Group-wise Partial Least Square Regression

José Camacho[1,*]

University of Granada

Edoardo Saccenti[2,**]

Wageningen University and Research

Abstract

This paper introduces the Group-wise Partial Least Squares (GPLS) regression. GPLS is a new Sparse PLS (SPLS) technique where the sparsity structure is defined in terms of groups of correlated variables, similarly to what is done in the related Group-wise Principal Component Analysis (GPCA). These groups are found in correlation maps derived from the data to be analyzed. GPLS is especially useful for exploratory data analysis, since suitable values for its metaparameters can be inferred upon visualization of the correlation maps. Following this approach, we show GPLS solves an inherent problem of SPLS: its tendency to confound the data structure as a result of setting its metaparameters using standard approaches for optimizing prediction, like cross-validation. Results are shown for both simulated and experimental data.

Keywords: Sparsity, Partial Least Squares, Sparse Partial Least Squares, Group-wise Principal Component Analysis, Exploratory Data Analysis

1. Introduction

Modern studies are characterized by the generation of large quantity of data such in the case of genomics, proteomics and metabolomics experiments [1]. However it is widely recognized that usually only a limited number of variables or of groups of variables are relevant to the problem being studied [1]. The challenge is to isolate the informative variables from the non-informative part,

*Corresponding author

**Co-corresponding author author

Email addresses: josecamacho@ugr.es (José Camacho), esaccenti@gmail.com (Edoardo Saccenti)

[1]Network Engineering and Security Group, Signal Theory, Networking and Communications Department, University of Granada, C/ Periodista Daniel Saucedo Aranda s/n 18071, Granada, Spain

[2]Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Stippeneng 4, 6708 WE, Wageningen, the Netherlands.

the latter consisting of variables showing random variation, not relevant systematic variation or accounting for technical variability such as sampling and measurement error.

Exploratory data analysis plays a critical role in the analysis of large data sets. One of the most used tools for data exploration is certainly Partial Least Squares (PLS) regression [2, 3] given its ability for dealing with those situations where the number of variables is much larger than the number of observations, which are often encountered in modern experiments. However, in PLS all variables are included in the model and this complicates model interpretability.

One possible solution is to use a PLS approach where some of the model parameters (such as the regression coefficients) are forced to zero. This is the realm of sparse methods, first introduced in the context of multiple linear regression [4]. This idea was later incorporated in PLS regression, usually by filtering out variables from the model obtained from the full data or imposing some sort of penalty (such as the LASSO) on model parameters while building the model. However, the sparsity induced by these approaches is closer in philosophy to variable selection than to exploratory data analysis: the main goal is to improve prediction performance by discarding a subset of predictor variables.

Due to its definition and fitting strategy, Sparse PLS (SPLS) presents several shortcomings that, to the best of our knowledge, have not been discussed elsewhere. First, SPLS shares with PLS a fundamental limitation that complicates the interpretation of the model: the same latent variable (LV) may include several sources of variability while the same source of variability may span the subspace of several LVs [5]. This same problem motivated the introduction of rotation methodologies in Principal Component Analysis. Second, a purely prediction-driven calibration while imposing simplicity may oversimplify the structure in the data, yielding a model where some predictors related to the response are in fact discarded. We will show these problems in the experimental part of this paper. As a result, when the goal is data exploration, sparse methodologies may not be the best choice. This is particular relevant in biological applications where for instance the interest is into retrieving groups of variables (as genes or metabolites) which are related to a given phenotypic trait with the aim of gaining knowledge into molecular mechanisms rather than optimizing prediction performance.

The idea of group-wise simplicity was adopted in the so called simplivariate models [6, 7] which aim to describe informative variation, under the assumption that a given biological or biochemical problem is not represented by all measured variables but only by a few variables or subsets of variables. These models aim to retain the comprehensiveness of a multivariate model together with the simplicity of interpretation of a univariate one. Related extensions of SPLS to group penalties, like the group LASSO, have also been defined [8]. In this approach, the penalties affect to predefined groups of variables, so that solutions are found that activate or deactivate all variables in the groups. From a similar perspective, Group-wise principal component analysis (GPCA) [9] was recently proposed as an exploratory tool. In GPCA, sparsity is defined in terms of groups of correlated variables found in correlation maps obtained from the data

to be analyzed. Each component contains non-zero loadings for a single group of correlated variables, which simplifies the interpretation according to the previous definition of group-wise simplicity. GPCA differs from equivalent group lasso PCA variants in the fact that only one group of variables can be active in a single LV and in the way meta-parameters are tuned. Reference [9] shows the advantages for data interpretation of this approach over sparse and group sparse PCA variants.

In this paper we propose Group-wise PLS, an extension of GPCA in the PLS setting. GPLS aims to model data under a group-wise simplicity approach where every component account for a group of correlated variables that are related with the response. This approach is especially suitable for exploratory data analysis. Like other chemometrics approaches aimed at improving data interpretation, for instance multi-block [10, 11] or orthogonal [12] extensions of PLS, GPLS is not defined to improve the prediction performance of PLS (or SPLS). However, there is a non-negligible link between the prediction capability of a regression model and our confidence on its appropriateness. For this reason, whilst not the main goal, we study the prediction performance of GPLS in comparison to that of PLS and SPLS in the supplementary materials attached to the paper.

