Bioinformatics, 32(1), 2016, 35–42
doi: 10.1093/bioinformatics/btv535
Advance Access Publication Date: 10 September 2015
Original Paper

OXFORD

Gene expression

Group and sparse group partial least square approaches applied in genomics context

Benoît Liquet[1,2,*], Pierre Lafaye de Micheaux[3], Boris P. Hejblum[4,5,6,7] and Rodolphe Thiébaut[4,5,6,7]

[1]School of Mathematics and Physics, The University of Queensland, Brisbane 4066, Australia, [2]ARC Centre of Excellence for Mathematical and Statistical Frontiers, QUT, Brisbane, Australia, [3]CREST, ENSAI, Campus de Ker-Lann, Rue Blaise Pascal, BP 37203, 35172 Bruz cedex, France, [4]Inria, SISTM, Talence and [5]Inserm, U897, Bordeaux, [6]Bordeaux University, Bordeaux and [7]Vaccine Research Institute, Creteil, France

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on March 14, 2015; revised on June 26, 2015; accepted on September 3, 2015

Abstract

Motivation: The association between two blocks of 'omics' data brings challenging issues in computational biology due to their size and complexity. Here, we focus on a class of multivariate statistical methods called partial least square (PLS). Sparse version of PLS (sPLS) operates integration of two datasets while simultaneously selecting the contributing variables. However, these methods do not take into account the important structural or group effects due to the relationship between markers among biological pathways. Hence, considering the predefined groups of markers (e.g. genesets), this could improve the relevance and the efficacy of the PLS approach.

Results: We propose two PLS extensions called group PLS (gPLS) and sparse gPLS (sgPLS). Our algorithm enables to study the relationship between two different types of omics data (e.g. SNP and gene expression) or between an omics dataset and multivariate phenotypes (e.g. cytokine secretion). We demonstrate the good performance of gPLS and sgPLS compared with the sPLS in the context of grouped data. Then, these methods are compared through an HIV therapeutic vaccine trial. Our approaches provide parsimonious models to reveal the relationship between gene abundance and the immunological response to the vaccine.

Availability and implementation: The approach is implemented in a comprehensive R package called sgPLS available on the CRAN.

Contact: b.liquet@uq.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recent advances in high-throughput 'omics' technologies enable quantitative measurements of expression or abundance of biological molecules of a whole biological system. Various popular 'omics' platforms in systems biology include transcriptomics, proteomics, cytomics and metabolomics. The integration of multi-layer information is required to fully unravel the complexities of a biological system, as each functional level is hypothesized to be related to each other (Jayawardana *et al.*, 2015; Kitano, 2002). Furthermore, multi-layer information is increasingly available such as in standard clinical trials. As an example, the evaluation of vaccines in phase I/II trials incorporates various measurements of the cell counts (tens of population of interest), of the cell functionality by many ways including the production of cytokines (intra and extracellular) and of the gene expression (Palermo *et al.*, 2011).

The integration of omics data is a challenging task. First, the high dimensionality of the data, i.e. the large number of measured biological entities (tens of thousands) makes it very difficult to

obtain a good overview or understanding of the system under study. The noisy characteristics of such high-throughput data require a filtering process to be able to identify a clear signal. Second, because of experimental or financial constraints, the small number of samples or patients (typically < 50) makes the statistical inference difficult and argue for using the maximum amount of available information. Third, the integration of heterogeneous data also represents an analytical and numerical challenge to try to find common patterns in data from different origins.

In recent years, several statistical integrative approaches have been proposed in the literature to combine two blocks of omics data, often in an unsupervised framework. These approaches aim at selecting correlated biological entities from two datasets (Chun and Keleş, 2010; Lê Cao *et al.*, 2008, 2009; Parkhomenko *et al.*, 2009; Waaijenborg *et al.*, 2008; Witten *et al.*, 2009) or more (Löfstedt *et al.*, 2014; Tenenhaus and Tenenhaus, 2011). This abundant literature clearly illustrates that the integrative analysis of two datasets poses significant statistical challenges to deal with the high dimensionality of the data. In particular, sparse partial least squares (sPLSs), using a L_1 penalty, has been developed for that purpose. With sPLS, it has been demonstrated that the integrative analysis of large scale omics datasets could generate new knowledge not accessible by the analysis of a single data type alone (Lê Cao *et al.*, 2008, 2009). Moreover, the biological relevance of the approach has been demonstrated in recent studies (Morine *et al.*, 2011; Rose *et al.*, 2011).

However, group structures often existing within such data have not yet been accounted for in these analyses. For example, genes within the same pathway have similar functions and act together in regulating a biological system. These genes can add up to have a larger effect and therefore can be detected as a group [i.e. at a pathway or gene set level (Tyekucheva *et al.*, 2011)]. This has been increasingly used thank to geneset enrichment analysis approaches (Subramanian *et al.*, 2005).

Considering a group of features instead of individual features has been found to be effective for biomarker identification (Meier *et al.*, 2008; Puig *et al.*, 2009; Simon and Tibshirani, 2012; Yuan and Lin, 2006). Yuan and Lin (2006) proposed group lasso for group variables selection. Meier *et al.* (2008) extended it to logistic regression. Puig *et al.* (2009) and Simon *et al.* (2013) modified group lasso to solve the non-orthonormal matrices problem. Although group lasso penalty can increase the power for variable selection, it requires a strong group-sparsity (Huang and Zhang, 2010) and cannot yield sparsity within a group. Ma *et al.* (2007) proposed a supervised group lasso which selects both significant gene clusters and significant genes within clusters for logistic binary classification and Cox survival analysis. Simon *et al.* (2013) proposed a sparse group lasso penalty by combining an L_1 penalty with group lasso to yield sparsity at both the group and individual feature level. Zhou (2010) applied it to genomic feature identification. Garcia *et al.* (2014) developed a sparse group-subgroup Lasso to accommodate selecting important groups, subgroups and individual predictors. In a regression context with a multivariate response variable, Li *et al.* (2015) have recently proposed a multivariate sparse group lasso.

Also some work has been reported to incorporate 'group effect' into a conventional canonical correlation analysis (CCA) model. Chen *et al.* (2013) studied structure-based CCA and proposed tree-based and network-based CCA (Chen *et al.*, 2013). Chen and Liu (2012) incorporated group effect into an association study of nutrient intake with human gut microbiome (Chen and Liu, 2012). Both papers show an improvement when incorporating group effect; however, a priori knowledge of group structure is needed and only the group effect of one type of data is discussed. More recently, Lin *et al.* (2013) developed a more general group sparse CCA method, which have been illustrated on genomics datasets (human gliomas data and NCI60 data). Witten *et al.* (2009) proposed a general penalized matrix decomposition (PMD) approach, which include sparse principal components and CCA. Sparsity have been realized by introducing different penalties forms such as L_1 penalty or fused lasso penalty to get smooth results in the context of ordered features. Based on generalized least square matrix decomposition, Allen *et al.* (2014) develop fast computational algorithms for generalized principal component analysis (PCA) and sparse PCA. In the same idea, a regularized PLS (RPLS) approach is proposed by Allen *et al.* (2013) to take into account the correlations between adjacent variables. However, none of the PMD and RPLS approaches have introduced group and sparse group lasso penalty.

