Classification using LS-PLS with logistic regression based on both clinical and gene expression variables

Caroline Bazzoli, Sophie Lambert-Lacroix

Classification using LS-PLS with logistic regression based on both clinical and gene expression variables

Caroline Bazzoli*, Sophie Lambert-Lacroix †

November 25, 2016

Abstract

Prediction from high-dimensional genomic data is an active field in today's medical research. Most of the proposed prediction methods make use of genomic data alone without considering established clinical data that often are available and known to have predictive value. Recent studies suggest that combining clinical and genomic information may improve predictions. We consider in this paper methods for classification purposeS that simultaneously use both types of variables, but applying dimension reduction only to the high-dimensional genomic ones. A usual way to deal with that is the use of a two-step approach. In step one, dimensionality reduction technique is just performed on the genomic dataset. In step two, the selected genomic variables are merged with the clinical variables to build a classification model on the combined dataset. Nevertheless, the reduction dimension is built without taking into account the link between the response variable and the clinical data. To address this issue, using Partial Least Squares (PLS) as reduction technique, we propose here a one step approach based on three extensions of LS-PLS (LS for Least Squares) method for logistic regression context. We perform a simulation study to evaluate these approaches compared to methods using only the clinical data or only genetic data. Then, we illustrate their performances to classify two real data sets containing both clinical information and gene expression.

Keywords : LS-PLS; clinico-genomic model; logistic regression; reduction dimension

1 Introduction

In the last 15 years, progress in the generation of high-dimensional genomic data has raised high expectations in biomedical research. In particular, large-scale gene expression data have been generated and analysed in numerous studies with the aim of predicting

*Laboratoire Jean Kuntzmann, Dpartement Statistique,Universite Grenoble Alpes, France : caroline.bazzoli@imag.fr

†TIMC-IMAG, Universite Grenoble Alpes France

a specific outcome [16]. In this article, we focus on binary class prediction where the outcome can be for instance alive versus dead. Most of these studies [26, 24, 19, 23] include clinical data in addition to genomic data using most of the proposed prediction methods with only genomic data, which involves some statistical issues. In genomic studies, the number n of samples is often relatively small compared to the number p of covariates and collinearity between measurements occurs. Unless a preliminary step of selection of variables is performed, standard classification methods are not appropriate. To address such a large p small n problem, reduction dimension methods can be used. The traditional approach is the Principal Component Regression (PCR) [15], an application of Principal Component Analysis (PCA) to regression model. PCA is applied without considering of the link between the outcome and the independent variables. An alternative method is the Partial Least Square (PLS) [11], that takes this link into account.

In recent studies [1, 22], it has been shown that most complex diseases are caused by the combined effects of many diverse factors, including genomic and clinical variables. This has led to an emerging research area of integrative studies of clinical and genomic data, which we will refer to as clinico-genomic models. Some strategies to combine these two kinds of data have been reviewed in a paper written by Boulesteix and Sauerbrei [3] to deal with predictive clinico-genomic models. More extensive overviews are available in Dey *et al.*[5] where advantages and disadvantages are given for each strategy. Regarding the dimension reduction strategy, one possible way to handle the high dimension of genomic data is to first perform dimensionality reduction techniques solely on the genomic dataset. In the second step, the selected genomic variables are merged with the clinical variables to build a classification model on the combined dataset. We will thus refer to it as the two-step approach. Beforehand most of the techniques select topmost discriminative genomic features and then combine those features into a combined score for future model development. In the same way, Boulesteix *et al.* [2] suggest an approach combining PLS dimension reduction with a pre-validation technique and Random Forests, applied with both the new components and the clinical variables as predictors. These papers mainly describe methods using PLS dimension reduction to treat high-dimensional data. Even if any type of dimensionality reduction methods can be incorporated, these two-step approaches cannot account for the relationship existing between two datasets. Indeed, this reduction is built without taking into account the link between the response variable and the clinical data.

An alternative could be to use an iterative procedure well suited to extract relevant information from the genomic data in combination with clinical variables. One idea is to use the principle of backfitting procedures developed in the context of multidimensional regression problems and derived for generalized additive models [10], estimating additive components successively in a non parametric manner. Precisely, this involves repeatedly fitting non parametric regression of some partial residuals on each covariate ; for each regression, a new additive component is estimated, which in turn gives new partial residuals, and the process is iterated until convergence. Then, relevant information from both types

of data updates takes place within the iterations. This approach has been developed by Jorgensen *et al.* [12] in the regression Gaussian context in chemometrics. In this context, non parametric regression is replaced by PLS regression for data to be compressed and Ordinary Least Squares regression (OLS) for other data, so-called LS-PLS. The PLS scores are thus incorporated into the OLS equations in an iterative fashion in order to obtain a model for both the clinical variables and the genomic ones. The authors conclude that the method seems to involve more information from the experiment and return lower variance in the parameter estimates.

The purpose of this paper is to adapt this one step procedure to logistic regression models. Some studies have been published proposing an adaptation of PLS for classification problems [17, 14, 8]. The focus will be on adapting these extensions to LS-PLS for logistic regression model. Section 2 describes the details of the linear logistic regression and the three proposed adaptations of LS-PLS for logistic regression models. The simulation study to evaluate these approaches is presented in Section 3 and the illustration on two real data sets containing both clinical information and gene expression data in Section 4.

2 LS-PLS for logistic regression

2.1 Linear logistic regression - ridge penalty and RIRLS

We consider situations where we have both collinear measurements such as high-dimensional genomic data and orthogonal (or near- orthogonal) design variables on one side that we want to relate to a response value on the other side. We denote by $\mathbf{X}$ the design matrix associated with collinear measurements. For instance in genetics, expression levels of the p genes for the n genomic samples are collected in a $n \times p$ data matrix $\mathbf{X}$. The clinical variables are stocked in matrix $\mathbf{D}$ of size $n \times q$. We denote $\mathbf{D}_{i\cdot}$ the i^{th} row of matrix $\mathbf{D}$. In logistic regression, the response variables are collected in a $\{0,1\}^n$-valued vector $\mathbf{Y}$.