The rest of the paper is organized as follows. We begin in Section 2 with a motivation example to illustrate the concept of group-wise sparsity. In Section 3 we introduce the GPLS approach. Section 4 describes the materials and methods used in the experimental sections. The performance of the methods is assessed by means of simulations and on experimental data in Section 5 and 6, respectively. Section 7 presents the conclusions of the work.

2. A motivating example

As a motivating example consider a simulated data set where the response $\mathbf{y}$ is related only to a subset of highly correlated predictors in $\mathbf{X}$, specifically the five first variables (denoted as $\mathbf{X}_{1-5}$ in Equation (1)), while the remaining predictors are not related to the response:

$$\begin{aligned} \mathbf{X}_{1-5} &= 0.1 \cdot \boldsymbol{\Delta} + \mathbf{x}_1 \mathbf{1}_5 \\ \mathbf{y} &= 0.1 \cdot \boldsymbol{\delta} + \sum_{i=1}^{5} \mathbf{x}_i \end{aligned} \tag{1}$$

where values in $\boldsymbol{\Delta}$ and $\boldsymbol{\delta}$ are drawn following a normal distribution with 0 mean and variance 1 and $\mathbf{1}_5$ is a vector of ones of size 1×5.

This is a simple case of group-wise sparsity which is commonly encountered: think for example of NMR spectra where different peaks, corresponding to the same molecule, are expected to exhibit a correlated behavior or when only a small subset of metabolites or genes is related to a phenotypic trait.

The simulated data sets are generated using the *simuleMV* algorithm [13] which allows the simulation of data matrices with different level of correlation

among the predictor variables. Specifically, the algorithm allows the generation of data with a randomly generated covariance/correlation matrix. For each simulation scheme data were simulated with low, medium and high correlation level among the predictor variables (correlation level in *simuleMV* of 5, 7 and 9, respectively). We restricted ourselves to simulate 20×100 data sets for the predictor variables, which is a typical size in chemometrics data sets. The correlation maps with low, medium and high correlation among predictors variables, are shown in Figure 1. A subset of highly correlated variables formed by the first five variables is present in the three correlation maps but the overall correlation structure among the predictors is different in the three cases. In the case of low correlation (Figure 1 panel A) only the first five predictor variables are correlated, while in the high correlation case (Figure 1 panel C) the variables in the group are correlated with many other predictor variables which in turn will be correlated with the response, following a more complex model than the one actually expressed in Equation (1).

Let us focus on the low correlation data set. When a standard PLS regression model with 1 latent variable (LV) is fitted to the data all variables contribute to the prediction (see Figure 2 at the top). The application of a sparse PLS approach (in this case the Sparse PLS method by Lê Cao [14], see Section 4.1) greatly simplifies the regression coefficients (see Figure 2 at the bottom). This sparse PLS model has also greater prediction ability than the standard PLS model ($Q^2 = 0.74$ versus $Q^2 = 0.46$ of the normal PLS model), which complies with the dogma that introducing in the model variables that are not relevant to the response reduces the predicting power of the model. This difference in prediction cannot be corrected by considering more components in the PLS model (not shown).

This example illustrates that *i*) even in very sparse data PLS introduces spurious information in the model and that *ii*) since sparsity is pursued with the aim of enhancing prediction, this can oversimplify the real structure in the data. Since the first five variables in the X-block are highly correlated, only one of them may constitute a good prediction model. In this specific case, due to random variability induced in the data set, the SPLS model with one prediction variable even outperforms a model with the five true predictors. However for data exploration this behavior is not desirable since relevant information is left out the model and this is may also be crucial when biochemical or biological interpretation is pursued.

To overcome this limitation we introduce here the Group-wise partial least square regression (GPLS) where sparsity is introduced in terms of groups of correlated predictor variables related to the response.

3. Group-wise partial least square regression

The group-wise PLS takes as input a set of K (possibly overlapping) groups $S_1, S_2, \ldots S_k, \ldots S_K$ of correlated variables that are obtained from a $M \times M$ correlation map $\mathbf{M}$ computed from the data. In principle, $\mathbf{M}$ can be any square symmetric matrix describing mutual relationships among the M variables. In

the GPLS algorithm we use a missing-data approach to construct this matrix which has the advantage of reducing the noise in the computation of the correlations. Details are given in Section 3.2; the m_{ij} elements of $\mathbf{M}$ are given by Equation (5).

The $S_1, S_2, \ldots S_k, \ldots S_K$ groups of correlated variables are determined using the group identification algorithm (GIA) proposed in [9] and available in the MEDA toolbox [15]. Briefly, let be $m_{i,j} \in [-1, 1]$ the i, j-th element of $\mathbf{M}$ and $|\gamma| < 1$ a threshold on the correlation values. The group S_k is built in such a way that all variables in S_k satisfy the conditions

$$\forall\ i, j \in S_k \rightarrow |m_{ij}| > \gamma \tag{2}$$

and

$$\forall\ j \notin S_k\ /\ \exists\ i \in S_k \rightarrow |m_{ij}| \leq \gamma \tag{3}$$

indicating that if the j-th variable is not in group S_k, it has a correlation magnitude $\leq \gamma$ with at least one of the other variables in the group. This is equivalent to define groups of variables with maximum cardinality where all variables within the group present a correlation larger than γ in absolute value.