Here, we develop in a more general framework a group PLS (gPLS) method and a sparse gPLS (sgPLS) method (see also Löfstedt *et al.* 2014). Both methods focus on sub-matrices decomposition taking into account the group structures. They could be used in 'regression' mode or in 'canonical mode'. The gPLS model aims at performing selection at group level while sgPLS enables selection at both group and single feature levels. Both irrelevant groups of features and individual features in the remaining groups will be simultaneously discarded with sgPLS.

Our article is organized as follows. The model and algorithm for group and sgPLS are described in Section 2 after introducing the main steps of sPLS. We also present our extension in a context of PLS discriminant analysis. The performances of our approaches are compared with sPLS via a simulation study in Section 3. This section also contains an illustration of our method with an HIV vaccine study. The results are compared with the one obtained by applying the multivariate sparse group lasso recently proposed by Li *et al.* (2015).

2 Methods

2.1 Notations

Let $\boldsymbol{X}$ and $\boldsymbol{Z}$ be two data matrices containing n observations (rows) of p predictors (gene expression) and q variables (cytokine secretion), respectively. The soft thresholding function is $g^{\text{soft}}(x, \lambda) = \text{sign}(x)(|x| - \lambda)_+$, where $(a)_+ = \max(a, 0)$. The Frobenius norm is denoted $||\cdot||_F$, while the Euclidean vector norm is $||\cdot||_2$ and the L_1 vector norm is $||\cdot||_1$.

2.2 PLS and sPLS for integrative analysis

2.2.1 Partial least square

PLS (Wold, 1966) is a well-known exploratory approach that was initially applied in chemometrics. It is particularly useful for analyzing noisy, collinear, even incomplete highly dimensional data; see Boulesteix and Strimmer (2007) for a review. It performs successive matrix decompositions of $\boldsymbol{X}$ and $\boldsymbol{Z}$ into new variables (called component scores or latent variables), denoted by $\xi_1, \ldots, \xi_H$ for the X-scores and $\omega_1, \ldots, \omega_H$ for the Z-scores. These scores should be few in number (H small) and orthogonal to each other within each dataset. They are estimated as linear combinations of the original variables in $\boldsymbol{X}$ and $\boldsymbol{Z}$, with their weight coefficients stored in the associated so-called loading vectors $\boldsymbol{u}_h$ and $\boldsymbol{v}_h$, $h = 1, \ldots, H$. In a matrix representation, we have

$$\boldsymbol{X} = \boldsymbol{\Xi}\boldsymbol{C}^{\mathrm{T}} + \boldsymbol{F}_{\mathrm{X}}, \quad \boldsymbol{Z} = \boldsymbol{\Omega}\boldsymbol{E}^{\mathrm{T}} + \boldsymbol{F}_{\mathrm{Z}},$$

where $\boldsymbol{F}_X$ and $\boldsymbol{F}_Z$ are the residual matrices and where the $(h+1)$th columns of $\boldsymbol{C}$ and $\boldsymbol{E}$ contain, respectively, the coefficients from the simple regressions of each column of the current deflated matrices $\boldsymbol{X}_h = \boldsymbol{X}_{h-1} - \xi_h \boldsymbol{c}_h^{\mathrm{T}}$ and $\boldsymbol{Z}_h = \boldsymbol{Z}_{h-1} - \omega_h \boldsymbol{e}_h^{\mathrm{T}}$ onto the score vectors ξ_{h+1} and ω_{h+1}.

PLS relates both matrices by maximizing the covariance between each pair of scores ($\xi_h = \boldsymbol{X}_{h-1}\boldsymbol{u}_h, \omega_h = \boldsymbol{Z}_{h-1}\boldsymbol{v}_h$):

$$\mathrm{argmax}_{||\boldsymbol{u}_h||_2=||\boldsymbol{v}_h||_2=1} \ \mathrm{Cov}(\boldsymbol{X}_h\boldsymbol{u}_h, \boldsymbol{Z}_h\boldsymbol{v}_h), \quad h = 1, \ldots, H. \qquad (1)$$

This PLS form is often referred to as 'PLS mode A' in the literature (Vinzi *et al.*, 2010) where, similarly to CCA, the relationship between the two datasets is symmetric. A variant is an asymmetric way ('PLS2', Wegelin, 2000; Wold *et al.*, 1983) of deflating $\boldsymbol{Z}$ and in this case the model consequently differs: $\boldsymbol{Z} = \boldsymbol{\Xi}\boldsymbol{D}^{\mathrm{T}} + \boldsymbol{F}_Z$, where $\boldsymbol{F}_Z$ is a residual matrix and where the $(h+1)$th column in $\boldsymbol{D}$ contains the coefficients from the local regressions of each column of the current deflated matrix $\boldsymbol{Z}_h = \boldsymbol{Z}_{h-1} - \xi_h \boldsymbol{d}_h^{\mathrm{T}}$ onto the score vector ξ_{h+1}.

2.2.2 Sparse PLS

The sPLS enables variable selection from both sets by including L_1 penalizations on both $\boldsymbol{u}_h$ and $\boldsymbol{v}_h$ simultaneously in (1), which is solved with a Lagrangian form (see Lê Cao *et al.* 2008, 2009). The result is a subset of correlated variables from both $\boldsymbol{X}$ and $\boldsymbol{Z}$ indicated through the non-zero elements of the loading vectors $\boldsymbol{u}_h$ and $\boldsymbol{v}_h$, respectively (for each PLS dimension h) and a set of score vectors (ξ_h, ω_h) which are useful for graphical representations. Let us consider the singular value decomposition of the r-rank matrix $\boldsymbol{M} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z}$: $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Delta}\boldsymbol{V}^{\mathrm{T}}$, where $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r] : p \times r$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r] : q \times r$ are orthonormal and where $\boldsymbol{\Delta}$ is a diagonal matrix containing the singular values δ_k. The column vectors of $\boldsymbol{U}$ and $\boldsymbol{V}$ are the PLS loadings of $\boldsymbol{X}$ and $\boldsymbol{Z}$. It is worth noting that $(\boldsymbol{u}_1, \boldsymbol{v}_1)$ is also solution of (1) when $h = 1$. Moreover, by Eckart–Young's theorem, these two vectors can also be obtained by first solving (for $\tilde{\boldsymbol{u}}$ and $\tilde{\boldsymbol{v}}$, without a norm constraint) the minimization problem

$$\min_{\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{v}}} ||\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z} - \tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}^{\mathrm{T}}||_F^2 = ||\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z} - \tilde{\boldsymbol{u}}_1\tilde{\boldsymbol{v}}_1^{\mathrm{T}}||_F^2 = ||\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z} - \delta_1\boldsymbol{u}_1\boldsymbol{v}_1^{\mathrm{T}}||_F^2,$$

