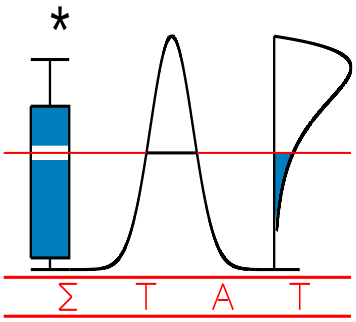
T E C H N I C A L
R E P O R T

0331

CLASSIFICATION USING PARTIAL LEAST SQUARES WITH PENALIZED LOGISTIC REGRESSION

FORT, G. and S. LAMBERT-LACROIX



I A P S T A T I S T I C S
N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Classification using Partial Least Squares with Penalized Logistic Regression

Gersende Fort *
CNRS and LMC-IMAG
BP 53
38041 Grenoble cedex 9
France
Gersende.Fort@imag.fr

Sophie Lambert-Lacroix
Laboratoire LMC-IMAG
BP 53
38041 Grenoble cedex 9
France
Sophie.Lambert@imag.fr

ABSTRACT

One important 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). Standard statistical methods in classification do not work because there are far more variables than observations. In this paper, the question of classification in such a high dimension 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 and Ridge penalized logistic regression. We review the existing methods based on PLS and / or penalized likelihood techniques, outline their interest in some case, explain theoretically their poor behavior. Our procedure is compared with these other classifiers. The predictive performance of the resulting classification rule is illustrated on two well known data sets: the Leukemia data set and the Colon data set.

Keywords

gene expression, supervised classification, generalized linear models, logistic regression, partial least squares, ridge penalty, Firth's procedure.

1. INTRODUCTION

The 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 data-mining of microarray data is to find out classification of different cell types, predominantly cancer types. To cite a few, Golub *et al.* [16] have considered classification of acute leukemia, Alon *et al.* [3] have addressed the cluster analysis of tumor and normal colon tissues, and Alizadeh *et al.* [2] for diffuse large B-cell lymphoma. The approaches developed in these papers consists in several discrimination methods and machine learning methods (see [11] for a comprehensive comparative study).

In microarray studies, the number of n of sample is relatively small compared to the number p of genes, usually in thousands. Unless a preliminary variable selection step is performed, standard statistical methods in classification do not work because there are far more variables than observations. One problem is the multicollinearity and estimating equations becomes 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 machine technique, 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 previous works mentioned, the authors have used univariate methods for reducing the number of genes. Alternative approaches to handle the dimension reduction problem can be used (see for instance [33; 17; 29; 5]).

Similar data structures have been seen in the field of chemometrics. The method of Partial Least Squares (PLS) has been found to be a useful dimension reduction technique [34; 28; 20] as well as Principal Component Regression (PCR, [26]) (see [14] for a statistical view of PLS and PCR). In the context of microarray, the purpose of PCR (see [33]) is to produce orthogonal tumor descriptors that reduce the dimension to only few gene component (super-genes). But the dimension reduction is achieved without regards to the response variation and may be inefficient. This is the reason why PLS looks more adapted than PCR in dimension reduction problem. Indeed, PLS components are chosen so that the sample covariance between the response and a linear combination of the p predictors (genes) is maximum.

Nguyen and Rocke [29] proposed using PLS method for a dimension reduction as a preliminary step to classification using linear logistic discrimination (LD), linear or quadratic discriminant analysis. However, it seems to be intuitively unappealing because PLS is really designed to handle continuous responses and especially for models that do not suffer from heteroscedasticity as it is the case for binary or multinomial data. Furthermore, in practice we have observed

*Corresponding author

problems in the convergence of the Iteratively Reweighted Least Squares (IRLS) algorithm, which is the usual procedure for solving the maximum likelihood in the field of the generalized linear model (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 (see [1]).

Marx [25] proposed an extension of PLS to categorical response variable and illustrates the developments from a spectroscopy example. Its approach embeds the usual PLS steps within the IRLS. Unfortunately, we have observed that this algorithm does not converge.

To deal with the high dimension problem, another approach consists in penalizing the likelihood. Eilers *et al.* [12] 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. Notice 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, that discourages regression coefficients to become large, unless they really contribute to the predictive performance of the model. All the genes contribute and that can inhibit and degrade the performances of the classification rules.

In this paper, we extend the PLS method to categorical response variable. To do that, we want to find a pseudo-response variable whose expected value has a linear relationship with the covariates, and apply PLS. In the IRLS algorithm, the pseudo-response variable seems to be a good candidate. Unfortunately in our situation "small n, large p", IRLS no longer works since the maximum likelihood does not admit a solution. The idea developed here is to penalize with Ridge penalty the maximum likelihood. That is our procedure combines Ridge penalty and PLS step and the dimension reduction step is incorporated in the classification step. Here we present classification rule for response variable as a binary vector indicating normal or ovarian tumor, for instance. Nevertheless, our approach remains valid for categorical response. But the binary case is the simplest case which allows us to point out that such a procedure works well or not and why. We review the existing methods based on PLS and / or penalized likelihood techniques, outline their interest and in some cases, explain theoretically their poor behavior. Our procedure is compared with these other classifiers. They are applied to two different microarray data sets: Acute Myeloid Leukemia (AML) versus Acute Lymphoblastic Leukemia (ALL) and normal colon tissues versus tumor colon tissues.

This paper is organized as follows.

Section 2 is the methodological part of this paper. Section 2.1 contains a description of the logistic regression and linear discrimination. Then, in Section 2.2, we recall and analyze some known regularization methods such as Partial Least Squares and penalized maximum likelihood methods. Section 2.3 is devoted to the analyses of the Nguyen and Rocke's algorithm and Marx's algorithm; finally, we end this methodological part by considering PLS extensions to GLM based on the Ridge's penalty, and derive a new classification method (section 2.4). Applications to disease classification through Microarrays are presented in Section 3.

2. METHODS

2.1 Logistic regression and linear logistic discrimination

After introducing some notations in the first subsection, we recall the principle of linear logistic discrimination (Subsection *2.1.2*) and some results on the existence of the maximum likelihood estimator and the classical algorithm used to compute it (Subsection *2.1.3*).

2.1.1 Notations

Expression level of p genes for 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. The *design* matrix $Z := (\mathbb{1}_n, X)$ of size $n \times (p+1)$, where $\mathbb{1}_n := (1, \cdots, 1)'$ stands for the row vector of length n, and the symbol $'$ denotes the transposition operator, is used when an intercept is included into the regression model. In supervised classification, each microarray sample is thought to originate from a specific class $k \in \{0, \ldots, g\}$, where the number of possible classes g is known and fixed. The data consists of n (statistically) independent observations of $(Y, \mathbf{X})$, stored in $(\mathbf{y}, X)$ where $X_{i\cdot} = (x_{i1}, \ldots, x_{ip})$ is the vector of a gene expression profile and $\mathbf{y}_i$ is a discrete-valued label variable. That is $(X_{i\cdot})_{1 \leq i \leq n}$ are the predictor variables and $(y_i)_{1 \leq i \leq n}$ the response variables. A classifier can be regarded as a function $G : \mathbb{R}^p \to \{0, \ldots, g\}$ that predicts the unknown class label of a new tissue sample $x \in \mathbb{R}^p$ by $G(x)$. In this contribution, we restrict our attention to the binary problem where the variable Y is $\{0, 1\}$-valued.

2.1.2 Linear Logistic Discrimination

In logistic regression, the regression function is given by the the conditional class probability $\pi := \mathrm{P}(Y = 1 | \mathbf{X} = x)$. The model consists in relating π to the linear predictor η

$$\eta := \alpha + x'\beta = z'\gamma \qquad \gamma' := (\alpha, \beta') \in \mathbb{R}^{p+1}, \tag{1}$$

through the response function h such that $\pi = h(\eta)$. We opt for the logit model where

$$\pi = h(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)} \Leftrightarrow \eta := h^{-1}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right). \tag{2}$$

The log-likelihood of the observations $\mathbf{y} = (\mathbf{y}_1, \cdots, \mathbf{y}_n)'$ for the value γ of the parameter is given by (the dependence upon the observations and the covariates is omitted)

$$l(\gamma) := \sum_{k=1}^{n} \left\{ \mathbf{y}_k \eta_k(\gamma) + \ln\left(1 + \exp(\eta_k(\gamma))\right) \right\}, \tag{3}$$

where for all $1 \leq k \leq n$,

$$\eta_k(\gamma) := (Z\gamma)_k. \tag{4}$$

This is a (univariate) generalized linear model (GLM) with canonical link [27].

Observe that the choice of a $\{0, 1\}$-code to model the dichotomic classification is not at all restrictive, and can be substituted by any binary code $\{a, b\}$, for some $a < b$. If Y is $\{a, b\}$-valued, then $\tilde{Y} := (Y - a)/(b - a)$ is $\{0, 1\}$-valued and the log-likelihood of Y is given by

$$\tilde{Y} \log(\pi) + \left(1 - \tilde{Y}\right) \log(1 - \pi).$$

This shows that in the case of a $\{a, b\}$-valued response variable, all the results of the present paper applies by replacing Y by the $\{0, 1\}$-valued response variable $\tilde{Y}$.

The vector of unknown parameter of the model γ is estimated by maximum log-likelihood. The predicted response probabilities are obtained by replacing the parameter γ with its estimate in formulas (1) and (2). The predicted class of each sample is $\hat{y} = \mathbb{1}_{(\hat{\pi} > 1-\hat{\pi})}$, where $\mathbb{1}_{(\cdot)}$ is the indicator function. This classification procedure is commonly called Logistic Discrimination (LD). In general, an optimal classification rule in terms of Bayes risk, could be determined by minimizing the expected cost of misclassification ([23]). It is given by

$$\hat{y} = \mathbb{1}_{(\hat{\pi} > \frac{c(1|0)}{c(0|1)}(1-\hat{\pi}))}, \tag{5}$$

where $c(1|0)$ (resp. $c(0|1)$) is the cost when an observation from the population 0 (resp. 1) is incorrectly classified as 1 (resp. 0). The LD procedure corresponds to a symmetric cost *i.e.* $c(1|0) = c(0|1)$.

2.1.3 *Maximum likelihood estimate and Iteratively Reweighted Least Squares (IRLS) algorithm*

Regression analysis in GLM is based on likelihoods and parameter inference in such a case relies on the maximum likelihood (ML) method. Nevertheless, existence of a ML estimate *i.e.* a vector $\hat{\gamma}^{\mathrm{ML}}$ such that $\|\hat{\gamma}^{\mathrm{ML}}\| < \infty$ which is a (local) maximizer of the log-likelihood l, is not guaranteed. We recall below known results on the existence of $\hat{\gamma}^{\mathrm{ML}}$ for logit model, and the classical algorithm used to compute $\hat{\gamma}^{\mathrm{ML}}$ - when it exists - in the classical statistical framework $n > p + 1$.