3.1. The GPLS algorithm

The GPLS algorithm consists of a set of nested PLS models together with a suitable deflation procedure. Given the data matrices $\mathbf{X}(N \times M)$ and $\mathbf{Y}(N \times O)$, the procedure is based in the first Kernel PLS algorithm proposed in [16], that makes use of matrices $\mathbf{X}$ and $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ and only deflates the latter:

Step 1: Initialize:

$$\mathbf{C} = \mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

$$\mathbf{B} = \mathbf{I}$$

where $\mathbf{I}$ is the identity matrix.

Step 2: For each latent variable (LV) a from 1 to A

Step 2.1: For each group S_k in the set of groups S

Step 2.1.1: Create $\mathbf{C}^k$ from $\mathbf{C}$ setting elements out of S_k to zero.

$$\mathbf{C}^k = \mathbf{C}$$

$$c^k_{lm} = 0,\ \forall l \notin S_k$$

Step 2.1.2: Compute $\mathbf{w}^k$, the first eigenvector of $(\mathbf{C}^k)^T\mathbf{C}^k$.

Step 2.1.3: Compute the corresponding scores as:

$$\mathbf{r}^k = \mathbf{B}\mathbf{w}^k$$

$$\mathbf{t}^k = \mathbf{X}\mathbf{r}^k$$

Step 2.2: Choose the coefficients of latent variable a from the group capturing the most correlation with $\mathbf{Y}$.

$$k^* = \arg\max_k(\text{corr}(\mathbf{t}^k, \mathbf{Y}))$$

$$\mathbf{w}^a = \mathbf{w}^{k^*}, \mathbf{r}^a = \mathbf{r}^{k^*}, \mathbf{t}^a = \mathbf{t}^{k^*}$$

Step 2.3: Perform the deflation steps:

$$\mathbf{q}^a = (\mathbf{r}^a)^T\mathbf{C}/((\mathbf{t}^a)^T\mathbf{t}^a)$$

$$\mathbf{p}^a = (\mathbf{t}^a)^T\mathbf{X}/((\mathbf{t}^a)^T\mathbf{t}^a)$$

$$\mathbf{C} = \mathbf{C} - \mathbf{p}^a(\mathbf{q}^a)^T((\mathbf{t}^a)^T\mathbf{t}^a)$$

$$\mathbf{B} = \mathbf{B}(\mathbf{I} - \mathbf{w}^a(\mathbf{p}^a)^T)$$

Step 3: Obtain the regression coefficients:

$$\mathbf{B}_{gpls} = \mathbf{R}^A(\mathbf{Q}^A)^T$$

The GPLS algorithm first computes the weights and scores of K PLS models of 1 LV, each of them considering only the set of variables corresponding to one of the groups $S_1, S_2, \ldots S_k, \ldots S_K$. From these, it chooses the one capturing the largest correlation with $\mathbf{Y}$, discarding the rest. Using the coefficients of this LV, $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ is deflated and $\mathbf{B}$ recomputed.

3.2. Defining the correlation map

As discussed, we first apply GIA to a given correlation map and then compute the GPLS model using the previous algorithm. Following [9], we use a technique referred to as the missing-data for exploratory data analysis (MEDA) [5] to construct the map. MEDA consists in a post-processing step after the PLS factorization to infer the relationships among variables using missing data imputation [17, 18]. The elements of the MEDA map $\mathbf{M}$ can be expressed as [19]:

$$m_{ij} = \left\{2\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j - (\mathbf{x}_i^A)^{\mathrm{T}}\mathbf{x}_j^A\right\} \cdot \frac{\text{abs}\left\{(\mathbf{x}_i^A)^{\mathrm{T}}\mathbf{x}_j^A\right\}}{\sigma^2_{\mathbf{x}_i}\sigma^2_{\mathbf{x}_j}} \tag{4}$$

where $\sigma^2_{\mathbf{x}_n}$ stands for the variance of the n-th variable in $\mathbf{X}$ and

$$\mathbf{x}_i^A = \mathbf{x}_i\mathbf{R}^A\mathbf{P}^A$$

and $\mathbf{R}^A(M \times A)$ and $\mathbf{P}^A(M \times A)$ are the coefficients of a standard PLS model with A latent variables. This equation can be simplified considering that

$$\left(\mathbf{x}_i^A\right)^{\mathrm{T}} \mathbf{x}_j^A = \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j^A$$

with the advantage of a more straightforward interpretability:

$$m_{ij} = \frac{\left\{\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j + (\mathbf{e}_i^A)^{\mathrm{T}}\mathbf{e}_j^A\right\} \cdot \mathrm{abs}\left\{\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j - (\mathbf{e}_i^A)^{\mathrm{T}}\mathbf{e}_j^A\right\}}{\sigma_{\mathbf{x}_i}^2 \sigma_{\mathbf{x}_j}^2} \tag{5}$$

where $\mathbf{e}_i^A$ is the vector of residuals for the i-th variable in the PLS model for the A latent variables.

This approach has the substantial advantage of filtering out the noise in the computation of correlations, reducing the risk of including spurious or chance associations among variables as often is the case in high dimensional data [20, 7].

3.3. Metaparameter selection

To obtain a GPLS model, suitable values for the number of LVs, A, in the MEDA step and for parameter γ in GIA need to be determined. If the goal is exploratory data analysis, γ and A can be obtained by visual inspection of the MEDA map, following the GPCA approach [9] which is consistent with the exploratory data analysis philosophy. Some hints on how to do this are included in Section 6. This is a main difference with sparse approaches like SPLS, where metaparameter selection is driven by prediction optimization.

