Arab. J. Math. (2020) 9:329–344
https://doi.org/10.1007/s40065-019-0248-6

Arabian Journal of Mathematics

Camilo Broc · **Borja Calvo** · **Benoit Liquet**

Penalized Partial Least Square applied to structured data

Received: 8 June 2018 / Accepted: 14 February 2019 / Published online: 4 March 2019

Abstract Nowadays, data analysis applied to high dimension has arisen. The edification of high-dimensional data can be achieved by the gathering of different independent data. However, each independent set can introduce its own bias. We can cope with this bias introducing the observation set structure into our model. The goal of this article is to build theoretical background for the dimension reduction method sparse Partial Least Square (sPLS) in the context of data presenting such an observation set structure. The innovation consists in building different sPLS models and linking them through a common-Lasso penalization. This theory could be applied to any field, where observation present this kind of structure and, therefore, improve the sPLS in domains, where it is competitive. Furthermore, it can be extended to the particular case, where variables can be gathered in given a priori groups, where sPLS is defined as a sparse group Partial Least Square.

Mathematics Subject Classification 62H99 · 62J07

1 Introduction

Since past years data analysis applied to high dimension in all domains has arisen [21]. Extracting information from ever larger data has become a trend in numerous fields and a large number of observation need to be gathered 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 has 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 that 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 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

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s40065-019-0248-6) contains supplementary material, which is available to authorized users.

C. Broc (✉) · B. Liquet
Laboratoire De Mathématiques et de leurs Applications de PAU Fédération MIRA, UMR5142, 64000 Pau, France
E-mail: camilo.broc@univ-pau.fr

B. Calvo
Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, 20018 Donostia, Spain

B. Liquet
Centre of Excellence for Mathematical and Statistical Frontiers and School of Mathematical, Sciences at Queensland University of Technology, Brisbane, Australia

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 [5]. 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. For instance, pleiotropy [11] is a field of genetics, where a gene (predictor) can have a particular effect on different phenotypes (independent variables). Data can be gathered from different studies, where the nature of the phenotype differs. Therefore, a gene can be highly correlated with each phenotype, but an overall model struggles to catch the particularity of those effects.

In the article, we tackle the problem of "batch effect" for dimension reduction such as Partial Least Square (PLS) method introduced by Wold [7]. Common dimension reduction techniques are Canonical Correlation Analysis (CCA) [14], Principal Component Analysis (PCA) [12], and PLS [1]. All these methods rely on the projection of the data into a subspace of lower dimension which represents most of the variation of the data. They are often posed as an eigen value problem [3]. PLS and CCA are both analysis two blocks of data and differ from the norm used, whereas PCA analyses one block. Aiming to apply our method to supervised analysis, PLS approach is considered in this article.

In these dimension reduction techniques, results are formulated with new variables that are linear combination of the original ones. These combinations can be hard to interpret due to the huge number of coefficient that they represent. To answer this problem, Lasso methods have been used. Introducing this penalization 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. These are are called sparse method and have been developed for linear regression [16,23], CCA [19], PCA [24], and Partial Least Square (sPLS). The sparse PLS (sPLS) has shown encouraging results [2,8] and is the object of analysis in this article. The PLS and sPLS methods have also been used to control "the batch effect" when related studies are combining to increase sample size combining independent but related studies ([4,13]). In particular, combining sPLS separating models and linking them can be an option like in the Multivariate INTegrative method (MINT) proposed in [13]. However, this approach cannot identify the true signal in the presence of different dynamics.

For high-dimensional regression problems, using problem-specific prior information improves the accuracy of the prediction and the interpretability of the model [10]. For example, in genomics, genes within the same pathway have similar functions and act together in regulating a biological system. Incorporation of this grouping structure is becoming increasingly common due to the success of gene set enrichment analysis approaches [17]. Using a model taking into account, this variable group structure allows to improve the performance and the readability of the results. To this end, sparse group Partial Least Square (sgPLS) has been developed [9], where two overlaid Lasso penalizations translate the group structure in the Partial Least Square formulation. A structure with group and sub-groups can also be handle by its generalization with three overlaid Lasso penalizations (sgsPLS [18]). Methods such as MINT do not take into account this kind of group structure.

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 to improve the performance of the analysis. Although this theory has 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 methods 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 is 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

Before going into further details, the notation used in this article are introduced. Data are represented by $X \in \mathbb{R}^{p\times n}$ and $Y \in \mathbb{R}^{q\times n}$, 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 $i \in \{1, \ldots, a\}$,

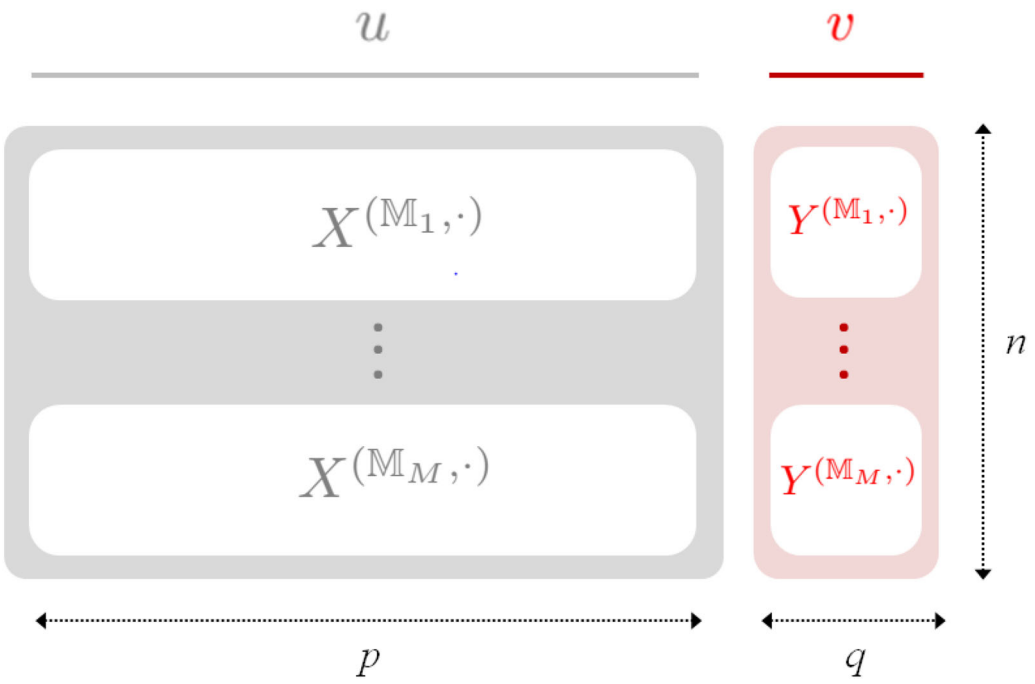

Fig. 1 Illustration of data structured by group of observation. Observations are assumed to be ordered by observation set

its rows are noted $A^{(i,\cdot)}$, and for $j \in \{1,\ldots,b\}$, its columns are noted $A^{(\cdot,j)}$, and for subsets $\tilde{a} \subset \{1,\ldots,a\}$ and $\tilde{b} \subset \{1,\ldots,b\}$ 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 $i \in \{1,\ldots,a\}$, its elements are noted $\omega^{(i)}$ and for subsets $\tilde{a} \subset \{1,\ldots,a\}$ $\omega^{(\tilde{a})}$ represents the elements of the vector corresponding to the positions in the subset. Matrices will always be in uppercase letters and vectors in lowercase letters to avoid any confusion.

