HAL
open science

A Comparison of Methods for Analysing Logistic Regression Models with both Clinical and Genomic Variables

Caroline Bazzoli, Sophie Lambert-Lacroix

To cite this version:

Caroline Bazzoli, Sophie Lambert-Lacroix. A Comparison of Methods for Analysing Logistic Regression Models with both Clinical and Genomic Variables. ERCIM 2014 - 7th International Conference of the ERCIM Working Group on Computational and Methodological Statistics, Dec 2014, Pise, Italy. hal-01905592

HAL Id: hal-01905592

https://hal.science/hal-01905592v1

Submitted on 25 Oct 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A COMPARISON OF METHODS FOR ANALYSING LOGISTIC REGRESSION MODELS WITH BOTH CLINICAL AND GENOMIC VARIABLES

C. Bazzoli[(1)], S. Lambert-Lacroix[(2)]

(1) LJK - Université de Grenoble BP 53, 38041 Grenoble cedex 9, France
(2)UJF-Grenoble 1 / CNRS / UPMF / TIMC-IMAG UMR 5525, Grenoble, F-38041, France

INTRODUCTION

- An active field in today medical research : prediction from high-dimensional genomic data
  - → Number of samples relatively small compared to the number of covariates
  - → Collinear measurements

⇒ Use of reduction dimension methods
  - Traditional approach : Principal Component Regression (PCR) [10]
  - Most useful : Partial Least Square (PLS) [6]
- Most of this studies included clinical data in addition to genomic data
  - However most of the proposed prediction methods used only genomic data
  - Recent studies consider situations combining both type of covariates improving predictions
    - → Development of Clinico-genomic models
    - → For survival prediction with Cox model's [3], for classification of patients [2]

  ⇒ Few of them used dimension reduction or applied dimension reduction only to the high-dimensional genomic variables
- Recently, prediction method have been proposed to combine both type of covariates [7]
  - Developed only in regression Gaussian context
    - ∗ Method based on combination of PLS and ordinary least square (OLS)=> LS-PLS

    ⇒ More information from the experiment

    ⇒ Lower variance in the parameter estimates
- Adaptation to logistic regression model
  - Not relevant to combine both type of variables without reduction dimension
  - An alternative → compress the information from genomic data into new few components before modelling
    - ∗ Partial Least Squares (PLS) adapted to logistic regression [11, 9, 4, 1]

OBJECTIVES

- **To adapt LS-PLS to combine clinical and genomic variables for logistic regression model**
- **To perform a comparison by simulation of the performance of these approaches**

LS-PLS FOR LOGISTIC REGRESSION

- **Linear Logistic Discrimination**
  - Y : binary response variable $\{0,1\}^n$
  - D : matrix of clinical variables of size $n \times q$
  - X : design matrix for expression levels of the p genes for the n microarray samples
    - ∗ $X = (x_{ij})$, $1 \le i \le n, 1 \le j \le p$
  - General statistical model
    - ∗ The conditional expectation of y_i given $D_{i\cdot}$, given by $\pi_i = P(y_i = 1|D_{i\cdot} = d_i)$
    - ∗ Related to linear predictor $\eta_i = [1; d_i]\gamma$
      - · with $\gamma \in \mathbb{R}^{q+1}$ and the non-linear relation $\pi_i = h(\eta_i)$
      - · where $h(\eta_i) = 1/(1+\exp(-\eta_i))$
  - Estimation of γ by maximum likelihood (ML) estimation method
  - Use of Iteratively Reweighted Least Squares (IRLS) algorithm [5]

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

$$\gamma^{(t+1)} \Leftarrow \textit{weighted regression by } W(\gamma^{(t)})^{(t)} \textit{ of } z^{(t)} \textit{ on } \tilde{D} \quad (2)$$

where $\tilde{D} = [1_n\ D]$, $1_n = (1, \cdots, 1)^T$
$W(\gamma^{(t)})$ the diagonal $n \times n$ matrix with entries $W_{i,i}(\gamma) = \pi_k(1-\pi_k)$.