In a typical designed experiment logistic model, the conditional class probability, i.e. the conditional expectation of Y_i given $\mathbf{D}_{i\cdot}$, given by $\pi_i = \mathbb{P}(Y_i = 1|\mathbf{D}_{i\cdot} = \mathbf{d}_i)$ is related to the linear predictor $\eta_i = [1\ \mathbf{d}_i^T]\boldsymbol{\gamma}$, with $\boldsymbol{\gamma} \in \mathbb{R}^{q+1}$ through the non-linear relation $\pi_i = h(\eta_i)$ where $h(\eta_i) = 1/(1+\exp(-\eta_i))$. The parameter $\boldsymbol{\gamma}$ is unknown and has to be estimated from the data. Let us notice that here we do not index vector $\boldsymbol{\pi}$ and $\boldsymbol{\eta}$ by $\boldsymbol{\gamma}$ for sake of clarity. In logistic discrimination, it is usually estimated by $\hat{\boldsymbol{\gamma}}^{\mathrm{ML}}$, the ML estimator. The log-likelihood of the observations for the value $\boldsymbol{\gamma}$ of the parameter, simply denoted by $\ell(\boldsymbol{\gamma})$, is given by

$$\ell(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \left\{ y_i \eta_i - \ln\left(1 + \exp(\eta_i)\right\}. \right. \tag{1}$$

Let $\mathbf{W}(\boldsymbol{\gamma})$ be the diagonal $n \times n$ matrix with entries $\{\mathbf{W}(\boldsymbol{\gamma})\}_{ii} = \pi_i(1-\pi_i)$. For a vector $\mathbf{d}_0$, the predicted class $\hat{Y_0}$ of the sample is given by $\hat{Y_0} = 1_{(\hat{\pi}_0 > 1-\hat{\pi}_0)}$, where $\hat{\pi}_0 = h([1\ \mathbf{d}_0^T]^T\hat{\boldsymbol{\gamma}}^{\mathrm{ML}})$ and $1_{(\cdot)}$ is the indicator function. When this estimate exists, it is computed as the limit of

a Newton-Raphson sequence; this algorithm is known as the Iteratively Reweighted Least Squares (IRLS($\mathbf{Y}, \tilde{\mathbf{D}}$)) algorithm (see [9]). From step t to $t+1$, we have :

$$\mathbf{z}^{(t)} = \tilde{\mathbf{D}}\boldsymbol{\gamma}^{(t)} + \left[\mathbf{W}^{(t)}\right]^{-1}\left(\mathbf{Y} - \boldsymbol{\pi}^{(t)}\right), \quad (2)$$

$$\boldsymbol{\gamma}^{(t+1)} = \left(\tilde{\mathbf{D}}^T\mathbf{W}^{(t)}\tilde{\mathbf{D}}\right)^{-1}\tilde{\mathbf{D}}^T\mathbf{W}^{(t)}\mathbf{z}^{(t)}, \quad (3)$$

where $\tilde{\mathbf{D}} = [\mathbb{1}_n \ \mathbf{D}]$, $\mathbb{1}_n = (1, \cdots, 1)^T$, and $\mathbf{W}^{(t)}$ is shorthand notation for $\mathbf{W}(\boldsymbol{\gamma}^{(t)})$. IRLS can thus be considered as iterative $\mathbf{W}(\boldsymbol{\gamma}^{(t)})$-weighted least square regression of a $\mathbb{R}^n$-valued pseudo-variable $\mathbf{z}^{(t)}$ onto the columns of $\tilde{\mathbf{D}}$. Let us note that in some cases, including in practice the case $n << p$, existence and unicity of $\hat{\boldsymbol{\gamma}}^{\mathrm{ML}}$ for logit models is not guaranteed. That calls for regularization methods such as ridge penalty. The ridge estimator [13], denoted by $\hat{\boldsymbol{\gamma}}^R$, is defined as the (unique) maximizer of the penalized likelihood $\ell^*(\boldsymbol{\gamma}) = \ell(\boldsymbol{\gamma}) - 0.5\lambda\boldsymbol{\gamma}^T\boldsymbol{\gamma}$, where $\lambda > 0$ is the shrinkage parameter. We denote by RIRLS($\mathbf{Y}, \tilde{\mathbf{D}}, \lambda$) (shorthand notation for Ridge-IRLS) this algorithm. It consists in replacing in IRLS, the weighted regression (3) by a weighted Ridge regression $\boldsymbol{\gamma}^{(t+1)} = (\tilde{\mathbf{D}}^T\mathbf{W}^{(t)}\tilde{\mathbf{D}} + \lambda\tilde{\mathbf{I}}_{q+1})^{-1}\tilde{\mathbf{D}}^T\mathbf{W}^{(t)}\mathbf{z}^{(t)}$, where $\mathbf{z}^{(t)}$ is built as in (2) and $\tilde{\mathbf{I}}_{q+1}$ is diagonal matrix of size $(q+1)\times(q+1)$ whose diagonal is equal to $(0, 1, \ldots, 1)$. Let us note that when the model does not contain the intercept term (use of $\mathbf{D}$ instead of $\tilde{\mathbf{D}}$), the matrix $\tilde{\mathbf{I}}_{q+1}$ is replaced by the identity matrix $\mathbf{I}_q$. The parameter λ controls the amount of shrinkage in the data and can be chosen from the data for instance by cross-validation procedure.

2.2 LS-PLS for logistic regression

Combination of least squares (LS) and PLS (called LS-PLS) has been introduced first in the Gaussian context by Jorgensen *et al.* [12]. This is an iterative procedure: the first step is to use OLS on $\tilde{\mathbf{D}}$ to predict $\mathbf{Y}$ and compute the residuals. Then PLS is performed between $\mathbf{X}$ and the residuals. The matrix of PLS scores $\mathbf{T}$, of size $n \times \kappa$, combined with $\tilde{\mathbf{D}}$ in a new OLS to predict $\mathbf{Y}$. New estimates for the residuals of $\mathbf{Y}$ on $\tilde{\mathbf{D}}$ are calculated from this regression and the algorithm is repeated until convergence. The authors suggest orthogonalising $\tilde{\mathbf{D}}$ on $\mathbf{X}$. The orthogonalised variant is better suited for situations where the focus is on identifying the unique information in each matrix. We propose here to project the matrix $\mathbf{X}$ into a space orthogonal to the space spanned by the design variables of $\tilde{\mathbf{D}}$:

$$\mathbf{X}_{Orth} = (\mathbf{I}_n - \tilde{\mathbf{D}}(\tilde{\mathbf{D}}^T\tilde{\mathbf{D}})^{-1}\tilde{\mathbf{D}}^T)\mathbf{X}. \quad (4)$$

The standard PLS regression is then used on $\mathbf{X}_{Orth}$ instead of $\mathbf{X}$. This avoids iterations in the algorithm since the residuals of $\mathbf{Y}$ on $\tilde{\mathbf{D}}$ of the OLS of $\mathbf{Y}$ on $[\tilde{\mathbf{D}}\ \mathbf{T}]$ are the same as in the first step. This procedure is denoted by LS-PLS($\mathbf{Y}, \mathbf{D}, \mathbf{X}, \kappa$).

Extending this approach to the framework of the logistic model is not straightforward. For instance, there are several ways to use PLS in classification context. In the following

section, we propose to consider three different approaches[17, 14, 8] to adapt LS-PLS to logistic regression.