The Frobenius norm on matrices is denoted $\|\ \|_F$. 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 $(x)_+ = \frac{|x|+x}{2}$.

2.1 Data with observation sets

Some data may present a structure among the observations gathered around groups of observations. For instance, data can be composed of different studies, each one presenting its own mechanisms and bias. Let us consider M different sets in the data. Noting, for $m \in \mathbb{N}$, $\mathbb{M}_m$ a subset of $\{1,\ldots,n\}$, let $\mathbb{M} = (\mathbb{M}_m)_{m=1..M}$ be a partition of $\{1, ..., n\}$ corresponding to the observation sets. We note $\#\mathbb{M}_m = n_m$. Row blocks are defined by these partitions in Fig. 1 (observations are assumed to be ordered by observation set).

2.2 Data with group of variables

Some data may present a structure among the variable gathered around groups. Let us consider that the variables are gather in K groups. 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 $\sum_{k=1}^{K} p_k = p$. This partition can define column blocks among the variables if variables are assumed to be ordered by variable group. Both observation set structure and variable group structure can be defined at the same time like in Fig. 2.

3 Formulation of the sparse Partial Least Square

In the literature, two formulations of the Partial Least Square exist, some extensions of the PLS follow a first one usually called PLS1 [22] and other extensions follow a second one called "PLS2" [2]. In the context of the article, we study exclusively the first one.

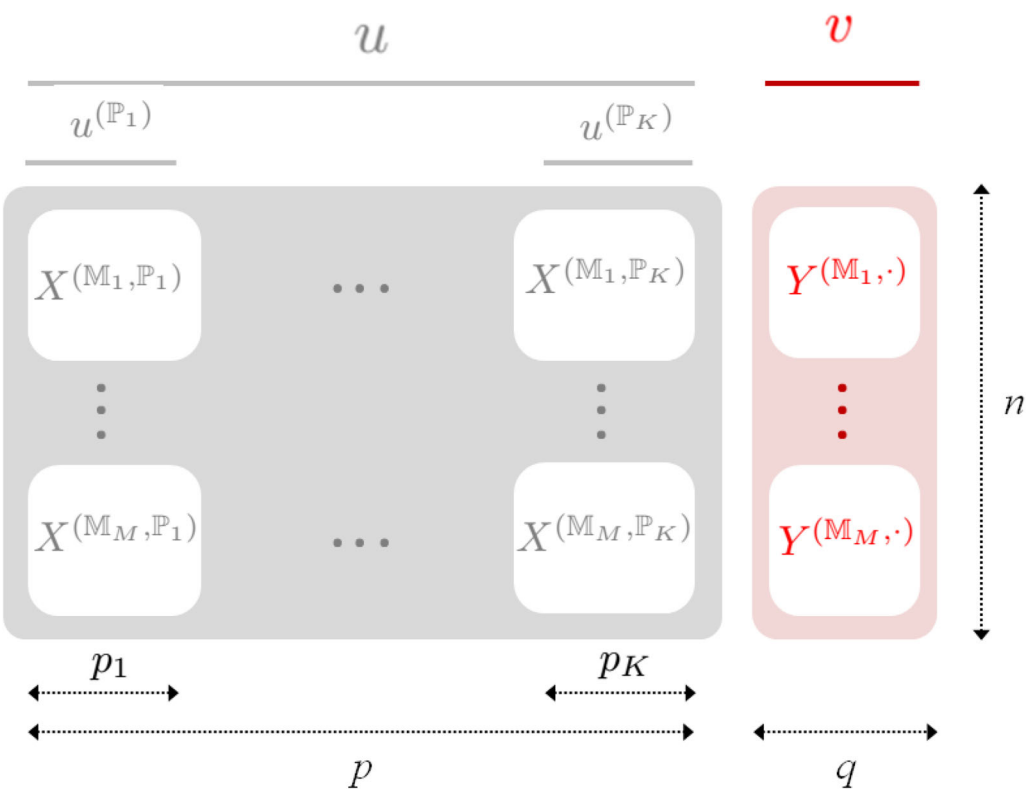

Fig. 2 Illustration of data structured by group of variables and group observation. Variables are assumed to be ordered by variable group

3.1 PLS and sPLS

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 $r < \min(p, q, r)$, where the couples are composed of vectors of length resp. p and q, maximizing $\mathrm{Cov}(Xu_i, Yv_i)$ 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 [7]. It can be solved considering successive minimization problems [15], for $h \in \{1, \ldots, r\}$:

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

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}, and Y_{h-2} for $h \in \{2, \ldots, n\}$. The deflation depends on the PLS mode that is chosen [7,20]. In this article, we focus on the enhancement of the optimization problem and its Lasso formulation in its hth step. According to [15], this step can be written as

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

where the notation h is removed 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 the 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 Eq. (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 [6]. Thus, the cross-product matrix $X^T Y$ is proportional to the empirical covariances between X- and Y -variables when the columns of X and Y are centered. When the columns are standardized, $X^T Y$ is proportional to the empirical correlations between X- and Y -variables. In this article, the standardization is an important step to overcome the issue of the "batch effect" or to aggregate observations from different studies.

3.2 Formulation of the penalized PLS

Six different formulations of the sPLS have been presented. The first four correspond to data presenting an observation set structure like in Fig. 1. The two last correspond to data presenting an observation set structure and a variable group structure like in Fig. 2 which correspond to sgPLS models (see [9]). We can note that problem 5 is a particular case of Fig. 2, where there is only one observation set ($M = 1$). Loading vectors introduced in those figures refer to vectors formulated in the following problems. The study of problems 4 and 6 is the main contribution of the article.

– Problem 1 (standard sPLS): This approach consists in simply considering all the observation set as one set. Data are standardized across all the sets, i.e., X and Y are standardized. The formulation is a standard sPLS problem:

$$\{u_{\text{opt}}, v_{\text{opt}}\} = \underset{||u||_2=||v||_2=1}{\text{argmin}} \left\| X^T Y - u v^T \right\|_F^2 + \lambda P(u). \tag{2}$$

In the model, the loading u is composed of p elements and the loading v is composed of q elements. The sparsity of u is controlled by the parameter λ: for a given λ, s_λ elements of u will be non-zero.

– Problem 2 (MINT): Introduced in [13], this approach consists in considering M different sPLS problems corresponding to each of the M observation sets. 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 sPLS problem is the same than in the previous problem in Eq. (2).
In the model, the loading u is composed of p elements and the loading v is composed of q elements. The sparsity of u is controlled by the parameter λ: for a given λ, s_λ elements of u will be non-zero.

– Problem 3 (multiple sPLS): This approach consists in considering all the observation set as one set, i.e., $X^{(\mathbb{M}_m,\cdot)}$ and $Y^{(\mathbb{M}_m,\cdot)}$ are standardized. 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. Formulation is a classic sPLS problem:

$$\{u_{m,\text{opt}}, v_{m,\text{opt}}\} = \underset{||u_m||_2=||v_m||_2=1}{\text{argmin}} \left\| X^{(\mathbb{M}_m,\cdot)T} Y^{(\mathbb{M}_m,\cdot)} - u_m v_m^T \right\|_F^2 + \lambda_m P(u_m). \tag{3}$$

In the model, the set of loading $\{u_m\}_{m\in\{1,\ldots,M\}}$ is composed of $p \times m$ elements (p elements per u_m). The set of loading $\{u_m\}_{m\in\{1,\ldots,M\}}$ is composed of $q \times m$ elements (q elements per v_m). The sparsity of u_m is controlled by the parameter λ_m: for a given λ_m, s_{m,λ_m} elements of u_m will be non-zero. Therefore, variables concerned by the shrinkage to zero will depend on the observation set m.