However, if the goal is to obtain a GPLS predictive model and asses its performance, the meta-parameters A and γ can be selected using a double cross-validation approach. In this case, GPLS is a simple variant of SPLS but it might provide good predictive results when the structure of the data is sparse in the sense of a GPLS model, *i.e.* when data presents one or more groups of correlated predictor variables related to the response. We stress here that GPLS is not proposed for increase prediction, but to enhance data understanding and interpretation. However, these two different goals can be considered to be interrelated: a number of experiments where the prediction performance of GPLS is compared to that of SPLS and PLS can be found in supplementary materials. We intentionally did not present and discuss these results in the main body of the paper to avoid the misleading message that the GPLS algorithm is introduced to achieve better prediction models.

It should be noted that the number of LVs in the final GPLS model can be different to A in MEDA and larger than the number of groups identified by GIA, since the data corresponding to each of the groups may be of rank higher than one. In this paper, we assume GPLS has A LVs for both interpretation and prediction results.

4. Material and Methods

4.1. SPLS

Given its excellent performance we choose the sparse PLS algorithm by Lê Cao *et al.* as a representative of SPLS in this paper. The algorithm starts by

solving the PLS problem using singular value decomposition [21]. Given $\mathbf{X}$ and $\mathbf{Y}$, the matrix

$$\mathbf{C} = \mathbf{X}^{\mathrm{T}}\mathbf{Y} \tag{6}$$

of rank r can be written as

$$\mathbf{C} = \mathbf{G}\mathbf{D}\mathbf{U}^{\mathrm{T}}$$

where the matrices $\mathbf{G}$ $(N \times r)$ and $\mathbf{U}$ $(L \times r)$ are orthonormal and $\mathbf{D}$ is $r \times r$ diagonal containing the s_k singular values with $k = 1, 2, \ldots, r$. In this setting the loading vectors $\mathbf{p}^k$ and $\mathbf{q}^k$ for $\mathbf{X}$ and $\mathbf{Y}$ are the first singular vectors $\mathbf{g}^k$ and $\mathbf{u}^k$ of $\mathbf{G}$ and $\mathbf{U}$, respectively. Since the loadings can be interpreted as a measure of the relative importance of the variables in the model [22], variable selection is performed by penalizing both loadings vectors $\mathbf{p}^k$ and $\mathbf{q}^k$ as in sparse principal component analysis [23]. The optimization problem to arrive to the sparse PLS solution is:

$$\underset{\mathbf{p},\mathbf{q}}{\arg\min} \|\mathbf{C} - \mathbf{p}\mathbf{q}^{\mathrm{T}}\|_F^2 + \lambda_1\|\mathbf{p}\|_1 + \lambda_w\|\mathbf{q}\|_1 \tag{7}$$

where the solution is found by the soft-thresholding $g_\lambda(x) = \mathrm{sign}(x)\left(|x| - \lambda\right)_+$ applied to the current (non-penalized) least squares estimates.

The SPLS according to Equation (7) needs the optimization of two meta-parameters: λ_1 and λ_2. However a more practical and equivalent alternative is to select the number of non zero components of the loadings $\mathcal{N}_x$ and $\mathcal{N}_y$ [14]. In this paper we restrict ourselves to set $\mathcal{N}_x$, so that only loadings in the x-block are sparse. Consistently with PLS practice [24], the optimal number of latent variables A and the number of on non-zero loadings $\mathcal{N}_x$ is selected using a double cross-validation approach. The goodness-of-prediction index used is:

$$Q^2 = 1 - \frac{\mathrm{PRESS}_{A,\mathcal{N}_x}}{\mathrm{PRESS}_0} \tag{8}$$

where $\mathrm{PRESS}_{A,\mathcal{N}_x}$ is the prediction error computed when $\mathcal{N}_x$ and A LVs are considered in the SPLS model.

4.2. Experimental data sets

The performance of GPLS and SPLS was investigated using two publicly available experimental data sets.

Slurry-Fed Ceramic Melter. This data set provided with the PLS-toolbox [27] consists of 450 observations on 21 variables corresponding to a vitrification process. The first 20 variables correspond to temperatures collected in two vertical thermowells. Variables 1 to 10 are taken from the bottom to the top in thermowell 1, and variables 11 to 20 from the bottom to the top in thermowell 2. Variable 21 is the level of molten glass.

NMR data (NMR). This data set contains NMR spectral profiles (noisy experiments, 416 bucketed spectral variables at 0.02 ppm width) measured on the plasma samples of 206 subjects together with HDL (high density lipoproteins) independently measured with a biochemistry assay. The data are described in [25, 26] and available at www.ebi.ac.uk/metabolights/MTBLS147.

4.3. Software

The Group-wise PLS algorithm is available in the MEDA Toolbox [15] at `github.com/josecamachop/MEDA-Toolbox`. PLS has been performed through the kernel implementation [16]. Sparse PLS has been performed using the approach of [14], with the algorithm implementation used in [28].

All mentioned routines and corresponding single and double cross-validation algorithms can be found in the MEDA Toolbox for proper reproducibility of results. The code to reproduce the simulation results is given as Supplementary material.