2.2.1 Nguyen and Rocke's approach.

To extend PLS to logistic regression, Nguyen and Rocke [17] substitute the data matrix $\mathbf{X}$ by a $n \times \kappa$ matrix $\mathbf{T}$, the columns of which are the first κ PLS-scores given by PLS regression of $\mathbf{Y}$ on $\mathbf{X}$. Then they estimate the parameter in the maximum likelihood sense by running IRLS($\mathbf{Y}, \mathbf{T}$). If we want to adopt this approach for LS-PLS, one can to replace the call to PLS step by LS-PLS($\mathbf{Y}, \mathbf{D}, \mathbf{X}, \kappa$) and then perform IRLS($\mathbf{Y}$, [$\mathbf{D}$ $\mathbf{T}$]). Let us note that applying PLS with a binary input $\mathbf{Y}$ is unappealing; in addition, the PLS-regression step does not take into account the heteroscedasticity of the response vector $\mathbf{Y}$. However this leads to relatively good results in practice. We call this method LS-PLS-IRLS. The choice of κ can be made by cross-validation.

2.2.2 Marx's approach.

Marx [14] introduces an algorithm that extends PLS to generalized linear models, so-called IRPLS. More precisely, IRPLS can be understood as an IRLS algorithm in which the weighted least square regression (3) is replaced with the PLS regression, PLS($[\mathbf{W}^{(t)}]^{1/2}\mathbf{z}^{(t)}, [\mathbf{W}^{(t)}]^{1/2}\mathbf{X}, \kappa$). The matrix $\mathbf{T}$ contains the first κ components "at convergence" of IRPLS. Let us notice that PLS applied with the maximal number of PLS components is nothing else than Least Square (note that Marx [14] chooses $\kappa = \text{rank}(\mathbf{X})$ while in theory). Hence when $\mathbf{X}$ is full row-rank (which is most often the case when $n << p$), this algorithm never converges. Some authors (see for instance [20, 18]) use similar algorithms but with $\kappa < \text{rank}(\mathbf{X})$. In this case nothing ensures that this algorithm converges. As previously, if we want to adopt this approach for LS-PLS, we can simply replace the call to PLS by LS-PLS. The drawback of this method is that there are often problems of convergence. We call this method IR-LS-PLS. The parameter κ can also be selected by cross-validation.

2.2.3 Ridge Partial Least Squares approach.

In order to extend PLS to logistic regression model, Fort *et al.* [8] suggest replacing the binary data by a pseudo-response variable whose expected value has a linear relationship with the covariates called R-PLS. The pseudo-response variable $\mathbf{z}^{\infty}$ at convergence of RIRLS algorithm verifies this condition: it can be written $\mathbf{z}^{\infty} = \mathbf{X}\hat{\boldsymbol{\gamma}}^{R} + \boldsymbol{\varepsilon}$, where, conditionally to $\hat{\boldsymbol{\gamma}}^{R}$ being the true value of the parameter, $\boldsymbol{\varepsilon}$ is a centred vector of covariance matrix $(\mathbf{W}^{\infty})^{-1}$. As a consequence, in the same spirit, to extend LS-PLS to logistic regression, we can propose a procedure which combines Ridge penalty and LS-PLS, called R-LS-PLS. Let λ be some positive real constant and κ be some positive integer. R-LS-PLS is divided in two steps:

1. $(\mathbf{z}^{\infty}, \mathbf{W}^{\infty}) \longleftarrow \text{RIRLS}(\mathbf{Y}, [\mathbf{D}\ \mathbf{X}], \lambda)$

2. $(\hat{\gamma}^{\tilde{\mathbf{D}}}, \hat{\gamma}^{\mathbf{X},\kappa}) \longleftarrow \text{LS-PLS}([\mathbf{W}^{\infty}]^{1/2}\mathbf{z}^{\infty}, [\mathbf{W}^{\infty}]^{1/2}\mathbf{D}, [\mathbf{W}^{\infty}]^{1/2}\mathbf{X}, \kappa)$,

where $[\mathbf{W}^{\infty}]^{1/2}$ is a square root matrix of $\mathbf{W}^{\infty}$ that satisfies $[\mathbf{W}^{\infty}]^{T/2}[\mathbf{W}^{\infty}]^{1/2} = \mathbf{I}_n$. The first step builds a continuous response variable $\mathbf{z}^{\infty}$ for the input of LS-PLS, the "dispersion matrix" of which is $[\mathbf{W}^{\infty}]^{-1}$. This explains, in the second step, the weight $[\mathbf{W}^{\infty}]^{1/2}$. Let us remark that in the Step 1, we do not chose to regularize $\mathbf{D}$. For the matrix X, when its dimension is low, we may decide to not regularize it by putting $\lambda = 0$ in the Step 1. R-LS-PLS depends on two parameters, λ and κ that can be selected by cross-validation.

These three approaches have been implemented in R software version 3.1.2.

3 Simulation study

The aim of the simulated study is to compare the different prediction methods developed previously based on clinical and/or gene expression variables. We simulate data sets with a range of predictor collinearity and with different functional relationships between the response, Y_i and the predictors $\mathbf{X}_{i\cdot}$ and $\mathbf{D}_{i\cdot}$ to mimic gene expression and clinical variable data. For an individual $i = 1, ..., n$ with $n = 100$, we simulated $Y_i \sim \mathcal{B}(\pi_i)$ with $\pi_i = [1\ \mathbf{D}_{i\cdot}\ \mathbf{X}_{i\cdot}]\boldsymbol{\gamma}$ and $\boldsymbol{\gamma} = [\gamma_1\ \boldsymbol{\gamma}_{\mathbf{D}}^T\ \boldsymbol{\gamma}_{\mathbf{X}}^T]^T$. We fixed $\gamma_1 = -2.5$, $\boldsymbol{\gamma}_{\mathbf{D}} = \{0.5\}^4$ and $\boldsymbol{\gamma}_{\mathbf{X}} = \{\{0\}^{475}, \{0\}^{475}, \{0.1\}^{25}, \{0.1\}^{25}\}$. The matrix $\mathbf{X}$ of size $n \times p$ with $p = 1000$ has been simulated such as $\mathbf{X} = (\mathbf{X}^1, \mathbf{X}^2, \mathbf{X}^3, \mathbf{X}^4)$ where $\mathbf{X}^k \sim N(0_{bs^k}, \boldsymbol{\Sigma}_{\mathbf{X}}^k)$ with $\{\boldsymbol{\Sigma}_{\mathbf{X}}^k\}_{ij} = c^k\rho^{|i-j|}$, $k = 1, ...4$, $i, j = 1, ..., bs^k$ where $c^1 = 8$, $c^2 = 4$, $c^3 = 2$, $c^4 = 1$, $bs^1 = bs^2 = 475$, $bs^3 = bs^4 = 25$, and $\rho = 0.9$. Regarding the matrix $\mathbf{D}$ of size n x q with $q = 4$, we used $N(0_q, \boldsymbol{\Sigma}_{\mathbf{D}})$ with $\{\boldsymbol{\Sigma}_{\mathbf{D}}\}_{ij} = \rho^{|i-j|}$ with $i,\ j = 1, ..., q$ and $\rho = 0.5$. According to this model, we generate 100 training sets of size $n = 50, 100$ and 100 training sets of size 450. Let us note that the context of this simulation is unfavourable for LS-PCR. Indeed since the variable blocks that are not active in the model are the ones that possess the strongest variability, they are going to stand out in the first κ components of the PCA.

