Clusterwise Sparse PLS

Stéphanie Bougeard[a], Ndeye Niang-Keita[b], Cristian Preda[c] and Gilbert Saporta[b]*

[a]ANSES, Ploufragan, France
[b]CEDRIC-CNAM, France
[c]Université de Lille 1, Lille, France

Keywords: PLS regression, sparse models, clusterwise methods

Introduction

PLS regression is a successful method when predictors are correlated as it provides stable regression coefficients. It has often been claimed that one of its advantages (like PCR or ridge regression) is that the model uses all predictors even in the case where the number of observations is smaller than the number of predictors: $p > n$. However in the case of high dimensional data where p is much larger than n, this advantage becomes a drawback due to the lack of interpretability of linear combinations of thousands of variables, as it is met in genomics. Sparse PLS (sPLS) and group sparse PLS have been proposed to overcome this problem by using Lasso type constraints on the parameters. On another hand, when one has a large number of observations, it is frequent that unobserved heterogeneity occurs, which means that there is no single model, but several local models: one for each latent class. Clusterwise methods optimize simultaneously the partition and the local models; they have been already extended to PLS regression. The originality of this paper is to present a combination of clusterwise PLS and sPLS which is well fitted for big data : large n , large p.

1 Sparse PLS

Sparse PLS has been developed to deal with a high-dimensional set of predictors, through the use of L_1 penalty in a similar way to sparse PCA. The two main techniques are described in [1] and [2]. Recently [3] introduce a second penalty for the block of responses, a case useful for omics data. Their criterion is:

$$\min\{\|X'Y - u'v\|^2 + \lambda_1|u| + \lambda_2|v|\}, \text{ subject to } \|u\| = \|v\| = 1. \tag{1}$$

where $|u| = \sum_{j=1}^{p} |u_j|$ is the L_1 norm of u .

Equation (1) is equivalent to equation (2) :

$$\max\{cov(Yv, Xu) + \lambda_1|u| + \lambda_2|v|\}, \text{ subject to } \|u\| = \|v\| = 1. \tag{2}$$

PLS components of higher order $Y_h v_h$ and $X_h u_h$ are obtained by deflating Y and X matrices.

2 Clusterwise PLS

2.1 Estimation

Clusterwise linear regression has been extended to local PLS regression models in [4] with a proof of convergence of the following algorithm, especially when the number of variables is infinite:

[*] Corresponding author. 292, rue Saint Martin, F-75003 Paris, E-mail: gilbert.saporta@cnam.fr

1. start from a partition into G clusters
2. estimate G PLS models, one for each cluster
3. For each observation, compute the G predictions of the response and move (if necessary) the observation into the cluster giving the minima prediction error
4. after a complete pass over all observations, go to step 2 until convergence

It must be noted that the number of PLS components was chosen beforehand and was the same for all clusters. A ”more stochastic” variant of the previous algorithm would consist in re-estimating the local models each time an observation is moved to another cluster: actually one needs only to re-estimate 2 models: one for the departure cluster, the other for the arrival cluster.

2.2 Prediction

Given a new data point for which we know only the values of the predictors, the prediction of y is done in 2 steps: first assign the observation to the nearest cluster by using some supervised classification algorithm and then apply the relevant model.

3 Clusterwise Sparse PLS

It consists in a modification of clusterwise PLS where the G local models are sPLS instead of PLS regressions. Here, we will deal with a single response Y, hence $\lambda_2 = 0$. For $g \in \{1, \dots G\}$, the regularization parameters λ_{1g} are optimized by cross validation. The allocation of a new observation to clusters is done by a supervised classification using either all the variables selected by the G sparse models, or if they are too numerous, the components of the PLS global model like in PLS-T [5], followed by the application of the relevant sPLS model. The prediction could also be a weighted average of the G predictions where the weights are the estimated probabilities that the new data point belongs to each cluster.

4 Application

We will present an application of the previous methodology to a real data set and compare to results from the literature.

References

[1] K. A. L. Cao, D. Rossow, C. Robert-Granié, and P. Besse, “Sparse pls: variable selection when integrating omics data,” *Stat. Appl. Mol. Biol* **7**, p. article 35, 2008.

[2] H. Chun and S. Keles, “Sparse partial least squares regression for simultaneous dimension reduction and variable selection.,” *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **72**, pp. 3–25, 2010.

[3] B. Liquet, P. L. de Micheaux, B. P. Hejblum, and R. Thiébaut, “Group and sparse group partial least square approaches applied in genomics context.,” *Bioinformatics* **32**, pp. 35–42, 2016.

[4] G. Saporta and C. Preda, “Clusterwise pls regression on a stochastic process.,” *Computational Statistics and Data Analysis* **49**, pp. 35–42, 2005.

[5] V. E. Vinzi, C. Lauro, and S. Amato, *New Developments in Classification and Data Analysis*, Springer, 134-140, 2005.