The log-likelihood of the logit model is twice continuously differentiable on $\mathbb{R}^{p+1}$ and

$$\nabla^2 l(\gamma) := -Z' W(\gamma) Z, \tag{6}$$

where ∇^2 denotes the Hessian operator and $W(\gamma)$ is a diagonal $n \times n$ matrix with positive entries given by

$$W(\gamma) := \mathrm{diag}\left(\pi_k(\gamma)\left(1 - \pi_k(\gamma)\right)\right), \tag{7}$$

and for all $1 \leq k \leq n$,

$$\pi_k(\gamma) := \left(1 + \exp(-\eta_k(\gamma))\right)^{-1}. \tag{8}$$

We denote by $\pi(\gamma)$ the vector of components $\pi_k(\gamma)$, $k = 1, \ldots, n$. Hence l is at least concave and is strictly concave if and only if Z is full rank and $n \geq (p+1)$. If $\hat{\gamma}^{\mathrm{ML}}$ exists, it is computed as the solution to the score equation

$$\nabla l(\gamma) = 0 \Longleftrightarrow Z'(\mathbf{y} - \pi(\gamma)) = 0, \tag{9}$$

where ∇ denotes the gradient operator.

When Z is full rank and $n > (p+1)$, Albert and Anderson [1] showed that the existence of the ML estimate depends on the configuration of the n samples points in the observation space. There are three following exclusive situations.

(i) The points are separated *i.e.* there exists γ such that

$$Z_{i}'\gamma > 0 \text{ if } \mathbf{y}_i = 0 \quad \text{and} \quad Z_{i}'\gamma < 0 \text{ if } \mathbf{y}_i = 1. \tag{10}$$

(ii) The points are quasi-separated *i.e.* (10) holds with large inequalities and equality for at least one i.

(iii) The points overlap each other.

In the first two cases, ML estimate does not exist since l reaches its maximum as $\|\gamma\|$ tends to $+\infty$, while it exists in the third one. Santner and Duffy [31] derive an algorithm to determine which of these three situations hold for given data $(\mathbf{y}, Z)$. In the overlap case, (9) possesses an unique solution, which in practice, is computed by the iterative Newton-Raphson method. It consists in the construction of a converging sequence $(\gamma^{(t)})_{t \geq 0}$ (with limit $\hat{\gamma}^{\mathrm{ML}}$) and such that $\gamma^{(t+1)}$ solves the equation in γ

$$Z' W^{(t)} Z \left(\gamma - \gamma^{(t)}\right) = Z'\left(\mathbf{y} - \pi^{(t)}\right), \tag{11}$$

where $W^{(t)} := W(\gamma^{(t)})$ and $\pi^{(t)} := \pi(\gamma^{(t)})$. Since the Fisher information matrix $Z'W^{(t)}Z$ is invertible, solving (11) is equivalent to regressing the pseudo-response variable $z^{(t)}$

$$z^{(t)} := Z\gamma^{(t)} + \left[W^{(t)}\right]^{-1}\left(\mathbf{y} - \pi^{(t)}\right), \tag{12}$$

onto the columns of Z with the weight matrix $W^{(t)}$ *i.e.*

$$\gamma^{(t+1)} := Z\left(Z' W^{(t)} Z\right)^{-1} Z' W^{(t)} z^{(t)}. \tag{13}$$

This algorithm is referenced in the literature as the Iteratively Reweighted Least Squares (IRLS) algorithm (see e.g. Green [18] and references therein).

When Z is full rank and $n \leq p+1$, it is readily seen from (2) that the unique solution to (9) is such that for all $1 \leq k \leq n$, $(Z\gamma)_k = \ln(\mathbf{y}_k) - \ln(1 - \mathbf{y}_k)$ so that $|(Z\gamma)_k| = +\infty$ for all k and the ML estimate can not exist.

Finally, when Z is not full rank (which is unlikely for real data sets), solutions to (9) are not unique.

2.2 Regularization methods

When $n \leq p + 1$, which is the case in the considered applications, ML is not unique when it exists, so inference of the parameter necessitates the introduction of new methods. A natural approach is to adapt the proposed methods in the case $n > p + 1$ to overcome the non-existence problem. A first solution consists in reducing the dimension of the problem, by replacing the p covariates of the initial data matrix by few appropriately defined "super-covariables". A second solution consists in maximizing the log-likelihood under constraints by introducing a penalty term; this yields a penalized maximum likelihood estimate [4; 24; 19]. Notice that these methods are not techniques of dimension reduction.

In this section, our goal is to describe some of theses techniques, that appear, as discussed in Section 2.3, to be basic ingredients for solving inference in GLM when $n \leq p + 1$. More precisely, we restrict our attention to three methods: Partial Least Squares (PLS), Firth's and Ridge penalized regression.

2.2.1 *Weighted Partial Least Squares (WPLS)*

The Partial Least Squares (PLS) method, first introduced in Chemometrics ([34; 28; 20]), can be read both as a tool for (weighted) linear regression and as a tool for dimension reduction.

For a given response vector $\mathbf{y}$, a data matrix X and a positive definite $n \times n$ matrix W, the PLS scope is to convey the relation between $\mathbf{y}$ and X through the definition of κ *scores* $(t_j)_{1 \leq j \leq \kappa}$, linear combinations of the columns of the design

matrix Z and such that

$$\mathbb{1}_n' W t_j = 0, \qquad t_j' W t_k = 0, \qquad \forall j,k = 1,\cdots,\kappa,\ j \neq k. \tag{14}$$

Hence, this allows the decomposition

$$\mathbf{y} = q_0 \mathbb{1}_n + q_1 t_1 + \cdots + q_\kappa t_\kappa + f_\kappa, \tag{15}$$

where $(q_j)_{0\leq j\leq \kappa}$ are real numbers, the remainder term f_κ is $\mathbb{R}^n$-valued and

$$\mathbb{1}_n' W f_\kappa = 0, \qquad t_j' W f_\kappa = 0, \qquad \forall j = 1,\cdots,\kappa. \tag{16}$$

Principal Component Regression (PCR, [26]) provides a similar decomposition, but in PCR, the scores are defined independently of the response vector $\mathbf{y}$. This is the reason why PLS looks more adapted than PCR to solve the dimension reduction problem in a regression framework. This method, derived in the literature in the unweighted case (*i.e.* $W = \mathrm{I}_n$) and thus simply called PLS, can naturally be extended to the weighted case; this extension will be referred as WPLS. As in the linear regression framework, the introduction of a weight matrix is to take into account the heteroscedasticity of the $\mathbf{y}$ variables. WPLS proceeds as follows (see Section A.1 for an algorithmic description)

1. (a) Project $\mathbf{y}$ and the columns of X on $\mathbb{1}_n$: set $q_0 := (\mathbf{y}' W \mathbb{1}_n)/(\mathbb{1}_n' W \mathbb{1}_n)$ and $p_0 := X' W \mathbb{1}_n / \mathbb{1}_n' W \mathbb{1}_n$.
   (b) Center the response vector and the data matrix and define $\mathbf{y}^c := \mathbf{y} - q_0 \mathbb{1}_n$ and $X^c := X - \mathbb{1}_n p_0'$.
2. (a) Choose t_1 as a linear combination of the columns of X^c *i.e.* on the form $X^c\omega_1$ where ω_1 is such that $\omega_1 \mapsto |(X^c\omega_1)' W \mathbf{y}^c|$ is maximal; this yields $\omega_1 := X^{c\prime} W \mathbf{y}^c$ and $t_1 := X^c X^{c\prime} W \mathbf{y}^c$.
   (b) Project $\mathbf{y}^c$ and the columns of X^c on t_1: set $q_1 := (\mathbf{y}^{c\prime} W t_1)/t_1' W t_1$ and $p_1 := (X^{c\prime} W t_1)/t_1' W t_1$.
3. For $j = 1,\cdots,\kappa-1$, repeat step 2 by replacing $(\mathbf{y}^c, X^c)$ with the deflated matrix $\mathbf{y}^c - q_1 t_1 - \cdots - q_j t_j$ and $X^c - t_1 p_1' - \cdots - t_j p_j'$.

The first score t_1 is thus chosen as the vector in the space spanned by the columns of X^c, denoted by $\mathrm{Sp}(X^c)$, maximizing the scalar product $|t' W \mathbf{y}^c|$ or, equivalently, maximizing the weighted empirical covariance $|\mathrm{Cov}(\sqrt{W}t, \sqrt{W}\mathbf{y}^c)|$ ($\sqrt{W}$ denotes the square root matrix of W). In this sense, t_1 is the vector of $\mathrm{Sp}(X^c)$ which is the most informative on the response variable $\mathbf{y}^c$. This information contained in t_1 is then subtracted to $\mathbf{y}^c$, and t_2 is chosen as the vector of $\mathrm{Sp}(X^c)$ which is the most informative on the unexplained part of $\mathbf{y}^c$, under the constraint that t_2 and t_1 are W-orthogonal. This process is repeated until the number of scores reaches κ. From this algorithmic description, it is easily seen that for given $(\mathbf{y}, X, W)$, the scores are unchanged if the response vector $\mathbf{y}$ is multiplied by a constant λ or translated by adding a vector collinear to $\mathbb{1}_n$.
As discussed in Helland [21], the maximal number of scores $\kappa_{\max}$ depends both of $\mathbf{y}$, W and X. It is the number of distinct eigenvalues of $X^{c\prime} W X^c$ such that there exists at least one corresponding eigenvector v such that $v' X^{c\prime} W \mathbf{y}^c \neq 0$. When $\kappa = \kappa_{\max}$, the vectors $(\mathbb{1}_n, t_1, \cdots, t_{\kappa_{\max}})$ form a basis of the subspace of $\mathrm{Sp}(Z)$ that contains the W-projection of $\mathbf{y}$ onto the columns of Z; in this case, the PLS algorithm is a method to find the weighted least squares (WLS) predictor of $\mathbf{y}$ with regressors Z:

$$\begin{aligned} \mathrm{WPLS}(\mathbf{y}, X, W, \kappa_{\max}) &\equiv \mathrm{WLS}(\mathbf{y}, Z, W) \\ &\equiv \mathrm{WLS}(\sqrt{W}\mathbf{y}, \sqrt{W}Z, \mathrm{I}_n), \end{aligned} \tag{17}$$

where I_n denotes the $n \times n$ identity matrix.
From the above description of WPLS, it is readily seen that the score t_k is on the form $X^c\psi_k$ where ψ_k is a linear combination of the $\mathbb{R}^p$-valued vectors ω_k: $\psi_k \in \mathrm{Sp}(\Omega_k)$, $\Omega_k := (\omega_1, \cdots, \omega_k)$. Hence, the regression coefficient vector (with respect to the matrix X^c) is a linear combinations of the columns of Ω_κ, which is full rank by construction (for $\kappa \leq \kappa_{\max}$). Hence $\hat\beta^{\mathrm{PLS},\kappa}$ is on the form $\Omega_\kappa v$. From (15) and the decomposition

$$X^c = t_1 p_1' + \cdots + t_\kappa p_\kappa' + E_\kappa,$$

where the columns of E_κ are W-orthogonal to the vector space $\mathrm{Sp}(\mathbb{1}_n, t_1, \cdots, t_\kappa)$, it is seen that for all $1 \leq k \leq \kappa$, $q_k = p_k' \Omega_\kappa v$ so that the WPLS predictor of $\mathbf{y}^c$ is given by $X^c\hat\beta^{\mathrm{PLS},\kappa}$ and

$$\hat\beta^{\mathrm{PLS},\kappa} := \Omega_\kappa (P_\kappa' \Omega_\kappa)^{-1} q, \qquad q' := (q_1, \cdots, q_\kappa), \tag{18}$$

where $P_\kappa := (p_1, \cdots, p_\kappa)$. The PLS estimate of γ, computed with κ PLS components is

$$\hat\gamma^{\mathrm{PLS},\kappa} := \left(q_0 - \frac{\mathbb{1}_n' W X \hat\beta^{\mathrm{PLS},\kappa}}{\mathbb{1}_n' W \mathbb{1}_n}, [\hat\beta^{\mathrm{PLS},\kappa}]' \right)'. \tag{19}$$

The choice of κ is, to our best knowledge, an open problem: the non linear dependence of $\hat\gamma^{\mathrm{PLS},\kappa}$ upon the observations $\mathbf{y}$, makes an explicit control of the error term $f_\kappa = \mathbf{y} - Z\hat\gamma^{\mathrm{PLS},\kappa}$ impossible. The results proposed in the literature all rely on crude simplifications of the dependence upon $\mathbf{y}$ ([8; 9; 30]). In practice, in the methods mentioned below as well as in our procedure, κ is either arbitrarily fixed to a small value or fixed to the maximal one.