Our proposed extensions, LS-PLS-IRLS, IR-LS-PLS, R-LS-PLS, are then applied on simulated data sets. For the sake of comparing accuracy and efficiency of the latter, GLM is implemented to clinical data alone and R-PLS to gene expression data alone. Usual method based on Principal Component Regression (PCR) is also considered. In our context, gene expression data is replaced by the first κ principal components of $\mathbf{X}$ (obtained by PCA); that constituted the directions of maximal variability in the data $\mathbf{X}$, without considering the response variable $\mathbf{Y}$. Let $\mathbf{T}$ be the matrix of columns which are the first κ PCA-scores associated with $\mathbf{X}$. The parameters are then estimated by running IRLS($\mathbf{Y}$, [$\mathbf{D}$ $\mathbf{T}$]). This approach is called LS-PCR. For all approaches, the optimal number of PLS or PCR components is selected by choosing the value of κ in the range 1, . . . , κ_{max}, with κ_{max}=1, 4 and 8, by a 5-fold cross validation on each of the 100 training sets. That is, each training set is split five fold into a test set with size equal to one fifth of the data

and a learning set size equal to four fifths of the remaining data. We retain the value of κ which minimizes the misclassification rate over these 5-fold cross validation. This is also employed for R-LS-PLS, where the κ value and λ for 6 $log_{10}-$linearly spaced points in the range $[10^{-3}; 100]$ are simultaneously determined by this cross-validation method.

As referenced in [2], although variable selection is not always necessary as a preliminary step to PLS-based classification, some authors argue that accuracy is improved in the high dimensional setting especially when there are indeed few relevant variables. Many variable selection procedures are available in the literature. In the present article, Sure Independence Screening (SIS) [7] has been performed to select relevant gene expression variables $p_{red} = 500$ such as $p_{red} < p$. SIS refers to ranking features according to marginal utility, namely, each feature is used independently as a predictor to decide its usefulness for predicting the response. Precisely SIS ranks the importance of features according to their magnitude of marginal regression coefficients.

To evaluate prediction performance, mean misclassification rates and the area under the receiver operating characteristic (ROC) curve, known as the AUC have been computed for each method. Rates of convergence are also assessed for LS-PCR and methods based on PLS algorithm. Simulations and analyses are performed using R software version 3.1.2.

The simulation results are summarized in Figure 1 and Table 1, which was produced on the basis of the 100 simulated data sets. They depict the distributions of both misclassification rate and AUC and the convergence rate in percent, respectively. For this simulation study, the two classes are much less discriminate by clinical data than gene expression data which is confirmed in Figure 1. Analysis of clinical features alone by GLM and genetic data alone with R-PLS are less informative to predict outcome than the approaches combining both type of variables. All approaches integrating clinical and genomic data, except LS-PCR, show comparable discrimination rates. The method using PCR increases the misclassification rate and decreases the AUC as κ_{max} decreases. Quite surprisingly, even with $\kappa_{max} = 4$ or 8, LS-PCR does not achieve the performance of the LS-PLS approaches. According to the model structure, we can thus expect LS-PCR to identify the two active components and so to lead to similar results. For each case of κ_{max}, R-LS-PLS seems to be better than the two other extensions of PLS (LS-PLS-IRLS and IR-LS-PLS) even if the median misclassification rates are very close. The analysis of the rate dispersion is also intuitive and follows the same trend as already described, i.e. R-LS-PLS leads to more precise results. These findings are supported by the convergence rate reported in Table 1. R-LS-PLS does not show convergence problems (all rates equal 100%). The results for LS-PLS-IRLS are roughly less good than R-LS-PLS, probably due to numerical instability of the methods when n is smaller than the number of variables. It must be noted that the interpretation of the convergence rate of IR-LS-PLS is seriously limited by the lack of optimum criterion in the approach. One explanation could be that when solving the weighted least square problem at each IRLS iteration with LS-PLS, the global problem cannot be rewritten as the optimization of a loss function.