- **LS-PLS in Gaussian context [7]**
  - Iterative procedure
    1. Perform ordinary least squares regression of Y on $(\tilde{D})$ and estimate the regression coefficients β and calculate residuals.
    2. Perform PLS regression of the residuals on centred X and combine the design $(\tilde{D})$
    3. Perform OLS regression on the combined matrix design $(\tilde{D})$ and the score T to predict Y.
    4. Calculate the predicted $\hat{y}$ based on $(\tilde{D})$ and calculate new residuals.
    5. Repeat steps 2 to 4 until convergence.

  ⇒ Method denoted LS-PLS(y, D, X, κ) where κ the PLS component number
- **Some extensions for logistic regression**
  - *NGUYEN and ROCKE's APPROACH [11]*
    - ∗ Extension of PLS to logistic regression substituting X by a $n \times \kappa$ matrix T
      - · Columns of which are the first κ PLS-scores given by PLS regression of y on X
      - · Estimation of the parameter with ML with IRLS(y, T).

    ⇒ Replace PLS by LS-PLS

    ⇒ κ obtained by cross-validation

  ⇒ Method denoted **LS-PLS**.
  - *MARX's APPROACH [9]*
    - ∗ Estimation of the parameter γ with ML using IRLS(y, T) algorithm
    - ∗ Where T defined by IRPLS, an algorithm that extends PLS to generalized linear models
    - ∗ Matrix T collects the first κ components "at convergence" of IRPLS

    ⇒ Replace PLS by LS-PLS to adopt this approach for LS-PLS

    ⇒ κ obtained by cross-validation

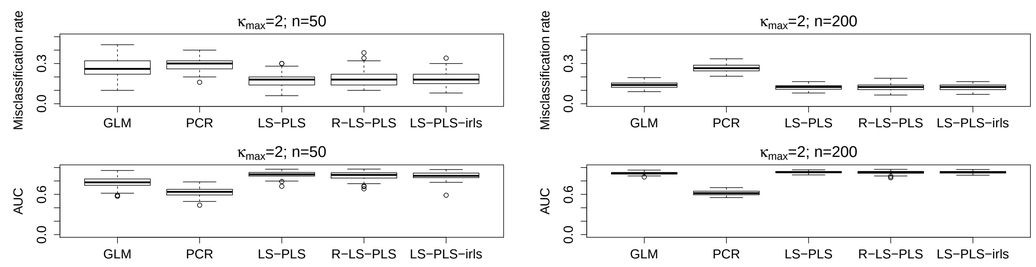
  ⇒ Method denoted **LS-PLS-irls**.
  - *RIDGE PARTIAL LEAST SQUARES APPROACH [4]*
    - ∗ Replace binary by a pseudo-response variable $z^{\inf}$ ⟵ linear relationship with the covariates
      - · At convergence of IRLS algorithm : $z^{\inf} = X\hat{\gamma}^R + \varepsilon$
      - · $\hat{\gamma}^R$: the true value of the parameter
      - · ε a centered vector of covariance matrix $(W^{\inf})^{-1}$.
    - ∗ When $n << p$ → Regularisation methods as Ridge Penalty withλ the shrinkage parameter [8]
    - ∗ Extension with LS-PLS

  ⇒ Method called **R-LS-PLS**

  ⇒ R-LS-PLS depends on two parameters : λ and κ obtained by cross-validation.
  - *BASTIEN's APPROACH [1]*
    - ∗ Not developed at the moment

