Sparse group PLS on data gathered from different studies : application to experimental bias correction and pleiotropy analysis

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

[1] camilo.broc@univ-pau.fr
[2] benoit.liquet@univ-pau.fr
[3] therese.truong@inserm.fr
[*]Laboratory of Mathematics and its Applications (LMAP), University of Pau & Pays de L'Adour [**]CESP, INSERM U1018, Villejuif.

Résumé. L'épidémiologie génétique a pour but de comprendre le rôle des gènes dans l'apparition de maladies. La mise en commun de différentes études cliniques donne accès un ensemble de données plus vaste rendant les résultats obtenus plus robustes. Cependant les méthodes multivariées telles la sparse group PLS (Partial Least Square), utilisée dans ce domaine en particulier pour étudier les structures gênes/pathway, bénéficient d'un cadre théorique peu développé lorsqu'il s'agit de mise en commun d'études. De nouveaux modèles sont développés dans cet article. Ils peuvent être mis en application pour la correction de biais d'expérimention ("batch effect") et pour la pléiotropie. Une formalisation du problème ainsi que de nouveaux résultats théoriques pour ces deux problèmes sont présentés dans cet article.

Mots-clés. Données de grande dimension, Épidémiologie génétique, Méthode sparse, Pléiotropy, Régression des moindres carrés partiels.

Abstract. Genetic epidemiology aims at understanding the role of genetic factors in the occurrence of disease. Combining studies provides larger data sets and more reliability on the result. However multivariates methods like the sparse group PLS (Partial Least Square), that is used for the identification of gene/pathway structures in genetic epidemiology, lacks of an established theory around multi-studies problems. New models can be used for batch effect correction and for pleiotropy. A formalisation of the problem is proposed and new theoretical results for these two problems.

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

1 Context

Genetic epidemiology aims at understanding the role of genetic factors in the occurrence of disease. The issue consists in knowing which genetic factors are associated with the appearance of diseases. For this purpose it is necessary to compare the genotype of people with the phenotype traits in question. A large number of clinical studies are performed. In each study, the accuracy and the reproducibility of the results relies for most on the sample size of the study. This is the reason why gathering different studies is a powerful way of improving the relevance of the result. The combining of independent but related study and pleiotropy (Yang et al. (2015)) are two approach that can be considered. However for some multvariates methods used in genetic epidemiology, such as the sparse group PLS (Liquet et al. 2016) (a method that can handle gene/pathway structures), present no established theory for a combining study approach. Problems like batch effects and pleiotropy can be handle with new models. The first problem consists in combining data from independent studies that are related to the same disease. Each studies have an intrinsic bias due to different population and different experimental protocols. This phenomenom is called batch effect (Gagnon-Bartsch and Speed (2012)). Conditioning the data per study can solve this problem (Rohart et al (2016)). The second problem, pleiotropy, consists in comparing the effect of genotypes on different phenotypes involved in the same biological processes in order to improve our gene analysis. Pleiotropy enriches our insights into the mechanisms involved in the appearance of the phenotypes. Gene set analysis with pleiotropy is expected to improve the chances of revealing the underlying genetic architecture of complex phenotypes. Highlighting pleiotropy provides opportunities for understanding the shared genetic underpinnings among associated diseases. The approach aims at developing novel statistical methods that will identify shared genes between different tumor types. In general for Genom Wide Association study (GWAS) and pathway study, multivariate methods like Partial Least Square, sparse PLS and sparse group PLS can be used for analysis. But no established theory exists for multi-study analysis. This article sets a theoretical framework for both problems : Batch effects correction and pleiotropy analysis.

2 Sparse group PLS and batch effect correction

The data available provides us the gene expressions and the phenotype of a population. Mathematically they are represented by X and Y, two data matrices, containing n observations of p predictors and q variables. X is a (n, p) matrix and Y a (n, q) matrix. The p predictors are the gene expression and the q variables are the diagnostic of a disease. A pathway structure is considered forming group of genes. Let $\mathbb{P} = (\mathbb{P}_k)_{k=1..K}$ be a partition of $\{1, ...p\}$. Noting $\#$ the cardinal of a set, we note $\#\mathbb{P}_k = p_k$. We then have $\sum_{k=1}^{K} \#\mathbb{P}_k = p$. We can then suppose that the columns of X are groups by pathway, in other worlds:

$$\{1,...,p\} = \{1,...max(\mathbb{P}_1), min(\mathbb{P}_2), ...max(\mathbb{P}_{K-1}), min(\mathbb{P}_K), ..., max(\mathbb{P}_K) = p\}.$$

With this notation we have K pathways. The idea of the PLS is to relate both matrices by maximazing the covariance between latent scores (noted Xu and Yv) defined as linear combinations of the original variables in X and Y. We have to find u and v, weight vector, of size respectively p and q. u can be partitioned according the pathway structure as sub-vectors of size p_k and X can be splited the same way among its columns in sub-matrices $X^{(k)}$ of size (n, p_k). As we want to know the relevance between the genes we can assume we are searching for normed vectors. Let's now consider that we want to take into account the fact that the data X and Y comes from M different studies. X and Y can be decomposed as M sub-matrices by rows X_m and Y_m $(m = 1, \ldots, M)$. Both decompositions (by study and by pathway) can be considered at the same time with $X_m^{(k)}$ and $Y_m^{(k)}$. Figure 1 summarizes the matrix notations.

