Joint-Lasso applied to sparse group Partial Least Square and application to pleiotropy

Camilo Broc [1*] & Thérèse Truong [2**] & Borja Calvo [3***] & Benoit Liquet [4*]

[1] *camilo.broc@univ-pau.fr* [2] *therese.truong@inserm.fr* [3] *borja.calvo@ehu.es* [4] *benoit.liquet@univ-pau.fr*

[*] *Laboratory of Mathematics and its Applications (LMAP), University of Pau & Pays de L'Adour*

[**] *CESP, INSERM U1018, Villejuif.*

[***] *Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, Donostia, 20018, Spain*

Résumé.

L'élaboration de données de grande dimension peut être menée en rassemblant des données provenant de différents jeu de données indépendants. Mais que se passe-t-il si l'on s'intéresse à l'effet global d'un prédicteur dans le cas où le type de variable ou la direction de l'effet dépend du jeu de données? Parmi les modèles parcimonieux, le "Joint Lasso" permet de construire un modèle spécifique à chaque jeu de données tout en liant les modèles par une pénalité Lasso. Cela permet de traiter les différents jeux de données de manière indépendante tout en ayant une sélection globale de prédicteurs. La Régression des moindres carrés partiels (PLS) est une méthode populaire dans l'étude des données Omics. Une de ses extensions parcimonieuses est la "sparse group Partial Least Square". Une application de l'idée du "Joint Lasso" à la sgPLS est proposée, ce qui permet d'ouvrir de nouvelles perspectives en pleiotropie, où une variable Omics peut avoir un effet sur plusieurs variables, et ce, même si dans différents jeux de données, la nature du phénotype ou la direction des effets varie d'un jeu de données à l'autre.

Mots-clés. Données de grande dimension, Épidémiologie génétique, Méthodes parcimonieuses, Pénalisation Lasso, Pléiotropy, Régression des moindres carrés partiels.

Abstract.

The edification of high dimensional data can be achieved by the gathering of different independent data. But what happens when we want to infer global properties about a predictor when different types of dependent variables or opposite effects exist among the different data sets? In sparse models, Joint Lasso allows to build one model specific to each data set while linking the models through a penalty. This allows to handle data sets different effects but with a unique overall selection of predictors. The Partial Least Square (PLS) is a popular dimension method in Omics data analysis. The sparse group Partial Least Square (sgPLS) is one of its sparse formulation where a priori grouping of variables are known. An application of the joint Lasso idea to the sgPLS is proposed. It

leads to novel perspectives in peliotropy where one Omic predictor is supposed to have an effect on several phenotype dependent variables. An interpretation of the overall data can be given on Omic features even if different types of dependent variables or opposite effects exist among the different data sets.

Keywords. Genetic epidemiology, High dimensional data, Lasso Penalization, Partial Least Square, Pleiotropy, Sparse methods.

1 Context

Since past years data analysis applied to high dimension in all domains has arisen. Extracting information from ever larger data has become a trend in numerous fields and a large number of observation need to be gathered in order to evaluate statistical models. When data are hard to retrieve, gathering existing data sets is an efficient way for assembling data of high dimension. However this technique have its drawbacks : existing independent data sets can present intrinsic bias which can decrease the performance of the models used.

Those biases imply an unwanted underlying structure that will interfere with the signal we want to find. Bias can come from a difference in the source of information or the process used during the recollection of the data. This set structure has to be taken into account in order to improve the efficiency of the models. For instance, in genomics, data can be gathered from different studies because of the cost of the experimentation. Each clinical study may have been performed with its own chemistry protocol, with its own experimental material and on its specific populations, and bias can arise among the different data sets obtained. This "batch effect" is known and can significantly decrease the power of the analysis as the article of Gagnon-Bartsch and Terence (2012) shows. Another bias can occur in particular analysis where different "dynamics" exist between the studies : a predictor can be highly correlated with independent variable, but the direction of the correlation depends on the study.

We tackle the problem of "batch effect" for the sparse group Partial Least Square (sgPLS) which is developped by Liquet et al. (2015) and which is an extension of the dimension reduction such as Partial Least Square (PLS) method introduced by Wold (1975). The sparse PLS (sPLS) , introduced by the article from Kim-han Le Cao (2008), adds a Lasso penalization to the PLS which shrink to zero the participation to the model of the least relevant variables. Results highlight a smaller number of variable that are easier to explain. In addition, noise of the signal is reduced and the power of the methods is boosted. The sgPLS uses problem-specific prior information in order to improve the accuracy of the prediction and the interpretability of the model : the variables are supposed to be gathered into groups of variables. Incorporation of this grouping structure is becoming increasingly common due to the success of gene set enrichment analysis approaches, like in Subramanian et al. (2005), and using a model taking into account this variable group

structure allow to improve the performance and the readability of the results.