– Problem 4 (" sparse PLS for structured data"): This approach consists in considering M different sPLS problems corresponding to each of the M observation sets. 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. All problems are solved at the same time with a common-Lasso.
The formulation of the problem is

$$\{U_{\text{opt}}, V_{\text{opt}}\} = \underset{U,V}{\text{argmin}} \sum_{m=1}^{M} \left\| Z_m - U^{(\cdot,m)} V^{(\cdot,m)T} \right\|_F^2 + \lambda P(U)$$

$$\text{with } P(U) = \sum_{i=1}^{p} \left\| U^{(i,\cdot)} \right\|_2 \text{ and } Z_m = X^{(\mathbb{M}_m,\cdot)T} Y^{(\mathbb{M}_m,\cdot)}. \tag{4}$$

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 $V^{(\cdot,m)}$). The sparsity of all $U^{(\cdot,m)}$ is controlled by the parameter λ: for a given λ, the same s_λ elements of each $U^{(\cdot,m)}$ will be non-zero.

– Problem 5 (classical sgPLS): When variables can be gathered in groups (Fig. 2), the 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

$$\begin{aligned} \{u_{\text{opt}}, v_{\text{opt}}\} &= \underset{u,v}{\text{argmin}} \| Z - uv^T \|_F^2 + \lambda\,(1-\alpha)\, P_{\text{group}}(u) + \lambda\alpha\, P_{\text{variable}}(u) \\ \text{with } P_{\text{group}}(u) &= \sum_{k=1}^{K} \sqrt{p_k} \|u^{(\mathbb{P}_k)}\|_2 \;,\; P_{\text{variable}}(u) = \sum_{i=1}^{p} \|u^{(i)}\|_2 \\ \text{and } Z &= X^T Y. \end{aligned} \tag{5}$$

In the model, the loading vectors u and v are 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.

– Problem 6 (" sgPLS for structured data"): In the same spirit of adapting problem 2 into problem 4, problem 5 can be adapted with a common-Lasso penalization. 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

$$\begin{aligned} \{U_{\text{opt}}, V_{\text{opt}}\} &= \underset{U,V}{\text{argmin}} \left\| Z_m - U^{(\cdot,m)} V^{(\cdot,m)T} \right\|_F^2 + \lambda\,(1-\alpha)\, P_{\text{group}}(U) + \lambda\alpha\, P_{\text{variable}}(U) \\ \text{with } P_{\text{group}}(U) &= \sum_{k=1}^{K} \sqrt{p_k} \left\| U^{(\mathbb{P}_k,\cdot)} \right\|_F \;,\; P_{\text{variable}}(U) = \sum_{i=1}^{p} \left\| U^{(i,\cdot)} \right\|_2 \\ \text{and } Z_m &= X^{(\mathbb{M}_m,\cdot)T} Y^{(\mathbb{M}_m,\cdot)}. \end{aligned} \tag{6}$$

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 $U^{(\cdot,m)}$, $m \in \{1, \ldots, M\}$.

4 Solutions of the penalized PLS

The classical sPLS can be seen as a biconvex optimization problem. It can be solved by successively optimizing the loading u and v [15]. For a given v, an optimized $\tilde{u}$ is computed and the value of u is updated. Then, the same is performed permuting the roles of u and v. This optimization process relies on solving the problems:

$$\begin{aligned} u_{\text{opt}} &= \underset{||u||_2=1}{\text{argmin}} \left\| X^T Y - uv^T \right\|_F^2 + \lambda P(u) \\ v_{\text{opt}} &= \underset{||v||_2=1}{\text{argmin}} \left\| X^T Y - uv^T \right\|_F^2 . \end{aligned} \tag{7}$$

The solution of problems 1–3 (composed of standard sPLS methods) is given by the following theorem:

Theorem 4.1 *The marginal optima in $\tilde{u}$ and $\tilde{v}$ in the sPLS* (Eq. (1)) *are: fixing v, the optimal u_{opt} for* (7) *is*

$$u_{\text{opt}}^{(i)} = u_0^{(i)} \left(1 - \frac{\lambda}{2 \left\| u_0^{(i)} \right\|_2} \right)_+ , \; u_0 = X^T Y v. \tag{8}$$

Fixing u, the optimal v_{opt} for (7) *is*

$$v_{\text{opt}} = Y^T X u. \tag{9}$$

In this formula, a soft-thresholding sets down to zero loadings corresponding to variables, whose scores are too low. Setting λ equal to zero, we find the formulation of the PLS problem without Lasso constraint. A proof can be find in [8].

For problems 4, 5, and 6, the solution is more complex. Problem 4 introduces a common-Lasso penalization, problem 5 introduces a variable group structure, and problem 6 introduces both common-Lasso penalization and variable group structure. We can note that problem 4 is a particular case of problem 6, where there is no group penalty, i.e., $\alpha = 1$. Problem 5 is a particular case of problem 6, where there is only one observation set, i.e., $M = 1$. The solution of problem 6 is given in Theorem 4.2 (presented in the following), whereas solutions of problems 4 and 5 are corollaries of this theorem and can be found after the proof (Corollaries 4.4 and 4.5).

Theorem 4.2 *The marginal optima in U and V in the "sparse group PLS for structured data"* [Eq. (6)] *are:*
Fixing V, the optimal U_{opt} for (6) is:

$$U_{\text{opt}}^{(\mathbb{P}_k,\cdot)} = U_1^{(\mathbb{P}_k,\cdot)} \left(1 - \frac{\lambda\,(1-\alpha)}{2\sqrt{\sum_{i\in\mathbb{P}_k}\left\|U_1^{(i,\cdot)}\right\|_2^2}}\right)_+ = U_1^{(\mathbb{P}_k,\cdot)} \left(1 - \frac{\lambda\,(1-\alpha)}{2\left\|U_1^{(\mathbb{P}_k,\cdot)}\right\|_F}\right)_+$$

$$\text{With } U_1^{(i,\cdot)} = U_0^{(i,\cdot)} \left(1 - \frac{\lambda\alpha}{2\left\|U_0^{(i,\cdot)}\right\|_2}\right)_+,\ U_0^{(\cdot,m)} = Z_m V^{(\cdot,m)} \text{ and } Z_m = X^{(\mathbb{M}_m,\cdot)T} Y^{(\mathbb{M}_m,\cdot)} \tag{10}$$

Fixing U, the optimal V_{opt} for (6) is:

$$V_{\text{opt}}^{(\cdot,m)} = Z_m^T U^{(\cdot,m)} \tag{11}$$

Proof The proof is composed of three steps. In Step 1, we settle the sub-gradient equation corresponding to the minimization problem. In Step 2, we make the sPLS thresholding emerge in the equation. In Step 3, we make emerge the group thresholding and prove the theorem.