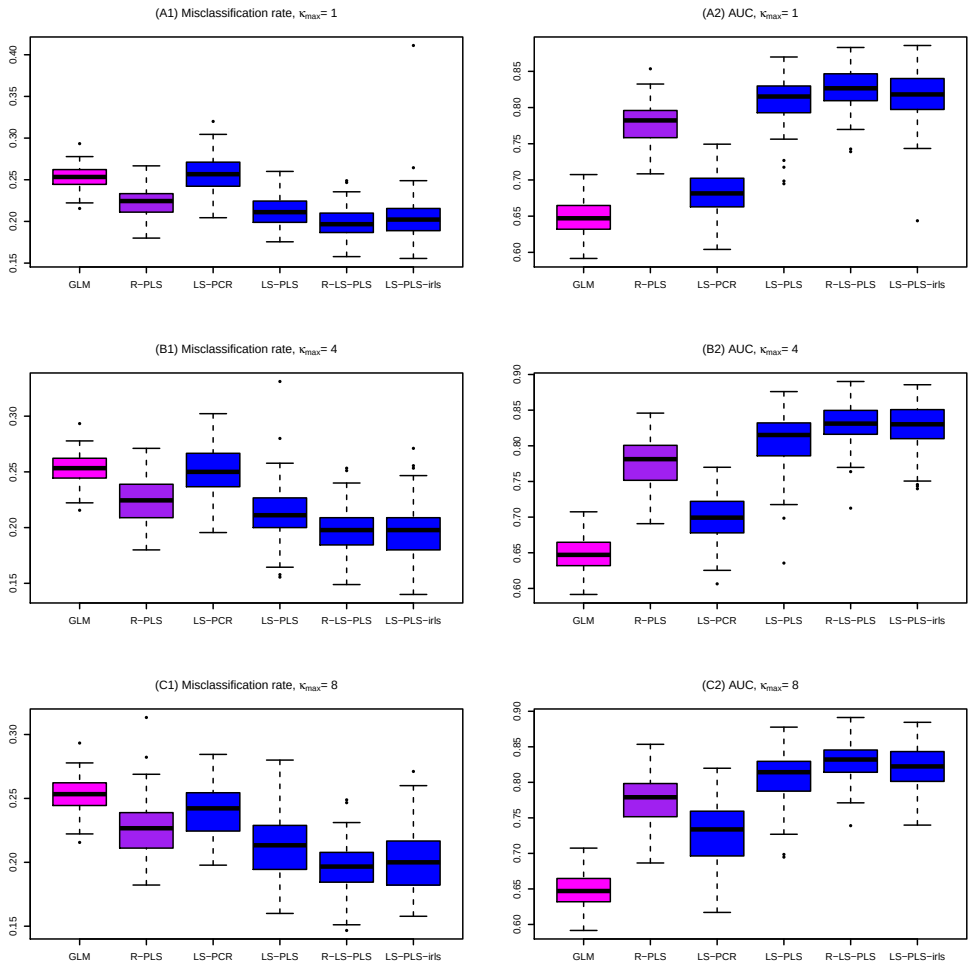

Figure 1: Boxplot of the misclassification rate (left part) and AUC (right part) from the 100 simulated data sets using the six methods, according to different κ_{max} : (A) $\kappa_{max} = 1$; (B) $\kappa_{max} = 4$; (C) $\kappa_{max} = 8$. GLM and R-PLS denote the misclassification rate and AUC obtained from GLM applied on clinical data alone and PLS to gene expression alone, respectively. LS-PCR denotes the approach derived to PCR where gene expression data are analysed using PCA and IRLS can thus applied on the merge data set of PCA-scores and clinical data. LS-PLS-IRLS, R-LS-PLS-IR-LS-PLS denote the misclassification rate and AUC obtained from the new proposed LS-PLS approaches combining expression and clinical data . For clarity of the figure, we use a code color to indicate the predictions : from clinical data alone in pink, from expression gene data alone in purple and the results from methods combining both type of variables in blue. From SIS procedure, the relevant gene expression variables p_{red} is set to 500.

Table 1: Rate of convergence (%) from the 100 simulated data sets, respectively, for the five methods, according to different $\kappa_{max} : 1, 4$ and 8. R-PLS denotes the results from the analysis of gene expression alone. LS-PCR denotes the approach derived to PCR where gene expression data are analysed using PCA and IRLS can thus applied on the merge data set of PCA-scores and clinical data. LS-PLS-IRLS, R-LS-PLS-IR-LS-PLS denote the rate of convergence from the new proposed approaches combining expression and clinical data. From SIS procedure, the relevant gene expression variables p_{red} is set to 500.

κ_{max}	R-PLS	LS-PCR	LS-PLS-IRLS	R-LS-PLS	IR-LS-PLS
1	100	100	71	100	22
4	100	100	41	100	76
8	100	99	44	100	78

4 Application to real data sets

We apply the extensions presented previously on two publicly available real data sets for which both clinical and gene expression variables are available. Similarly to the simulation study, to validate procedures of the clinico-genomic models, we compare the combined clinico-genomic model's accuracy and AUC with the ones from models built either with genomic data or clinical data alone. We apply and compare all the methods considered in the simulation study. On both real data sets, we perform a re-randomization study on 100 random subdivisions of the data set into a learning set and a test set. We choose a test set size equal to one third of the data (2:1 scheme of [6]). Pre-filtering method has been applied on gene expression data, as in the simulation study, considering different numbers of relevant genes : p_{red} =50, 100, 500 and 750. For the real data the κ range is $\{1, 2, ..., 5\}$ and λ range is given by 6 log10- linearly spaced points in the range $[10^{-3}; 100]$.

4.1 Breast Cancer data

The first original data set [24], already used in [2], contains information on 78 primary breast cancers (34 from patients who developed metastases within 5 years and 44 from patients who continue to be disease-free after a period of at least 5 years) which have been selected from patients who were lymph node negative and under 55 years of age at diagnosis. The data set gives expression of 25000 human genes and clinical variables : age (nominal), tumor grade (ordinal), oestrogen receptor status (binary), tumor size (numeric), progesterone receptor status (binary) and angioinvasion (binary). The goal here is to predict the presence of subclinical metastases in order to provide a strategy to select patients who would benefit from adjuvant therapy. concerning gene expression, the data set used in this paper has been prepared as described in the original manuscript, yielding 4348 genes.

Figure 2 shows the boxplot of the misclassification rate and the AUC for $p_{red} = 100$

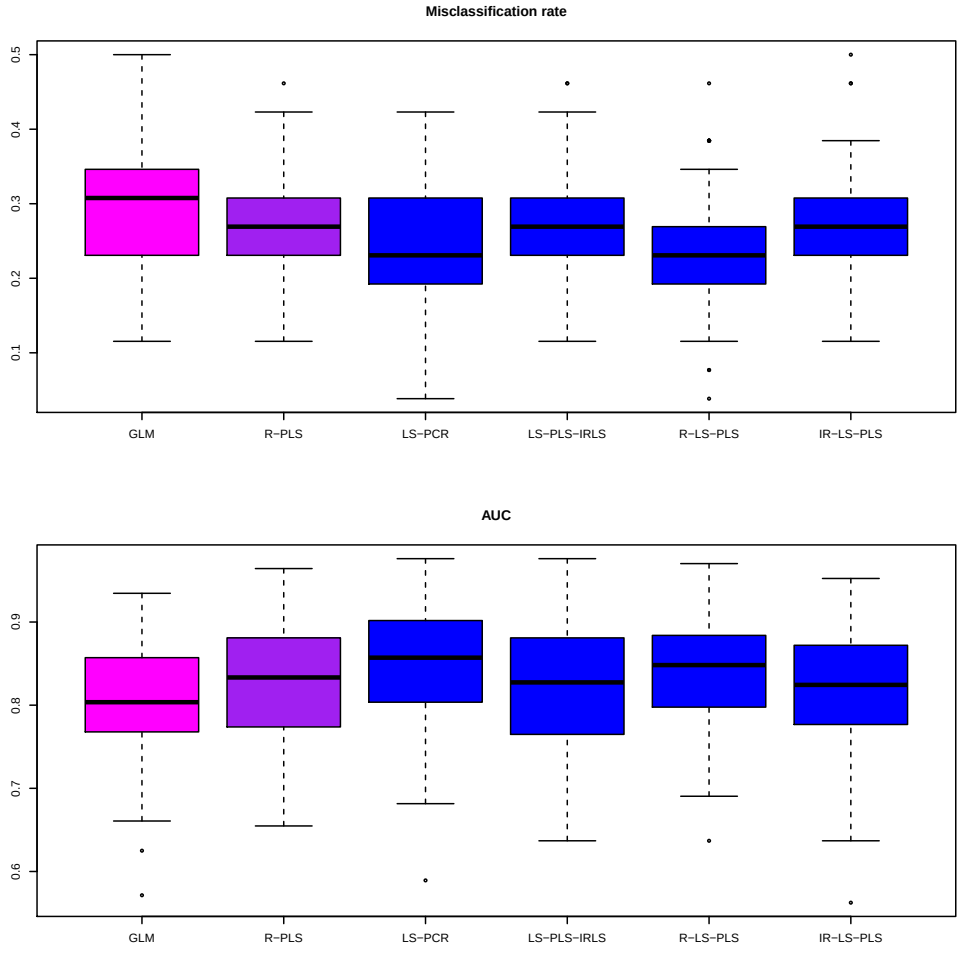

Figure 2: Distribution of misclassification rate and AUC for the Breast Cancer data estimated by 100 sampling using the six methods. GLM and R-PLS denote the misclassification rate and AUC obtained from GLM applied on clinical data alone and PLS to gene expression alone, respectively. LS-PCR denotes the approach derived to PCR where gene expression data are analysed using PCA and IRLS can thus applied on the merge data set of PCA-scores and clinical data. LS-PLS-IRLS, R-LS-PLS-IR-LS-PLS denote the misclassification rate and AUC obtained from the new proposed LS-PLS approaches combining expression and clinical data from Brest Cancer data set. From SIS procedure, the relevant gene expression variables p_{red} is set to 100. The code color for the methods is similar to Figure 1 according to the different methods.