2.2.2 Firth's penalty

To solve the non-existence question of the ML estimate in logistic regression when $n \geq p+1$, Heinze and Schemper [19] advocates the use of the Firth's modified score procedure. The work by Firth [13] was originally introduced to reduce the bias of the ML estimate when exists, in the case where Z is full rank $n \geq p+1$. Observe that in that case, the Fisher information matrix $\mathcal{I}(\gamma)$ for a canonical GLM exists and is invertible for all $\gamma \in \mathbb{R}^{p+1}$. ML estimate in GLM is known to be biased when exists; the bias is divided into two terms $b_1(\gamma) + O(n^{-3/2})$ where, for canonical model, the $O(n^{-1/2})$-term $b_1(\gamma)$ is equal to $-\mathcal{I}(\gamma)^{-1}\ \tau(\gamma)/2$, and $\tau(\gamma) := (\tau_1(\gamma), \cdots, \tau_{p+1}(\gamma))'$ with

$$\begin{aligned} \tau_k(\gamma) &= \sum_{t,u} \left[\mathcal{I}(\gamma)^{-1}\right]_{t,u} \partial_{\gamma_k} \left[\mathcal{I}(\gamma)^{-1}\right]_{t,u} \\ &= \mathrm{Trace}\left(\mathcal{I}(\gamma)^{-1} \partial_{\gamma_k} [\mathcal{I}(\gamma)]\right), \end{aligned}$$

where ∂_t denotes the partial derivative wrt the variable t. To remove the first-order term, Firth [13] proposes a suitable modification of the score function; denote by $\hat\gamma^*$ the solution to the modified score equation $\nabla l(\gamma) + A(\gamma) = 0$ for some A to be specified that may depend upon the observations. Then the first-order term in the bias of $\hat\gamma^*$ is equal to

$$b_1(\gamma) + \mathcal{I}(\gamma)^{-1}\ \mathrm{E}_\gamma\left[A(\gamma)\right],$$

where E_γ denotes the expectation when the true value of the parameter is γ. To correct $\hat{\gamma}^{\mathrm{ML}}$ in an estimate with reduced bias (by removing the first-order term) it is thus sufficient to choose

$$A(\gamma) := -\mathcal{I}(\gamma) b_1(\gamma) = \frac{1}{2}\tau(\gamma). \tag{20}$$

By using classical linear algebra results (see e.g. [6]), A may be expressed as a differential term and one has $A(\gamma) = 0.5\ \nabla \log|\mathcal{I}(\gamma)|$, where $|\cdot|$ denotes the determinant. As a consequence, the estimate $\hat{\gamma}^*$ can be seen as the maximum of the penalized maximum likelihood

$$l^*(\gamma) := l(\gamma) + \frac{1}{2}\log|\mathcal{I}(\gamma)|; \tag{21}$$

in a Bayesian framework, this penalty is known as the Jeffreys invariant prior.
Firth [13] (see also [19]) asserts that, when the design matrix Z is full rank and $n \geq p+1$, $\hat{\gamma}^*$ always exists and is unique. In addition, this maximum can be computed by slightly modifying the IRLS algorithm as follows ([13; 7]). Since $\mathcal{I}(\gamma) = Z'W(\gamma)Z$, it is trivial to prove that

$$\begin{aligned}\tau_k(\gamma) &= \mathrm{Trace}\left\{(Z'W(\gamma)Z)^{-1}Z'\,\partial_{\gamma_k}[W(\gamma)]Z\right\}\\ &= \mathrm{Trace}\left\{Z(Z'W(\gamma)Z)^{-1}Z'\,\partial_{\gamma_k}[W(\gamma)]\right\}\\ &= -\sum_{j=1}^{n} Z_{jk}\,(2\pi_j(\gamma)-1)\,H_{jj}(\gamma)\end{aligned}$$

where $H(\gamma) := Z(Z'W(\gamma)Z)^{-1}Z'W(\gamma)$. Hence, by using (9) and (20), we obtain

$$\nabla l^*(\gamma) = Z'\left(\tilde{\mathbf{y}}(\gamma) - \tilde{\pi}(\gamma)\right),$$

where $\tilde{\mathbf{y}}_k(\gamma) := \mathbf{y}_k + 0.5H_{kk}(\gamma)$ and $\tilde{\pi}_k(\gamma) := \pi_k(\gamma)(1 + H_{kk}(\gamma))$ for all $1 \leq k \leq n$. In the literature, it is thus proposed to exhibit $\hat{\gamma}^*$ as the limiting value of the sequence $(\tilde{\gamma}^{(t)})_{t\in\mathbb{N}}$, where $\tilde{\gamma}^{(t+1)}$ solves the equation

$$Z'W^{(t)}Z\,\left(\gamma - \tilde{\gamma}^{(t)}\right) = Z'\left(\tilde{\mathbf{y}}(\tilde{\gamma}^{(t)}) - \tilde{\pi}(\tilde{\gamma}^{(t)})\right), \tag{22}$$

and $W^{(t)} := W(\tilde{\gamma}^{(t)})$. This algorithm, based on a simplification of the Hessian $\nabla^2 l^*$, is thus similar to the IRLS algorithm, in which the response variables $\mathbf{y}$ (resp. the mean π) are replaced with the current values of $\tilde{\mathbf{y}}$ (resp. $\tilde{\pi}$).
We do not think it is sensible to penalize the intercept parameter α. A penalty on the whole parameter γ is a kind of boundedness condition around zero on the linear predictor η and is thus restrictive. To avoid this, we introduce the penalty only on the parameter β. This leads to the following penalized log-likelihood,

$$l^*(\gamma) := l(\gamma) + \frac{1}{2}\log|\mathcal{I}(\beta)| = l(\gamma) + \frac{1}{2}\log|-\nabla^2_\beta l(\gamma)|,$$

the maximum of which can be computed by applying the above algorithm where the matrix $H(\gamma)$ is replaced by

$$X(X'W(\gamma)X)^{-1}X'W(\gamma).$$

2.2.3 Ridge penalty

In linear regression, the regression coefficient vector γ obtained with the Ordinary Least Squares (OLS) is estimated by minimizing the sum of squares $S := \|\mathbf{y} - Z\gamma\|^2$ where $\|\cdot\|$ is the Euclidean norm. When multicollinearity is present *i.e.* $\mathrm{rank}(Z) < n \wedge (p+1)$, the usual OLS estimator will not be well-defined. That is the system has no unique solution and some regression coefficients will be very large. Hoerl and Kennard [22] propose to add to S the square of the norm of the regression coefficients, weighted by a positive *shrinkage parameter* λ: $S^* = S + \lambda/2\ \|\beta\|^2$. The estimator $\hat{\gamma}$ is given by

$$\hat{\gamma} = (Z'Z + \lambda R)^{-1}Z'\mathbf{y},$$

where R is an $(p+1)\times(p+1)$ identity matrix with R_{11} set to zero. The second term of S, that can be read as a constraint with Lagrange multiplier $\lambda/2$, is called the penalty and discourages high values for the elements of β. This method is called the Ridge regression.
As commented in the subsection *2.2.2*, there is no reason to penalize the intercept parameter α, which explains that the penalty only applies on β. The larger λ, the stronger its influence and the smaller the elements of β are forced to be: λ controls the amount of shrinkage in the data. Concerning the choice of λ, a very common approach is to use cross-validation ([32]).
In the same spirit, we find ridge estimators in logistic regression ([24]). The authors propose to estimate γ as a maximizer of the likelihood under the constraint that $\|\beta\| < \infty$ *i.e.* they introduce a ridge penalty term in the criterion l^* to maximize. As previously, this penalty does not apply to the location parameter α but only concerns the last p components of γ so that

$$l^*(\gamma) := l(\gamma) - \frac{\lambda}{2}\|\beta\|^2, \tag{23}$$

for some shrinkage parameter $\lambda > 0$. A second practical interest of the penalty term is that the maximum of l^* always exists and is unique, whatever the size and the rank of Z. Algorithms for research of an extremum to l^* are never vain as they may be for research of an extremum to l (see Section *2.1.3*). l^* is twice-continuously differentiable and strictly concave since for any non null vector $\mu = (\mu_1, \cdots, \mu_{p+1})' \in \mathbb{R}^{p+1}$, it holds

$$-\mu'\,\nabla^2 l^*(\gamma)\,\mu \geq \lambda\|\mu\|^2 + (\mathbb{1}_n'W(\gamma)\mathbb{1}_n - \lambda)\mu_1^2 > 0.$$

In addition, l^* tends to $-\infty$ when $\|\gamma\| \to +\infty$ since

$$l^*(\gamma) \leq -\frac{\lambda}{2}\|\beta\|^2 - \ln(1+\exp(\eta_1(\gamma))) - \ln(1+\exp(-\eta_n(\gamma))),$$

where we assumed that $\mathbf{y}_1 = 0$, and $\mathbf{y}_n = 1$ (which can be done without loss of generality). As a consequence, l^* possesses an unique maximum γ^* that can be computed as the limit of a Newton-Raphson sequence $(\gamma^{(t)})_{t\geq 0}$: this leads to an iterative algorithm that mimics IRLS. (11) is replaced by

$$\left(Z'W^{(t)}Z + \lambda R\right)\,\left(\gamma - \gamma^{(t)}\right) = Z'\left(\mathbf{y} - \pi^{(t)}\right) - \lambda R\gamma^{(t)},$$

so that (13) gets into

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

and the definition (12) of the pseudo-response variable $z^{(t)}$ is unchanged. The ridge parameter λ can be estimated by cross-validation.

2.3 Procedures using PLS

For solving the classification problem in the case $n << p$, different algorithms based on the procedures detailed above have been proposed. These methods consist in algorithms

that first derive a dimension reduction step to exhibit "super-covariates" using PLS, followed by an inference step in which the initial covariates are replaced by few super-covariates. We now review these methods, outline their interest and in some cases, explain theoretically their poor behavior.

Some of the discussions below are illustrated by applying the methods to the classification of *AML-ALL Leukemia* microarrays (see Section 3 for a complete description of this data set): the data are divided into a learning set with 38 samples and a test set with 34 samples. The data matrix is of size $38 \times p$, for some p to be precised, and the p genes are selected as suggested in Dudoit *et al.* [11]. This data set also contains a test set with 34 samples.

