RIDGE-PARTIAL LEAST SQUARES FOR GENERALIZED LINEAR MODELS WITH BINARY RESPONSE

Gersende Fort and Sophie Lambert-Lacroix

Address: CNRS/LMC, 51, rue des Mathématiques, BP 53, 38041 Grenoble Cedex 9,France

E-mail: Gersende.Fort,Sophie.Lambert@imag.fr

Key words: Partial Least Squares, Ridge-Penalized Logistic Regression, Classification of Microarrays

COMPSTAT 2004 section: Partial Least Squares.

Acknowledgement: We are really grateful to A. Antoniadis for constructive and fruitful discussions. This work is supported by the project ASBGEN and the IAP research network P5/24.

Abstract An extension of PLS for regression and dimension reduction in logit models is derived, an extension that still works when the number of covariates is far larger than the number of observations. It is applied to classification of Microarray.

1 Introduction

Partial Least Squares (PLS), first introduced in chemometrics [12, 9] is both used as a dimension reduction tool and as a linear regression tool. The goal of the present contribution is to extend its application to regression in univariate Generalized Linear Models (GLM) with binary response, an extension that covers the case where the length p of the covariate vector is larger or equal to the number of observations n.

PLS constructs predictive models by exhibiting latent covariates (or scores) that account for most of the variation in the response. Unlike Principal Component Regression (PCR), the definition of the scores is based both on the covariates and on the response variable $\mathbf{Y}$, and in that sense, PLS looks more appropriate than PCR to overcome the problems involved by the large number of covariates and their high collinearity. Nguyen and Rocke [10] combines PLS and Iteratively Reweighted Least Squares (IRLS, [6])

i.e. PLS and a regression analysis based on the Maximum Likelihood (ML) method; they determine the first κ PLS components from $\mathbf{Y}$ and the initial design matrix; then a regression onto these scores is performed in the ML sense. Besides the question on the pertinence of applying the PLS machinery with a categorical response vector, this algorithm has convergence weaknesses since the ML estimate does not necessarily exist. A second try for extension of PLS to GLM can be found in Marx [8] in a two step procedure. The first step is to exhibit κ PLS scores at convergence of a PLS-within-IRLS algorithm; the second one runs a ML regression onto these scores. In many applications, the Marx algorithm is nothing else than the Nguyen and Rocke algorithm and thus inherits its drawbacks, as discussed in [4].
This is the reason why we introduce an extension, called *Ridge-PLS algorithm*, that can be summarized as a weighted PLS algorithm in which the categorical response variable $\mathbf{Y}$ is replaced with a continuous-valued *pseudo-variable* that captures the information contained in $\mathbf{Y}$. Roughly speaking, *Ridge* fights the multicollinearity while *PLS* is the dimension reduction part. In this contribution, the method is derived for logit models. We show how Ridge-PLS can be used for supervised classification of Microarray data, characterized by a number of covariates far larger than the number of observations.

2 Heuristic of the Ridge-PLS algorithm

We postpone the algorithmic description to Section 4, and start with a naive description. Ridge-PLS is based on the following observation : Least Squares inference and ML inference coincide for regression in a normal linear model which is both a linear model and a GLM. For canonical GLM, the ML estimate $\hat{\theta}^{\mathrm{ML}}$ is the weighted least squares estimate when regressing a pseudo-response variable ψ onto the columns of the design matrix; ψ is obtained at convergence of an IRLS procedure, and for normal models, is equal to $\mathbf{Y}$ [6]. As a consequence, our extension of PLS consists in applying PLS by replacing $\mathbf{Y}$ with the pseudo-variable at convergence of IRLS. Nevertheless, this rough idea

has to be made robust in order (i) to be valid when $\hat{\theta}^{\mathrm{ML}}$ does not exist and (ii) to take into account the heteroscedasticity of the pseudo-variable. This is done by respectively (i) substituting $\hat{\theta}^{\mathrm{ML}}$ for a penalized ML estimator, namely the Ridge one, and (ii) introducing a Weighted PLS (WPLS) algorithm. Before deriving Ridge-PLS for logit models, we introduce notations and basic algorithmic ingredients.

