$ for Leukemia data, $p \in \mathcal{P}_c = \{100, 500, 1000, p_{\max}\}$ for Colon data and $p \in \mathcal{P}_p = \{100, 500, 1000, 1500\}$ for Prostate data.

Assessing prediction methods

It is common to assess the performance of the classification rules for a selected subset of genes by their errors on the test set and also by their leave-one-out cross-validated errors. Due to the instability of leave-one-out error rates, we moreover perform re-randomization study *i.e.* an out of sample analysis on 100 random subdivisions of the data set into a learning set and a test set. When a test set is available, we randomly split the original data set into a training

[1] They can be downloaded from `http: //sdmc.lit.org.sg/GEDatasets/Datasets.html`

set and a test set of the same size as the original ones. Otherwise, we choose a test set size equal to one third of the data (2:1 scheme of [6]). Each subdivision yields a test set error rate for each predictor; boxplots are used to summarize these error rates over the runs.

The optimal number of PLS or PCR components (for RPLS, FPLS or RPCR) is selected by choosing the value of κ minimizing leave-one-out error rates for the training set. This is also employed for other procedures that involve hyperparameters, such as MAVE or KNN. In practice, on the Leave One Out analyzes performed on the Colon data sets and on the Prostate Data sets, we observed many cases of indecisions for even values of k. This is the reason why, as suggested in Devroye, Gyorki and Lugosi [4], we run KNN for odd values of k. We really believe that the frequent occurrence of the indecision case shows that KNN is not a pertinent method (for this kind of data sets). The weakness of this classical statistical method is clearly illustrated by the numerical results.

The κ range is given by $\mathcal{K}_l = \{1, \ldots, 8\}$ for Leukemia data, $\mathcal{K}_c = \{1, \ldots, 9\}$ for Colon data and $\mathcal{K}_p = \{1, \ldots, 14\}$ for Prostate data. Moreover, the shrinkage parameter (for RIRLS, RPLS or RPCR) is determined as mentioned above on 51 $\log_{10}$-linearly spaced points in the range $[10^{-2}; 10^3]$. Note that, to fairly evaluate and compare the test or leave-one-out cross-validated errors, pre-processing, gene selection and (hyper)-parameters estimations are performed on the training set (at each step of the cross-validation process).

Discussion

Different numerical results are reported in Tables 1 to 3 and boxplots are plotted in Figures 1 to 3. In the tables, the number in brackets for RPLS, RPCR, FPLS or MAVE are the optimal numbers of components chosen as previously indicated and those for KNN are the optimal numbers of nearest neighbors. The numerical results and graphics show the necessity of the dimension reduction step. This is particularly evident from the Colon and prostate data results. Indeed note that most of the classifiers proposed in the literature well behave on the Leukemia data set though the other data set are known to be more 'problematic'. In particular, the boxplots suggest that errors rates for RPLS, RPCR and FPLS are typically lower and less variable. There is no obvious difference between the distributions of error rates for these three methods. However, we can mention that for Colon and Prostate data FPLS has converged only for small κ values; and that RPCR needs κ values greater than the one of RPLS. Otherwise these methods are robust to the growth of p, thanks to the dimension reduction step (the larger p is, the larger κ has to be chosen to reach the best classification result except for FPLS which does not converge for large κ), and to an increasing value of the shrinkage parameter. The good performance of these methods when $p = p_{\max}$ (Table 2) is particularly interesting when applied to Microarrays, since it can allow the practitioner to avoid a pre-selection step and thus makes the classification result independent of the criterion applied in this preliminary selection. On the other hand, the methods such as RIRLS, DLDA, DQDA or KNN have very poor performances when p gets large. Note that MAVE stands between the both although being a dimension reduction method.

Concerning the comparison between RPLS and RIRLS, as mentioned above we do not trust RIRLS due to the non-scaling of the design matrix that makes the interest of their method problem specific. It may be read in the tables

and figures below that RPLS and RIRLS have an equivalent behavior for "small" p values. Nevertheless the later is not robust to large p. This legitimately suggests to add to this method a dimension reduction step; we observed on the three data sets, in the resampling analysis, that this would improve RIRLS method.