2.3.1 Nguyen and Rocke's (NR) approach

In [29], the authors use the PLS method with entries the response vector and the data matrix X with κ PLS components for dimension reduction, as a preliminary step before classification using LD (see the algorithmic description in Appendix A.1). That is the matrix X used in the LD is replaced by a matrix with few columns formed by the first κ PLS components. They have also compared the Quadratic discriminant analysis to the LD analysis. Their method is applied to various data sets and the results appear good. Nevertheless it seems to be intuitively unappealing to extract PLS components in a first place, since this method is really designed to handle continuous response and especially for models that do not really suffer from conditional heteroscedasticity.
In practice, we observed problems in the convergence of the NR algorithm. When using the LD on the *Leukemia* data set, for example, the PLS components form a new data matrix such that the 38 samples points in the observation space are separated. In Figure 1, we plot the (normalized) second component $t_{\cdot,2}$ vs the (normalized) first one $t_{\cdot,1}$; this separation exists on the 38 coordinates of the first score $t_{\cdot,1}$, and hence in any κ-dimensional space $\mathrm{Sp}(t_{\cdot,1}, \cdots, t_{\cdot,\kappa})$, $\kappa \leq \kappa_{\max}$.

Insert Figure 1 approximatively here.

As commented in Section *2.1.3*, this shows that the IRLS step of the NR algorithm can not converge: the linear predictor, linear combinations of the PLS scores, tends (in norm) to infinity and hence so does the parameter.
In practice, the authors advocates to stop the IRLS step after an arbitrary fixed number of iterations. This rule yields an "unstable" classification method. When applied on the *Leukemia* data set, we observed that the number of misclassified samples of the test set depends upon the number of iterations of the IRLS step. On Figure 2, we plot the estimates $\hat{\gamma}^{\mathrm{NR}}$ obtained respectively after $\iota = 7$ and $\iota = 10$ iterations, when $p = 150$ and $\kappa = 3$; we also plot the classification results on the test set (Out Of Sample analysis). In the case $\iota = 7$, the sample 66 is misclassified while in the case $\iota = 10$, the samples $60, 66$ are misclassified.

Insert Figure 2 approximatively here.

2.3.2 Marx's approach

Marx [25] proposes an extension of the concept of PLS into the framework of generalized linear models and illustrates the developments from a spectroscopy example. His approach is based on an algorithm called Iteratively Reweighted Partial Least Squares (IRPLS, see the algorithmic description in Appendix B.2). The key idea is first to try to find a pseudo-response variable whose expected value has linear relationship with the covariates; and then to apply WPLS. To that goal, IRPLS differs from IRLS in the regression step (13). Since $Z'W^{(t)}Z$ is not invertible, the weighted regression of the pseudo-response variable $z^{(t)}$ onto the columns of Z is replaced by a weighted partial least squares step applied to $(\mathbf{y}, X, W^{(t)})$ and run until the components are $\kappa = \mathrm{rank}(X^c)$. The first s WPLS components obtained at "convergence" of this step are then used as columns of the new data matrix X^{new}. A classical IRLS step is then performed to regress $\mathbf{y}$ onto the columns of $[\mathbb{1}_n, X^{\mathrm{new}}]$ in the ML sense. Some problems appear in the implementation of this algorithm.
The first objection is about the choice of κ in the calls to the WPLS function. As mentioned in Section *2.2.1*, the maximal number of components $\kappa_{\max}$ is lower or equal to $\mathrm{rank}(X^c)$. Hence, the last $(\mathrm{rank}(X^c) - \kappa_{\max})$ PLS components, if non null, are nothing else than "noisy null vectors".
The second objection is about the convergence of the IRPLS step. Equation (11) means that the new value $\gamma^{(t+1)}$ is chosen such that $\sqrt{W^{(t)}}Z\gamma^{(t+1)}$ is the orthogonal projection of $\sqrt{W^{(t)}}z^{(t)}$ onto the columns of $\sqrt{W^{(t)}}Z$. In addition, as discussed in Section *2.2.1*, WPLS applied with $\kappa_{\max}$ components is a way to solve (11) without "inverting the non-invertible" Hessian matrix. Hence, IRPLS can be read as a robust implementation of the Newton-Raphson's procedure, that produces a sequence $(\gamma^{(t)})_{t\in\mathbb{N}}$ such that $\gamma^{(t+1)}$ solves (11). Unfortunately, if the Hessian $Z'W^{(t)}Z$ is nowhere invertible, the sequence $(\gamma^{(t)})_{t\in\mathbb{N}}$ does not necessarily converge. This phenomenon can be observed on the *Leukemia* data set [$p = 100$]. On Figure 3, we plot some components of $\eta^{(t)} = Z\gamma^{(t)}$ and of the estimate $\gamma^{(t)}$ vs the number t of IRPLS loops.

Insert Figure 3 approximatively here.

The third objection is the sensibility of the IRPLS step to the initial value. If convergence, the limiting value γ^{∞} is a solution to (9). Since this solution is not unique, the limiting point depends upon the initialization: different initial values lead to different WPLS components and hence may lead to different classification rules.
We finally conclude this analysis by showing that, in some cases, a strict implementation of the algorithm described in Marx can not run since the iterative procedure IRPLS does not converge; and if stopped after a fixed number of iterations, the Marx's algorithm provides the same result as the NR algorithm. This occurs when, as suggested by Marx, the IRPLS is initialized by choosing a linear predictor on the form

$$\eta^{(0)} = c_0\mathbf{y} - c_0(\mathbb{1}_n - \mathbf{y}); \tag{24}$$

and for data sets $(\mathbf{y}, X)$ such that the maximal number of WPLS components relative to $(\mathbf{y}, X, \mathrm{I}_n)$ is equal to $\mathrm{rank}(X^c)$. In the classical statistical framework $n >> p + 1$, the literature advocates an initialization such as (24): usually, $\pi^{(0)} = (\mathbf{y} + 0.5\mathbb{1}_n)/2$, which leads to $c_0 = \ln(3)$. A triv-

ial induction shows that for all $t \geq 0$,

$$W^{(t)} = \exp(c_t)/(1+\exp(c_t))^2 \mathrm{I}_n \qquad \text{is a scalar matrix,}$$
$$z^{(t)} = c_{t+1}\mathbf{y} - c_{t+1}(\mathbb{1}_n - \mathbf{y}), \quad c_{t+1} := 1 + c_t + \exp(-c_t),$$
$$\eta^{(t+1)} = z^{(t)}.$$

The last assertion results from the fact that $\kappa_{\max} = \mathrm{rank}(X^c)$ and $W^{(t)}$ is a scalar matrix, so that the projection of $z^{(t)}$ on $\mathrm{Sp}(Z)$ and the variable $z^{(t)}$ coincide. This shows that $|c_t|$ tends to infinity and so does $\|\eta^{(t)}\|$. Finally, the pseudo-response variable $z^{(t)}$ inherits the structure of $\mathbf{y}$ (*i.e.* $z_k^{(t)} = c_t$ if $y_k = 1$ and $-c_t$ otherwise). From the PLS description (Section *2.2.1*), we see that since $z^{(t)} = 2c_t\mathbf{y} - c_t\mathbb{1}_n$ and $W^{(t)}$ is proportional to the identity matrix I_n, the WPLS components computed from $(z^{(t)}, X, W^{(t)})$ are the same as the WPLS components computed from $(\mathbf{y}, X, \mathrm{I}_n)$. Hence NR's algorithm and Marx's algorithm lead to the same prediction/classification analysis.
The condition $\kappa_{\max} = \mathrm{rank}(X^c)$ is satisfied for example when (a) $X^c X^{c\prime}$ has $(n-1)$ distinct eigenvalues, and (b) for any eigenvectors ν_j of $X^{c\prime}X^c$, $\mathbf{y}'X^c\nu_j \neq 0$. This is the case for example for the *Leukemia* data set [$p = 100$].

Insert Figure 4 approximatively here.

The equivalence of the two procedures can be visualized: each of the first four WPLS components are drawn in Figure 3 [$p = 100$].

Insert Figure 5 approximatively here.

Hence, the Marx's classification algorithm inherits the instability of the NR's classification algorithm.
In a recent work, Ding and Gentleman ([10], private communication) propose an algorithm based on Firth's penalty in order to make the Marx's algorithm "robust". We have some troubles with the preliminary version of their paper, in particular because they formulate a condition in terms of the determinant of the Fisher information matrix which, even if, in the case $n > p+1$ this quantity is not defined. A solution could consist in changing the parameterization of the model - till now based on γ-, since this parameter is not identifiable when $n < p+1$, which leads to a non-invertible Fisher matrix. This approach will be treated in a forthcoming paper.

2.4 A new PLS extension based on Ridge penalty

In this subsection, a new algorithm based on Ridge penalty extending PLS is proposed that seems to overcome some of the mentioned problems. Before that, we review the Eilers *et al.* 's approach based on the Ridge penalized maximum log-likelihood.

2.4.1 Eilers *et al.* 's approach

Eilers *et al.* [12] propose to use the Ridge penalized logistic regression approach in order to both stabilize the statistical problem and remove numerical degeneracy due to multicollinearity. Furthermore they have shown that this method appears to work well with microarray data. The algorithm derived in Eilers *et al.* [12] is presented in Appendix B.3.
The Eilers *et al.* approach consists in using the regression coefficient vector $\hat{\gamma}^{\mathrm{E}}$ that maximizes the penalized maximum likelihood criterion (23) in order to estimate the predicted response probability $\hat{\pi}$ using (1) and (2). The classification rule differs from LD and may be retrieved by using (5). LD consists in choosing the cost such that the conditional error probabilities $c(0|1)$ and $c(1|0)$ are equal. The Eilers's approach consists in choosing the costs such that the error probabilities are equal: $c(1|0)\mathrm{P}(\mathbf{y}=0) = c(0|1)\mathrm{P}(\mathbf{y}=1)$. Since $\mathrm{P}(\mathbf{y}=0)$ and $\mathrm{P}(\mathbf{y}=1)$ are unknown, they are respectively estimated by $1-\overline{y}$ and $\overline{y}$ where $\overline{y}$ denotes the empirical mean of the response variable $\mathbf{y}$. This yields the classification rule

$$\hat{y} = \mathbb{1}_{(\hat{\pi} > \frac{\overline{y}}{1-\overline{y}}(1-\hat{\pi}))} = \mathbb{1}_{(\hat{\pi} > \overline{y})}.$$

In their paper, the authors propose different approaches to choose an optimal shrinkage parameter λ, based either on cross validation methods or on Akaike's Information Criterion; they use the second approach in their illustrations.

The Eilers's *et al.* method does not reduce the dimension. In particular, all the explanatory variables are allowed and included into the regression model, which can deteriorate the performances of the classifier. This is the reason why we prefer to introduce a dimension reduction step as it is now derived in the following subsection.