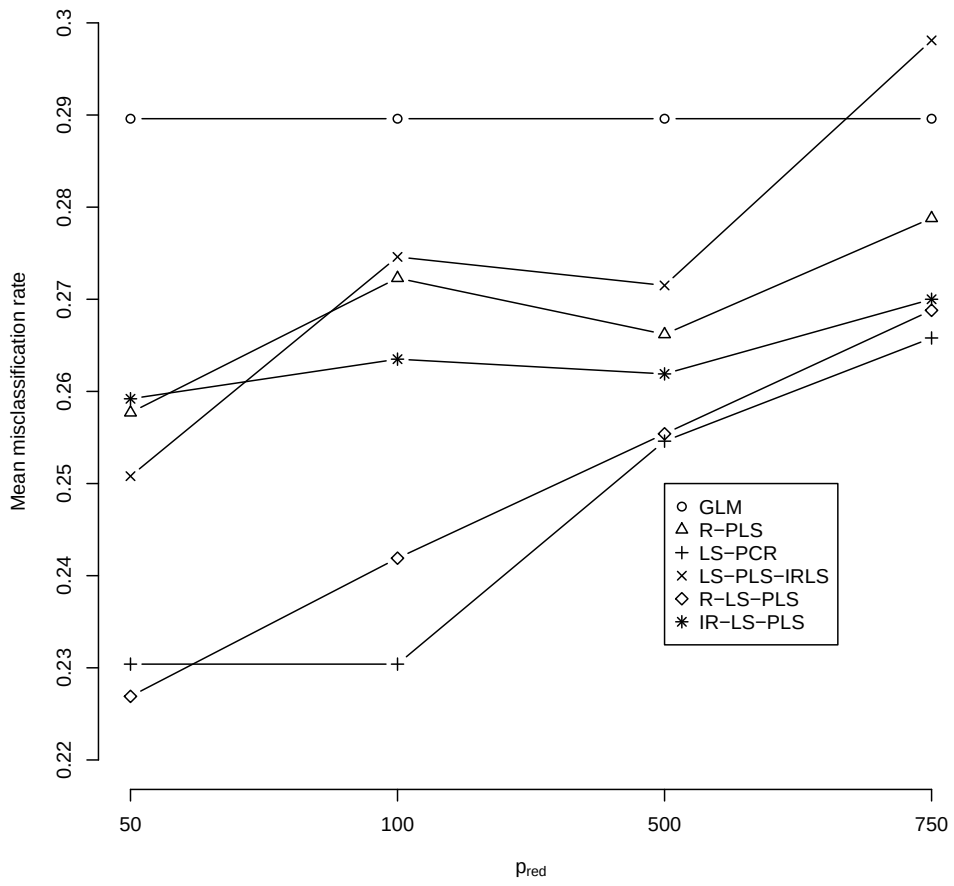

Figure 3: Evolution of mean misclassification rate from Breast Cancer data set using the six methods considering different number of relevant genes : p_{red} =50, 100, 500 and 750. GLM and R-PLS denote the misclassification rate and AUC obtained from GLM applied on clinical data alone and PLS to gene expression alone, respectively. LS-PCR denotes the approach derived to PCR where gene expression data are analysed using PCA and IRLS can thus applied on the merge data set of PCA-scores and clinical data. LS-PLS-IRLS, R-LS-PLS-IR-LS-PLS denote the misclassification rate and AUC obtained from the new proposed LS-PLS approaches combining expression and clinical data. For each method, a line is drawn to connect symbols to improve readability.

obtained from this data set. The analysis of gene expression data only slighlty improves the prediction accuracy yielded by clinical variables alone. The misclassification rate median obtained with R-PLS is smaller than that of GLM. The four methods combining clinical and genomic data provide significantly better misclassification rates as well as AUC than GLM, which are less pronounced for R-PLS using gene expression only. Note that LS-PCR and R-LS-PLS yield better prediction accuracy even if we notice the large variability in the result of LS-PCR as in the simulation study. These findings suggest that gene expression data performs slightly better than clinical data but the integration of both features seems be more interesting to predict the response. We report in Figure 3, the evolution of the mean misclassification rate according to the p_{red} most pertinent covariates from SIS procedure for all the methods. Taken together the misclassification rates increase when p_{red} grows. This can be explained by the fact that for $p_{red} > 100$ too many features are selected from genomic data, overfitting may occurs for the clinico-genomic model, giving back poorer mean misclassification. R-LS-PLS and LS-PCR stand out from all methods. They may have an equivalent behaviour except for $p_{red} = 100$ where LS-PCR shows more accurate predictions. Note that proposed combining approaches provide better results than GLM or R-PLS, except for LS-PLS-IRLS when the number of selected genes is too high.

4.2 Central Nervous System data

The second data set was obtained from [21] which has been used to predict the response of childhood malignant embryonal tumors of Central Nervous System (CNS) to therapy. The data set is composed of 60 patient samples, 21 patients died and 39 survived within 24 months described by gene and clinical data. There are 7129 genes and clinical features are sex(binary), age(nominal), chemo CX (binary) and chemo VP (binary). The original data set contains clinical variable Chang stage which has been omitted due to the high number of categories.

Figure 4 illustrates the accuracy of prediction approaches for CNS using only 500 selected genes. This data set presents a different situation from the previous one, where clinical data performed better than genomic data. R-LS-PLS attains the highest accuracy, close from the misclassification rate analysing only gene expression data with PLS (R-PLS). The accuracy from LS-PLS-IRLS and IR-LS-PLS are very similar and somewhat higher than the one from R-LS-PLS. LS-PCR is the least proficient to predict the response underlying the poor performance of PCR to treat information in high-dimensional data. Unsurprisingly, as can be seen from the mean misclassification rate according to the number of selected genes in Figure 5, the proposed procedures integrating clinical and genetic features show overall good performance which convince us that information to predict correctly the response could be concentrated in only a set of 50 genes. As provided, the performance of R-PLS increases with the size of p_{red} contrary to LS-PLS-IRLS as seen above.