followed by a norming step of the vectors found. We thus have $(\boldsymbol{u}_1, \boldsymbol{v}_1) = (\tilde{\boldsymbol{u}}_1/||\tilde{\boldsymbol{u}}_1||_2, \tilde{\boldsymbol{v}}_1/||\tilde{\boldsymbol{v}}_1||_2)$. This is equivalent to solve

$$\min_{||\tilde{\boldsymbol{u}}||_2=1, \tilde{\boldsymbol{v}}} ||\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z} - \tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}^{\mathrm{T}}||_F^2 \qquad (\text{respectively}, \min_{\tilde{\boldsymbol{u}}, ||\tilde{\boldsymbol{v}}||_2=1} ||\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Z} - \tilde{\boldsymbol{u}}\tilde{\boldsymbol{v}}^{\mathrm{T}}||_F^2)$$

followed by norming $\tilde{\boldsymbol{v}}$ (respectively, $\tilde{\boldsymbol{u}}$). To obtain sPLS loadings, Lê Cao *et al.* (2008) followed this idea, similar to the one implemented by Shen and Huang (2008) to develop sparse PCA. In sPLS, one tries to optimize

$$\min_{\boldsymbol{u}_h, \boldsymbol{v}_h} ||\boldsymbol{M}_h - \boldsymbol{u}_h\boldsymbol{v}_h^{\mathrm{T}}||_F^2 + P_{\lambda_{1,h}}(\boldsymbol{u}_h) + P_{\lambda_{2,h}}(\boldsymbol{v}_h), \qquad (2)$$

using an iterative algorithm (see Supplementary Material) in which at each iteration, $\boldsymbol{u}_h$ (respectively, $\boldsymbol{v}_h$) is alternatively fixed, while $\boldsymbol{v}_h$ (respectively, $\boldsymbol{u}_h$) is constrained to be of unit-norm and where $\boldsymbol{M}_h = (m_{ij,h})_{i,j} = \boldsymbol{X}_h^{\mathrm{T}}\boldsymbol{Y}_h$, $\boldsymbol{u}_h = (u_{i,h})_i$ and $\boldsymbol{v}_h = (v_{j,h})_j$, $h = 1, \ldots, H$. The penalizations $P_{\lambda_{1,h}}(\boldsymbol{u}_h) = \sum_{i=1}^{p} 2\lambda_1^h|u_{i,h}|$ and $P_{\lambda_{2,h}}(\boldsymbol{v}_h) = \sum_{j=1}^{q} 2\lambda_2^h|v_{j,h}|$ are introduced to penalize the loading vectors $\boldsymbol{u}_h$ and $\boldsymbol{v}_h$. This procedure leads to normed sparse loading vectors.

2.3 Group PLS and sparse group PLS

2.3.1 Group PLS

Let us consider that both matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ can be divided, respectively, into K and L sub-matrices (groups) $\boldsymbol{X}^{(k)} : n \times p_k$ and $\boldsymbol{Z}^{(l)} : n \times q_l$, where p_k (respectively, q_l) is the number of covariates in group k (respectively, l). For example, for gene expression data, these sub-matrices may be gene pathways or factor level indicators in categorical data. The aim is to select only a few groups of $\boldsymbol{X}$ which are related to a few groups of $\boldsymbol{Z}$. For each dimension h, following the same idea as in (Yuan and Lin, 2006), we propose to use group lasso penalties in the optimization problem (2):

$$P_{\lambda_1}(\boldsymbol{u}) = \lambda_1 \sum_{k=1}^{K} \sqrt{p_k}||\boldsymbol{u}^{(k)}||_2 \quad \text{and} \quad P_{\lambda_2}(\boldsymbol{v}) = \lambda_2 \sum_{l=1}^{L} \sqrt{q_l}||\boldsymbol{v}^{(l)}||_2,$$

where $\boldsymbol{u}^{(k)}$ (respectively, $\boldsymbol{v}^{(l)}$) is the loading vector associated to the kth (respectively, lth) block. The subscript h has been removed to improve readability. The minimization criterion (2) can thus be rewritten as

$$\sum_{k=1}^{K}\sum_{l=1}^{L} ||\boldsymbol{M}^{(k,l)} - \boldsymbol{u}^{(k)}\boldsymbol{v}^{(l)^{\mathrm{T}}}||_F^2 + P_{\lambda_1}(\boldsymbol{u}) + P_{\lambda_2}(\boldsymbol{v}), \qquad (3)$$

where $\boldsymbol{M}^{(k,l)} = \boldsymbol{X}^{(k)}\boldsymbol{Z}^{(l)^{\mathrm{T}}}$. Next, we discuss optimization over $\boldsymbol{u}$ for a fixed $\boldsymbol{v}$. The minimization criterion (3) can be rewritten as

$$\sum_{k=1}^{K} \left\{ ||\boldsymbol{M}^{(k,\cdot)} - \boldsymbol{u}^{(k)}\boldsymbol{v}^{\mathrm{T}}||_F^2 + \lambda_1\sqrt{p_k}||\boldsymbol{u}^{(k)}||_2 \right\} + P_{\lambda_2}(\boldsymbol{v}),$$

where $\boldsymbol{M}^{(k,\cdot)} = \boldsymbol{X}^{(k)}\boldsymbol{Z}^{\mathrm{T}}$. Therefore, we can optimize over group wise components of $\boldsymbol{u}$ separately. The first term in the above equation expands as

$$\mathrm{trace}[\boldsymbol{M}^{(k,\cdot)}\boldsymbol{M}^{(k,\cdot)^{\mathrm{T}}}] - 2\mathrm{trace}[\boldsymbol{u}^{(k)}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{M}^{(k,\cdot)^{\mathrm{T}}}] + \mathrm{trace}[\boldsymbol{u}^{(k)}\boldsymbol{u}^{(k)^{\mathrm{T}}}].$$

Hence, the optimal $\boldsymbol{u}^{(k)}$ minimizes:

$$\mathrm{trace}[\boldsymbol{u}^{(k)}\boldsymbol{u}^{(k)^{\mathrm{T}}}] - 2\mathrm{trace}[\boldsymbol{u}^{(k)}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{M}^{(k,\cdot)^{\mathrm{T}}}] + \lambda_1\sqrt{p_k}||\boldsymbol{u}^{(k)}||_2.$$

The objective function is convex, so the optimal solution is characterized by subgradient equations. For group k, $\boldsymbol{u}^{(k)}$ must satisfy

$$-2\boldsymbol{u}^{(k)} + 2\boldsymbol{M}^{(k,\cdot)}\boldsymbol{v} = \lambda_1\sqrt{p_k}\theta, \qquad (4)$$

where θ is the subgradient of $||\boldsymbol{u}^{(k)}||_2$ evaluated at $\boldsymbol{u}^{(k)}$. So,

$$\theta = \begin{cases} \dfrac{\boldsymbol{u}^{(k)}}{||\boldsymbol{u}^{(k)}||_2} & \text{if } \boldsymbol{u}^{(k)} \neq 0; \\ \in \{\theta : ||\theta||_2 \leq 1\} & \text{if } \boldsymbol{u}^{(k)} = 0. \end{cases}$$