RPLS confirms different analyzes in the literature. For example, it is known that in the Leukemia data set, samples $\#28, 66, 67$ have a high misclassification rate (henceforth denoted MR[i] for sample $\#i$) [6]. In the resampling study, for $\kappa \in \mathcal{K}_l$ and $p \in \mathcal{P}_l$, RPLS systematically misclassifies sample #66, whereas MR[28] and MR[67] decrease when κ and p both increase: for $p = 1000$ and $\kappa = 3$ (resp. $p = 50$ and $\kappa = 1$), MR[28]=7.02% and MR[67]=7.55% (resp. 38.60% and 39.62%). Another example is given by the Colon data set, for which samples $N8, 34$ and $T30, 33, 36$ are misclassified by both the contributions [2, 10]; in the resampling analysis, performed for $\kappa \in \mathcal{K}_c$, $p \in \mathcal{P}_c$, $N8, 34$ and $T36$ are systematically misclassified, MR[T30] $\geq$ 88.89% and MR[T33] $\geq$ 96.87%. In addition, RPLS always misclassifies $N36$ (a sample pointed out in [10]), and behaves poorly for samples $T2, 33$ (samples pointed out in [2]). For the Prostate Data Set, in the leave one out study, the minimal number of misclassified samples is 5 (and is reached by RPLS), namely samples $32, 64, 68, 84, 92$. The samples $32, 84, 92$ are misclassified for all of the LO analysis ($p \in \mathcal{P}_p, \kappa \in \mathcal{K}_p$); samples 64 and 68 are misclassified in more than 96% and 91% of the LO analysis. In the resampling study, MR[32] = 1, MR[64] ≥ 0.41, MR[68] ≥ 0.72, MR[84]= 1 and MR[92]≥ 0.90.

Conclusions

We have proposed a statistical dimension reduction approach for the classification of tumor based on microarray gene expression data. Our method is designed to address the curse of dimensionality to overcome the problem of a high dimensional gene expression space so common in such type of problems. We have provided a new extension of Partial Least Squares to binary response data, that seems to have better properties than some of the currently used methods. We restricted our attention to the binary case, but the methodology can be extended to cover multi-class problems and we are interested in making it. Indeed the structure of the algorithm for the binary case and multi-class case is the same, but the choice of the parameter λ necessitates more attention in the multi-class case than in the binary case. Future research will also concern the variable and model selection themes in order to determine optimal values for (λ, κ).

Acknowledgements

The authors are really grateful to A. Antoniadis for constructive and fruitful discussions and to the referees for their constructive comments and criticisms which have substantially improved this article. They would like also thank I. De Feis for helpful comments and B. Ding and R. Gentleman for providing a preprint of their paper prior to publication.

Part of this work was supported by the research project ASBGEN and the Interuniversity Attraction Pole (IAP)

research network in Statistics P5/24.

Appendix: RPLS

For given $(\underline{\mathbf{y}}, X)$, $\lambda > 0$ and $\kappa \geq 1$.

1. Compute $Z \longleftarrow [\mathbb{1}_n \; X]$ and Σ as in (4).

2. RIRLS step

   a. Initialize $\gamma^{(0)} \in \mathbb{R}^{p+1}$.

      $t \longleftarrow 0$.

   b. Until convergence, do

      $\eta^{(t)} \longleftarrow Z\gamma^{(t)}$.

      $\pi^{(t)} \longleftarrow \left((1+\exp(-\eta_k^{(t)}))^{-1}, 1 \leq k \leq n\right)'$.

      $W^{(t)} \longleftarrow \mathrm{diag}\left(\pi^{(t)}(1-\pi^{(t)})\right)$.

      $z^{(t)} \longleftarrow \eta^{(t)} + \left(W^{(t)}\right)^{-1}\left(\underline{\mathbf{y}} - \pi^{(t)}\right)$.