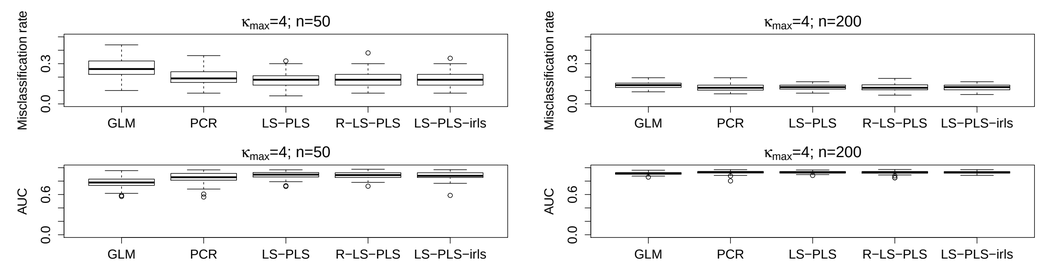
SIMULATION STUDY - A "STANDARD" EXAMPLE

- **Methods**
  - Simulation framework (R software)
    - ∗ Simulation of 100 data sets of size n (learning set=100 and test set = 100)
    - ∗ Statistical model
      - · For an individual i $(i = 1, ..., n)$, $y_i \sim \mathcal{B}(\pi_i)$
      - · $\beta_X = \{\{0\}^8, \{2\}^8\}$ and $\beta_D = \{0.5\}^4$ ($p = 16$ and $q = 4$)
      - · X such as $X = (X^1, X^2, X^3, X^4)$ where $X^k \sim N(0_4, \Sigma_X^k)$ with $\Sigma_{i,j}^k = c^k\rho^{|i-j|}$ with $k = 1, ...4$, $i = 1, ..., 4$, $j = 1, ..., 4$ where $c^1 = 8$, $c^2 = 4$, $c^3 = 2$, $c^4 = 1$ and $\rho = 0.9999$
      - · $D \sim N(0_q, \Sigma_D)$ with $\Sigma_D = \rho^{\{|i-j|}$ with $i = 1, ..., q$ and $j = 1, ..., q$ and $\rho = 0.5$
  - Comparison of the different prediction methods to analyse both X and D matrices :
    - ∗ **1) GLM 2) PCR on X then GLM on $[DX_{PCR}]$ 3) LS-PLS 4) LS-PLS-irls 5) R-LS-PLS**
    - ∗ Choice for hyperparameters :
      - · κ (for all approaches) obtained by cross-validation over a grid 1 to κ_{max} (for κ_{max} =2,4, 8)
      - · λ (for R-LS-PLS) obtained by cross-validation
    - ∗ $n = 50, 200$
    - ∗ Criterion of comparison
      - · Computation of AUC and Missclassification rate errror for each method
- **Results**



  - No optimal conditions for PCR → Poor performance especially when n increase
  - Similar comportment for GLM only when n=50

  ⇒ Relevance of method using reduction
  - Closed distribution of Missclassification rate and AUC for new approaches and GLM when $n = 200$



  - Slightly poor performance of GLM when $n = 50$
  - ⇒ Correct and closed distribution of missclassification rate and AUC when n and κ increase

  ⇒ Relevance of all the approaches as well as PCR

CONCLUSION

- Comparison by simulation of prediction methods developed for logistic regression model to analyse both clinical and genomic data based on LS-PLS approach
  - Poor performance of GLM compared to other approaches especially n small
  - Similar relevance for proposed extensions of LS-PLS and PCR with ideal number of components

  ⇒ Interest of reduction even if reasonable case of "high dimensional" case

  ⇒ Need of extensive simulations to explore high-dimensional situations
- Some applications of these approaches to real data without relevant results

  ⇒ Need further exploration
- Extension of LS-PLS approach to survival prediction model Cox'models

REFERENCES

[1] P. Bastien, V. E. Vinzi, and M. Tenenhaus. PLS generalised linear regression. *Computational Statistics & Data Analysis*, 48(1):17–46, January 2005.

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

[3] N. S. Bovelstad H.M and B. O. Survival prediction from clinico-genomic models–a comparative study. *BMC Bioinformatics*, 10(413), 2009.

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

[5] P. J. Green. Iteratively reweighted least squares for maximum likelihood estimation and some robust and resistant alternatives. *Journal of the Royal Statistical Society, B*, 46(2):149–192, 1984.

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

[7] K. T. 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.

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

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

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

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