We can see that subgradient Equation (4) is satisfied with $\boldsymbol{u}^{(k)} = 0$ if

$$||\boldsymbol{M}^{(k,\cdot)}\boldsymbol{v}||_2 \leq 2^{-1}\lambda_1\sqrt{p_k}. \qquad (5)$$

For $\boldsymbol{u}^{(k)} \neq 0$, Equation (4) gives

$$-2\boldsymbol{u}^{(k)} + 2\boldsymbol{M}^{(k,\cdot)}\boldsymbol{v} = \lambda_1\sqrt{p_k}\frac{\boldsymbol{u}^{(k)}}{||\boldsymbol{u}^{(k)}||_2}. \qquad (6)$$

Combining Equations (5) and (6), we find:

$$\boldsymbol{u}^{(k)} = \left(1 - \frac{\lambda_1}{2}\frac{\sqrt{p_k}}{||\boldsymbol{M}^{(k,\cdot)}\boldsymbol{v}||_2}\right)_+ \boldsymbol{M}^{(k,\cdot)}\boldsymbol{v}. \qquad (7)$$

In the same vein, optimization over $\boldsymbol{v}$ for a fixed $\boldsymbol{u}$ is also obtained by optimizing over group wise components:

$$\boldsymbol{v}^{(l)} = \left(1 - \frac{\lambda_2}{2}\frac{\sqrt{q_l}}{||\boldsymbol{M}^{(\cdot,l)^{\mathrm{T}}}\boldsymbol{u}||_2}\right)_+ \boldsymbol{M}^{(\cdot,l)^{\mathrm{T}}}\boldsymbol{u}. \qquad (8)$$

The iterative procedure for gPLS is similar to that of sPLS. Only the steps (1) and (2) in 2.c. of the algorithm (see Supplementary Material) are modified:

1. For k in $1,\ldots,K$: apply (7)
   norm $\boldsymbol{u}_{\text{new}}$
2. For l in $1,\ldots,L$: apply (8)
   norm $\boldsymbol{v}_{\text{new}}$

2.3.2 Sparse group PLS

One potential drawback of gPLS is to include a group in the model when all loadings in that group are non-zero. However, sometimes we would like to combine both sparsity of groups and within each group. For example, if the predictor matrix contains genes, we might be interested in identifying particularly important genes in pathways of interest. To achieve this, in the same spirit as Simon *et al.* (2013), we introduce sparse group lasso penalties in the optimization problem (2):

$$P_{\lambda_1}(\boldsymbol{u}) = (1-\alpha_1)\lambda_1\sum_{k=1}^{K}\sqrt{p_k}||\boldsymbol{u}^{(k)}||_2 + \alpha_1\lambda_1||\boldsymbol{u}||_1,$$

$$P_{\lambda_2}(\boldsymbol{v}) = (1-\alpha_2)\lambda_2\sum_{l=1}^{L}\sqrt{q_l}||\boldsymbol{v}^{(l)}||_2 + \alpha_2\lambda_2||\boldsymbol{v}||_1.$$

Next, we discuss optimization over $\boldsymbol{u}$ for a fixed $\boldsymbol{v}$. The minimization criterion can be rewritten as

$$\sum_{k=1}^{K}\left\{||\boldsymbol{M}^{(k,\cdot)} - \boldsymbol{u}^{(k)}\boldsymbol{v}^{\mathrm{T}}||_F^2 + (1-\alpha_1)\lambda_1\sqrt{p_k}||\boldsymbol{u}^{(k)}||_2 + \alpha_1\lambda_1||\boldsymbol{u}^{(k)}||_1\right\} + P_{\lambda_2}(\boldsymbol{v}).$$

Therefore, we can optimize over group wise components of $\boldsymbol{u}$ separately. Hence, the optimal $\boldsymbol{u}^{(k)}$ minimizes:

$$\text{trace}[\boldsymbol{u}^{(k)}\boldsymbol{u}^{(k)^{\mathrm{T}}}] - 2\text{trace}[\boldsymbol{u}^{(k)}\boldsymbol{v}^{\mathrm{T}}\boldsymbol{M}^{(k,\cdot)^{\mathrm{T}}}] + (1-\alpha_1)\lambda_1\sqrt{p_k}||\boldsymbol{u}^{(k)}||_2 + \alpha_1\lambda_1||\boldsymbol{u}^{(k)}||_1.$$

Using similar tools as before, we define the subgradient equations:

$$-2\boldsymbol{u}^{(k)} + 2\boldsymbol{M}^{(k,\cdot)}\boldsymbol{v} = \lambda_1(1-\alpha_1)\sqrt{p_k}\theta + \lambda_1\alpha_1\gamma, \tag{9}$$

where θ is the subgradient of $||\boldsymbol{u}^{(k)}||_1$ evaluated at $\boldsymbol{u}^{(k)}$. So,

$$\gamma_j = \begin{cases} \text{sign}\left(u_j^{(k)}\right) & \text{if } u_j^{(k)} \neq 0; \\ \in \{\gamma_j : |\gamma_j| \leq 1\} & \text{if } u_j^{(k)} = 0. \end{cases}$$

We can see that subgradient Equation (9) is satisfied with $\boldsymbol{u}^{(k)} = 0$ if

$$||g^{\text{soft}}\left(\boldsymbol{M}^{(k,\cdot)}\boldsymbol{v}, \lambda_1\alpha_1/2\right)||_2 \leq \lambda_1(1-\alpha_1)\sqrt{p_k}, \tag{10}$$

where the thresholding function $g^{\text{soft}}(\cdot,\lambda)$ is applied to the vector $\boldsymbol{M}^{(k,\cdot)}\boldsymbol{v}$ componentwise. The subgradient equations can also give insight into the sparsity within a group which is at least partially non-zero. For $\boldsymbol{u}^{(k)} \neq 0$, the subgradient conditions for a particular $u_j^{(k)}$ become

$$-2u_j^{(k)} + 2(\boldsymbol{M}^{(k,\cdot)}\boldsymbol{v})_j = \lambda_1(1-\alpha_1)\sqrt{p_k}\frac{u_j^{(k)}}{||\boldsymbol{u}^{(k)}||_2} + \lambda_1\alpha_1\gamma_j. \tag{11}$$