Let us settle the sub-gradient equation. The optimal U for a given V is

$$\min_U \sum_{m=1}^{M} \left\|Z_m - U^{(\cdot,m)} V^{(\cdot,m)T}\right\|_F^2 + \lambda\,(1-\alpha) \sum_{k=1}^{K} \sqrt{p_k} \left\|U^{(\mathbb{P}_k,\cdot)}\right\|_F + \lambda \sum_{i=1}^{p} \left\|U^{(i,\cdot)}\right\|_2.$$

We note that the problem can be formulated making appearing the column blocks corresponding to the variable groups. A second formulation of the problem would be

$$\min_U \sum_{k=1}^{K}\sum_{m=1}^{M} \left\|Z_m^{(\mathbb{P}_k,\cdot)} - U^{(\mathbb{P}_k,m)} V^{(\cdot,m)T}\right\|_F^2 + \lambda\,(1-\alpha) \sum_{k=1}^{K} \sqrt{p_k} \left\|U^{(\mathbb{P}_k,\cdot)}\right\|_F + \lambda \sum_{k\in 1}^{K} \sum_{i\in\mathbb{P}_k} \left\|U^{(i,\cdot)}\right\|_2.$$

We can see that the problem can be separated in K distinct problems for every $k \in \{1,\ldots,K\}$:

$$\min_{U^{(\mathbb{P}_k,\cdot)}} \sum_{m=1}^{M} \left\|Z_m^{(\mathbb{P}_k,\cdot)} - U^{(\mathbb{P}_k,m)} V^{(\cdot,m)T}\right\|_F^2 + \lambda\,(1-\alpha)\sqrt{p_k} \left\|U^{(\mathbb{P}_k,\cdot)}\right\|_F + \lambda \sum_{i\in\mathbb{P}_k} \left\|U^{(i,\cdot)}\right\|_2.$$

To solve this problem, let us consider the kth problem developing the Frobenius norm:

$$\min_{U^{(\mathbb{P}_k,\cdot)}} \sum_{m=1}^{M} \left[\text{Trace}\left(Z_m^{(\mathbb{P}_k,\cdot)} Z_m^{(\mathbb{P}_k,\cdot)T}\right) - 2\text{Trace}\left(Z_m^{(\mathbb{P}_k,\cdot)} V^{(\cdot,m)} U^{(\mathbb{P}_k,m)T}\right) + \text{Trace}\left(U^{(\mathbb{P}_k,m)} U^{(\mathbb{P}_k,m)T}\right)\right]$$
$$+\lambda\,(1-\alpha)\sqrt{p_k}\left\|U^{(\mathbb{P}_k,\cdot)}\right\|_F + \lambda\alpha \sum_{i\in\mathbb{P}_k} \left\|U^{(i,\cdot)}\right\|_2.$$

Taking the sub-gradient, the optimal U_{opt} verify for $m \in \{1, \ldots, M\}$:

$$
\begin{aligned}
& - U_{\text{opt}}^{(\mathbb{P}_k,m)} + U_0^{(\mathbb{P}_k,m)} = \frac{\lambda\,(1-\alpha)\,\sqrt{p_k}}{2}\Theta_g^{(\mathbb{P}_k,m)} + \frac{\lambda\alpha}{2}\Theta_v^{(\mathbb{P}_k,m)} \\
& \text{with the } (p \times M) \text{ matrix } U_0 \text{ such that } U_0^{(\mathbb{P}_k,m)} = Z_m^{(\mathbb{P}_k,\cdot)} V^{(\cdot,m)} \\
& \text{with the } (p \times M) \text{ matrix } \Theta_g \text{ such that } \Theta_g^{(\mathbb{P}_k,m)} = \begin{cases} \dfrac{U_{\text{opt}}^{(\mathbb{P}_k,\cdot)}}{\left\| U_{\text{opt}}^{(\mathbb{P}_k,\cdot)} \right\|_F} & \text{if } U^{(\mathbb{P}_k,\cdot)} \neq 0 \\ \Theta \in \{\Theta \in \mathbb{R}^{p_k \times M}, \|\Theta\|_F \leq 1\} \text{ if } U_{\text{opt}}^{(\mathbb{P}_k,\cdot)} = 0 \end{cases} \\
& \text{and with the } (p \times m) \text{ matrix } \Theta_v \text{ such that } \Theta_v^{(i,\cdot)} = \begin{cases} \dfrac{U_{\text{opt}}^{(i,\cdot)}}{\left\| U_{\text{opt}}^{(i,\cdot)} \right\|_2} & \text{if } U_{\text{opt}}^{(i,\cdot)} \neq 0 \\ \Theta \in \{\Theta \in \mathbb{R}^{M}, \|\Theta\|_2 \leq 1\} \text{ if } U_{\text{opt}}^{(i,\cdot)} = 0 \end{cases}.
\end{aligned}
\tag{12}
$$

We can note that when there is no penalty (i.e., $\lambda = 0$), $U_{\text{opt}} = U_0$ is the solution of the non-sparse problem.

The sub-gradient equation is settle (Step 1). Let us now make emerge the thresholding of sPLS.

We investigate in which case $U_{\text{opt}}^{(\mathbb{P}_k,\cdot)} = 0$, i.e., when loading corresponding to a group of variables is set to zero. If $U_{\text{opt}}^{(\mathbb{P}_k,\cdot)} = 0$, then $U_{\text{opt}}^{(i,\cdot)} = 0$ for every $i \in \mathbb{P}_k$. Hence, we have

$$
\begin{aligned}
U_0^{(\mathbb{P}_k,m)} &= \frac{\lambda\,(1-\alpha)\,\sqrt{p_k}}{2}\Theta_g^{(\mathbb{P}_k,m)} + \frac{\lambda\alpha\Theta_v^{(\mathbb{P}_k,m)}}{2} \\
&\text{with } \left\|\Theta_g^{(\mathbb{P}_k,\cdot)}\right\|_2^2 \leq 1 \\
&\text{and with } \left\|\Theta_v^{(i,\cdot)}\right\|_2^2 \leq 1.
\end{aligned}
\tag{13}
$$

and for $i \in \mathbb{P}_k$, we have also

$$
\begin{aligned}
U_0^{(i,\cdot)} - \tfrac{\lambda\alpha\Theta_v^{(i,\cdot)}}{2} &= \tfrac{\lambda(1-\alpha)\sqrt{p_k}}{2}\Theta_g^{(i,\cdot)} \\
&\text{with } \left\|\Theta_g^{(i,\cdot)}\right\|_2^2 \leq 1 \\
&\text{and with } \left\|\Theta_v^{(i,\cdot)}\right\|_2^2 \leq 1.
\end{aligned}
$$

Let us define

$$
U_1^{(i,\cdot)} = \left(1 - \frac{\lambda\alpha}{2\left\|U_0^{(i,\cdot)}\right\|_2}\right)_+ U_0^{(i,\cdot)}. \tag{14}
$$

We can establish in the following lemma, which makes emerge the variable thresholding term of sPLS like in (1) in Eq. (12).

Lemma 4.3

$$
\left\|U_1^{(i,\cdot)}\right\|_2 \leq \left\|U_0^{(i,\cdot)} - \frac{\lambda\alpha\Theta_v^{(i,\cdot)}}{2}\right\|_2 \tag{15}
$$

and there is Θ_v, such that

$$
U_1^{(i,\cdot)} = U_0^{(i,\cdot)} - \frac{\lambda\alpha\Theta_v^{(i,\cdot)}}{2}. \tag{16}
$$

Proof If $\left\|U_0^{(i,\cdot)}\right\|_2 \leq \left\|\frac{\lambda\alpha}{2}\right\|_2$, then $\left(1-\frac{\lambda\alpha}{2\left\|U_0^{(i,\cdot)}\right\|_2}\right)_+ = 0$ and then $U_0^{(i,\cdot)} = 0$. The inequality is then true. Furthermore, there is a $\Theta_v^{(i,\cdot)} = U_0^{(i,\cdot)}$ reach the equality (16). Otherwise, $U_1^{(i,\cdot)} \neq 0$ and

$$U_0^{(i,\cdot)} - \frac{\lambda\alpha\Theta_v^{(i,\cdot)}}{2} = U_0^{(i,\cdot)} - \frac{\lambda\alpha U_0^{(i,\cdot)}}{2\left\|U_0^{(i,\cdot)}\right\|_2} = U_1^{(i,\cdot)}$$

The inequality (15) is true, because the equality (16) is reached. In any case, the Lemma 4.3 is proved. □

From lemma 4.3, we can infer that in (12)

$$\left\|U_1^{(i,\cdot)}\right\|_2 \leq \left\|\frac{\lambda\,(1-\alpha)\,\sqrt{p_k}}{2}\Theta_g^{(i,\cdot)}\right\|_2$$

and the inequality can be reached as an equality.