5. Performance of Group-wise PLS on simulated data

We go back to the motivation example introduced in Section 2 and presented in Figure 1: Panel A shows the MEDA map (see Equation (5)) calculated for this simulated data set, where the group of five highly correlated variables is evident. To select that group of variables and analyze it with the GPLS approach, we only need to set γ (see Equations (2) and (3)) to the appropriate value, which can be done upon inspection of the MEDA map consistently with the philosophy of exploratory analysis. For this particular case we set $\gamma = 0.4$ and by doing so we are able to select that particular group of variables of interest. Together with this group the GIA procedure may select other groups of variables, in this case of cardinality one since some variables may show a variance higher than γ. This defines a set of $S_1, S_2, \ldots, S_k$ possibly overlapping groups of variables that are then passed to the GPLS algorithm. In this case we fit the model with one latent variable. The corresponding regression coefficients are shown in Figure 3. The GPLS algorithm is able to fully recover the original data structure avoiding the oversimplification introduced by the SPLS approach in Figure 2. In GPLS the sparsity is controlled by setting the threshold on the strength of the relationship among variables rather than by setting *a priori* the number of non-zero elements as in SPLS. Since we can select the correlation threshold upon inspection of the map, this approach is more coherent with the goal of data exploration. On the contrary, in SPLS we need to relay on the metaparameter that minimizes the prediction error, and this has the risk of confounding part of the structure in the data.

In Figure 4 we repeat the same experiment for high correlation, with the map in Figure 1 panel C. Recall that in this case there are many variables that are correlated with the subset of predictor variables, and therefore to the response. Again, SPLS oversimplifies the structure while GPLS includes those many more variables, offering a more accurate representation of the data structure. In this example, PLS presents a good result in terms of prediction. In general, as shown in the Supplementary Materials, we expect sparse methodologies to outperform common PLS only in the low correlation setting [13].

6. Performance of Group-wise PLS on experimental data

Let us start with the Slurry-Fed Ceramic Melter data. Figure 5 panel A presents the MEDA map of 4 LVs from this data showing how some temperatures of both thermowells are correlated. Also, as expected, there is specific correlation within each thermowell for close sensors. Upon the inspection of the correlation map we select $\gamma = 0.6$ as plausible value to decide on the number and size of the possibly overlapping sets of correlated variables.

The resulting GPLS model with 4 LVs is compared (weights and regression coefficients, β) to a 4 LVs SPLS model (where the number of non-zero loadings is selected by double cross-validation) in Figure 6. In the GPLS model, each LV is restricted to a single source of variability: latent variables 1 to 3 capture the inter-correlation between thermowells while LV 4 is restricted to the intra-correlation in one of the thermowells. On the contrary, SPLS provides a model which is sparse but where everything is mixed: each LV contains information from different sources of variability; this also makes the regression coefficients of SPLS to be less sparse than those of GPLS. This shows that fitting model metaparameters to optimize prediction implies the risk of confounding the true structure in the data. Both models, however, yield similar performance in terms of prediction, as shown in Figure 7 panel A where a scatter plot of the predicted vs measured response is presented.

Now let us investigate the sensitivity of GPLS to the meta-parameters. In Figure 8 we can see three examples of inadequately chosen meta-parameters: Panel B shows the case when the number of latent variables A is underestimated when defining the MEDA map and γ is maintained to 0.6. The corresponding MEDA plot is shown in Figure 5 panel B. We can see that part of the correlation structure in the data is missing in the MEDA plot due to the underestimation, but this does not affect the resulting GPLS model in this example. In general, we may consider that the overestimation of A is less harmful than the underestimation, see detailed results in [5]. In Figure 8 panel B we show the case when $A = 4$ but γ is underestimated. The resulting GPLS model is less sparse, but still we maintain one source of variability per LV. However, in other examples, the underestimation of γ may lead to include several sources of variability in a single LV. To avoid this problem, the analyst should always check that the resulting loadings are coherent with the structure seen in the MEDA plot. In panel C of Figure 8 we show the opposite case, when γ is too high. Then, the GPLS misses part of the structure that we can see in the original MEDA plot. Again, by inspecting the MEDA plot we can always check whether the choice of γ is adequate.

As a second example we consider the NMR-HDL data sets. Here the problem is to predict the HDL concentration in blood from the full NMR spectra while selecting for HDL related NMR peaks. The MEDA map for this data set is shown in Figure 5 panel B. Upon inspection of the map we select $\gamma = 0.9$ as input for the GPLS algorithm. Also in this case SPLS model parameters are set by double cross validation. The regression coefficients for the PLS, SPLS and GPLS models are given in Figure 9. GPLS produces a sparser model than

SPLS and selects only variables belonging to the region relevant for HDL signals in the NMR spectrum, with variables corresponding to ppm in the region 0.45-1.00, 1.13-1.45, 1.51-1.69, 1.87-2.2 ppm as reported in previous publications [25]. In particular the larger regression coefficient correspond to 0.84 ppm bucket comprising the LDL CH_3 signal. Both models yield similar performance in terms of prediction, as shown in Figure 7 panel B.

7. Conclusions

We have presented here a new algorithm for sparse partial least squares modeling referred to as Group-wise PLS, which is a PLS extension of the recently proposed Group-wise Principal Component Analysis. GPLS is based on two steps: *i*) identification of groups of correlated variables in correlation maps obtained from the data and. *ii*) the fitting of group-wise models where each latent variable corresponds only to one group of variables. The main advantage of GPLS in comparison to other sparse variants is that the choice of metaparameters can be decided upon data visualization. This approach is more coherent with exploratory data analysis than using techniques to optimize prediction, which are the state-of-the-art in the PLS setting. This makes the Group-wise PLS algorithm well suited for data exploration avoiding problems arising in standard PLS and other sparse variants.