This is satisfied for $u_j^{(k)} = 0$ if

$$|(\boldsymbol{M}^{(k,\cdot)}\boldsymbol{v})_j| \leq 2^{-1}\lambda_1\alpha_1.$$

For $u_j^{(k)} \neq 0$, we find by that $u_j^{(k)}$ satisfies

$$g^{\text{soft}}\left((\boldsymbol{M}^{(k,\cdot)}\boldsymbol{v})_j, \lambda_1\alpha_1/2\right) = \lambda_1(1-\alpha_1)\sqrt{p_k}\frac{u_j^{(k)}}{||\boldsymbol{u}^{(k)}||_2} + 2u_j^{(k)} \tag{12}$$

by noting that $\text{sign}((\boldsymbol{M}^{(k,\cdot)}\boldsymbol{v})_j) = \text{sign}(u_j^{(k)})$. This equation holds for each element of the vector $\boldsymbol{u}^{(k)}$ and we get, in vector representation:

$$g^{\text{soft}}\left(\boldsymbol{M}^{(k,\cdot)}\boldsymbol{v}), \lambda_1\alpha_1/2\right) = \lambda_1(1-\alpha_1)\sqrt{p_k}\frac{\boldsymbol{u}^{(k)}}{||\boldsymbol{u}^{(k)}||_2} + 2\boldsymbol{u}^{(k)} \tag{13}$$

Taking the L_2 norm of both sides of (13), we get

$$||\boldsymbol{u}^{(k)}||_2 = \frac{1}{2}\left[||g^{\text{soft}}\left(\boldsymbol{M}^{(k,\cdot)}\boldsymbol{v}, \lambda_1\alpha_1/2\right)||_2 - \lambda_1(1-\alpha_1)\sqrt{p_k}\right]. \tag{14}$$

By substituting (14) into (13), we find

$$\boldsymbol{u}^{(k)} = \frac{1}{2}\left[g^{\text{soft}}\left(\boldsymbol{M}^{(k,\cdot)}\boldsymbol{v}, \lambda_1\alpha_1/2\right) - \lambda_1(1-\alpha_1)\sqrt{p_k}\frac{g^{\text{soft}}\left(\boldsymbol{M}^{(k,\cdot)}\boldsymbol{v}, \lambda_1\alpha_1/2\right)}{||g^{\text{soft}}\left(\boldsymbol{M}^{(k,\cdot)}\boldsymbol{v}, \lambda_1\alpha_1/2\right)||_2}\right]. \tag{15}$$

In the same vein, optimization over $\boldsymbol{v}$ for a fixed $\boldsymbol{u}$ is also obtained by optimizing over group wise components. We find that $\boldsymbol{v}^{(l)} = 0$ if

$$||g^{\text{soft}}\left(\boldsymbol{M}^{(\cdot,l)^{\mathrm{T}}}\boldsymbol{u}, \lambda_2\alpha_2/2\right)||_2 \leq \lambda_2(1-\alpha_2)\sqrt{q_l} \tag{16}$$

and

$$\boldsymbol{v}^{(l)} = \frac{1}{2}\left[g^{\text{soft}}\left(\boldsymbol{M}^{(\cdot,l)^{\mathrm{T}}}\boldsymbol{u}, \lambda_1\alpha_1/2\right) - \lambda_2(1-\alpha_2)\sqrt{q_l}\frac{g^{\text{soft}}\left(\boldsymbol{M}^{(\cdot,l)^{\mathrm{T}}}\boldsymbol{u}, \lambda_2\alpha_2/2\right)}{||g^{\text{soft}}\left(\boldsymbol{M}^{(\cdot,l)^{\mathrm{T}}}\boldsymbol{u}, \lambda_2\alpha_2/2\right)||_2}\right] \tag{17}$$

otherwise. The iterative procedure for sgPLS is similar to that of sPLS. Only the steps (1) and (2) in 2.c. of the algorithm (see Supplementary Material) are modified:

- For k in $1,\ldots,K$: if (10) is true $\boldsymbol{u}_{\text{new}} = \boldsymbol{0}$ else apply (15)
  norm $\boldsymbol{u}_{\text{new}}$
- For l in $1,\ldots,L$: apply (16) is true $\boldsymbol{v}_{\text{new}} = \boldsymbol{0}$ else apply (17)
  norm $\boldsymbol{v}_{\text{new}}$

2.3.3 Remark

Calibration of the different tuning parameters is discussed in the Supplementary Material.

2.3.4 Extension to discriminant analysis of one dataset

PLSs has often been used for discrimination problems (Nguyen and Rocke 2002) by recoding the qualitative response as a dummy block matrix $\boldsymbol{Y} : n \times g$ indicating the group of each sample (g being the number of groups). Barker and Rayens (2003) give some theoretical justification for this approach. One can also directly apply PLS regression on the data as if $\boldsymbol{Y}$ was a continuous matrix (from now on called PLS-DA). A sparse version has been proposed by Lê Cao *et al.* (2011a, b) using the Lagrangian form of PLS-DA to include a L_1 constraint. Our algorithm for the gPLS-DA is obtained by replacing $\boldsymbol{Z}$ with $\boldsymbol{Y}$ and by using only a group penalty on the loading related to the $\boldsymbol{X}$ matrix ($\lambda_2 = 0$ in Equation 3). Algorithm for sgPLS-DA is also obtained by replacing $\boldsymbol{Z}$ with $\boldsymbol{Y}$ and by using only penalties on

the loading related to the $\boldsymbol{X}$ matrix ($\lambda_2 = 0$ and $\alpha_2 = 0$ in Equation 3). The tuning parameters (λ_1^h and α_1^h) are calibrated sequentially by evaluating the classification error rate using leave-one-out or k-folds cross-validation.

3 Results and discussion

A simulation study is performed to demonstrate the good performance of gPLS and sgPLS when compared with sPLS. Then, the three methods are applied on an HIV vaccine evaluation study.

3.1 Simulation study

Different simulation studies are performed to investigate the detection power of gPLS and sgPLS when group effect exists. We compare the results with the popular sPLS method under several conditions such as different values of the standard deviation of noise, the number of groups associated with the multivariate responses and different sample sizes. We focus on the regression mode of PLS. For all cases described below, we generated $p = 400$ variables stored in the dataset $\boldsymbol{X}$, which was divided into G_X groups. We considered two situations for the q variables of dataset $\boldsymbol{Z}$: (*i*) $q = 10$ variables which are not divided into groups and (*ii*) $q = 500$ variables which are divided into G_Z groups. The link between $\boldsymbol{X}$ and $\boldsymbol{Z}$ is specified by the following models:

$$\boldsymbol{X} = \boldsymbol{\Xi}\boldsymbol{C}^{\mathrm{T}} + \boldsymbol{F}_{\mathrm{X}}, \qquad \boldsymbol{Z} = \boldsymbol{\Xi}\boldsymbol{D}^{\mathrm{T}} + \boldsymbol{F}_{\mathrm{Z}},$$