2.4.2 RIDGE-PLS procedure

In order to extend PLS to GLM, we want to try to find a pseudo-response variable whose expected value has a linear relationship with the covariates, and then to apply PLS. In the classical case when $n > p+1$ and the ML admits a solution, one can choose the pseudo-response variable at convergence z^∞ and apply WPLS $(z^\infty, X, W^\infty, \kappa)$. This yields an estimate of γ, denoted $\hat{\gamma}^{\mathrm{R-PLS},\kappa}$ When we have a new sample, the predicted response probability is obtained by $(1+\exp(-\hat{\eta}))^{-1}$, where $\hat{\eta} = [1, x^{new\prime}]\hat{\gamma}^{\mathrm{PLS},\kappa}$. Indeed, roughly speaking, the pseudo-response variable z^∞ can be seen as $z^\infty = Z\gamma^\infty + \varepsilon$, where conditionally to $\hat{\gamma}^{\mathrm{ML}}$ being the true value of the parameter, ε is a centered vector of covariance matrix equal to $(W^\infty)^{-1}$.
When $n < p+1$ the ML does not admit a solution and we can penalize the likelihood in order to avoid this problem. When applying the Ridge penalty, we have the same interpretation for the pseudo-response variable z^∞ at convergence. As a consequence, we propose a new procedure which combines Ridge penalty and PLS, so called RIDGE-PLS. Let λ be some positive real constant and κ be some positive integer. This algorithm divides into three steps.

1. IRRLS Step

   $(z^\infty, W^\infty) \longleftarrow \mathrm{IRRLS}(\mathbf{y}, X, \lambda, 1)$

2. WPLS Step

   $(T, \Psi, \hat{\gamma}^{R-PLS,\kappa}) \longleftarrow \mathrm{WPLS}(z^\infty, X, W^\infty, \kappa)$

3. Classification Step

   For a new sample x, compute

   $$\hat{\eta} = [1, x']\hat{\gamma}^{R-PLS,\kappa}, \quad \hat{\pi} = (1+\exp(-\hat{\eta}))^{-1}.$$

   Affect $\hat{\mathbf{y}} = 0$ if $\hat{\pi} \leq \frac{1}{2}$, otherwise set $\hat{\mathbf{y}} = 1$.

In this method we need a procedure to estimate an "optimal" value of the parameter λ from the data. We opt for a leave-one-out cross-validation approach as described in the following section.
Notice that the choice of κ is beyond the scope of this paper and will be addressed in a forthcoming work.

3. APPLICATIONS TO CLASSIFICATION OF MICROARRAY DATA

We illustrate the interest of the approaches described above by considering applications to classification of Microarray data. More precisely, we will consider in turn the Leukemia data set, and the Colon data set : the data matrix consists of gene expression intensities obtained from Affymetrix high density oligonucleotide arrays. They can be both downloaded from

$$http://sdmc.lit.org.sg/GEDatasets/Datasets.html.$$

The Leukemia data set, initially analyzed in Golub *et al.* [16], contains 72 tissue samples with 7129 genes: 47 cases of acute lymphoblastic leukemia (ALL), coded 0, and 25 cases of acute myeloid leukemia (AML), coded 1. In [16], this data set is divided into a learning set formed with 27 ALL samples and 11 AML samples; and a test set that collects the 34 remainding samples. (see [16] for a complete description of the samples).
The Colon data set, initially analyzed in Alon *et al.* [3], contains 62 tissue samples with 2000 gene expressions: 40 tumors tissues, coded 0, and 22 normal tissues, coded 1 (see [3] for a complete description of the samples) .
The MATLAB codes that implement the procedures on which the paper focuses, are available upon request from the corresponding author.

3.1 Pre-selection of genes

The collected data are first pre-processed by tresholding (floor at 100 and ceil at 16000); filtering (exclusion of genes such that max/min$\leq$ 5 and max - min $\leq$ 500, where the extremum values are computed on the *learning* set); base 10-logarithmic transformation. After this pre-processing step, the number of genes remains large (for example, 3051 for the Golub subdivision of the Leukemia data set), and many genes have similar expression pattern. Before applying a classification algorithm, a method of pre-selection of the p most informative genes for disease discrimination is run. Different pre-selection methods are proposed in the litterature, [11; 16; 29; 5], and we choose the protocol of Dudoit *et al.* [11] which is based on the ratio of the between-groups to within-groups sum of squares of expression levels of a given gene. The genes are then sorted by the value of this statistic. In the following, p refers to the number of pre-selected genes, collected in a $n \times p$ data matrix where n is the length of the learning set.

3.2 Assessing prediction methods

We restrict our study to the Ridge-PLS algorithm for $\kappa = 1, 2$, and the Eilers *et al.* 's algorithm, and to that goal, follow the NR's framework [29].
Out of Sample (OS), Leave one out (LO), Resampling: The performance of the classification rule is assessed by three kind of analyses, the first two ones being narrowly related. In the OS approach, the parameters γ of the classifier are determinded on the learning set; and the error rate od classification is computed on the test set. In the LO approach, the homogeneity of the learning set is analyzed: the learning set (of length n) is successively divided into a learning set of length $(n-1)$ and a test set with a single sample. An OS approach is applied and the leave-one-out error rate is the mean of the n out of sample error rates. For the Leukemia data set, we choose the learning set introduced in Golub [16]; for the Colon data set, we define at random a learning set that contains 14 tumor samples and 28 normal samples:

$$43, 12, 14, 10, 4, 50, 16, 2, 54, 18, 55, 60, 20, 8, 58, 19, 61, 49, 34,$$
$$44, 26, 29, 40, 25, 33, 56, 15, 41, 32, 23, 17, 21, 36, 47, 37, 46, 57,$$
$$31, 35, 52, 53, 28.$$

For both data sets, we investigate the case $p = 50$.
In the Resampling approach, the data set is divided at random into a learning set and a test set, and an OS analysis is performed. This procedure is (independently) repeated N times and the resampling error rate is the mean of the N out-of-sample error rates. In the following, we choose $N = 50$; for the Leukemia data set, each learning set contains 27 ALL samples and 11 AML samples; for the Colon data set, each learning set contains 14 tumor samples and 28 normal samples. For both data sets, we investigate the case $p = 50$.

Cross- Validation (CV): The Eilers *et al.* 's algorithm and the Ridge-PLS algorithm both necessitate the determination of a parameter λ. This is done by cross-validation and the criterion to be optimized is one of the criteria mentioned in Eilers *et al.* : we choose λ that minimizes

$$\sum_{k=1}^{n} (\mathbf{y}_k - \hat{\pi}_{-k})^2, \tag{25}$$

where $-k$ means that the sample k is left out and the probability is estimated with the remaining $(n-1)$ samples of the learning set. In the LD case, $\mathbf{y}_k - \hat{\pi}_{-k}$ is lower or equal to 0.5 when observation k is correctly classified; and the smaller it is, the more accurate the prediction is. Hence, by minimizing this quantity, one wants to minimize the number of misclassification, to correctly classify a sample with "high probability" (*i.e.* $\hat{\pi}_{-k}$ close to $\mathbf{y}_k$) and to make an error of classification with "low probability" (*i.e.* $\hat{\pi}_{-k}$ close to 0.5).
In the LO analysis, the optimal value of λ, denoted λ_{opt}, is determined by using N_l linearly spaced values of $\log_{10}\lambda$ between $l_{\min}$ and $l_{\max}$. For the Leukemia data set, $N_l = 6$, $l_{\min} = 70$ and $l_{\max} = 100$. For the Colon data set, $N_l = 14$, $l_{\min} = 35$ and $l_{\max} = 700$ when applying LO analysis to the Eilers classification rule and Ridge-PLS-1; $N_l = 15$, $l_{\min} = 270$ and $l_{\max} = 340$ when applying LO analysis to the Ridge-PLS-2. In the Resampling analysis, $N_l = 15$, $l_{\min} = 70$ and $l_{\max} = 1000$ for the Leukemia data set and $N_l = 15$, $l_{\min} = 100$ and $l_{\max} = 2000$ for the Colon data set.

A remark on the Cross- Validation step λ_{opt} is determined as the minimum of the criterion (25), over a wide range of values. In some cases, the minimal value is reached while the IRPLS step did not converge; indeed, even if the penalized likelihood is strictly concave as shown in Section 2.2.3, the concavity may be very "small" (for small values of λ) so the Newton-Raphson's algorithm do not converge.
When the minimum is reached while IRPLS did not con-

verge, we do not accept this value and choose the minimum over the values λ such that all the IRPLS steps used in the computation of the criterion converge. This phenomenon is illustrated in Figure 6: for the Leukemia data set and the OS analysis, we plot the evolution of the CV criterion vs λ; the points drawn with a $\circ$ are not considered since they are computed with non-converging IRPLS steps and hence are not significant; while the points drawn with a $\times$-mark are taken into account. Hence, we say that the minimum is not reached at $\lambda = 65 = 10^{1.81}$ but at $\lambda_{\rm opt} = 75 = 10^{1.87}$.

Implementation of Eilers et al.ś algorithm: We run the Eilers' algorithm, for different values of p. The IRPLS step is initialized as suggested in (24) with $c_0 = \ln(3)$, and stops either if

$$\sup_{1\leq j\leq p+1} \left| \gamma_j^{(t+1)} / \gamma_j^{(t)} - 1 \right| \leq 0.01, \qquad (26)$$

or when the maximal number of iterations, fixed at 20, is reached.

Implementation of the Ridge-PLS algorithm: We run the Ridge-PLS algorithm for different values of p and of the number of PLS components, κ. The initialization and the stopping rule of the IRPLS step are the same as in the Eilers' algorithm.

3.3 Results

Eilers et al. vs Ridge-PLS: The assertions of this paragraph are illustrated by considering an OS analysis on the Leukemia data set, when $p = 50$, and for the CV criterion (25) -except when specified-.

When comparing the Eilers's method with the Ridge-PLS one, we first observe that for the CV criterion (25), $\lambda_{\rm opt}^{\rm E} \sim \lambda_{\rm opt}^{\rm R-PLS,1}$ (see Table 1 and Figure 9). This property is not true in general and depends upon the CV criterion. For example, for some of the mentioned analysis, we also considered a CV criterion based on the log-likelihood: we choose λ that maximizes

$$\sum_{k=1}^{n} \{ \mathbf{y}_k \log(\hat{\pi}_{-k}) + (1-\mathbf{y}_k) \log(1-\hat{\pi}_{-k}) \}. \qquad (27)$$

The results, given in Table 4, illustrate that $\lambda_{\rm opt}^{\rm E} \neq \lambda_{\rm opt}^{\rm R-PLS,1}$.

We observed a similarity in the vector of regression coefficients γ given by the Eilers's algorithm $\hat{\gamma}^{\rm E}$ and the one by Ridge-PLS-1 $\hat{\gamma}^{\rm R-PLS,1}$: the last p components are more or less equal and the two estimates differ from the intercept parameter. In Figure 7, we plot $\hat{\gamma}^{\rm E}$, $\hat{\gamma}^{\rm R-PLS,1}$ and $\hat{\gamma}^{\rm R-PLS,2}$; in this case, $\hat{\gamma}_1^{\rm E} = -2.2765$ and $\hat{\gamma}_1^{\rm R-PLS,1} = 0.2183$. This observation is not specific to the CV criterion in use, since the same behavior can be observed when the $\lambda_{\rm opt}$ is determined by using (27), as shown in Figure 10.

The different estimation of the intercept implies a significant difference in the estimates $\hat{\pi}$, and a significant difference in the classification result even if the classification rules are different. In Figure 8, we plot the true class label of the test samples vs their estimated probability $\hat{\pi}$; as in Eilers *et al.* [12], the labels are plotted with a random vertical shift. It may be seen that the correctly classified samples (according to the Ridge-PLS method) are classified with "high accuracy" *i.e.* $\hat{\pi}$ is close to $\mathbf{y}$. The "sample of exception" is the #67 which is known in the Leukemia data set literature to be a problematic sample (and the sample #66 too; we report in Tables 1 and 4, the estimated probability of being in Class 1 of these two special samples).