We have

$$\sum_{i\in\mathbb{P}_k}\|\Theta_g^{(i,\cdot)}\|_2^2 = \|\Theta_g^{(\mathbb{P}_k,\cdot)}\|_2^2$$

and we have also

$$\sum_{i\in\mathbb{P}_k}\|U_1^{(i,\cdot)}\|_2^2 \leq \left\|\Theta_g^{(\mathbb{P}_k,\cdot)}\right\|_2^2$$

and the inequality can be reached.

Therefore

$$\|\Theta_g^{(\mathbb{P}_k,\cdot)}\|_2^2 \leq 1$$

stand if and only if

$$\sum_{i\in\mathbb{P}_k}\|U_1^{(i,\cdot)}\|_2^2 \leq 1.$$

In the end, we have $U^{(i,\cdot)} = 0$ if

$$\sum_{i\in\mathbb{P}_k}\|U_1^{(i,\cdot)}\|_2^2 \leq 1.$$

Let us now consider that $U_{\text{opt}}^{(\mathbb{P}_k,\cdot)} \neq 0$ or $U_{\text{opt}}^{(i,\cdot)} \neq 0$ for at least one $i \in \mathbb{P}_k$ then

$$U_{\text{opt}}^{(i,\cdot)} = U_0^{(i,\cdot)} - \frac{\lambda\alpha}{2}\Theta_v^{(i,\cdot)} - \frac{\lambda\,(1-\alpha)}{2}\frac{U_{\text{opt}}^{(i,\cdot)}}{\|U_{\text{opt}}^{(\mathbb{P}_k,\cdot)}\|_F}.$$

If $\left\|U_0^{(i,\cdot)}\right\|_2 \leq \|\frac{\lambda\alpha}{2}\|$, then we can set $\Theta_v^{(i,\cdot)}$, such that $U_0^{(i,\cdot)} - \frac{\lambda\alpha}{2}\Theta_v^{(i,\cdot)} = 0$. Otherwise, $U_0^{(i,\cdot)} - \frac{\lambda\alpha}{2}\Theta_v^{(i,\cdot)} = U_0^{(i,\cdot)} - \frac{\lambda\alpha}{2}\frac{U^{(i,\cdot)}}{\|U^{(i,\cdot)}\|_2}$. In both cases, we can consider that $U_0^{(i,\cdot)} - \frac{\lambda\alpha}{2}\Theta_v^{(i,\cdot)} = U_1^{(i,\cdot)}$.

From this point, we find successively

$$U_{\text{opt}}^{(i,\cdot)} = U_1^{(i,\cdot)} - \frac{\lambda\,(1-\alpha)\,U_{\text{opt}}^{(i,\cdot)}}{2\left\|U_{\text{opt}}^{(\mathbb{P}_k,\cdot)}\right\|_F},$$

$$U_{\text{opt}}^{(i,\cdot)}\left(1+\frac{\lambda\,(1-\alpha)}{2\left\|U_{\text{opt}}^{(i,\cdot)}\right\|_2}\right) = U_1^{(i,\cdot)}$$

and

$$U_{\text{opt}}^{(i,\cdot)}\left(1+\frac{\lambda\,(1-\alpha)}{2\left\|U_{\text{opt}}^{(i,\cdot)}\right\|_2}\right)=U_1^{(i,\cdot)}.$$

Summing the square for every element of $\mathbb{P}_k$, we have

$$\sum_{i\in\mathbb{P}_k}\left\|U_{\text{opt}}^{(i,\cdot)}\right\|_2^2\left(1+\frac{\lambda\,(1-\alpha)}{2\left\|U_{\text{opt}}^{(i,\cdot)}\right\|_2}\right)^2=\sum_{i\in\mathbb{P}_k}\left\|U_1^{(i,\cdot)}\right\|_2^2.$$

and hence, $\left\|U_{\text{opt}}^{(i,\cdot)}\right\|_F^2\left(1+\frac{\lambda(1-\alpha)}{2\left\|U_{\text{opt}}^{(i,\cdot)}\right\|_2}\right)^2=\sum_{i\in\mathbb{P}_k}\left\|U_1^{(i,\cdot)}\right\|_F^2=\left\|U_1^{(\mathbb{P}_k,\cdot)}\right\|_F^2$.

After extracting the value of $\left\|U^{(\mathbb{P}_k,\cdot)}\right\|_F$ from this equation, we finally find that

$$U_{\text{opt}}^{(\mathbb{P}_k,\cdot)}=U_1^{(\mathbb{P}_k,\cdot)}\left(1-\frac{\lambda\,(1-\alpha)}{2\left\|U_1^{(\mathbb{P}_k,\cdot)}\right\|_F}\right).$$

□

Corollary 4.4 *The solution to Eq. (4) can be seen as a biconvex optimization problem.*
Fixing V, the optimal U_{opt} for (4) is:

$$U_{\text{opt}}^{(i,\cdot)}=U_0^{(i,\cdot)}\left(1-\frac{\lambda\alpha}{2\left\|U_0^{(i,\cdot)}\right\|_2}\right)_+,\ U_0^{(\cdot,m)}=Z_mV^{(\cdot,m)}$$
$$\text{and } Z_m=X^{(\mathbb{M}_m,\cdot)T}Y^{(\mathbb{M}_m,\cdot)}. \tag{17}$$

Fixing U, the optimal V_{opt} for (4) is:

$$V_{\text{opt}}^{(\cdot,m)}=Z_m^TU^{(\cdot,m)}. \tag{18}$$

Corollary 4.5 *The solution to Eq. (5) can be seen as a biconvex optimization problem is:*
Fixing v, the optimal u_{opt} for (5) is

$$u^{(\mathbb{P}_k)}=u_1^{(\mathbb{P}_k)}\left(1-\frac{\lambda\,(1-\alpha)}{2\sqrt{\sum_{i\in\mathbb{P}_k}\left\|u_1^{(i)}\right\|_2^2}}\right)_+=u_1^{(\mathbb{P}_k)}\left(1-\frac{\lambda\,(1-\alpha)}{2\left\|u_1^{(\mathbb{P}_k)}\right\|_F}\right)_+$$
$$\text{With } u_1^{(i)}=u_0^{(i)}\left(1-\frac{\lambda\alpha}{2\left\|u_0^{(i)}\right\|_2}\right)_+,\ u_0=Zv$$
$$\text{and } Z=XY^T. \tag{19}$$

Fixing u, the optimal v_{opt} for (5) is

$$v_{\text{opt}}=Z^Tu. \tag{20}$$

5 Discussion

Problems 1–4 are discussed in this part, but we consider that the following remarks can be transposed to problems 5 and 6, problems 5 and 6 being resp. the equivalents of problems 2 and 4 for data with a variable group structure.