      $\gamma^{(t+1)} \longleftarrow \left(Z'W^{(t)}Z + \lambda\Sigma^2\right)^{-1} Z'W^{(t)}z^{(t)}$.

      $t \longleftarrow t+1$.

   c. Set

      $z^{\infty} \longleftarrow z^{(t-1)}$.

      $W^{\infty} \longleftarrow W^{(t-1)}$.

3. WPLS step

   a. $\tilde{\Sigma} \longleftarrow \Sigma_{2:p+1,2:p+1}$, $X^s \longleftarrow X\tilde{\Sigma}^{-1}$.

   b. $t_0 \longleftarrow \mathbb{1}_n$, $E_0 \longleftarrow X^s$, $f_0 \longleftarrow z^{\infty}$, $\omega_0 \longleftarrow 0_{\mathbb{R}^p}$, $\psi \longleftarrow \mathrm{I}_p$.

   c. For $k = 0, \cdots, \kappa$,

      $q_k \longleftarrow t_k'W^{\infty}f_k/(t_k'W^{\infty}t_k)$.

      $p_k \longleftarrow E_k'W^{\infty}t_k/(t_k'W^{\infty}t_k)$.

      $f_{k+1} \longleftarrow f_k - q_k t_k$.

      $E_{k+1} \longleftarrow E_k - t_k p_k'$.

      $\psi \longleftarrow \psi(\mathrm{I}_p - \omega_k p_k')$.

      $\omega_{k+1} \longleftarrow E_{k+1}'W^{\infty}f_{k+1}$.

      $\tilde{\psi}_{k+1} \longleftarrow \psi\omega_{k+1}$.

$t_{k+1} \longleftarrow E_{k+1}\omega_{k+1}$.

d. Set $\tilde{\Psi} \longleftarrow [\tilde{\psi}_1 \dots \tilde{\psi}_\kappa]$ and $q \longleftarrow [q_1 \dots q_\kappa]'$.

e. Conclude

$\hat{\alpha} \longleftarrow q_0 - p_0' \tilde{\Psi} q$.

$\hat{\beta} \longleftarrow \tilde{\Sigma}^{-1} \tilde{\Psi} q$.

The procedure, presently derived in $\mathbb{R}^{p+1}$ can be equivalently derived in $\mathbb{R}^{r+1}$ where $r+1 = \text{rank}(Z) \leq n$. To that goal, compute UDV' the singular values decomposition (svd) of $(X - \mathbb{1}_n \mathbb{1}_n' X/n)\tilde{\Sigma}^{-1}$, the scaled covariate matrix; collect the first r columns of UD in $\Xi = (UD)_{\cdot,1:r}$ so that $Z\gamma = [\mathbb{1}_n \ \Xi]\ \theta$ for some $\theta \in \mathbb{R}^{r+1}$. We denote by $J^{(r)}$ a diagonal matrix with $J^{(r)}_{1,1} = 0$ and $J^{(r)}_{k,k} = 1$, $k = 2, \dots, r+1$. It is readily seen that RPLS, run by replacing (X, Σ^2) by $(\Xi, J^{(r)})$, yields an estimate $\hat{\theta}^{\text{PLS},\kappa}$ uniquely related to $\hat{\gamma}^{\text{PLS},\kappa}$ by the formulas

$$\hat{\gamma}_{2:p+1} = \tilde{\Sigma}^{-1} V_{\cdot,1:r} \hat{\theta}_{2:r+1}, \qquad \hat{\gamma}_1 = \hat{\theta}_1 - \mathbb{1}_n' X \hat{\gamma}_{2:p+1}/n.$$

Hence, up to a single svd, the procedure is independent of p which is of computational importance.

References

[1] A. Albert and J. Anderson. On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, 71(1):1–10, 1984.

[2] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine. 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(12):6745–6750, 1999.

[3] A. Antoniadis, S. Lambert-Lacroix, and F. Leblanc. Effective Dimension Reduction Methods for Tumor Classification using gene Expression Data. *Bioinformatics*, 19(5):563–570, 2003.