The sparse group PLS minimization problem is :

$$\begin{gathered} \min_{u,v} \left\| X^T Y - uv^T \right\|_F^2 + (1-\alpha) P_\lambda^{(1)}(u) + \alpha P_\lambda^{(2)}(u) \\ \text{with } P_\lambda^{(1)}(u) = \lambda \sum_{i=1}^{p} |u_i| \;\text{ and }\; P_\lambda^{(2)}(u) = \lambda \sum_{k=1}^{K} \sqrt{p_k} \left\| u^{(k)} \right\|_2, \end{gathered} \tag{1}$$

where λ and α are parameters of the penalization and the Frobenius norm on matrices is denoted $\|\|_F$.
$P_\lambda^{(1)}$ drives the gene selection, whereas $P_\lambda^{(2)}$ drives the pathway selection. α drives the priority between these two selections. We can see it as a biconvex optimization problem. Fixing $\|v\|_2 = 1$ the optimal $u_{(k)}$ is

$$\begin{gathered} u^{(k)} = \frac{1}{2}\left[1 - \frac{\lambda(1-\alpha)\sqrt{p_k}}{2\left\| g_{soft}(Z^{(k,.)}v, \lambda\alpha/2) \right\|_2}\right]_+ g_{soft}(Z^{(k,.)}v, \lambda\alpha/2) \\ \text{with } \; Z^{(k,.)} = X^{(k)T} Y, \end{gathered} \tag{2}$$

and fixing $\|u\|_2 = 1$ the optimal v is

$$\begin{gathered} v = Z^T u \\ \text{with } \; Z = X^T Y, \end{gathered} \tag{3}$$

where g_{soft} is a thresholding function:

$$\begin{gathered} g_{soft} : \mathbb{R}^r \times \mathbb{R}^+ \to \mathbb{R}^r \\ (u, \lambda) \mapsto g_{soft}(u, \lambda) = (sign(u_i)(|u_i| - \lambda)_+)_{i \in \{1,\ldots,r\}} \end{gathered} \tag{4}$$