8. Acknowledgements

Dr. Ewa Szymańska in gratefully acknowledged for making available the Matlab implementation of the SPLS algorithm. This work is partly supported by the Spanish Ministry of Economy and Competitiveness and FEDER funds through project TIN2014-60346-R and by the EU Commission through the FP7 project INFECT (Contract No. 305340).

References

[1] E. Saccenti, H. C. Hoefsloot, A. K. Smilde, J. A. Westerhuis, and M. M. Hendriks, "Reflections on univariate and multivariate analysis of metabolomics data," *Metabolomics*, vol. 10, no. 3, p. 0, 2014.

[2] H. Wold, *Multivariate Analysis*, ch. Estimation of principal components and related models by iterative least squares. Academic Press, New York, 1966.

[3] S. Wold, E. Johansson, and M. Cocchi, "Pls-partial least squares projections to latent structures," *3D QSAR in drug design*, vol. 1, pp. 523–550, 1993.

[4] P. Filzmoser, M. Gschwandtner, and V. Todorov, "Review of sparse methods in regression and classification with application to chemometrics," *Journal of Chemometrics*, vol. 26, no. 3-4, pp. 42–51, 2012.

[5] J. Camacho, "Missing-data theory in the context of exploratory data analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 103, pp. 8–18, 2010.

[6] J. A. Hageman, M. M. Hendriks, J. A. Westerhuis, M. J. Van Der Werf, R. Berger, and A. K. Smilde, "Simplivariate models: Ideas and first examples," *PLoS One*, vol. 3, no. 9, p. e3259, 2008.

[7] E. Saccenti, J. A. Westerhuis, A. K. Smilde, M. J. van der Werf, J. A. Hageman, and M. M. W. B. Hendriks, "Simplivariate models: Uncovering the underlying biology in functional genomics data," *PLoS ONE*, vol. 6, p. e20747, 06 2011.

[8] B. Liquet, P. L. de Micheaux, B. P. Hejblum, and R. ThiÃľbaut, "Group and sparse group partial least square approaches applied in genomics context," *Bioinformatics*, vol. 32, no. 1, pp. 35–42, 2016.

[9] J. Camacho, R. A. Rodríguez-Gómez, and E. Saccenti, "Group-wise principal component analysis for exploratory data analysis," *To appear in Journal of Computational and Graphical Statistics*, 2017.

[10] L. Wangen and B. Kowalski, "A multiblock partial least squares algorithm for investigating complex chemical systems," *Journal of chemometrics*, vol. 3, no. 1, pp. 3–20, 1989.

[11] J. A. Westerhuis, T. Kourti, and J. F. MacGregor, "Analysis of multiblock and hierarchical pca and pls models," *Journal of chemometrics*, vol. 12, no. 5, pp. 301–321, 1998.

[12] J. Trygg and S. Wold, "Orthogonal projections to latent structures (o-pls)," *Journal of chemometrics*, vol. 16, no. 3, pp. 119–128, 2002.

[13] J. Camacho, "On the generation of random multivariate data," *Chemometrics and Intelligent Laboratory Systems*, vol. 160, pp. 40 – 51, 2017.

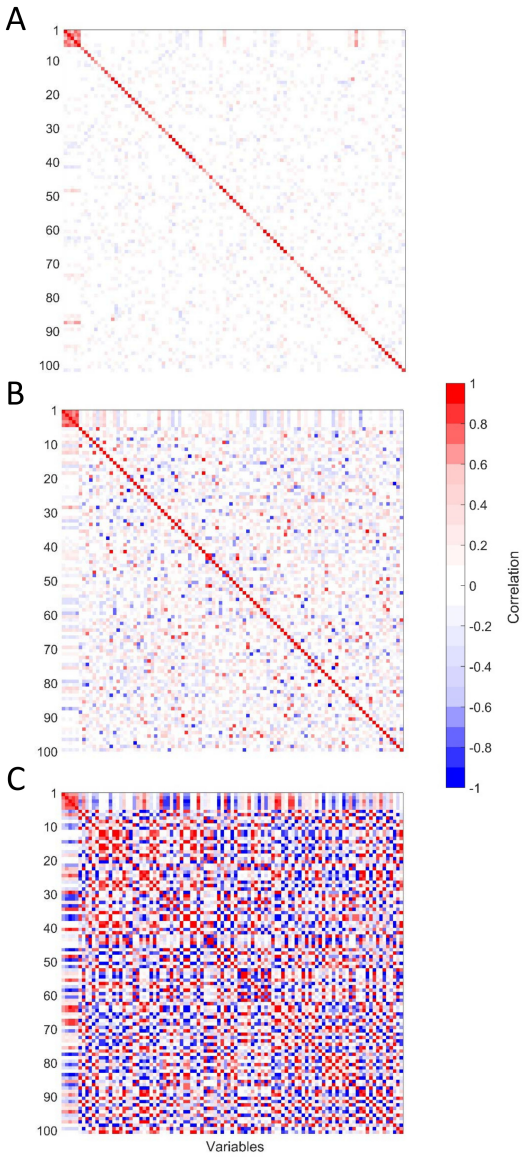

Figure 1: MEDA maps for the motivating example (see Equation (1)) with different levels of correlation in the X-block. A) low correlation, B) medium correlation and C) high correlation. See text for more details.