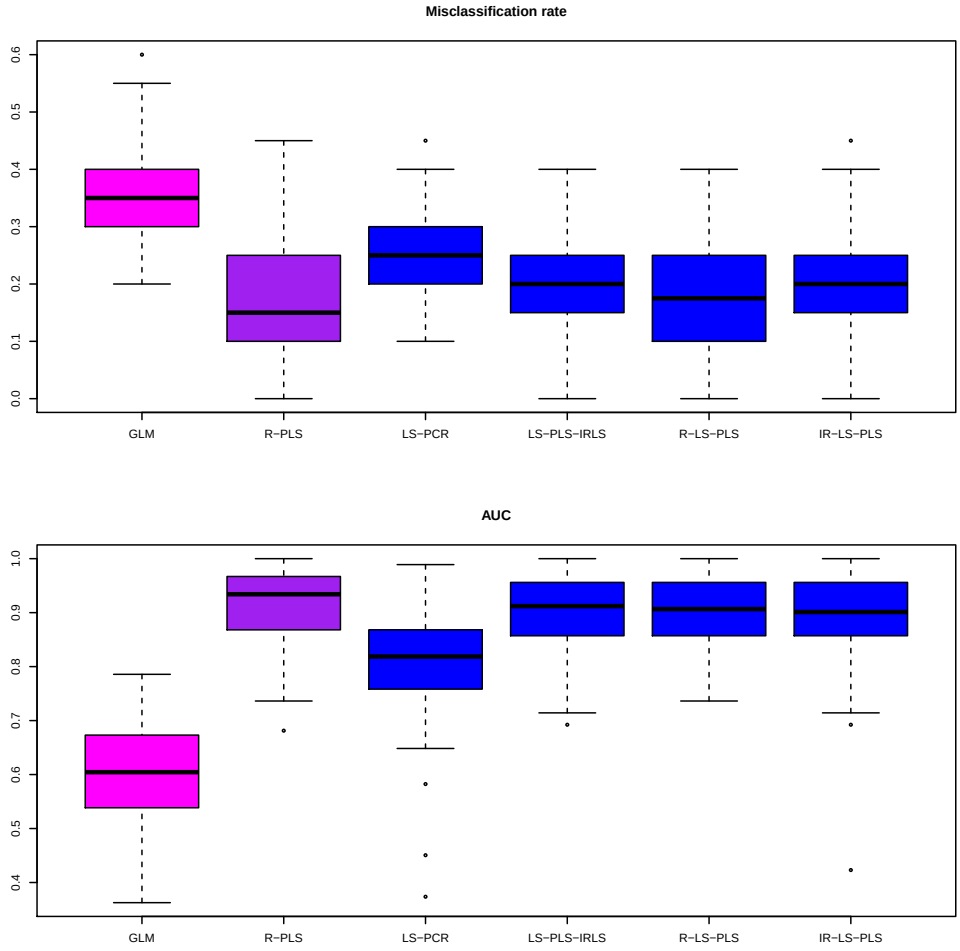

Figure 4: Distribution of misclassification rate and AUC for Central Nervous System data estimated by 100 sampling using the six methods. GLM and R-PLS denote the misclassification rate and AUC obtained from GLM applied on clinical data alone and PLS to gene expression alone, respectively. LS-PCR denotes the approach derived to PCR where gene expression data are analysed using PCA and IRLS can thus applied on the merge data set of PCA-scores and clinical data. LS-PLS-IRLS, R-LS-PLS-IR-LS-PLS denote the misclassification rate and AUC obtained from the new proposed LS-PLS approaches combining expression and clinical data from Central Nervous System data set. From SIS procedure, the relevant gene expression variables p_{red} is set to 500. The code color for the methods is similar to Figure 1 according to the different methods.

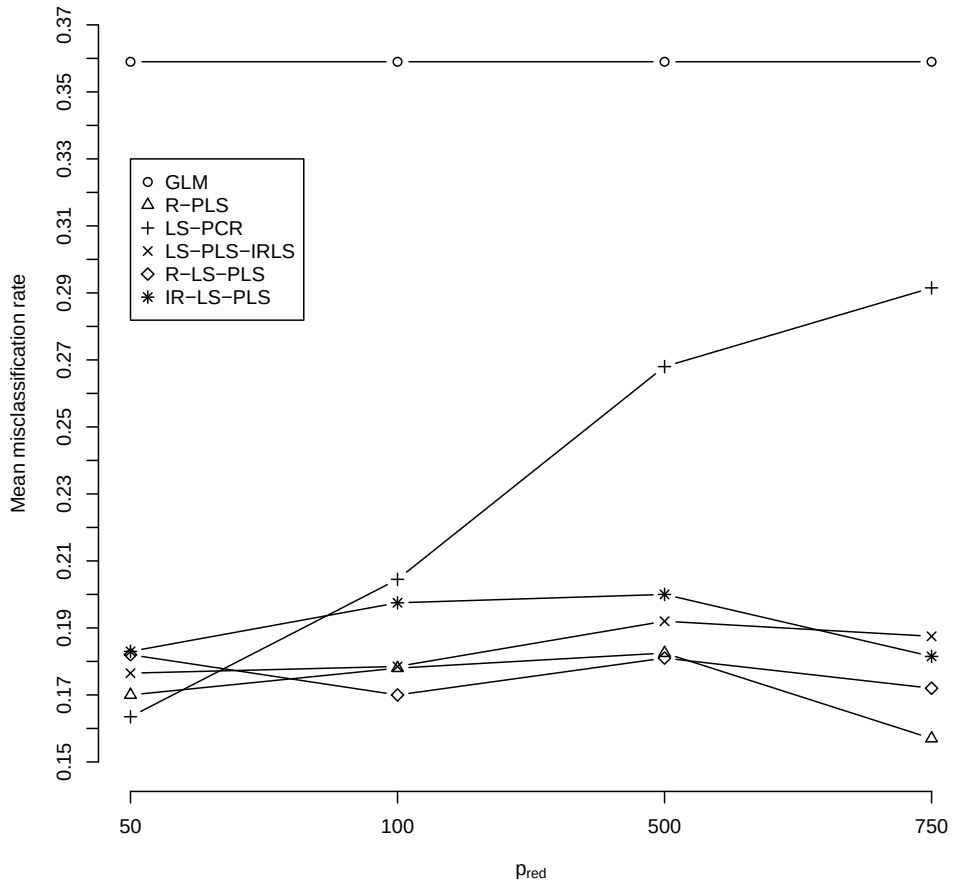

Figure 5: Evolution of mean misclassification rate from Central Nervous System data set using the six methods considering different number of relevant genes : p_{red} =50, 100, 500 and 750. GLM and R-PLS denote the misclassification rate and AUC obtained from GLM applied on clinical data alone and PLS to gene expression alone, respectively. LS-PCR denotes the approach derived to PCR where gene expression data are analysed using PCA and IRLS can thus applied on the merge data set of PCA-scores and clinical data. LS-PLS-IRLS, R-LS-PLS-IR-LS-PLS denote the misclassification rate and AUC obtained from the new proposed LS-PLS approaches combining expression and clinical data. For each method, a line is drawn to connect symbols to improve readability.