where the $(n \times H)$ matrix $\boldsymbol{\Xi}$ contains H latent variables $\xi_1, \ldots, \xi_H$. The components of these vectors have all been independently generated from a standard normal distribution. The rows of the residual matrix $\boldsymbol{F}_X$ (respectively, $\boldsymbol{F}_Z$) have been generated from a multivariate normal distribution with zero mean μ_X (respectively, μ_Z) and covariance matrix $\boldsymbol{\Sigma}_{\mathrm{X}} = \sigma \boldsymbol{I}_p$ (respectively, $\boldsymbol{\Sigma}_{\mathrm{Z}} = \sigma \boldsymbol{I}_q$). The orthogonal matrix of regression coefficients $\boldsymbol{C} = [\boldsymbol{C}_1, \ldots, \boldsymbol{C}_H]$ enables us to specify the 'true' (i.e. active) X-variables linked to the response Z-variables. The p-vector $\boldsymbol{C}_j = (c_j^1, \ldots, c_j^p)^{\mathrm{T}}$ includes non-zero values $c_j^i \neq 0$ if the corresponding variable $\boldsymbol{X}^i$ (jth column of $\boldsymbol{X}$) is a true variable (i.e. associated to one of the latent variables ξ_h) and zero values otherwise. The matrix orthogonal $\boldsymbol{D} = [\boldsymbol{D}_1, \ldots, \boldsymbol{D}_H]$ enables us to specify the association of the response variables to each latent variable. All components of the column vectors $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_H$ have been independently generated from a uniform distribution $\mathcal{U}[0.5, 2]$ when the dataset $\boldsymbol{Z}$ is not divided into groups. When the dataset $\boldsymbol{Z}$ is divided into groups, the column vectors $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_H$ include non-zero values $d_j^i \neq 0$ if the corresponding variable $\boldsymbol{Z}^j$ (jth column of $\boldsymbol{Z}$) is a true variable and zero values otherwise. Matrix $\boldsymbol{C}$ (respectively, $\boldsymbol{D}$) should be made orthogonal. To understand and compare the behavior of all methods, we used an optimal value for tuning parameters. Thus, the sPLS tuning parameter is set to the number of true variables, and the gPLS and sgPLS tuning parameter relative to λ is set to the number of active groups. The mixing parameter α in the sgPLS approach is selected using a 5-folds cross-validation on a grid of 15 values between 0.05 and 0.95 since it is not possible to set an optimal value for α.

We evaluate the performance of each method by presenting the true-positive rate (TPR) which reflects the number of correctly identified true variables and the total discordance (TD) which is the number of incorrectly identified variables. These measures are defined as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad \mathrm{TD} = \mathrm{FP} + \mathrm{FN},$$

where TP, FP and FN are the number of true positives, false positives and false negatives, respectively.

3.2 Recovering the signal

The first simulation corresponds to the case when the $q = 10$ variables from dataset $\boldsymbol{Z}$ have not been divided into groups. The sample size is $n = 100$ and we consider a model with only one latent variable ξ_1. Dataset $\boldsymbol{X}$ is divided into $G_X = 20$ groups of 20 variables. Among them, only 4 groups each containing 15 true variables and 5 noise variables are associated to the response variables $\boldsymbol{Z}$. We set the p-vector $\boldsymbol{C}_1$ to have 15 1s, 30 -1s and 15 1.5s, the other components being set to 0. All 15 non-zero coefficients are assigned randomly into one group along with the remaining 5 zero coefficients corresponding to noise variables (see top left panel of Supplementary Fig. S1). In the top four panels of Supplementary Figure S1, the standard deviation σ has been set to 0.5. A standard deviation $\sigma = 2.5$ is set for the other panels. Supplementary Figure S1 displays the results of the loading vectors $\boldsymbol{u}_1$ recovered by sPLS, gPLS and sgPLS. Notice that sPLS and sgPLS recover satisfactorily the true $\boldsymbol{u}_1$ when the noise is small ($\sigma = 0.5$). However, when $\sigma = 2.5$, sPLS selects more noise variables than the true $\boldsymbol{u}_1$ and misses out some true variables while sgPLS still works well. For both situations, gPLS recovers all groups but gives false positives within the groups. We arrive at the same conclusion when the $q = 500$ variables of $\boldsymbol{Z}$ are divided into $G_Y = 25$ groups (see Supplementary Figs S2–S5). In this situation, we consider a model with two latent variables ξ_1 and ξ_2. The vector $\boldsymbol{C}_2$ is chosen in the same way as previously described for $\boldsymbol{C}_1$. The two columns of $\boldsymbol{D}$ are q-vectors containing 15 -1s, 15 -1.5s and 30 1s and the rest are 0 s. For a small standard deviation $\sigma = 0.5$ (Supplementary Figs S2–S3), all methods perform well to recover the loading vectors $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ related to the first component and the loading vectors $\boldsymbol{u}_2$ and $\boldsymbol{v}_2$ related to the second component. However, gPLS gives false positives within the groups. For a high standard deviation $\sigma = 2.5$ (Supplementary Figs S4–S5), sPLS selects more noise variables, while sgPLS still performs well by selecting only relevant variables into the true groups.

3.3 Effect of noise

We investigate the performance of our methods with respect to noise. This simulation corresponds to the situation when the $q = 10$ variables from dataset $\boldsymbol{Z}$ have not been divided into groups and when one latent variable has been generated. We use a sequence of 10 equispaced standard deviation values σ, from 1.5 to 3. Supplementary Figure S7 presents the average of TPR and TD over 50 replications for each method. For all noise levels, we found that gPLS and sgPLS manage to recover the true groups. Thus, TD for gPLS is equal to 20 as each of the 4 true groups contains 5 noise variables (FP). However, for high values of noise, the sparsity introduced within the group by sgPLS misses out some true variables. Note that sPLS also misses out true variables when the noise is increased and gets the highest TD due to the high number of false positives selected. We present in Supplementary Figure S6 the average, over 50 replications, of the signal recovered (original loading $\boldsymbol{u}_1$) by the three methods when $\sigma = 3$. This figure highlights the number of false positives selected by sPLS.

3.4 Effect of the number of true variables in the group

We investigate the performance of our methods when the number of true variables within each group is varied. We are still in the framework of one latent variable generated and only the $p = 400$ variables

from dataset $\boldsymbol{X}$ have been divided into groups. Among them, 40 variables are associated to the latent variable. Dataset $\boldsymbol{X}$ contains $G_X = 40$ groups, each with the same group size of 10. We set the sample size $n = 100$ and the standard deviation of the noise $\sigma = 2$. The number of groups containing true variables is varied in $\{4, 5, 8, 10, 20, 40\}$ while each group contains 10, 8, 5, 4, 2 and 1 true variables, respectively. Supplementary Figure S8 presents the average results over 50 replications. When true variables are distributed in 4 and 5 groups, both gPLS and sgPLS give a much higher value of TPR and a lower value of TD than sPLS. When the true variables are more sparsely distributed into different groups (number of groups increased), gPLS still gives high TPR but TD is increased. Note that sgPLS is the best method with high TPR and low TD values.

3.5 Effect of sample size

We investigate the performance of our methods when the sample size is increased by steps of 50 from $n = 50$ to $n = 500$. Dataset $\boldsymbol{X}$ contains $G_X = 40$ groups, each with the same group size of 10. Among the $p = 400$ variables in dataset $\boldsymbol{X}$, 60 are true variables, which are distributed evenly into 6 groups. This clearly advantages gPLS. Supplementary Figure S9 presents the average results over 50 replications of TPR and TD with respect to different sample sizes. It can be seen that gPLS gives better results than the two other methods by finding the true groups whatever the sample size. For both sPLS and sgPLS, TPR increases while TD decreases when sample size increases. The TPR of sPLS is most of the time lower than that of sgPLS. Moreover, the TD of sPLS is higher than both those of gPLS and sgPLS.