In this article we consider data that are composed of independent observation sets. The observation sets are assumed to be known and are expected to introduce bias in the data. The presented methods allow us to use the information about the edification of the data set in order to improve the performance of the analysis. Although this theory have been developed with the aim to answer a problem occurring in genomic public data sets, it can be applied to any field where a certain observation set structure exists. Different method using Lasso penalization on data structured toward observation sets are discussed. In particular a "penalized PLS for structured data" is defined where separate PLS model are linked together with a common-Lasso penalization. In the end variables selected by the model are the same for all observation sets but the underlying model computes separated models for each observation set, giving both readability and flexibility to the model. We present the theoretical background for this method. Especially, we can show that the common-Lasso constraint that is used (i.e. a penalization across studies) can be be written as a standard Lasso with an overlaid group structure in an equivalent formulation of the PLS problems. We extend also this idea of common-Lasso constraint to a case where an a priori structure is known, where the variables are gathered into groups.

2 Notations

Data are represented by $X \in \mathbb{R}^{p\times n}$ and $Y \in \mathbb{R}^{qn}$, two matrices, representing n observations of p predictors and q independent variables. Then X is a (n, p) matrix and Y a (n, q) matrix. For any matrix A of size (a, b) , for $i \in \{1, \dots, a\}$ its rows are noted $A^{(i,\cdot)}$ and for $j \in \{1, \dots, b\}$ its columns are noted $A^{(\cdot,j)}$ and for subsets $\tilde{a} \subset \{1, \dots, a\}$ and $\tilde{b} \subset \{1, \dots, b\}$ resp. row and column sub-matrices are noted $A^{(\tilde{a},\cdot)}$ and $A^{(\cdot,\tilde{b})}$. For any vector ω of size a , for $i \in \{1, \dots, a\}$ its elements are noted $\omega^{(i)}$ and for subsets $\tilde{a} \subset \{1, \dots, a\}$ $\omega^{(\tilde{a})}$ represents the elements of the vector corresponding to the positions in the subset.

Let us consider M different sets in the data. Noting, for $m \in \mathbb{N}$, $\mathbb{M}_m$ a subset of $\{1, \dots, n\}$, let $\mathbb{M} = (\mathbb{M}_m)_{m=1..M}$ be a partition of $\{1, ..., n\}$ corresponding to the observation sets. We note $\#\mathbb{M}_m = n_m$. Let $\mathbb{P} = (\mathbb{P}_k)_{k=1..K}$ be a partition of $\{1, ..., p\}$ corresponding to this variable group structure. We note $\#\mathbb{P}_k = p_k$. We then have $\sum_{k=1}^{K} p_k = p$.

Let us consider that the variables are gather in K groups and observations are gathered in M studies. The partitions $\mathbb{M}$ and $\mathbb{P}$ can define resp. row blocks and column blocks of both the matrices X and Y as shown in 1.

The Frobenius norm on matrices is noted $\| \ \|_F$. We note X^T the transpose matrix of X. The cardinal of a set S is noted $\#S$. The positive value of a real number x is noted $(x)_+ = \frac{|x|+x}{2}$.

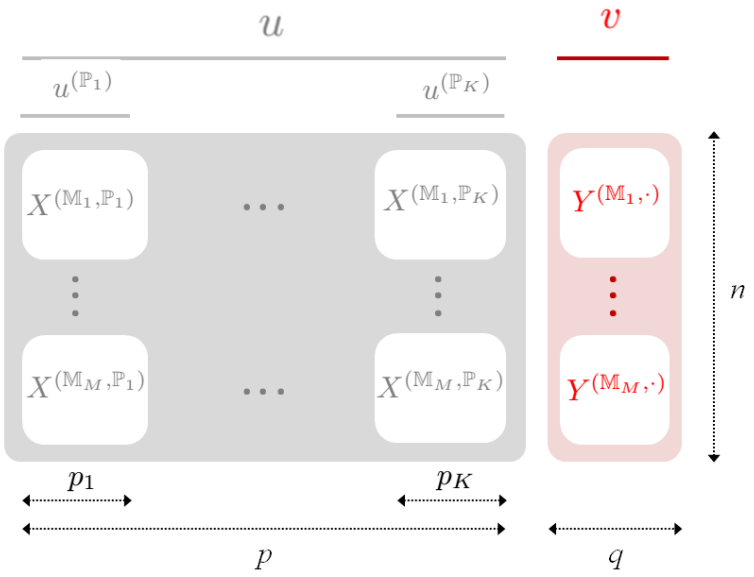

Figure 1: Illustration of data structured by group of variables and group observation. Variables are assumed to be ordered by variable group.