5 Conclusion

In spite of the great potential of clinico-genomic integration, the topic is still in its elaboration phase. In general integrating heterogeneous datasets like clinical and genomic data is an important issue. We have proposed three extensions of LS-PLS approaches for logistic regression models to analyse both clinical and genomic data. A comparison of the performance of prediction by simulation and on real data have been performed. In general the methods using only the clinical data or only genetic data perform less well. We observed that it is not always advisable to use the PCR-type method which can lead to wrong results depending on the data type. We believe this is because the variable response is not used in the construction of the compression of genetic data. Regarding the three proposed extension of LS-PLS, the improved performance of the extension R-LS-PLS is noted.

As you recall, that the results obtained from LS-PLS approaches presented here are different to the findings of Boulesteix et al. [2] where data are analysed using a two-step approach based on Random Forests and PLS reduction. Even if the sampling protocol for Breast Cancer data is not quite the same as this article, we can still compare the classification error rate. From this study, regarding methods based on PLS, the best rate of misclassification is 0.30 on average while the worst is 0.43. In our study, (3), the best one is 0.2269 on average while the worst is 0.2981. As a result, it would appear that the one-step approach using the two data sets simultaneously seems better than the two-step approach using the two data sets separately.

Determining the appropriate number of genomic features in the first step is hard. The number of features may impact the comparison between the additive performances of clinical and genomic variables. For example, if too many features are selected from genomic data, it may overfit the clinico-genomic model in the second phase. On the other hand, if too few genomic factors are retained, then the predictive capability of the genomic factor can be underestimated. We may conclude that model's performance was not improved by the addition of large numbers of genes but was improved by the addition of significant clinical features and genomic profiles.

This work constitutes a first step towards the extension of LS-PLS. In the present study, we consider only the case of LS-PLS for classification problems. Due to the large number of studies modelling survival using gene expression [4, 25], another natural extension of this work is to explore LS-PLS approach to survival prediction model. The outcome would be a right-censored time-to-event such as the time to death or the time to next relapse and Cox regression models have to be considered.

Lastly, the comparison of misclassification rates predicted by approaches combining both clinical and gene expression data to misclassification rates computed with only clinical or genetic data as well as to misclassification rates obtained from the simulation study, supports the appropriateness of the extensions of LS-PLS for classification using the logistic regression model. Its implementation in an R package could be a useful computing tool for integrating clinical and gene expression data to get a clinico-genomic model to predict

binary outcome.

References

[1] J. Beane, T. P. Sebastiani, K. Whitfield, Y. Steiling, M. Dumas, Lenburg, and A. Spira. A prediction model for lung cancer diagnosis that integrates genomic and clinical features. *Cancer Prevention Research*, 1(1):56, 2008.

[2] Anne-Laure Boulesteix, Christine Porzelius, and Martin Daumer. Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics*, 24(15):1698–1706, 2008.

[3] Anne-Laure Boulesteix and Willi Sauerbrei. Added predictive value of high-throughput molecular data to clinical data and its validation. *Briefings in Bioinformatics*, 12(3):215–229, 2011.

[4] Hege M Bøvelstad, Ståle Nygård, and Ørnulf Borgan. Survival prediction from clinico-genomic models-a comparative study. *BMC bioinformatics*, 10(1):1, 2009.

[5] Sanjoy Dey, Rohit Gupta, Michael Steinbach, and Vipin Kumar. Integration of clinical and genomic data: a methodological survey. Technical report, Technical Report, Department of Computer Science and Engineering University of Minnesota, 2013.

[6] Sandrine Dudoit, Jane Fridlyand, and Terence P Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97(457):77–87, 2002.

[7] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society*, 70:849–911, 2008.

[8] Gersende Fort and Sophie Lambert-Lacroix. Classification using partial least squares with penalized logistic regression. *Bioinformatics*, 21(7):1104–1111, 2005.

[9] Peter J Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 149–192, 1984.

[10] Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1:297–310, 1986.

[11] Inge S Helland. On the structure of partial least squares regression. *Communications in statistics-Simulation and Computation*, 17(2):581–607, 1988.

[12] K. Thyholt K. Jorgensen, V. Segtnan and T. Naes. A comparison of methods for analysing regression models with both spectral and designed variables. *Journal of Chemometrics*, 18:451–464, 2004.

[13] Saskia Le Cessie and Johannes C Van Houwelingen. Ridge estimators in logistic regression. *Applied statistics*, pages 191–201, 1992.

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

[15] William F Massy. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309):234–256, 1965.

[16] Bent Müller, Arndt Wilcke, Anne-Laure Boulesteix, Jens Brauer, Eberhard Passarge, Johannes Boltze, and Holger Kirsten. Improved prediction of complex diseases by common genetic markers: state of the art and further perspectives. *Human genetics*, 135(3):259–272, 2016.

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

[18] Ståle Nygård, Ørnulf Borgan, Ole Christian Lingjærde, and Hege Leite Størvold. Partial least squares cox regression for genome-wide data. *Lifetime Data Analysis*, 14(2):179–195, 2008.

[19] S. Paik, S. Shak, G. Tang, C. Kim, J. Baker, M. Baehner Cronin, F. Walker, M., D.R. Watson, T. Park, W. Hiller, E. Fisher, D. Wickerham, J. Bryant, and N. Wolmark. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New England Journal of Medicine*, 351(27):2817–2826, December 2004.

[20] P. J. Park, L. Tian, and I.S. Kohane. Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, 18:S120–S127, 2002.

[21] S. L. Pomeroy, P. Tamayo, and M. Gaasenbeek. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442, 2002.

[22] A. J. Stephenson, A. Smith, M. W. Kattan, J. Satagopan, V. E. Reuter, P. T. Scardino, and W. L. Gerald. Integration of gene expression profiling and clinical variables to predict prostate carcinoma recurrence after radical prostatectomy. *Cancer*, 2(104):290–298, 2005.

[23] M. Van de Vijver, Y. D. He, and L. Vant Veer. A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*, 347(25):1999–2009, 2002.

[24] L. J. van Veer, H. Dai, M.J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[25] W. N. Van Wieringen, D. Kun, R. Hampel, and Anne-Laure Boulesteix. Survival prediction using gene expression data: a review and comparison. *Computational statistics & data analysis*, 53(5):1590–1603, 2009.

[26] Y. Wang, J. Klijn, Y. Zhang, A. Sieuwerts, M. Look, F. Yang, D. Talantov, M. Timmermans, M. Meijer-van Gelder, J. Yu, T. Jatkoe, E. Berns, D. Atkins, and J. Foekens. Gene-expression pro-files to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet*, 365(9460, pages=671-679, month=February, keywords=,), 2005.