3.6 Effect of the dimension p

We investigate the performance of our methods when the dimension p of $\boldsymbol{X}$ is increasing ($p \in \{500, 1000, 2000, 3000, 4000\}$).Dataset $\boldsymbol{X}$ contains $G_X = p/50$ groups, each with the same group size of 50. Among the G_X groups, $p/100$ groups contains 35 true variables. This simulation corresponds to the situation when the $q = 10$ variables from dataset $\boldsymbol{Z}$ have not been divided into groups and when one latent variable has been generated. We set different sample size $n = 25,\ 50$ and 100. The standard deviation of the noise is $\sigma = 3.5$. Supplementary Table S4 presents the average results over 50 replications of TPR, false-positive rate (FPR) and TD for the different values of p and for different sample sizes. In this simulation setting gPLS manages to find the true groups whatever the dimension p and the sample size. As 15 noise variables are included in each true group, gPLS method gets higher FPR and TD than sgPLS and sPLS. Almost every time, sgPLS outperforms sPLS by getting higher TPR and lower TD and FPR.

3.7 Application for an HIV vaccine evaluation study

3.7.1 Description of the study

The method has been applied to an HIV vaccine trial: the DALIA trial (Lévy *et al.*, 2014). In this trial, 19 HIV-infected patients have been included for an evaluation of the safety and the immunogenicity of a dendritic-cell-based vaccine. The vaccine was injected on weeks 0, 4, 8 and 12, while patients received an antiretroviral therapy. An interruption of the antiretrovirals was performed at week 24 and the treatment was resumed if the CD4 + T cell count dropped below 350 cells/ÂμL. After vaccination, a deep evaluation of the immune response was performed at week 16 while repeated measurements of the main immune markers and gene expression were performed every 4 weeks until the end of the trial at 48 weeks. The analyses of the immune markers showed a strong immunological response especially among the CD4 + T cell populations with an increase of cells producing various cytokines. After the antiretroviral treatment interruption, there was a strong viral replication that was heterogeneous among the study population and the immune response appeared to be associated with the observed maximum value of viral load during the rebound (Lévy *et al.*, 2014). One of the question that follows these results was whether the immune response and its impact on the viral dynamics could be explained and predicted with the observation of the change of gene expression during vaccination. To answer this question, we first analyzed the repeated measurements of gene abundance performed by microarrays (Illumina HumanHT-12 v4). Because of the sparsity of the measurements, only geneset analyses allow to detect significant changes over time of group of genes. Using previously defined group of genes, so called modules (Chaussabel *et al.*, 2008), we reported a significant change of gene expression among 69 modules over time before antiretroviral treatment interruption. Then, we asked how the gene abundance of the 5399 genes (p) from these 69 modules as measured at week 16 correlated with immune markers measured at the same time point. The immune markers were either direct measurement of cytokine concentrations on supernatants (IL21, IL2, IL13, IFNg) or a combination of markers as measured by multiplex and intracellular staining (Luminex score, TH1 score, CD4 polyfunctionnality). Seven (q) different immune markers have been used in the current application. Measurements and calculations are detailed elsewhere (Lévy *et al.*, 2014). Of note, the modules as defined in Chaussabel *et al.* (2008) were not overlapping: each gene contributed to only one module. To use the present approach with overlapping groups of genes (such as in Gene Ontology), an extension in the spirit of Jacob *et al.* (2009) would be required.

3.7.2 Results

The selection process of the genes according to the mean square error of prediction criteria for each method is shown in Supplementary Figures S11–S13. Supplementary Table S5 shows the cumulative percentage of variation of the responses variables explained by the components according to the method. With three components, more than 80% of the variance could be explained. The sgPLS methods selected slightly more genes than the sPLS (respectively, 487 and 420 genes selected), but sgPLS selected fewer modules than the sPLS (respectively, 21 and 64 groups of genes selected). Of note, all the 21 groups of genes selected by the sgPLS were included in those selected by the sPLS method. sgPLS selected slightly more modules than gPLS (4 more, 14/21 in common). However, gPLS led to more genes selected than sgPLS (944). Supplementary Figure S10 displays Venn diagrams of the selected sets by the three methods, both at the gene and module levels. Therefore, in this context of hierarchical data, sgPLS ends up being more parsimonious than sPLS in term of modules and than gPLS in term of genes by taking the most predictive genes within a selected module (see Supplementary Table S6). Such results are consistent with those of the simulation study. The selection of the most predictive genes inside a module improves the prediction capacity and does not avoid any interpretation of the biological significance of the remaining genes inside a module based on the biological knowledge available for this module.

Interestingly, the modules commonly selected by the three approaches were the most biologically relevant ones, such as the modules associated to inflammation (M3.2, M4.13, M4.2, M4.6, M7.1) and cellular responses (M3.6 cytotoxic/NK cell, 4.15 on

T cells). A significant number of additional modules have been selected by sPLS but not by the other methods. Although, some are biologically sound (e.g. M3.1 Erythrocytes), many were not annotated or are unrelated to the immune system making the hypothesis of false signals likely. Regarding the modules differentially selected by sgPLS and gPLS, three modules were selected by gPLS but not by sgPLS (M4.11 Plasma cells and undetermined modules 4.8 and 7.35), whereas seven modules were selected by sgPLS and not by gPLS (M3.5 and M4.7 Cell cycle, M4.1 T cell, M5.1 and M5.7 inflammation, M6.7 and M5.2 undetermined). Clearly, the selection by sgPLS sounds more biologically relevant as M4.1 is known to be associated to the other T-cell module M4.15 that was selected by both methods (see www.biir.net/public_wikis/module_annotation/V2_Trial_8_Modules). This is the same for the inflammatory modules M5.1 and M5.7 that are known to be related. In regards of the module M4.11 (plasma cells), we have already notice that its selection was not robust in bootstrap analyses performed with sPLS (data not shown). Therefore, in this application, the sgPLS approach led to a parsimonious selection of modules and genes that sound very relevant biologically and is in agreement with the simulation results.

An illustration of the results is shown through the correlation matrix in Supplementary Figure S14 according to the genes selected in common by the three methods. Genes and modules negatively or positively associated to the immune response appeared very clearly. As expected, inflammatory modules were negatively correlated to the immune response whereas modules related to the cellular immune response (M3.6, M4.15) were positively correlated to the immune response.