5.1 Size of the data

The larger data (in terms of number of observations) are, the better models are supposed to perform. We can see that Problems 1 and 2 have the merit of performing an sPLS on data containing n observations, whereas problems 3 and 4 perform M different sPLS methods on data with resp. $\mathbb{M}_m$ observations for $m \in \{1, \ldots, M\}$. For some observation set, the number of observation can be significantly smaller than the size of the hole data which can have a negative impact on the result.

5.2 Number of loading elements in the model

Number of loading elements is an important parameter to control. From one side, the bigger the number is, the more information can be stored by the model, from the other side having too much loading elements can give results harder to interpret and there is higher risk of over-fitting. Problems 1 and 2 have only p loadings for u, whereas problems 3 and 4 have $M \times p$ ones. For problem 4, the number of loadings is important, but the number of non-zero variable will vary between 1 and p in the same way than in problem 1 and problem 2. The problem 4 gives readable results while keeping the flexibility of a model with higher number of loading elements. For problem 3, the non-zero variables can be different from one study to another, we cannot control weather a variable will be null for all studies and the number of non-zero variable will be significantly higher than for other problems.

5.3 Sensibility to batch effect

Batch effect can arise when data provided from different source present a bias. This effect can happen when observation sets are expected to introduce its intrinsic error. Therefore, the cross-product matrices could be represented by a model like:

$$Z_m = Z + E_m \text{ for } m \in \{1, \ldots, M\},$$

where Z follows a given law and E_m are Gaussian noise with parameters depending on m. Under this hypothesis, the standardization within studies can bypass this bias. Therefore, problems 3–4 can correct this kind of batch effect.

However, more complex bias can exist. For instance, what happens if different observation sets have different dynamics? Let us consider a variable that is positively correlated in some observation sets and negatively correlated in others. In problems 1 and 2 and their overall sPLS, the variable will have a small corresponding loading, because positive and negative effects compensate each other and the variable will be cut because of the sparsity heuristic. In problem 3, the distinction between all dynamics will be highlighted by the model. Finally, problem 4 will select the same variable, because it has a significant loading on every observation set. In the end, problem 4 can handle more cases, where the observation sets introduce bias.

5.4 Relation between problem 4 and classic sgPLS

We can establish also that problem 4 (sPLS method with a common-Lasso penalization) applied to matrices X and Y of size resp. $(n, p)$and (n, q) can be equivalent to a classical sgPLS without a standard Lasso on well chosen-matrices $\tilde{X}$ and $\tilde{Y}$ of size resp. $(n, p \times M)$ and $(n, q \times M)$. Those matrices are constructed by shifting the row blocks of X and Y: they are diagonal bloc matrices whose blocs are resp. $X^{(\mathbb{M}_m,\cdot)}$ and $Y^{(\mathbb{M}_m,\cdot)}$ for $m \in \{1, \ldots, M)\}$. The corresponding loading vectors of size resp. $p \times M$ and $q \times M$ are called here resp. u_e and v_e. The representation of those objects is shown in Fig. 3.

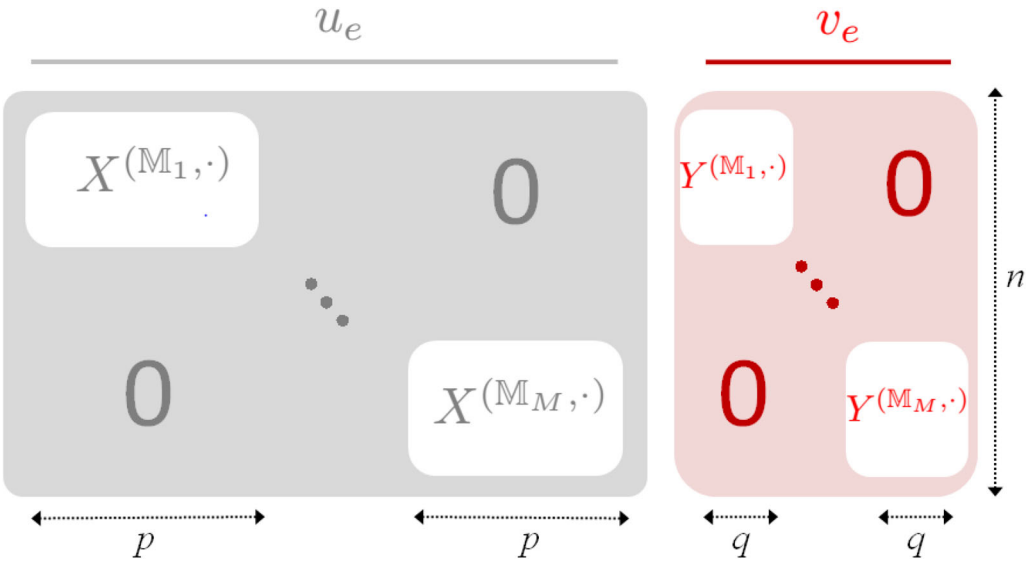

Fig. 3 Notation of $\tilde{X}$ (grey rectangle) and $\tilde{Y}$ (red rectangle) to write the sPLS for structured data as a sgPLS

On those basis, the formulation of the sgPLS problem searching for optimal $u_{e,\text{opt}}$ and $u_{e,\text{opt}}$ would be

$$\begin{aligned} \{u_{e,\text{opt}}, v_{e,\text{opt}}\} &= \underset{u_e, v_e}{\text{argmin}} \left\| \tilde{Z} - u_e {v_e}^T \right\|_F^2 + \lambda\,(1-\alpha)\, P_{\text{group}}(u_e) + \lambda\alpha\, P_{\text{variable}}(u_e) \\ \text{with } P_{\text{group}}(u_e) &= \sum_{k=1}^{K} \sqrt{p_k} \left\| {u_e}^{(\mathbb{P}_k)} \right\|_2, \end{aligned} \tag{21}$$

where $\mathbb{P}_k$ refers here to the variables of $\tilde{X}$ associated to the variables corresponding to the group k

$$P_{\text{variable}}(u_e) = \sum_{i=1}^{p} \left\| {u_e}^{(i)} \right\|_2 \text{ and } \tilde{Z} = \tilde{X}^T \tilde{Y} \tag{22}$$

In this formulation, loading vectors u_e and v_e can be seen as the concatenation of the rows of resp. U and V in a unique uni-dimensional vector. This notation is interesting from a theoretical point of view, because it ensure that problem 4 can inherit properties from sPLS. However, this notation is not wise for computational efficiency, because the matrices $\tilde{X}$ and $\tilde{Y}$ are M times bigger than X and Y, where M is the number of observation set. For implementation, computing directly the solution from Eqs. (10) and (11) seems wiser.

6 Application on simulated data

Presented methods are illustrated on simulated data. A first simulation case presents data, where a batch effect exists and a second simulation case presents data, where different dynamics exist among different observation sets. For each case, different noise levels are considered. Every simulation is performed 50 times. The code can be found at https://github.com/camilobroc/sgPLS_for_structured_data.

6.1 Design of the simulated data

In the following, a training data set of 900 observations gathered in 3 observation sets of 300 observations is generated and then a test data set of 300 observations gathered in 3 observation sets of 100 observations (for the training data $M = 3$ and $n_1 = n_2 = n_3 = 300$, for the test data $n_1 = n_2 = n_3 = 100$).