Results of Tables 1 and 4 depend, by nature, upon the homogeneity of the learning set/test set. It is known that the samples #32, 35, 66, 67 of the Leukemia data set are often problematic : the first (resp. last) two ones are in the learning set (resp. test set) of the OS analysis. Hence, the presence of samples #32 and #35 in the learning set, induce a bias which is visible on the LO analysis 2. Whatever the classification rule, these two samples are misclassified when the inference of the parameter $\hat{\gamma}$ is based on the 37 remaining samples.

This bias can be weakened by running a Resampling analysis. Tables 3 and 7 show that, for the CV criterion (25), the Ridge-PLS algorithm is more outstanding, and the case $\kappa = 2$ appears slightly better than the case $\kappa = 1$.

Some comments on the numerical results We end this section with some comments of the results collected in Tables 1 to 7.

As commented above, samples #66 and 67 in the Leukemia data set are often misclassified. In our case, #67 is always correctly satisfied in the OS analysis. A strict reading of Table 1 shows that the Ridge-PLS algorithm seems to be a better classifier than the Eilers' one. Nevertheless, we must mention that the poor behavior of the Eilers' algorithm on this OS analysis may come from the choice of the CV criterion; in their paper, Eilers *et al.* use a criterion based on the Akaike's Information criterion which must be more favorable to their approach.

In the LO analysis, we see that #32, 35 are misclassified whatever the methods; here again, the Eilers' algorithm and the Ridge-PLS-1 algorithm have a similar behavior, which is slightly less accurate than the Ridge-PLS-2.

Finally, the Resampling approach confirms that the good performance of the Ridge-PLS.

Concerning the Colon data set, the OS analysis shows that the three methods have a similar behavior; observe that the two or three misclassified samples are also misclassified by Alon *et al.* [3] (more precisely samples $\#N34, T2, T30$) and by Furey *et al.* [15] (more precisely samples $\#N34, T30$). As said above, the learning set was drawn at random, and we observed that it contains 50% of the misclassified samples of Alon *et al.* (more precisely, $\#N8, T33, T36, T37$) and 50% of the misclassified samples of Furey *et al.* (more precisely, $\#T36, N8, N36, N34$). Nevertheless, the methods are quite robust and lead to results as nice as or better than what is obtained in earlier works ([3; 5; 15]).

4. CONCLUSIONS

We have proposed a statistical dimension reduction approach for the classification of tumor based on Microarrays 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 commun in such type of problems. We have extended the Partial Least Squares to binary response variable. The results on two real data

sets show that such an approach is successful. While we have not illustrated the methodology for multi-class problems, we believe that our approach can be adapted for such situations.

5. ACKNOWLEDGEMENTS

We are really grateful to A. Antoniadis for constructive comments and fruitful discussions that substantially improved this article. We would like also thank 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 "Statistical techniques and modeling for complex substantive questions with complex data".

APPENDIX

A. BASIC PROCEDURES

The procedures are described for the logit model. The input variables of the algorithms described below are chosen through

$\mathbf{y}$ Response vector, matrix $n \times 1$.

X Data matrix, matrix $n \times p$.

W Weight matrix, matrix $n \times n$.

κ Number of components, matrix 1×1.

λ Shrinkage positive parameter, matrix 1×1.

A.1 Function WPLS: $(T, \Psi, \gamma_\kappa) = \mathrm{WPLS}(\mathbf{y}, X, W, \kappa)$

The following procedure is a slight modification of the original one [20]. The instructions $(*)$ are added to determine a $(p+1) \times \kappa$ matrix Ψ such that $T = Z\Psi$ where $Z := [\mathbb{1}_n, X]$.

1. meanY $\longleftarrow \mathbf{y}'W\mathbb{1}_n/(\mathbb{1}_n'W\mathbb{1}_n)$.
   meanX $\longleftarrow \mathbb{1}_n'WX/(\mathbb{1}_n'W\mathbb{1}_n)$.
   $f_0 \longleftarrow \mathbf{y} - \mathrm{meanY}\ \mathbb{1}_n$.
   $E_0 \longleftarrow X - \mathbb{1}_n\ \mathrm{meanX}$.
   $(*)\ \psi \longleftarrow \mathrm{I}_p$.
2. For $k = 1, \cdots, \kappa$,
   $\omega(k) \longleftarrow E_{k-1}'Wf_{k-1}$.
   $T(:,k) \longleftarrow E_{k-1}\omega(k)$.
   $c(k) \longleftarrow T(:,k)'WT(:,k)$.
   $P(:,k) \longleftarrow E_{k-1}'WT(:,k)/c(k)$.
   $q(k) \longleftarrow f_{k-1}WT(:,k)/c(k)$.
   $E_k \longleftarrow E_{k-1} - T(:,k)P(:,k)'$.
   $f_k \longleftarrow f_{k-1} - q(k)T(:,k)$.
   $(*)\ \tilde{\Psi}(:,k) \longleftarrow \psi\omega(k)$.
   $(*)\ \psi \longleftarrow \psi\,(\mathrm{I}_p - \omega(k)P(:,k)')$.
3. $\beta \longleftarrow P(P'P)^{-1}q$ and $\gamma_\kappa \longleftarrow (\mathrm{meanY} - \mathrm{meanX}\ \beta, \beta')'$.
4. $\Psi(1,:) \longleftarrow -\mathrm{meanX}\ \tilde{\Psi}$ and $\Psi(2:p+1,:) \longleftarrow \tilde{\Psi}$.
5. Return
   T, matrix $n \times \kappa$.
   Ψ, matrix $(p+1) \times \kappa$.
   γ_κ, matrix $(p+1) \times 1$.

A.2 Function IRLS: $\gamma = \mathrm{IRLS}(\mathbf{y}, X)$

1. Choose $\gamma^{(0)} \in \mathbb{R}^{p+1}$.
   $t \longleftarrow 0$.
2. While $\|\Delta\gamma\| \geq$ treshold,
   $\eta^{(t)} \longleftarrow Z\gamma^{(t)}$.
   $\pi^{(t)} \longleftarrow \left((1+\exp(-\eta_k^{(t)}))^{-1}, 1 \leq k \leq n\right)'$.
   $W^{(t)} \longleftarrow \mathrm{diag}\left(\pi^{(t)}(1-\pi^{(t)})\right)$.
   $z^{(t)} \longleftarrow \eta^{(t)} + \left(W^{(t)}\right)^{-1}\left(\mathbf{y} - \pi^{(t)}\right)$.
   $\gamma^{(t+1)} \longleftarrow \left(Z'W^{(t)}Z\right)^{-1}Z'W^{(t)}z^{(t)}$.
   $t \longleftarrow t+1$.
3. Return
   $\gamma^{(t+1)}$, matrix $(p+1) \times 1$.

Each loop of this algorithm is presented as an update of the variable γ. It is equivalent to define the methods as iterative loops producing the parameters η. In that case, it must be adapted as follows: initialization concerns $\eta^{(0)}$, the loop starts with the definition of $\pi^{(t)}$ and ends with the definition of $\eta^{(t+1)}$. Finally, the stopping rule relies on the variation of η.

A.3 Function IRPLS: $T = \mathrm{IRPLS}(\mathbf{y}, X, \kappa)$

1. Choose $\eta^{(0)} = (\eta_1^{(0)}, \cdots, \eta_n^{(0)})' \in \mathbb{R}^n$.
   $t \longleftarrow 0$.
2. While $\|\Delta\eta\| \geq$ treshold,
   $\pi^{(t)} \longleftarrow \left((1+\exp(-\eta_k^{(t)}))^{-1}, 1 \leq k \leq n\right)'$.
   $W^{(t)} \longleftarrow \mathrm{diag}\left(\pi^{(t)}(1-\pi^{(t)})\right)$.
   $z^{(t)} \longleftarrow \eta^{(t)} + \left(W^{(t)}\right)^{-1}\left(\mathbf{y} - \pi^{(t)}\right)$.
   $(T^{(t+1)}, \Psi^{(t+1)}, \gamma_\kappa^{(t+1)}) \longleftarrow \mathrm{WPLS}(z^{(t)}, X, W^{(t)}, \kappa)$.
   $\eta^{(t+1)} \longleftarrow Z\gamma_\kappa^{(t+1)}$.
   $t \longleftarrow t+1$.
3. Return
   $T^{(t+1)}$, matrix $n \times \kappa$.
   $\Psi^{(t+1)}$, matrix $(p+1) \times \kappa$.

A.4 Function IRRLS: $\gamma = \mathrm{IRRLS}(\mathbf{y}, X, \lambda, switch)$

Remind that by definition, R_p is a $(p+1) \times (p+1)$ matrix such that $R(2:p+1, 2:p+1) = \mathrm{I}_p$ and the first column and first row are null vectors.

1. Choose $\gamma^{(0)} \in \mathbb{R}^{p+1}$.
   $t \longleftarrow 0$.
2. While $\|\Delta\gamma\| \geq$ treshold,
   $\eta^{(t)} \longleftarrow Z\gamma^{(t)}$.

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

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

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

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

$t \longleftarrow t+1$.

3. Return

If $switch == 0$

$\gamma^{(t+1)}$, matrix $(p+1)\times 1$.

else

$z^{(t)}$, matrix $n \times 1$.

$W^{(t)}$, matrix $n \times n$.

The update of $\gamma^{(t+1)}$ in line $(*)$ requires an inversion of a $(p+1)\times(p+1)$ matrix, which is of large cost as discussed in Eilers *et al.* [12]. By using SVD decomposition of the data matrix X, this can be substituted by instructions requiring the inversion of a $(n+1)\times(n+1)$ matrix. More precisely, set $X = UDV'$ and $\Xi := [\mathbb{1}_n, UD]$ where U (resp. V) is a $n\times n$ (resp. $p\times n$) matrix and D is a $n\times n$ diagonal matrix; the instruction $(*)$ can be replaced with the two instructions

$$\delta^{(t+1)} \longleftarrow \left(\Xi'W^{(t)}\Xi + \lambda R_n\right)^{-1}\Xi'W^{(t)}z^{(t)}.$$

$$\gamma_1^{(t+1)} \longleftarrow \delta_1^{(t+1)} \text{ and } \gamma_{2:p+1}^{(t+1)} \longleftarrow V\delta_{2:n+1}^{(t+1)}.$$

B. ALGORITHMS

B.1 Nguyen and Rocke's algorithm

1. WPLS step

   $(T, \Psi) \longleftarrow \text{WPLS}(\mathbf{y}, X, \mathrm{I}_n, \kappa)$ where κ is chosen by the user.

2. IRLS step

   $\gamma_{s+1} \longleftarrow \text{IRLS}(\mathbf{y}, T(:, 1:s))$ where $s \le \kappa$ is chosen by the user.

   $\hat{\gamma}^{\text{NR}} \longleftarrow [e_1, \Psi(:, 1:s)]\,\gamma_{s+1}$ where e_1 is the first vector of the canonical basis of $\mathbb{R}^{p+1}$.