3 sparse Partial Least Square : Partial Least Square with Lasso penalization

In the literature, two formulations of the Partial Least Square exist, some extensions of the PLS follow a first one usually called PLS1 (an article from Wang (2009) is an example) and other extensions follow a second one called "PLS2" (an article from Chun and Kele (2010) is an example). We study here exclusively the first one.

Let X be a predictor matrix of size (n, p) and Y a matrix of independent variables of size (n, q). PLS finds successively couples of vector $\{u_1, v_1\}, \ldots, \{u_r, v_r\}$ for $r < \min(p, q)$ where the couples are composed of vectors of length resp. p and q, maximizing $Cov(Xu_i, Yv_i)$ for any $i \in \{1, \ldots, r\}$, under the constraint that the family of vectors $u_1, \ldots, u_r$ and $v_1, \ldots, \ldots, v_r$ are both of them orthogonal families (see Wold (1975)). It can be solved considering successive minimization problems (see Shen (2008)), for $h \in \{1, \ldots, r\}$

$$Cov(X_{h-1}u_h, Y_{h-1}v_h) \text{ for any } h \in \{1, \ldots, r\},$$

where $X_0 = X$, $Y_0 = Y$ and X_{h-1}, Y_{h-1} are deflated matrices computed from u_{h-1},v_{h-1}, X_{h-2}, Y_{h-2} for $h \in \{2, \ldots, n\}$. The deflation depends on the PLS mode that is chosen (see for instance the article from Esposito Vinzi et al. (2010) or the article from Wold (1975)). In this article we focus on the enhancement of the optimization problem and its Lasso formulation in its h-th step. According to the article from Shen (2008) this step

can be written as

$$\{u_{opt}, v_{opt}\} = \underset{||u||_2=||v||_2=1}{\operatorname{argmin}} \left\|X^T Y - uv^T\right\|_F^2 + \underbrace{\lambda P(u)}_{\substack{\text{Lasso Penalty term} \\ \text{for sparse PLS}}} . \quad (1)$$

where the notation h is removed in order to simplify the formulation because we are interested in only one of the r steps of the PLS.

The sparse PLS introduces a penalization in this formulation of the problem. The penalty $P(\cdot)$ forces lowest values of u to be set to zero. The parameter controlling the degree of sparsity in the model is λ. In the presented formula the sparsity is applied only to the vector u, but a similar penalization can be define for v. In the context of this article we treat only the penalization of u but all the results stand also for a v penalization. The following sections compare different ways of writing the sPLS optimization problem presented in Equation (1) taking into account an observation or/and variable set structure.

Remark: Before analysis, the X and Y matrices are transformed by subtracting their column averages. Scaling each column by their mean and standard deviation is also often recommended as ot is shown in the article from Geladi and Kowalski(1986).

4 Joint Lasso

When variables can be gathered in groups (Figure 1), the classical sgPLS propose to add a group-Lasso penalization to the classical PLS. Data are standardized within each observation set, i.e. for every $m \in \{1, \ldots, M\}$, $X^{(\mathbb{M}_m,\cdot)}$ and $Y^{(\mathbb{M}_m,\cdot)}$ are standardized instead of X and Y. The formulation of the problem is

$$\begin{gathered}\{u_{opt}, v_{opt}\} = \underset{u,v}{\operatorname{argmin}} \left\|Z - uv^T\right\|_F^2 + \lambda\left(1-\alpha\right) P_{group}(u) + \lambda\alpha P_{variable}(u) \\ \text{with } P_{group}(u) = \sum_{k=1}^{K} \sqrt{p_k} \left\|u^{(\mathbb{P}_k)}\right\|_2 \ , \ P_{variable}(u) = \sum_{i=1}^{p} \left\|u^{(i)}\right\|_2 \\ \text{and } Z = X^T Y.\end{gathered} \quad (2)$$

In the model the loading vectors u and v is composed of resp. p and q elements. The penalization $P_{variable}$ forces single variables to be set to zero whereas the penalization P_{group} forces sets of variables to be set to zero. The degree of sparsity in general in the model is λ whereas the parameter controlling the balance between both kind of sparsity is α. In this model elements of u corresponding to least relevant variables and least relevant group of variables are set to zero.

The idea of the joint Lasso relies on minimizing M sgPLS problems linking the Lasso penalty term of the different studies.