6.1.1 Batch effect cases

In first the simulation case, applications of the methods with data presenting a batch effect are performed. The simulation is performed with different noise levels. Data have an observation set structure and a group of variable structure, as shown in Fig. 2. A batch effect implies that one same physical process is observed but the methods of measurement vary among the different group of observation. We represent this difference of measurements by a bias in depending on the observation set.

A matrix X with 1000 variables gathered in 50 groups of 20 variables ($K = 50$ and $p_1 = \cdots = p_K = 20$) and a matrix Y with 3 variables ($q = 3$) are generated. To mimic a batch effect, the generation procedure of the matrices resp. X and Y has different parameters depending on different sub-matrices of resp. X and Y. Those matrices are composed of a signal and a noise. The signal corresponds to a PLS model with one latent variable. For $m \in \{1, \ldots, M\}$:

$$\begin{aligned} X^{(\mathbb{M}_m,\cdot)} &= \underbrace{H^{(\mathbb{M}_m)} C^T}_{\text{Signal}} \underbrace{\lambda_m^{X,B} + \mu_m^{X,B} \mathbb{1}_{n_m \times p}}_{\text{Batch}} + \underbrace{E_X^{(\mathbb{M}_m,\cdot)}}_{\text{Noise}} \\ Y^{(\mathbb{M}_m,\cdot)} &= \underbrace{H^{(\mathbb{M}_m)} D^T}_{\text{Signal}} \underbrace{\lambda_m^{Y,B} + \mu_m^{Y,B} \mathbb{1}_{n_m \times q}}_{\text{Batch}} + \underbrace{E_Y^{(\mathbb{M}_m,\cdot)}}_{\text{Noise}} \end{aligned} \tag{23}$$

Signal The latent variable H is a $(n \times 1)$ column vector, where each element follows a normal distribution of mean 0 and standard deviation 1. The loadings associated with this latent variable are resp. C a $p \times 1$ column vector and D a $q \times 1$ column vector corresponding resp. to X and Y.

Batch effect
The signal is blurred by a batch effect. The parameters $\lambda_m^{(X,B)}$, $\mu_m^{(X,B)}$, $\mu_m^{(Y,B)}$, and $\lambda_m^{(Y,B)}$ are real numbers depending on the observation set m. They control the shape of the batch effect. The notations $\mathbb{1}_{n_m,p}$ and $\mathbb{1}_{n_m,q}$ correspond to the matrices which elements are equal to 1 and of respective size $n_m \times p$ and $n_m \times q$.

Noise
The noise is represented by E_X, a $(n \times p)$ matrix, and E_Y, a $(n \times q)$ matrix. The matrix E_X is constructed by group of variables: for $k \in \{1, \ldots, K\}$, the rows of $E_X^{(\cdot,\mathbb{P}_k)}$ follow a multivariate normal distribution $N_{p_k}(\mathbb{0}_{p_k}, \lambda^{X,E} \Sigma_{p_k,\rho})$, where ρ and $\lambda^{X,E}$ are real parameters and $\Sigma_{p_k,\rho}$ is a $(p_k \times p_k)$ matrix which diagonal elements are equal to $1 - \rho$ and non-diagonal elements are equal to ρ. The notations $_{p_k}$ stands for the vector of size p_k and which elements are all equal to 0. The rows of the matrix E_Y follow a multivariate normal distribution $N_q(\mathbb{1}_q \times 0, \lambda^{Y,E} \Sigma_{q,\rho})$, where $\lambda^{Y,E}$ is a real parameter and $\Sigma_{q,\rho}$ is a $(q \times q)$ matrix which diagonal elements are equal $1 - \rho$ and non-diagonal elements are equal to ρ. The notations $_q$ to the vector of size q and which elements are all equal to 1. The parameter ρ represents a correlation between variables of a same group and $\lambda^{Y,E}$ and $\lambda^{Y,E}$ represents the noise levels.

The non-null parameters of C are the 15 first variables of the first 4 group of variables. Among those elements resp. 15, 30 and 15 are equal to resp. 1, -1, 1.5, and the values are randomly distributed. Other parameters are given in Table 1 and the noise levels are indicated in Table 2.

6.1.2 Effects of different magnitudes among group of observations

This simulation case mimic data presenting different dynamics among observation sets. The generation process follows the same formulas as the previous one but with different parameters. The main difference with the previous cases is that the parameters $\lambda_m^{X,B}$ for $m \in \{1, \ldots, M\}$ can have opposite signs. While in the first case, a batch could be represented by a difference of magnitude, the effects can have here opposite directions. In this simulation case, we are not interested in a bias concerning $\mu_m^{(X,B)}$ or $\mu_m^{(Y,B)}$, and the parameters are set to zero. The non-null parameters of C are the 15 first variables of the first 4 group of variables. Among those elements, resp. 15, 30m and 15 are equal to resp. 1, -1, 1.5 and the values are randomly distributed. Other parameters are given in Table 1 and the noise levels are indicated in Table 2.

6.2 Compared methods

In the first simulation case, methods corresponding to problems 1, 2, and 5 are compared, whereas in the second simulation case, methods corresponding to problems 1, 2, 4, and 6 are compared. For the methods

Table 1 Table of the parameters used in first and second simulation cases

	First simulation case	Second simulation case
ρ	0.05	0.05
$\mu_1^{X,B}$	2	0
$\mu_2^{X,B}$	-1	0
$\mu_3^{X,B}$	-1	0
$\lambda_1^{X,B}$	1	1
$\lambda_2^{X,B}$	0.8	-0.8
$\lambda_3^{X,B}$	1.5	1.05
$\mu_1^{Y,B}$	2	0
$\mu_2^{Y,B}$	0	0
$\mu_3^{Y,B}$	-2	0
$\lambda_1^{Y,B}$	0.6	0.6
$\lambda_2^{Y,B}$	1.4	1.4
$\lambda_3^{Y,B}$	1	1
D	$\{1, -1, 1.5\}$	$\{1, -1, 1.5\}$

Table 2 Results for the first and second simulation cases. Results in terms of MSEP, TPR, and TD are presented for each noise level

Simulation case 1									
Noise level	$\lambda^{X,E} = \lambda^{X,E} = 2$			$\lambda^{X,E} = \lambda^{X,E} = 20$			$\lambda^{X,E} = \lambda^{X,E} = 30$		
	MSEP	TPR	TD	MSEP	TPR	TD	MSEP	TPR	TD
sPLS	3.99	0.5	60	22.85	0.27	87.76	33.35	0.18	98.92
MINT	2.20	1.0	0.0	20.97	0.76	28.88	31.59	0.54	55.00
sgPLS	2.20	1.0	7.4	20.83	0.99	15.44	31.16	0.98	17.88
Simulation case 2									
Noise level	$\lambda^{X,E} = \lambda^{X,E} = 2$			$\lambda^{X,E} = \lambda^{X,E} = 10$			$\lambda^{X,E} = \lambda^{X,E} = 20$		
	MSEP	TPR	TD	MSEP	TPR	TD	MSEP	TPR	TD
sPLS	3.58	0.61	46.84	12.38	0.15	101.68	23.19	0.08	110.08
MINT	3.67	0.88	14.12	12.40	0.17	99.84	23.16	0.08	110.12
sPLS for structured data	2.85	1.00	0.00	11.11	0.79	24.84	21.76	0.43	67.96
sgPLS for structured data	2.85	1.00	5.56	11.06	0.98	13.60	21.39	0.93	20.28

corresponding to problems 1, 2, and 4, the penalization parameter λ is set, such that the number of variables is equal to the true number of variables having an effect (in this case 60). For the method corresponding to problems 5 and 6, the penalization parameter λ is set, such that the number of groups of variable is equal to the true one (in this case 4), while α is chosen from the set $\{0.1, \ldots, 0.9\}$ by cross validation: the value giving the best Mean Square Error Prediction is kept.