3. Classification Step

   For a new sample x, compute

   $$\hat{\eta} = [1, x']\hat{\gamma}^{\text{NR}}, \quad \hat{\pi} = (1+\exp(-\hat{\eta}))^{-1}.$$

   Affect $\hat{\mathbf{y}} = 0$ if $\hat{\pi} \le 0.5$, otherwise set $\hat{\mathbf{y}} = 1$.

B.2 Marx's algorithm

1. IRPLS Step

   $(T, \Psi) \longleftarrow \text{IRPLS}(\mathbf{y}, X, \kappa_{\max})$.

2. IRLS Step

   $\gamma_{s+1} \longleftarrow \text{IRLS}(\mathbf{y}, T(:, 1:s))$, where $s \le \kappa$ is chosen by the user.

   Set $\hat{\gamma}^{\text{M}} \longleftarrow [e_1, \Psi(:, 1:s)]\,\gamma_{s+1}$ where e_1 is the first vector of the canonical basis of $\mathbb{R}^{p+1}$.

3. Classification Step

   For a new sample x, compute

   $$\hat{\eta} = [1, x']\hat{\gamma}^{M}, \quad \hat{\pi} = (1+\exp(-\hat{\eta}))^{-1}.$$

   Affect $\hat{\mathbf{y}} = 0$ if $\hat{\pi} \le 0.5$, otherwise set $\hat{\mathbf{y}} = 1$.

B.3 Eilers's algorithm

1. IRRLS Step

   $\hat{\gamma}^{E} \longleftarrow \text{IRRLS}(\mathbf{y}, X, \lambda, 0)$ for some positive λ.

2. Classification Step

   For a new sample x, compute

   $$\hat{\eta} = [1, x']\hat{\gamma}^{E}, \quad \hat{\pi} = (1+\exp(-\hat{\eta}))^{-1}.$$

   Affect $\hat{\mathbf{y}} = 0$ if $\hat{\pi} \le \mathbb{1}_n'\mathbf{y}/n$, otherwise set $\hat{\mathbf{y}} = 1$.

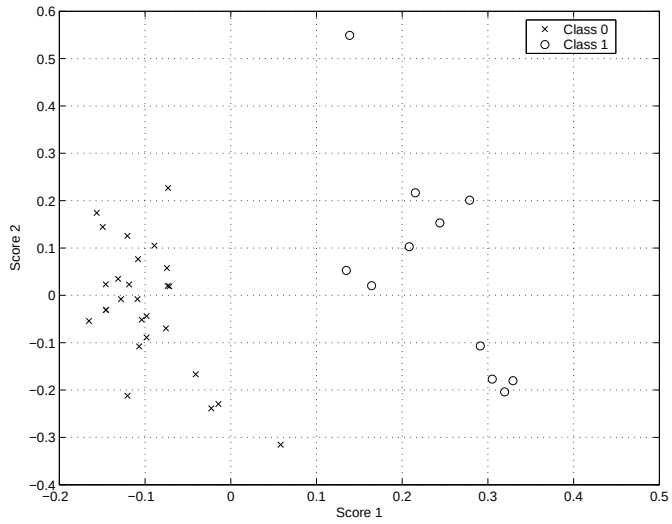


Figure 1: 38 points with coordinates $(t_{1,k}, t_{2,k})$ where $t_{j,k}$ stands for the k-th component of the normalized j-th PLS component. This shows that the points are separated. This discrimination is induced by the first PLS component so the IRLS step can not converge, whatever the number of "super-covariates" used in the IRLS step.

C. FIGURES AND TABLES

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

[3] 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*, 96:6745–6750, 1999.

[4] J. Anderson and V. Blair. Penalized maximum likelihood estimation in logistic regression and discrimination. *Biometrika*, 69:123–136, 1982.

[5] 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.

[6] D. Bates. The derivative of $|X'X|$ and its uses. *Technometrics*, 25(4):373–376, 1983.

[7] D. Collett. *Modelling Survival Data in Medical Research*. Chapman and Hall, London, 1994.

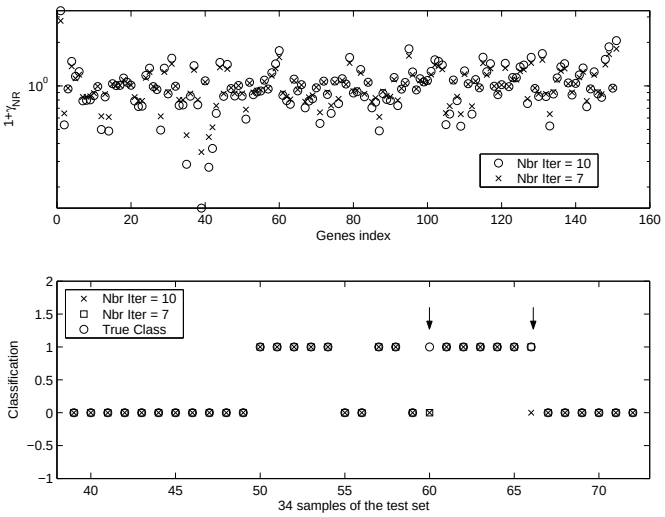


Figure 2: (Top) Plot (on a logarithmic scale) of $1 + \hat{\gamma}^{\mathrm{NR}}$ when IRLS (that does not converge) is stopped after 7 and 10 iterations. (Bottom) Out of Sample classification : it depends upon the number of iterations of IRLS, showing the instability of this classification algorithm.

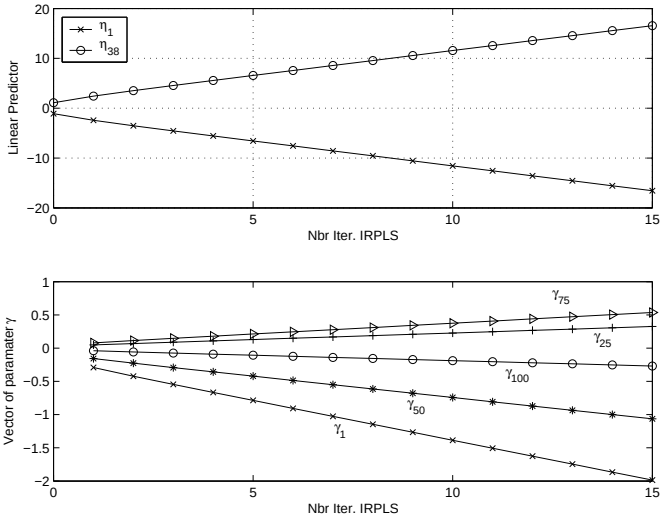


Figure 3: IRPLS step: (Top) Evolution of two components of the linear predictor: η_1 (sample from Class 0) and η_{37} (sample from Class 1). This shows that $\eta_1^{(t)} = -\eta_{37}^{(t)}$ for all t, and $\eta_1^{(t+1)} = 1 + \eta_1^{(t)} + \exp(-\eta_1^{(t)})$. (Bottom) Evolution of five components of the vector γ: this step can not converge and $\|\gamma\| \to +\infty$.

Table 1: Leukemia: Out of Sample

		λ_{opt}	# of misclassified samples	Comments
$p = 50$	$\kappa = 1$	75	66	$\hat{\pi}_{66} = 0.3158 \quad \hat{\pi}_{67} = 0.4611$
	$\kappa = 2$	79	57,60,61,66	$\hat{\pi}_{66} = 0.2289 \quad \hat{\pi}_{67} = 0.2112$
	Eilers	75	54,60,61,64,66	$\hat{\pi}_{66} = 0.0336 \quad \hat{\pi}_{67} = 0.0812$
$p = 150$	$\kappa = 1$	166	66	$\hat{\pi}_{66} = 0.2942 \quad \hat{\pi}_{67} = 0.4667$
	$\kappa = 2$	500	66	$\hat{\pi}_{66} = 0.0294 \quad \hat{\pi}_{67} = 0.3188$
	Eilers	150	54,60,61,62,64,66	$\hat{\pi}_{66} = 0.0338 \quad \hat{\pi}_{67} = 0.0840$
$p = 300$	$\kappa = 1$	240	61,66	$\hat{\pi}_{66} = 0.2689 \quad \hat{\pi}_{67} = 0.3595$
	$\kappa = 2$	627	66	$\hat{\pi}_{66} = 0.2909 \quad \hat{\pi}_{67} = 0.2656$
	Eilers	240	54,60,61,62,64,66	$\hat{\pi}_{66} = 0.0393 \quad \hat{\pi}_{67} = 0.0721$

Table 2: Leukemia: Leave One Out

		# of misclassified samples
$p = 50$	$\kappa = 1$	28,32,35
	$\kappa = 2$	32,35
	Eilers	28,32,35

Table 3: Leukemia: Resampling

	$\kappa = 1$	$\kappa = 2$	Eilers
Mean	0.0523	0.0518	0.1047
Std	0.0461	0.0349	0.0476

Table 4: Leukemia: Out of Sample, CV criterion (27)

		λ_{opt}	# of misclassified samples	Comments
$p = 50$	$\kappa = 1$	66	66,67	$\hat{\pi}_{66} = 0.4226 \quad \hat{\pi}_{67} = 0.5355$
	$\kappa = 2$	74	57,60,61,66	$\hat{\pi}_{66} = 0.2403 \quad \hat{\pi}_{67} = 0.2227$
	Eilers	101	54,60,61,64,66	$\hat{\pi}_{66} = 0.0584 \quad \hat{\pi}_{67} = 0.1143$
$p = 150$	$\kappa = 1$	180	66	$\hat{\pi}_{66} = 0.2540 \quad \hat{\pi}_{67} = 0.4229$
	$\kappa = 2$	245	54,60,62,66	$\hat{\pi}_{66} = 0.2284 \quad \hat{\pi}_{67} = 0.2489$
	Eilers	185	54,60,61,62,64,66	$\hat{\pi}_{66} = 0.0521 \quad \hat{\pi}_{67} = 0.1086$
$p = 300$	$\kappa = 1$	250	61,66	$\hat{\pi}_{66} = 0.2544 \quad \hat{\pi}_{67} = 0.3454$
	$\kappa = 2$	356	54,60,62,66	$\hat{\pi}_{66} = 0.1757 \quad \hat{\pi}_{67} = 0.1735$
	Eilers	356	54,60,61,62,64,66	$\hat{\pi}_{66} = 0.0784 \quad \hat{\pi}_{67} = 0.1183$

Table 5: Colon: Out of Sample

		λ_{opt}	# of misclassified samples	Identification as in Alon *et al.*
$p = 50$	$\kappa = 1$	40	45,51	N34 T30
	$\kappa = 2$	300	3,45,51	N34 T2,30
	Eilers	40	3,45,51	N34 T2,30
$p = 150$	$\kappa = 1$	100	3,45,51	N34 T2,30
	$\kappa = 2$	397	3,45,51	N34 T2,30
	Eilers	100	3,45,51	N34 T2,30

Table 6: Colon: Leave One Out

		# of misclassified samples	Identification as in Alon *et al.*
$p = 50$	$\kappa = 1$	16,18,49,55,56	N8,9,36 T33,36
	$\kappa = 2$	15,16,43,49,50,55,56,57	N8,29,33,36 T8,33,36,37
	Eilers	16,18,49,55,56,57	N8,9,36 T33,36,37

Table 7: Colon: Resampling

	$\kappa = 1$	$\kappa = 2$	Eilers
Mean	0.1430	0.1500	0.1440
Std	0.0693	0.0631	0.0787

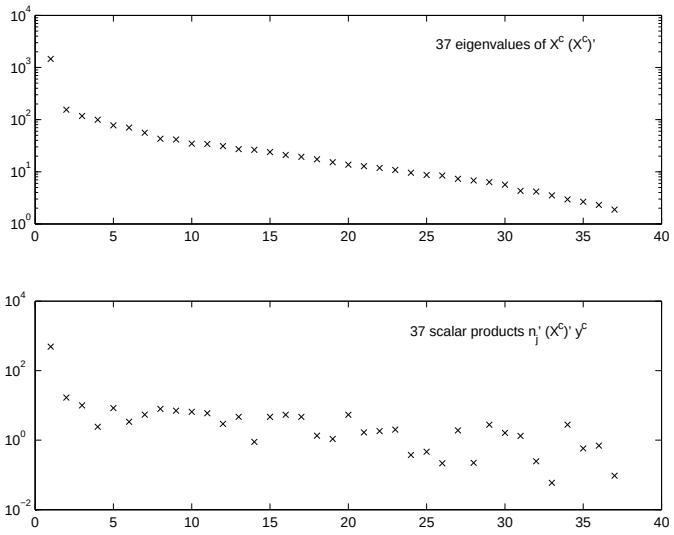


Figure 4: (Top) 37 distinct eigenvalues of the matrix $X^c X^{c\prime}$. (Bottom) 37 scalar product $\mathbf{y}' X^c \nu_j$ where ν_j is the j-th eigenvector of $X^{c\prime} X^c$. For this data set, the maximal number of PLS components is $\kappa_{\max} = 37$.