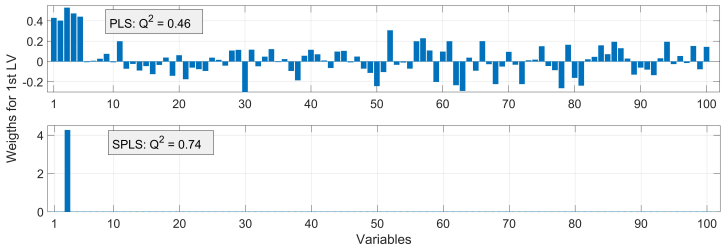

Figure 2: Weights for the first latent variable for the PLS and SPLS regression models for simulated data in the motivating example with low correlation. See Equation (1) and Figure 1 panel A for the corresponding MEDA map.

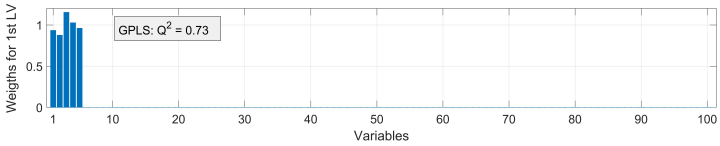

Figure 3: Weights for the first latent variable in the GPLS regression model for simulated data with low correlation in the motivating example. See Equation (1) and Figure 1 panel A for the corresponding MEDA map.

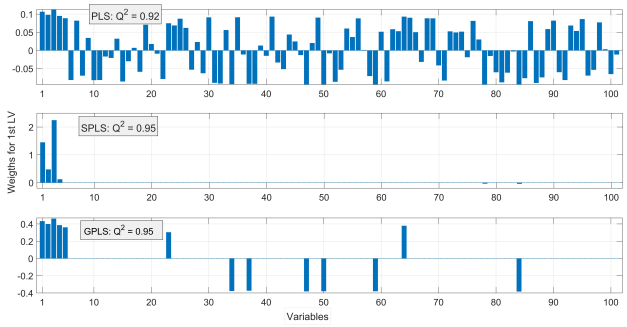

Figure 4: First latent variable from PLS, SPLS and GPLS regression models for simulated data with high correlation in the motivating example. See Equation (1) and Figure 1 panel C for the corresponding MEDA map.

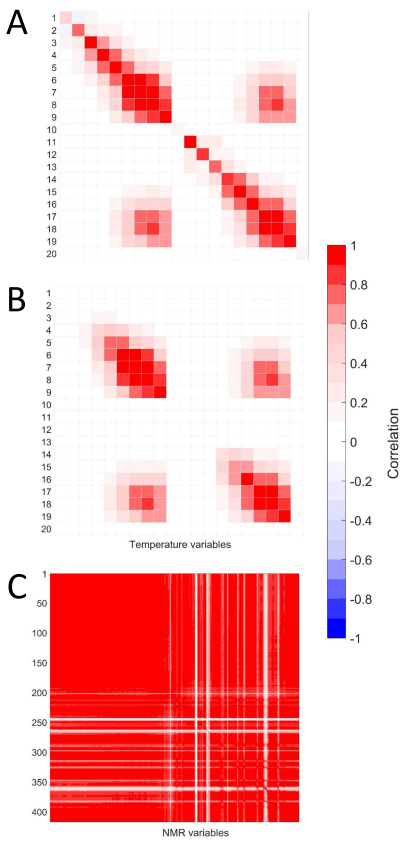

Figure 5: MEDA maps for the experimental data sets. A and B) Slurry-Fed Ceramic Melter data for 4 LVs and 1 LV; C) NMR HDL data.

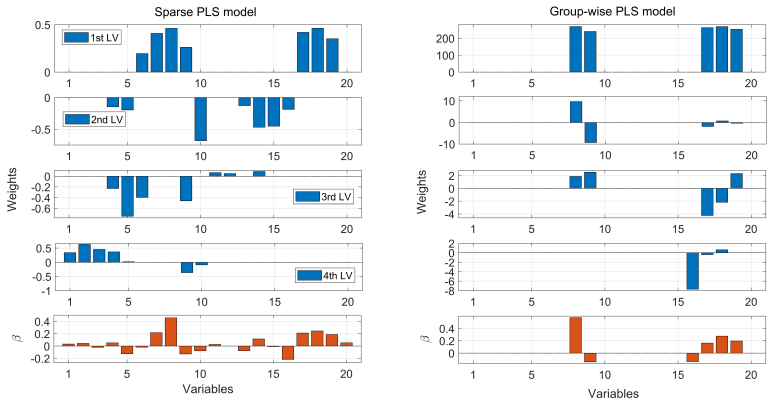

Figure 6: Weights for the first four latent variables (LVs) and regression coefficients (β) for the Sparse PLS and Group-wise PLS models for the Slurry-Fed Ceramic Melter data.