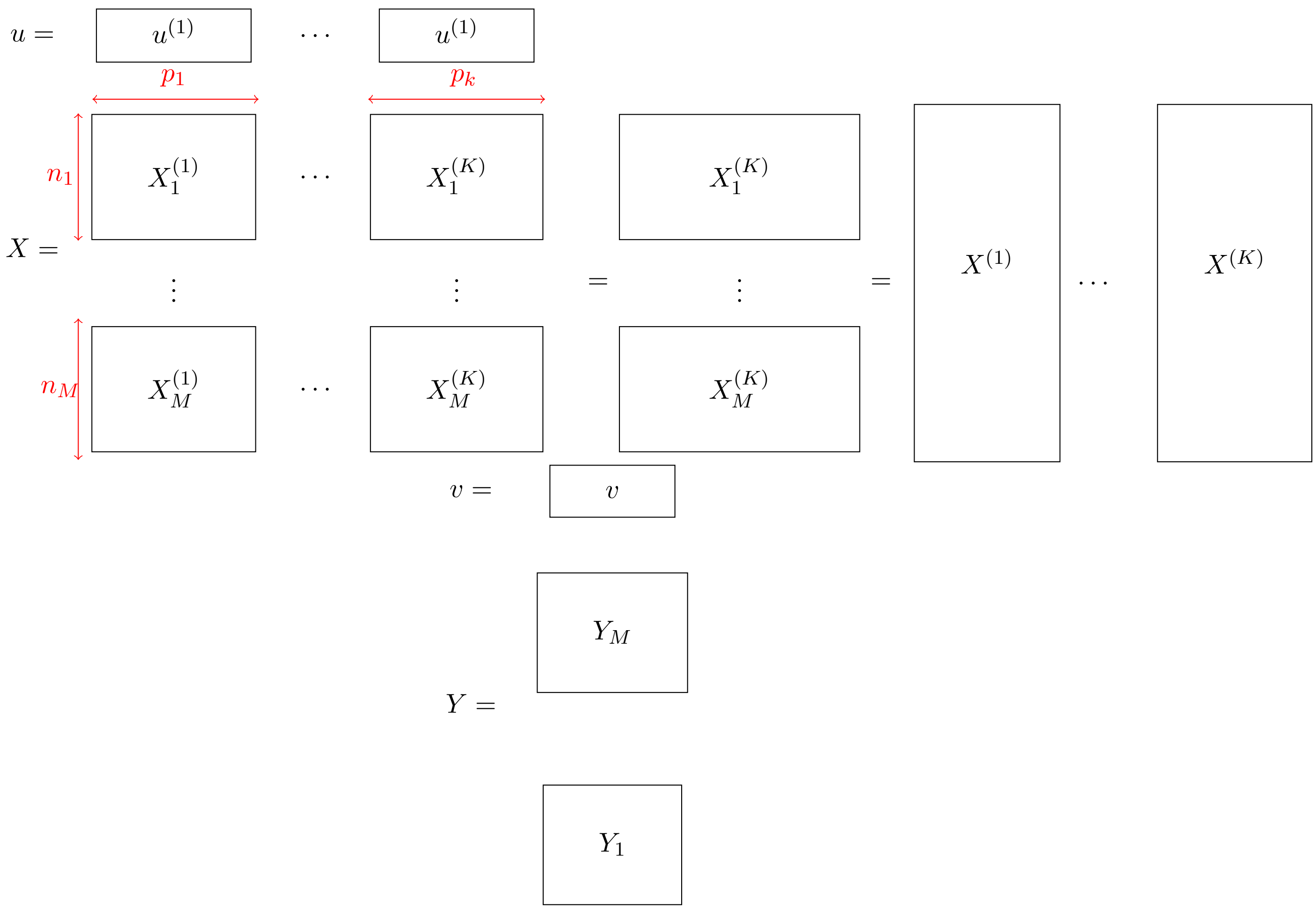

Figure 1: Representation of the data for sparse group PLS and PLS problem with study structure

In PLS algorithms the matrix X is deflated following u and the procedure is reiterated. This weight vectors give the importance of each gene in each phenotype. The sparse group PLS model enables a gene and pathway selection in the analysis by adding an L_1 and L_2 penalization in the model. At each step of the biconvex optimization resolution we see that the lowest values of the weight vector u will be set to zero. This is how the selection is performed.

Usually in order to get rid of bias, the data are normalized : before the application of the PLS algorithm each column of X and Y is centered according to its mean and scaled according to its standard deviation. But in multi study data sets the accuracy of the results is impacted by the different experimental bias between each each study. In multi study models the normalization have to be changed to take this diversity into account. The normalization is performed on the matrices X_m and Y_m. Another benefit of the decomposition by study of the data is that examining the covariances $cov(X_m u, Y_m v)$ gives us the participation to the covariance of each study.

3 Sparse group PLS for pleiotropy

The splitting of the raw data by study can be used for pleiotropy. In this case, we have several studies that provide information on different phenotypes. The aim is to select genes and pathways involved in biological mechanisms related to all the observed phenotypes. For simplicity, we present in this document only the selection of genes but the pathway selection procedure for pleiotropy has also been developed. A relevant gene can be a gene involved in all phenotypes. However, the gene/phenotype correlation for one gene can be different from one phenotype to another. For instance a gene that is positively correlated for one phenotype and negatively correlated to another phenotype would be highly interesting for our purpose, but a rough analysis without taking into account the study structure can miss that part of the information. The model proposed consists in solving M sparse group PLS optimization models but linking them in the penalization procedure.

For each study m, u_m and v_m optimal weight vectors are computed to maximize the covariance $cov(X_m u_m, Y_m v_m)$. Let's gather in a matrix those weight vectors. Weight vectors are gathered in U a (p, M) matrix and V a (q, M) matrix which columns are respectively the vectors u_m and v_m.

Let's note $U^{(1)}, ..., U^{(p)}$ and $V^{(1)}, ..., V^{(q)}$ the rows of those matrices. The matrices U and V are explained in figure 2. The rows corresponds to all the weights related to a same gene. Then, we want to set to zero all the weights related to a same gene at the same time. This is why a L_2 penalization is introduced on the rows $\left(U^{(i,.)}\right)$ of U.

$$\begin{gathered}\min_{u_1,...u_M,v_1,...,v_M} \sum_{m=1}^{M} \left\| X_m^T Y_m - u_m v_m^T \right\|_F^2 + P_\lambda(u) \\ \text{with } P_\lambda(u) = \lambda \sum_{i=1}^{p} \left\| U^{(i,.)} \right\|_2\end{gathered} \tag{5}$$