3 Some basic ingredients

For a column-vector u, $\|u\|$ is the Euclidean norm, $u_{1:p}$ collects the first p components of u. For a matrix A, A' is the transpose, A_{ij} denotes the entry (i,j), and $A_{.,1:r}$ the matrix that contains the first r columns of A. $\mathbb{1}_n$ is the vector $(1,\cdots,1)'$ of length n and $J^{(r)}$ is a diagonal $(r+1)\times(r+1)$-matrix with $J^{(r)}_{11}=0$ and $J^{(r)}_{kk}=1$ otherwise.

Logit model and Logistic discrimination rule The observations consist of n independent $\{0,1\}\times\mathbb{R}^p$-valued pairs $(\mathbf{y}_i, X_{i.})$ where given $X_{i.}$, the conditional mean of $\mathbf{y}_i$ is π_i, which is related to the linear predictor η_i by $\pi_i = (1+\exp(-\eta_i))^{-1}$, or equivalently $\eta_i = \ln(\pi_i/(1-\pi_i))$. η_i depends on the design vector $Z_{i.} := [1\ X_{i.}']'$ through the relation $\eta_i = Z_{i.}'\theta$, where $\theta\in\mathbb{R}^{p+1}$ is the unknown parameter. The n response variables (resp. conditional means) are collected in the vector $\mathbf{Y}$ (resp. Π). The $n\times(p+1)$ design matrix is denoted by $Z=[\mathbb{1}_n\ X]$.
For a given estimate $\hat{\theta}$, and a new design vector z, the binary variable $\hat{\mathbf{y}}$ is predicted by applying the logistic discrimination rule, *i.e.* $\hat{\mathbf{y}}=1$ if $\hat{\eta} := z'\hat{\theta}\geq 0$, and $\hat{\mathbf{y}}=0$ otherwise.

The Ridge-ML estimator When $n>\mathrm{rank}(Z)$, $\hat{\theta}^{\mathrm{ML}}$ is unique when it exists. Unfortunately, the likelihood may be maximal on the boundary of $\mathbb{R}^{p+1}$ so that $\|\hat{\theta}^{\mathrm{ML}}\|=+\infty$ [11]. When $n=\mathrm{rank}(Z)$ - which occurs if and only if $n\leq(p+1)$ and Z has full rank - the solution to the normal equation yields $\|\hat{\theta}^{\mathrm{ML}}\|=+\infty$. Hence, inference of the parameter necessitates the introduction of a regularization method; we opt for a *Ridge*-penalized ML ap-

proach, which shrinks the coefficients towards zero (except the intercept one θ_1). The Ridge estimator $\hat{\theta}^{\mathrm{R}}$ is defined as the maximum of the penalized log-likelihood l^*

$$l^*(\theta) = \sum_{k=1}^{n} \{\mathbf{y}_k Z'_{k.}\theta - \ln(1+\exp(Z'_{k.}\theta))\} - \frac{\lambda}{2}\theta'\Sigma^2\theta, \qquad (1)$$

where $\lambda > 0$ is a *shrinkage* parameter, and Σ is a diagonal matrix taking into account the non-standardization of the covariate matrix : $\Sigma^2_{11} = 0$ and $\Sigma^2_{kk} = \sum_{j=1}^{n}(Z_{j,k} - \mathbb{1}'_n Z_{.,k}/n)^2$ for $k \in [2, p+1]$. $\hat{\theta}^{\mathrm{R}}$ exists, is unique and is computed by the (iterative) Newton-Raphson algorithm, each iteration of which is a weighted Ridge-regression of a pseudo-variable onto the columns of Z.

WPLS algorithm For a given $\mathbb{R}^n$-valued observation ψ, a covariate matrix X, and a positive-definite symmetric weight matrix W, the PLS scope is to convey the relation between ψ and X through the definition of κ scores $(t_j)_{1 \leq j \leq \kappa}$. These are linear combinations of the columns of the design matrix Z such that for all j, $\mathbb{1}'_n W t_j = 0$ and for all $j \neq k$, $t'_j W t_k = 0$. This yields the decomposition $\psi = q_0 \mathbb{1}_n + q_1 t_1 + \cdots + q_\kappa t_\kappa + f_{\kappa+1}$ where $f_{\kappa+1}$ is W-orthogonal to the vectors $(\mathbb{1}_n, t_1, \cdots, t_\kappa)$. The pairs (q_j, t_j) are recursively computed as follows

1. $t_0 = \mathbb{1}_n$; $E_0 = X$; $f_0 = \psi$.