Stability of the module selection has been assessed for sPLS, gPLS and sgPLS (see Supplementary Fig. S15) in the spirit of Bach (2008), Meinshausen and Bühlmann (2010) and Lê Cao *et al.* (2011a, b). It highlights the insights gained from incorporating the grouping structure into the analysis, comforting the above biological conclusions. Also, it reveals the instability of the selection of the module 4.11 by sPLS and gPLS, module that was not selected by sgPLS. Furthermore, we compared the proposed novel approach with the multivariate lasso implemented in `glmnet` R package (Friedman *et al.*, 2010) and a multivariate (sparse) group lasso proposed by Li *et al.* (2015). The two alternative approaches selected less modules and genes but the modules selected were most often the same across the different methods (see Supplementary Materials, Section S1.7).

Funding

This research was supported by the NSERC of Canada (to P.L.d.M.). The second author also thanks the GENES.

Conflict of Interest: none declared.

References

Allen,G.I. *et al.* (2013) Regularized partial least squares with an application to NMR spectroscopy. *Stat. Anal. Data Mining*, **6**, 302–314.

Allen,G.I. *et al.* (2014) A generalized least-square matrix decomposition. *J. Am. Stat. Assoc.*, **109**, 145–159.

Bach,F. (2008) Bolasso: model consistent Lasso estimation through the bootstrap. In: Cohen,W.W. *et al.* (eds) *ICML '08 Proceedings of the 25th International Conference on Machine Learning*, pp. 33–40.

Barker,M. and Rayens,W. (2003) Partial least squares for discrimination. *J. Chemometrics*, **17**, 166–173.

Boulesteix,A. and Strimmer,K. (2007) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief. Bioinform.*, **8**, 32.

Chaussabel,D. *et al.* (2008) A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity*, **29**, 150–164.

Chen,J. *et al.* (2013) Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, **14**, 244–258.

Chen,X. and Liu,H. (2012) An efficient optimization algorithm for structured sparse CCA, with applications to eqtl mapping. *Stat. Biosci.*, **4**, 3–26.

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

Friedman,J. *et al.* (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.

Garcia,T.P. *et al.* (2014) Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data. *Bioinformatics*, **30**, 831–837.

Huang,J. and Zhang,T. (2010) The benefit of group sparsity. *Ann. Stat.*, **38**, 1978–2004.

Jacob,L. *et al.* (2009) Group lasso with overlap and graph lasso. In: *Proceedings of the 26th Annual International Conference on Machine Learning ICML '09* ACM, New York, pp. 433–440.

Jayawardana,K. *et al.* (2015) Determination of prognosis in metastatic melanoma through integration of clinico-pathologic, mutation, mRNA, microRNA, and protein information. *Int. J. Cancer*, **136**, 863–874.

Kitano,H. (2002) Computational systems biology. *Nature*, **420**, 206–210.

Lê Cao,K.A. *et al.* (2008) Sparse PLS: variable selection when integrating omics data. *Stat. Appl. Mol. Biol.*, **7**, 37.

Lê Cao,K.A. *et al.* (2009) Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, **10**, 1–17.

Lê Cao,K. *et al.* (2011a) Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, **12**, 1–16.

Lê Cao,K.A. *et al.* (2011b) Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, **12**, 253.

Lévy,Y. *et al.* (2014) Dendritic cell-based therapeutic vaccine elicits polyfunctional HIV-specific T-cell immunity associated with control of viral load. *Eur. J. Immunol.*, **44**, 2802–2810.

Li,Y. *et al.* (2015) Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, **71**, 354–363.

Lin,D. *et al.* (2013) Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinformatics*, **14**, 245.

Löfstedt,T. *et al.* (2014) Structured variable selection for generalized canonical correlation analysis. In: *PLS 2014 8th International Conference on Partial Least Squares and Related Methods*, pp. 127–128.

Ma,S. *et al.* (2007) Supervised group lasso with applications to microarray data analysis. *BMC Bioinformatics*, **8**, 60.

Meier,L. *et al.* (2008) The group lasso for logistic regression. *J. R. Stat. Soc. Ser. B Methodol.*, **70**, 53–71.

Meinshausen,N. and Bühlmann,P. (2010) Stability selection. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **72**, 417–473.

Morine,M. *et al.* (2011) Transcriptomic coordination in the human metabolic network reveals links between n-3 fat intake, adipose tissue gene expression and metabolic health. *PLoS Comput. Biol.*, **7**, e1002223.

Nguyen,D. and Rocke,D. (2002) Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18**, 39–50.

Palermo,R.E. *et al.* (2011) Genomic analysis reveals pre- and postchallenge differences in a rhesus macaque aids vaccine trial: insights into mechanisms of vaccine efficacy. *J. Virol.*, **85**, 1099–1116.

Parkhomenko,E. *et al.* (2009) Sparse canonical correlation analysis with application to genomic data integration. *Stat. Appl. Genet. Mol. Biol.*, **8**, 1–36.

Puig,A. *et al.* (2009) A multidimensional shrinkage-thresholding operator. In: *Statistical Signal Processing, 2009. IEEE/SP 15th Workshop on SSP '09*. IEEE, pp. 113–116.

Rose,M. *et al*. (2011) Revisiting the role of organic acids in the bicarbonate tolerance of zinc-efficient rice genotypes. *Funct. Plant Biol*., **38**, 493–504.

Shen,H. and Huang,J.Z. (2008) Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivariate Anal*., **99**, 1015–1034.

Simon,N. and Tibshirani,R. (2012) Standardization and the group lasso penalty. *Stat. Sin*., **22**, 983–1001.

Simon,N. *et al*. (2013) A sparse-group lasso. *J. Comput. Graph. Stat*., **22**, 231–245.

Subramanian,A. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Tenenhaus,A. and Tenenhaus,M. (2011) Regularized generalized canonical correlation analysis. *Psychometrika*, **76**, 257–284.

Tyekucheva,S. *et al*. (2011) Integrating diverse genomic data using gene sets. *Genome Biol*., **12**, R105.

Vinzi,V. *et al*. (2010) PLS path modeling: from foundations to recent developments and open issues for model assessment and improvement. In: Esposito Vinzi,V. *et al*. (eds.) *Handbook of Partial Least Squares*. Springer Handbooks of Computational Statistics, Springer Berlin Heidelberg, pp. 47–82.

Waaijenborg,S. *et al*. (2008) Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Stat. Appl. Genet. Mol. Biol*., **7**, 1–29.

Wegelin,J.A. (2000). A survey of partial least squares (PLS) methods, with emphasis on the two-block case. *Technical report*. University of Washington.

Witten,D.M. *et al*. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515–534.

Wold,H. (1966) Estimation of principal components and related models by iterative least squares. In: Krishnaiah,P.R. (ed.) *Multivariate Analysis*. Academic Press, New York, pp. 391–420.

Wold,S. *et al*. (1983) The multivariate calibration problem in chemistry solved by the PLS methods. In: Ruhe,A. and Kågstrøm,B. (eds) *Proc. Conf. Matrix Pencils, March 1982, Lecture Notes in Mathematics*. Springer Verlag, Heidelberg, pp. 286–293.

Yuan,M. and Lin,Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **68**, 49–67.

Zhou,H. (2010) Association screening of common and rare genetic variants by penalized regression. *Bioinformatics*, **26**, 2375–2382.