Data are standardized within each observation set, i.e. for every $m \in \{1, \dots, M\}$, $X^{(\mathbb{M}_m,\cdot)}$ and $Y^{(\mathbb{M}_m,\cdot)}$ are standardized instead of X and Y. The formulation of the " sgPLS for structured data" is then :

$$\{U_{opt}, V_{opt}\} = \underset{U,V}{\operatorname{argmin}} \left\| Z_m - U^{(\cdot,m)} V^{(\cdot,m)^T} \right\|_F^2 + \lambda (1-\alpha) P_{group}(U) + \lambda \alpha P_{variable}(U)$$
$$\text{with } P_{group}(U) = \sum_{k=1}^{K} \sqrt{p_k} \left\| U^{(\mathbb{P}_k,\cdot)} \right\|_F \ , \ P_{variable}(U) = \sum_{i=1}^{p} \left\| U^{(i,\cdot)} \right\|_2 \tag{3}$$
$$\text{and } Z = X^T Y.$$

In the model the set of loading U is composed of $p \times m$ elements (p elements per $U^{(\cdot,m)}$). The set of loading V is composed of $q \times m$ elements (q elements per $V^{(\cdot,m)}$). In this model elements of U corresponding to least relevant variables and least relevant group of variables are set to zero. In this model the same variables and variable groups corresponding to least significant variables are set to zero for all $U^{(\cdot,m)}$, $m \in \{1, \dots, M\}$.

The solution is given by the following theorem :

Theorem 1. *The marginal optima in $\tilde{u}$ and $\tilde{v}$ in the sgPLS (Equation (1)) are : Fixing v, the optimal u_{opt} for " sgPLS for structured data" is*

$$u^{(\mathbb{P}_k)} = u_1^{(\mathbb{P}_k)} \left(1 - \frac{\lambda (1-\alpha)}{2\sqrt{\sum_{i \in \mathbb{P}_k} \left\| u_1^{(i)} \right\|_2^2}} \right)_+ = u_1^{(\mathbb{P}_k)} \left(1 - \frac{\lambda (1-\alpha)}{2 \left\| u_1^{(\mathbb{P}_k)} \right\|_F} \right)_+$$
$$\textit{With } u_1^{(i)} = u_0^{(i)} \left(1 - \frac{\lambda \alpha}{2 \left\| u_0^{(i)} \right\|_2} \right)_+ \ , \ u_0 = Zv \tag{4}$$
$$\textit{and } Z = XY^T.$$

5 Field of applications

The method can be useful on any data presenting a group of variables and a set of observations structure. It takes into account cases where : (i) Observations can present biases due to the methods of experimentation (ii) The dependent variables can be different (iii) The effect of a feature can be significant in different sets but of difference size and hence can be unnoticed. The " sgPLS for structured data" can scope with all of those aspect.

Pleiotropy (Paaby et al. (2013)) is a field of genetics where a genomic feature can have an effect on several phenotype traits. In an application to cancer data sets, if we search

for a genomic feature related to different type of cancers, all of the preceding problems can arise : (i) Set of observation can be obtained with different experimental protocols, with different instruments of measures, with different populations and a bias can be introduced (ii) The dependent variables may not be comparable (different cancers) (iii) A genomic feature can have a positive effect on one type of cancer and a negative one on another cancer, both effect can undermine each other in the overall analysis.

Bibliographie

Chun, H. and Kele, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection, *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 72(1) , pp. 325.

Esposito Vinzi, V. ; Trinchera, L. and Amato, S. (2010). Pls path modeling from foundations to recent developments and open issues for model assessment and improvement, *In Handbook of partial least squares*, Springer, pp. 4782.

Gagnon-Bartsch, A. J. and Terence, P. S. (2012). Using control genes to correct for unwanted variation in microarray data, *Biostatistics*, 13(3), pp. 539552.

Geladi, P. and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial, *Analytica chimica acta*, 185, pp. 117.

L Cao, K. ; Rossouw, D. ; Robert-Grani, C. and Besse P. (2008). A sparse pls for variable selection when integrating omics data, *Statistical applications in genetics and molecular biology*, 7(1), pp. 35.

Liquet, B ; Lafaye de Micheaux, P. ; Hejblum, B. P. and Thibaut R. (2015). Group and sparse group partial least square approaches applied in genomics context, *Bioinformatics*, 32(1) , pp. 3542.

Paaby, A. B. and Rockman, M. V. (2013). The many faces of pleiotropy, *Trends in Genetics*, 29(2), pp. 6673.

Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation, *Journal of multivariate analysis*, 99(6) , pp. 10151034.

Subramanian, A. et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proceedings of the National Academy of Sciences*, 102(43), pp. 1554515550.

Wang, T. ; Ho, G. ; Ye, K. ; Strickler, H. and Elston, R. C. (2009). A partial least-square approach for modeling gene-gene and gene-environment interactions, when multiple markers are genotyped. *Genetic epidemiology*, 33(1) , pp. 615.

Wold, H. (1975). Path models with latent variables: The nipals approach *Quantitative sociology*, Elsevier, pp. 307357.