[4] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, New-York, 1996.

[5] B. Ding and R. Gentleman. Classification Using Generalized Partial Least Squares. Technical Report 5, Bioconductor Project Working Papers, 2004.

[6] S. Dudoit, J. Fridlyand, and T. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Stat. Assoc.*, 97:77–87, 2002.

[7] P. Eilers, J. Boer, G. Van Ommen, and H. Van Houwelingen. Classification of Microarray Data with Penalized Logistic Regression. In *Proceedings of SPIE. Progress in biomedical optics and images*, volume 4266, pages 187–198, 2001.

[8] L. Fahrmeir and G. Tutz. *Multivariate statistical modelling based on generalized linear models. 2nd ed.* Springer Series in Statistics. New York, 2001.

[9] I. Frank and J. Friedman. A statistical view of some chemometrics regression tools,with discussion. *Technometrics*, 35(2):109–148, 1993.

[10] T. Furey, N. Cristianini, N. Duffy, D. Bednarsky, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16:906–914, 2000.

[11] D. Ghosh. Singular value decomposition regression models for classification of tumors from microarray experiments. *Pac. Symp. Biocomput.*, 98:18–29, 2002.

[12] D. Ghosh. Penalized discriminant methods for the classification of tumors from gene expression data. *Biometrics*, 59(4):992–1000, 2003.

[13] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439):531–537, 1999.

[14] P. Green. Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and some Robust and Resistant Alternatives. *J.R. Statist.Soc. B*, 46(2):149–192, 1984.

[15] I. Helland. On the structure of Partial Least Squares Regression. *Commun. Stat., Simulation Comput.*, 17(2):581–607, 1988.

[16] X. Huang and W. Pan. Linear regression and two-class classification with gene expression data. *Bioinformatics*, 19:2072–2078, 2003.

[17] R. Kass and A. Raftery. Bayes factor. *J. Amer. Stat. Assoc.*, 90:733–795, 1995.

[18] S. Le Cessie and J. Van Houwelingen. Ridge estimators in logistic regression. *J. R. Stat. Soc., Ser. C*, 41(1):191–201, 1992.

[19] B. D. Marx. Iteratively Reweighted Partial Least Squares estimation for Generalized Linear Regression. *Technometrics*, 38(4):374–381, 1996.

[20] W. F. Massy. Principal components regression in exploratory statistical research. *J. Amer. Stat. Assoc.*, 60:234–246, 1965.

[21] T. Naes and H. Martens. Comparison of prediction methods for multicollinear data. *Commun. Stat., Simulation Comput.*, 14:545–576, 1985.

[22] D. Nguyen and D. Rocke. Tumor classification by Partial Least Squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50, 2002.

[23] T. Santner and D. Duffy. A note on A. Albert and J.A. Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, 73(3):755–758, 1986.

[24] D. Singh, P. Febbo, K. Ross, D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. Renshaw, A. D'Amico, J. Richie, E. Lander, M. Loda, P. Kantoff, T. Golub, and W. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209, 2002.

[25] N. Timm. *Applied Multivariate Analysis*. Springer-Verlag. New York, 2002.

[26] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. Olson, J. Marks, and J. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci.*, 98(20):11462–11467, 2001.

[27] H. Wold. Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach. In *Perspect. Probab. Stat., Pap. Honour M. S. Bartlett Occas. 65th Birthday*, pages 117–142, 1975.

	RIRLS		RPLS		RPCR		FPLS		MAVE		DLDA		DQDA		KNN	
p	LO	OS	LO	OS	LO	OS	LO	OS	LO	OS	LO	OS	LO	OS	LO	OS
50	0	1	0 (3)	1	1 (1)	1	0 (2)	1	4 (3)	1	1	1	1	1	1 (1)	2
300	2	3	0 (1)	3	0 (2)	1	0 (2)	0	2 (1)	0	1	2	1	1	1 (1)	1
500	2	3	0 (1)	3	0 (3)	2	0 (2)	0	0 (1)	0	0	2	0	1	0 (1)	1
1000	2	3	0 (2)	2	0 (4)	2	0 (2)	0	1 (1)	0	0	2	0	2	0 (1)	1