2. For $j = 0, \cdots, \kappa$, set $q_j = t'_j W f_j / (t'_j W t_j)$, $f_{j+1} = f_j - q_j t_j$, $E_{j+1} = E_j - t_j t'_j W E_j / (t'_j W t_j)$, $t_{j+1} = E_{j+1} E'_{j+1} W f_{j+1}$.

We refer to the literature for an interpretation of the above algorithm and a discussion on the maximal number of W-orthogonal scores $\kappa_{\max}$ [7]. WPLS, read as a regression method, yields a PLS estimate $\hat{\theta}^{\mathrm{PLS},\kappa}$ through the relation $\hat{\psi}_\kappa = \psi - f_{\kappa+1} = Z\hat{\theta}^{\mathrm{PLS},\kappa}$.

4 The Ridge-PLS algorithm, $n \leq p+1$

Given $(\mathbf{Y}, X)$, for the parameters (λ, κ),

A. Determine ψ : compute $\hat{\theta}^{\mathrm{R}}$, the limiting value of $(\theta^{(t)})_t$ where

$$\theta^{(t+1)} := \left(Z'W^{(t)}Z + \lambda\Sigma^2\right)^{-1} Z'W^{(t)}\psi(\theta^{(t)}), \qquad (2)$$

$$\psi(\theta^{(t)}) := Z\theta^{(t)} + \left[W^{(t)}\right]^{-1}\left(\mathbf{Y} - \Pi^{(t)}\right), \qquad (3)$$

$Z := [\mathbb{1}_n \; X]$, $\Pi^{(t)}$ is the mean vector Π computed at the current value of the parameter and $W^{(t)}$ is a diagonal matrix with $W^{(t)}_{kk} := \Pi^{(t)}_k(1-\Pi^{(t)}_k)$. Set $\psi := \psi(\hat{\theta}^{\mathrm{R}})$ and $W := W^{(\infty)}$.

B. Run the WPLS with κ components for the variables (ψ, X, W) and compute $\hat{\theta}^{\mathrm{PLS},\kappa}$ as described in Section 3.

Step A builds a continuous response variable ψ whose expected value has linear relationship with the covariates, for the input of PLS; conditionally to $\hat{\theta}^R$, the dispersion matrix of ψ is W^{-1}, which explains the call, in Step B, to a weighted PLS procedure with weight W.

Implementation The procedure, presently derived in $\mathbb{R}^{p+1}$ can be equivalently derived in $\mathbb{R}^{r+1}$ where $r+1 := \mathrm{rank}(Z) \leq n$. To that goal, compute UDV', the singular values decomposition (svd) of $(X - \mathbb{1}_n\mathbb{1}_n'X/n)\Sigma^{-1}$, the standardized covariate matrix, and set $\Xi := (UD)_{\cdot,1:r}$ so that $Z\theta = [\mathbb{1}_n \; \Xi]\gamma$ for some $\gamma \in \mathbb{R}^{r+1}$; it is readily seen that the above procedure, run by replacing (X, Σ^2) by $(\Xi, J^{(r)})$, yields an estimate $\hat{\gamma}^{\mathrm{PLS},\kappa}$ uniquely related to $\hat{\theta}^{\mathrm{PLS},\kappa}$ by the formulas

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

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

In the application, λ is chosen as the value λ_{opt} in a given range $\mathcal{R}$ minimizing the BIC criterion $-2\hat{l} + \log(n)\mathrm{Dim}$ where $\hat{l}$ is the log-likelihood for the value $\hat{\theta}^R$ of the parameter, and Dim is the trace of $Z\left(Z'WZ + \lambda\Sigma^2\right)^{-1}Z'W$.