The biconvex resolution use the following formulas. Fixing each $\|v_m\|_2 = 1$ the optimal $u_m^{(k)}$ is

$$\begin{gathered}\left(u_m^{(k)}\right)_i = \left(1 - \frac{\lambda}{2\sqrt{\sum_{\ell=1}^{M}\left((Z_\ell v_\ell)_i\right)^2}}\right)_+ \left(Z_m^{(k,.)} v_m\right)_i \\ \text{with } \ (.)_i \text{ is the i-th term of the vector}\end{gathered} \tag{6}$$

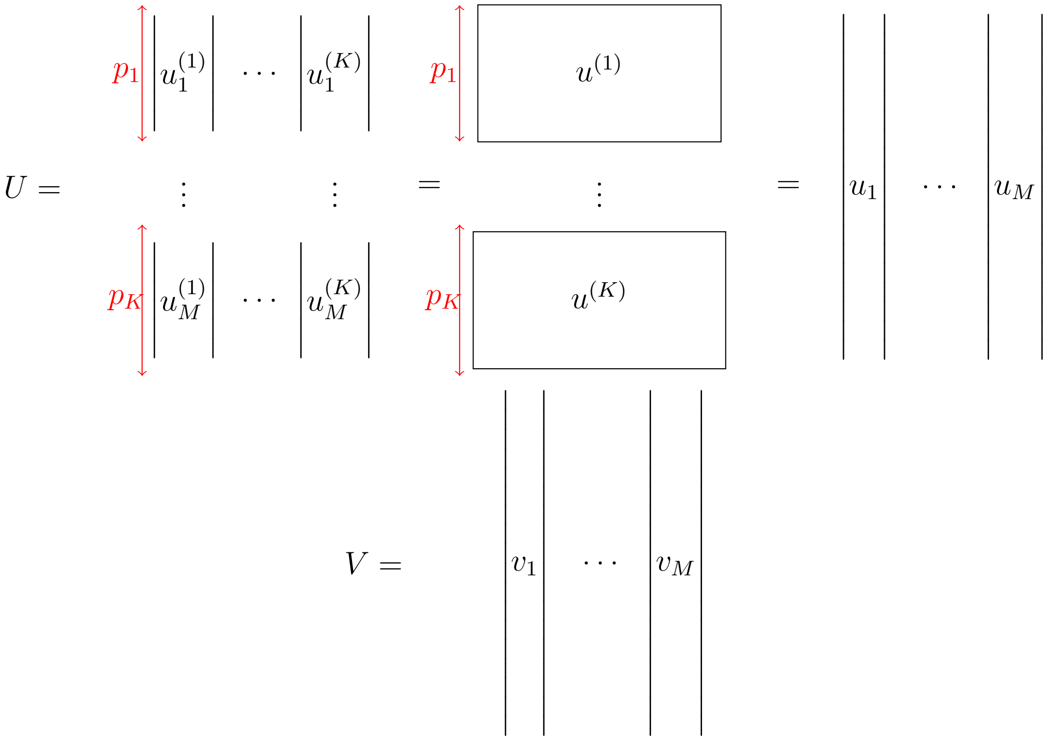

Figure 2: Study structure of weight vectors for pleiotropy analysis purpose.

and fixing $\|u\|_2 = 1$ the optimal v is

$$\begin{aligned} v &= Z^T u \\ \text{with } \; Z &= X^T Y. \end{aligned} \tag{7}$$

We can see that u_m implies another thresholding function. The thresholding function sets to zero all the weights of a same gene at the same time. In the end a gene selection is performed across all the study.

4 Conclusion

The combination of independent data sets enriches our insights about the genetic mechanisms in the appearance of disease by increasing the amount of data available for performing our models. Maintaining the coherence within data is an issue in combining related studies or in pleiotropy. For related-study analysis the efficiency of the Sparse group PLS model can by improved with a batch effect correction. For pleiotropy a new Sparse group PLS problem provides us a gene and pathway selection tool taking into account the study structure of the problem.

References

[1] Gagnon-Bartsch, J. A. and Speed, T. P. (2012). Using control genes to correct for unwanted variation in microarray data. *Biostatistics, 13(3)*, 539-552.
[2] Rohart, F., Eslami, A., Matigian, N., Bougeard, S. and Le Cao, K. A. (2016). MINT: A multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *bioRxiv*, 070813.
[3] Liquet, B., de Micheaux, P. L., Hejblum, B. P. and Thiébaut, R. (2016). Group and sparse group partial least square approaches applied in genomics context. *Bioinformatics*, 32(1), 35-42.
[3] Yang, C., Li, C., Wang, Q., Chung, D. and Zhao, H. (2014). Implications of pleiotropy: challenges and opportunities for mining Big Data in biomedicine. *Frontiers in genetics*, 6, 229-229.