Table 1: Comparison of misclassification for Leukemia Data: Leave One Out and Out Of Sample analyzes performed on the Learning/Test set of the Golub's subdivision.

p	RIRLS	RPLS	RPCR	FPLS	MAVE	DLDA	DQDA	KNN
100	9	9 (1)	7 (6)	8 (1)*	12 (1)	17	17	7 (5)
500	10	8 (3)	8 (5)	8 (1)*	7 (6)	18	22	9 (5)
1000	15	7 (3)	7 (6)	8 (1)*	15 (1)	20	23	8 (7)
$p_{\max}$	17	7 (3)	7 (6)	8 (1)*	6 (4)	22	25	8 (7)

Table 2: Comparison of misclassification for Colon Data: Leave One Out analysis performed on 62 subdivisions of the data set into a learning set (resp. test set) of cardinal 61 (resp. cardinal 1). * means that during the Leave One Out procedure, for a given κ in the range $\mathcal{K}_c$, some FPLS algorithms did not converge. The optimal value of κ is chosen among the values for which all the FPLS steps converged.

p	RIRLS	RPLS	RPCR	FPLS	MAVE	DLDA	DQDA	KNN
100	9	7 (3)	6 (8)	8 (2)*	51 (1)	11	11	7 (3)
500	10	8 (2)	9 (6)	8 (2)*	7 (2)	21	18	8 (13)
1000	10	5 (3)	5 (13)	8 (2)*	8 (4)	28	24	10 (3)
1500	10	7 (4)	5 (12)	10 (2)*	14 (4)	31	28	12 (3)

Table 3: Comparison of misclassification for Prostate Data: Leave One Out analysis performed on 102 subdivisions of the data set into a learning set (resp. test set) of cardinal 101 (resp. cardinal 1). * has the same meaning as in Table 2.

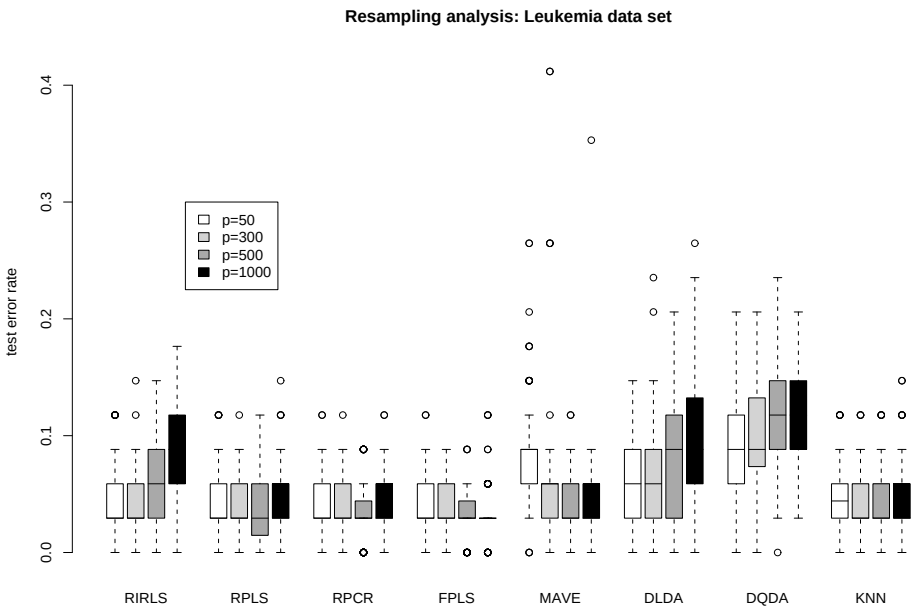

Figure 1: Resampling analysis for Leukemia data: Boxplots of test error rates for classifiers with 50 (white), 300 (light grey), 500 (dark grey) and 1000 (black) genes.