6.3 results

The performances of the method are measured through the True Positive Rate (TPR), the Total Discordance (TD), and the Mean Square Error Prediction. The TPR is defined as

$$\text{TPR} = \frac{\text{True Positives}}{\text{True Positive} + \text{False Negatives}}$$

and TN is defined as:

$$\text{TD} = \text{False Postives} + \text{False Negatives}.$$

Results of the first and second simulations are given in Table 2.

In the first simulation case, data present a bias that depends on the observation set. Different noise levels are generated ($\lambda^{X,E} = 2, 20, 30$). We can see that when noise is small ($\lambda^{X,E} = 2$), MINT and "sparse group PLS for structured data" can retrieve the true variables, whereas a classic sPLS cannot. We can also see that the MSEP is better for the "sparse group PLS for structured data" than for MINT. We can note that "sparse group PLS for structured data" misses a few true variables which gives a non-null TD. This is due to the fact that the calibration of the method does not seek for a selection of the true number of variables; hence, a small number of true variables can be missed. When noise is greater ($\lambda^{X,E} = 20$), a difference in terms of detection of true variables is observed. The classical PLS have a much worse TPR and TD, while "sparse group PLS for structured data" is above MINT. When noise is even greater ($\lambda^{X,E} = 30$), "sparse group PLS for structured data" clearly outperforms MINT.

In the second simulation case, data present a magnitude in the latent variables that depends on the observation set. Different noise levels are generated ($\lambda^{X,E} = 2, 10, 20$). We can see that when noise is small ($\lambda^{X,E} = 2$), we can see that to retrieve the true variables "s(g)PLS for structured data" is better than MINT which is better than sPLS. In the same way that in the first simulation case, "sgPLS for structured data" miss a few number of the true variables, whereas "sPLS for structured data" do not because of the specificity of the calibration. We can see at noise level $\lambda^{X,E} = 10$, that only the methods calibrated "for structure data" are able to retrieve the true variables. At the highest noise level ($\lambda^{X,E} = 20$), we can see that sgPLS stands clearly above "sPLS for structured data", and those two methods outperform the existing ones.

7 Conclusion

In the end, different ways of formulating an sPLS problem on data presenting an observation set structure have been discussed. The MINT formulation has the merit of being easy to implement and correct the batch effect. The novel method “sparse PLS for structured data” can also correct it. Furthermore, it allows to take into account a lot of different bias, especially when the different observation set do not have the same dynamics. Despite its high number of parameters, the common-Lasso penalization ensures that the result is readable with a small number of selected variables in the overall analysis.

This article proved its ability to inherit properties of sPLS. Its adaptation to variable groups developed in this article, called “sparse group PLS for structured data”, is a notable example of “ sparse PLS for structured data” benefiting from an extension of the sPLS. We can note also that it can be applied on either quantitative or qualitative variables as any sPLS can.

A simulation shows that the new methods can outperform existing methods for detecting a small signal in a large noise. Because its requirements on the nature of the data are very general, we are confident that the method can be applied to the wide area of domains, where sPLS is competitive.

References

1. Boulesteix, A.-L.; Strimmer, K.: Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief. Bioinform. **8**(1), 32–44 (2006)
2. Chun, H.; Keleş, S.: Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **72**(1), 3–25 (2010)
3. De Bie, T.; Cristianini, N.; Rosipal, R.: Eigenproblems in pattern recognition. In: Handbook of Geometric Computing, pp. 129–167. Springer (2005)
4. Eslami, A.; Qannari, E.M.; Kohler, A.; Bougeard, S.: Algorithms for multi-group pls. J. Chemom. **28**(3), 192–201 (2014)
5. Gagnon-Bartsch, J.A.; Speed, T.P.: Using control genes to correct for unwanted variation in microarray data. Biostatistics **13**(3), 539–552 (2012)
6. Geladi, P.; Kowalski, B.R.: Partial least-squares regression: a tutorial. Anal. Chim. Acta **185**, 1–17 (1986)
7. Herman, W.: Path models with latent variables: the nipals approach. In: Quantitative Sociology, pp. 307–357. Elsevier (1975)
8. Lê Cao, K.-A.; Rossouw, D.; Robert-Granié, C.; Besse, P.: A sparse pls for variable selection when integrating omics data. Stat. Appl. Genet. Mol. Biol. **7**(1), 35 (2008)

9. Liquet, B.; de Micheaux, P.L.; Hejblum, B.P.; Thiébaut, R.: Group and sparse group partial least square approaches applied in genomics context. Bioinformatics **32**(1), 35–42 (2015)
10. Liquet, B.; Mengersen, K.; Pettitt, A.N.; Sutton, M.: Bayesian variable selection regression of multivariate responses for group data. Bayesian Anal. **12**(4), 1039–1067 (2017)
11. Paaby, A.B.; Rockman, M.V.: The many faces of pleiotropy. Trends Genet. **29**(2), 66–73 (2013)
12. Price, A.L.; Patterson, N.J.; Plenge, R.M.; Weinblatt, M.E.; Shadick, N.A.; Reich, D.: Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. **38**(8), 904 (2006)
13. Rohart, F.; Eslami, A.; Matigian, N.; Bougeard, S.; Le Cao, K.-A.: Mint: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. BMC Bioinform. **18**(1), 128 (2017)
14. Seoane, J.A.; Campbell, C.; Day, I.N.M.; Casas, J.P.; Gaunt, T.R.: Canonical correlation analysis for gene-based pleiotropy discovery. PLoS Comput. Biol. **10**(10), e1003876 (2014)
15. Shen, H.; Huang, J.Z.: Sparse principal component analysis via regularized low rank matrix approximation. J. Multivar. Anal. **99**(6), 1015–1034 (2008)
16. Simon, N.; Friedman, J.; Hastie, T.; Tibshirani, R.: A sparse-group lasso. J. Comput. Graph. Stat. **22**(2), 231–245 (2013)
17. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. **102**(43), 15545–15550 (2005)
18. Sutton, M.; Thiébaut, R.; Liquet, B.: Sparse partial least squares with group and subgroup structure. Stat. Med. (2018) **37**(23), 3338–3356
19. Tenenhaus, A.; Philippe, C.; Guillemot, V.; Le Cao, K.-A.; Grill, J.; Frouin, V.: Variable selection for generalized canonical correlation analysis. Biostatistics **15**(3), 569–583 (2014)
20. Vinzi ,VE; Trinchera, L; Amato, S: Pls path modeling from foundations to recent developments and open issues for model assessment and improvement. In: Handbook of Partial Least Squares, pp. 47–82. Springer (2010)
21. Walker, S.J.: Big data: a revolution that will transform how we live, work, and think. Int. J. Advert. **33**(1), 181–183 (2014)
22. Wang, T.; Ho, G.; Ye, K.; Strickler, H.; Elston, R.C.: A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped. Genet. Epidemiol. **33**(1), 6–15 (2009)
23. Yuan, M.; Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **68**(1), 49–67 (2006)
24. Zou, H.; Hastie, T.; Tibshirani, R.: Sparse principal component analysis. J. Comput. Graph. Stat. **15**(2), 265–286 (2006)