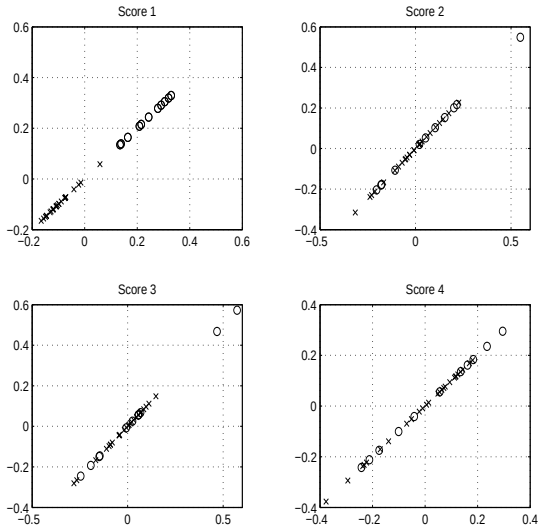


Figure 5: Normalized PLS components obtained "at convergence" of the IRPLS step of Marx's algorithm (y-axis) vs normalized PLS components given by the PLS step of the NR algorithm (x-axis). Coordinates corresponding to a sample from Class 0 (resp. 1) are drawn with a circle (resp. $\times$-mark). The components are equal so the classification step of the two algorithms coincide. In addition, the observation of the Score 1 shows that the classes are separated so the IRLS step of both algorithms can not converge.

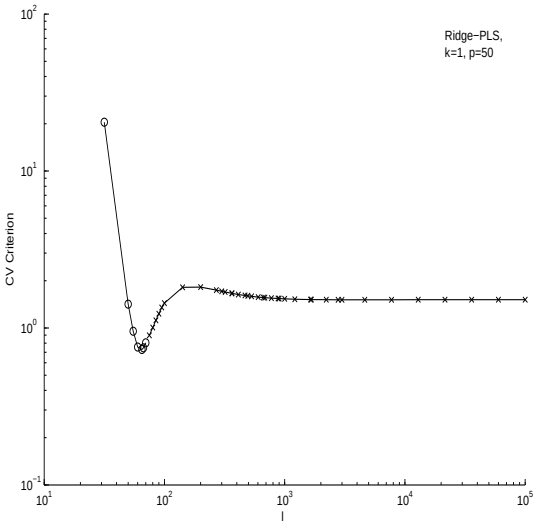


Figure 6: Leukemia data set, OS analysis. Evolution of the CV criterion (25) vs λ. The points drawn with a circle (resp. $\times$-mark) correspond to a criterion computed with non-converging (resp. converging) IRPLS steps.

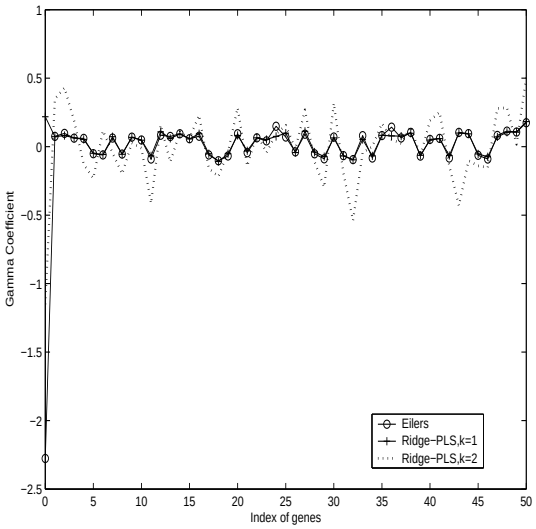


Figure 7: Leukemia data set, OS analysis, $p = 50$. Plot of the estimated coefficients $\hat{\gamma}$ vs the index of the genes.

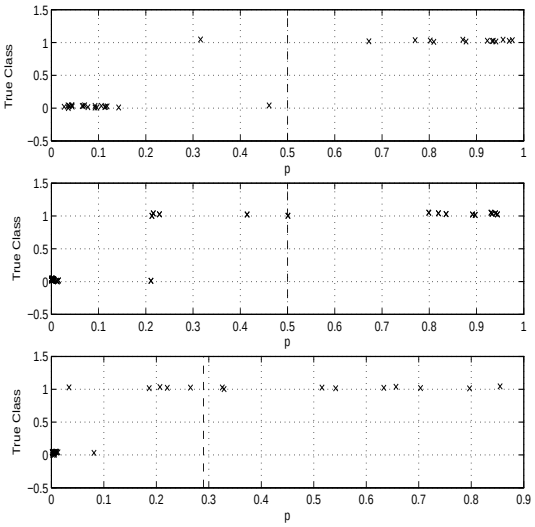


Figure 8: Leukemia data set, OS analysis, $p = 50$. True Class of the test samples vs the estimated probability $\hat{\pi}$. (Top) Ridge-PLS, $\kappa = 1$, (Middle) Ridge-PLS, $\kappa = 2$, (Bottom) Eilers. In the three cases, the samples are classified as 0 if on the left of the dash line, and 1 otherwise. Upper left and lower right samples are thus misclassified.

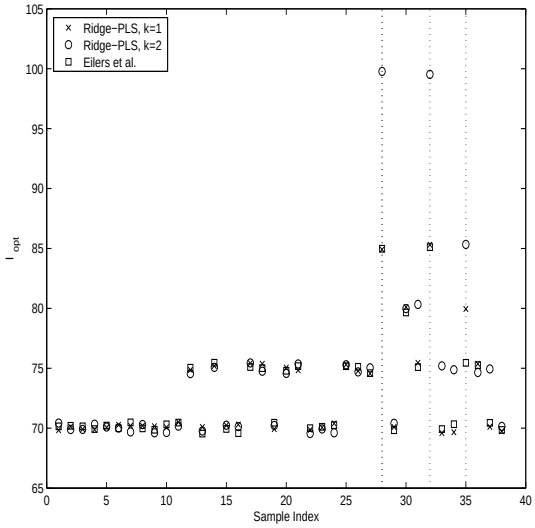


Figure 9: Leukemia data set, LO analysis, $p = 50$. λ_{opt} is plotted vs the index of the genes in the learning set; the extremum is computed over 7 log-linearly spaced points between 70 and 100. The values are drawn with a random vertical shift (lower than 0.5 in absolute value).

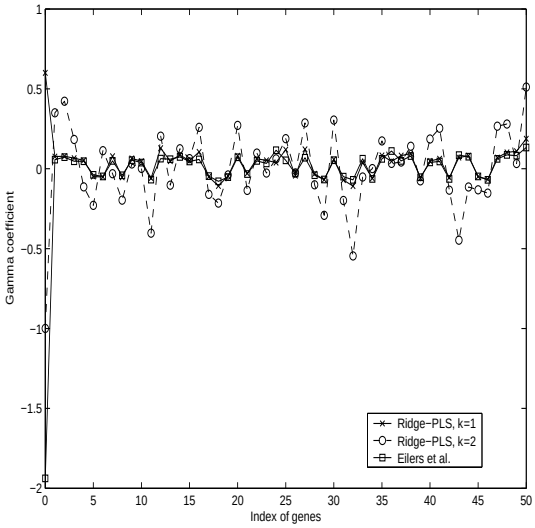


Figure 10: Leukemia data set, LO analysis, $p = 50$. Plot of the estimated coefficients $\hat{\gamma}$ vs the index of the genes when the optimal value λ_{opt} is computed by using the CV criterion (27).

[8] M. Denham. Prediction intervals in Partial Least Squares. *Journal of Chemometrics*, 11:39–52, 1997.

[9] M. Denham. Choosing the number of factors in partial least squares regression : estimating and minimizing the mean squared error of prediction. *Journal of Chemometrics*, 14:351–361, 2000.

[10] B. Ding and R. Gentleman. Classification Using Generalized Partial Least Squares. Work in progress, 2003.

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

[12] P. Eilers, J. Boer, V. O. G.J., 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.

[13] D. Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, 1993.

[14] I. Frank and J. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–135, 1993.

[15] 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.

[16] 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*, 286:531–537, 1999.

[17] D. Gosh. Singular value decomposition regression modelling for classifiaction of tumors from microarray experiments. *Procceding of the Pacific Symposium on Biocomputing*, 98:11462–11467, 2002.

[18] 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.

[19] G. Heinze and M. Schemper. A solution to the probelm of separation in logistic regression. *Statistics in Medicine*, 21:2409–2419, 2002.

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

[21] I. Helland. Partial Least Squares Regression and Statistical Models. *Scand. J. Statist.*, 17:97–114, 1990.

[22] A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.

[23] R. A. Johnson and D. W. Wichern. *Applied multivariate statistical analysis. 3rd ed.* Prentice-Hall International Editions, 1992.

[24] S. le Cessie and J. Van Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41:191–201, 1992.

[25] B. D. Marx. Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics*, 38(4):374–381, 1996.

[26] W. F. Massy. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60:234–246, 1965.

[27] P. McCullagh and J. Nelder. *Generalized Linear Models. 2nd ed.* New-York : Chapman & Hall, 1989.

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

[29] D. Nguyen and D. Rocke. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50, 2002.

[30] A. Phatak, P. Reilly, and A. Penlidis. The asymptotic variance of the univariate PLS estimator. *Linear Algebra and its Applications*, 354:245–253, 2002.

[31] 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.

[32] M. Stone. Cross-validatory choice and assessment of statistical predictions. Discussion. *J. R. Stat. Soc., Ser. B*, 36:111–147, 1974.

[33] 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 using gene expression profiles. *P.N.A.S*, 2002.

[34] H. Wold. Soft modelling by latent variables: The non-linear iterative partial least squares (NIPALS) approach. pages 117–142, 1975.