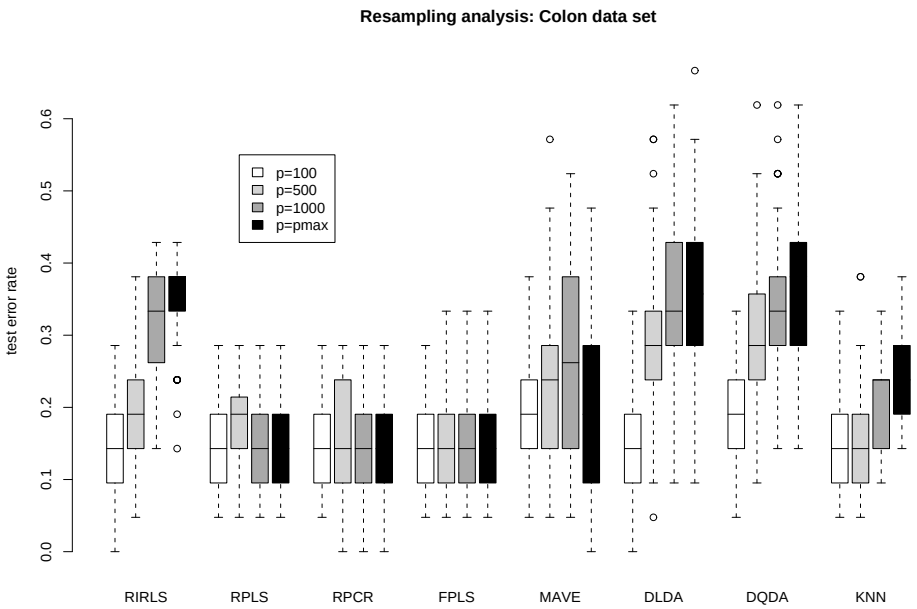

Figure 2: Resampling analysis for Colon data: Boxplots of test error rates for classifiers with 100 (white), 500 (light grey), 1000 (dark grey) and $p_{\max}$ (black) genes (2:1 scheme).

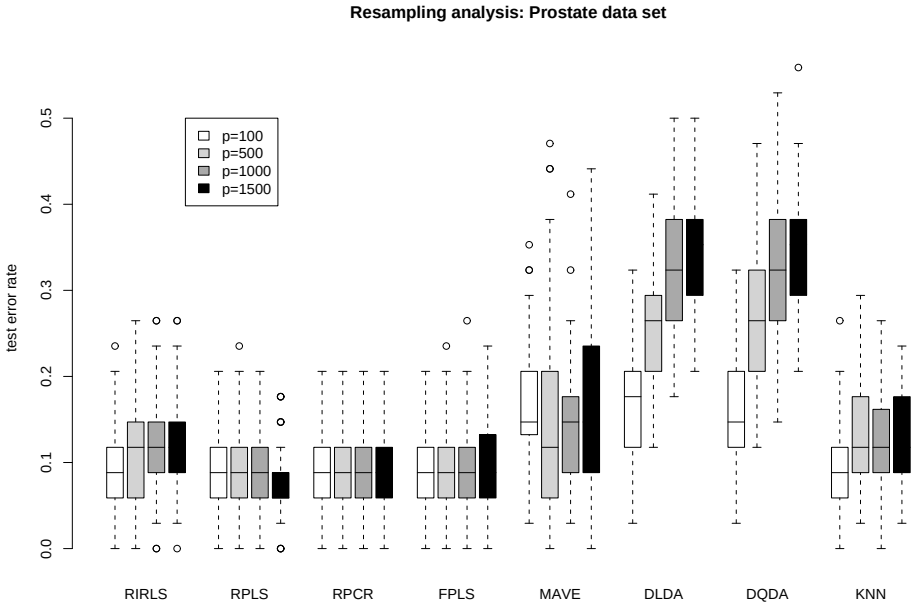

Figure 3: Resampling analysis for Prostate data: Boxplots of test error rates for classifiers with 100 (white), 500 (light grey), 1000 (dark grey) and 1500 (black) genes (2:1 scheme).