5 Application to binary classification

We apply the above procedure to supervised classification of Microarray data; the data set *Leukemia*[1], contains 72 samples divided into 47 cases of acute lymphoblastic leukemia, labeled 0, and 25 cases of acute myeloid leukemia, labeled 1. Each sample consists in a $\{0,1\}$-valued label and 7129 gene expression levels (see Golub *et al.* [5] for a description of the data set). We perform an out of sample (OS) analysis on 100 random partitions of the data set into a learning set and a test set. The learning set contains 27 samples type 0 and 11 samples type 1. We report in Table 1, row "*RPLS* κ" the mean number (and the standard deviation) of misclassified samples in the test set, when the classification rule is determined on the learning set $[\kappa = 1, \cdots, 6]$. Regression is not performed with the 7129 initial covariates; some of them are irrelevant and are deleted following the pre-processing method described in Dudoit *et al.* [2]. We stress that this filtering and the number of remaining genes depend on the learning set. We test the procedures by considering different values of p ($> n = 38$) and select the p most pertinent covariates as advocated in Dudoit *et al.* [2]. We run the OS analysis for the classification rule induced by the Ridge estimator $\hat{\theta}^{\mathrm{R}}$ (row "*Ridge*", Table 1); the results outline the interest of a dimension reduction step after the regularization one. Eilers *et al.* [3] propose a method quite similar to the Ridge analysis. They compute $\hat{\theta}$ as maximizing the criterion (1) in which Σ is replaced by $J^{(p)}$ (although Z is not standardized); then, their classification rule is based on a Bayes risk : $\hat{\mathbf{y}} = 1$ iff $\hat{\pi}$ is greater than the empirical mean of the observations in the learning set. We run their algorithm and report the results in row "*Eilers*", Table 1. Ridge-PLS yields better results; nevertheless, this assertion has to be nuanced since for less "regular" data sets, Ridge-PLS and the Eilers *et al.* 's method may have an equivalent behavior.

For each partition, λ_{opt} is determined as described above, over 51 $\log_{10}$-linearly spaced points in $\mathcal{R} = [10^{-2}, 10^{3}]$. The mean value of λ_{opt} over the 100 partitions is given in Table 2 for the Eilers

[1] available at http://www.broad.mit.edu/cgi-bin/cancer/publications

et al. 's algorithm (λ_{E}) and the Ridge and Ridge-PLS algorithms (λ_{R}). Whatever p, $\lambda_{\mathrm{E}} > \lambda_{\mathrm{R}}$, which is due to the standardization of the design Z.

method	p=50	p=100	p=300	p=500	p=1000
Ridge	1.52 (1.11)	1.35 (1.09)	1.62 (1.05)	1.89 (1.21)	2.83 (1.37)
RPLS 1	1.24 (0.93)	1.18 (0.98)	1.12 (0.86)	1.20 (0.97)	1.45 (1.07)
RPLS 2	1.36 (0.98)	1.24 (0.91)	1.15 (0.93)	1.08 (0.79)	1.27 (0.96)
RPLS 3	1.43 (1.01)	1.32 (0.91)	1.10 (0.77)	1.06 (0.79)	1.14 (0.82)
RPLS 4	1.40 (0.94)	1.34 (0.93)	1.09 (0.79)	1.12 (0.85)	1.39 (0.94)
RPLS 5	1.40 (0.95)	1.33 (0.96)	1.08 (0.80)	1.12 (0.77)	1.21 (0.74)
RPLS 6	1.43 (0.97)	1.27 (0.89)	1.12 (0.79)	1.13 (0.79)	1.25 (0.77)
Eilers	1.44 (1.00)	1.52 (1.00)	1.48 (0.94)	1.42 (0.90)	1.45 (0.95)

Table 1: Mean number of misclassified samples (standard deviation between parentheses).

	p=50	p=100	p=300	p=500	p=1000
λ_{E}	12.16	26.06	78.60	131.20	269.60
λ_{R}	0.38	0.94	3.76	7.30	18.42

Table 2: Mean value of λ_{opt}.

6 Conclusion

We derived an extension of PLS to GLM for logit models. The numerical results show the pertinence of the combination of a regularization step and a dimension reduction step. The technique can be easily adapted to other GLM models such as the multivariate ones, and this will be done in a forthcoming paper. Future research will concern the choice of the regularization method (based for example on the Firth's penalty, as proposed in [1], private communication), and the variable selection and the model selection themes in order to determine optimal values for (λ, κ).

References

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

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

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

[4] G. Fort and S. Lambert-Lacroix. Classification using Partial Least Squares with Penalized Logistic Regression. Technical report, IAP Network, TR 0331, 2003.

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

[6] 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.

[7] I. Helland. Partial Least Squares Regression and Statistical Models. *Scand. J. Stat.*, 17(2):97–114, 1990.

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

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

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

[11] 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.

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