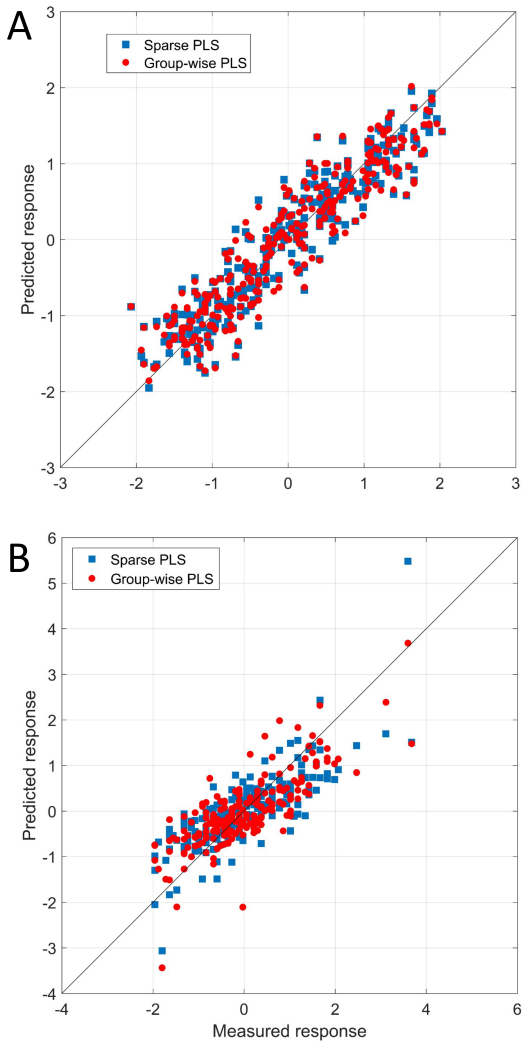

Figure 7: Predicted *vs* measured values of the response for the SPLS and GPLS models for the A) Slurry-Fed Ceramic Melter data and B) NMR data.

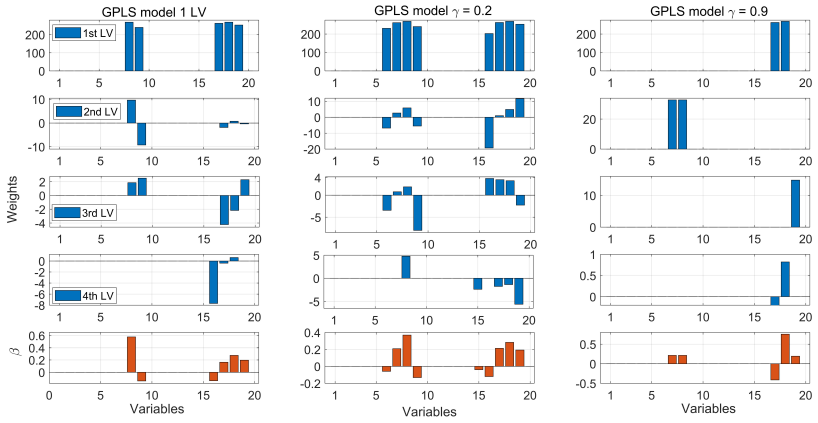

Figure 8: Weights (first 4 Latent variables (LV)) and regression coefficients (β) for the GPLS model for the Slurry-Fed Ceramic Melter data obtained using different model parameters. The first column corresponds to a model using 1 latent variable to define the MEDA map and $\gamma = 0.6$. The second and third columns correspond to a model using 4 latent variables to define the MEDA map and γ equal to 0.2 and 0.9, respectively.

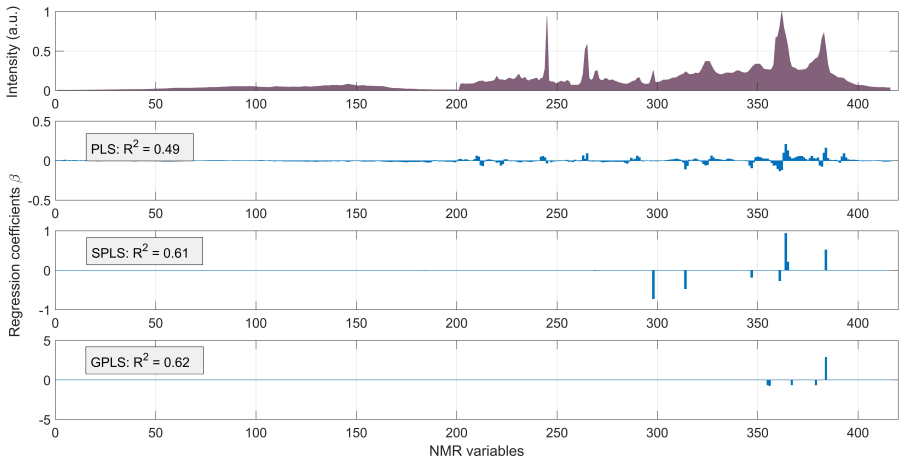

Figure 9: Regression coefficients (β) for the first 3 Latent variables (LV) for PLS, SPLS and GPLS regression models on the NMR-HDL data set. The NMR intensity profile (in arbitrary units (a.u) is given as a reference; peaks are normalized to the maximum intensity for visualization.

[14] K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse, "A sparse pls for variable selection when integrating omics data," *Statistical applications in genetics and molecular biology*, vol. 7, no. 1, 2008.

[15] J. Camacho, A. Pérez-Villegas, R. A. Rodríguez-Gómez, and E. Jiménez-Manas, "Multivariate exploratory data analysis (meda) toolbox for matlab," *Chemometrics and Intelligent Laboratory Systems*, vol. 143, pp. 49 – 57, 2015.

[16] B. Dayal and J. MacGregor, "Improved PLS algorithms," *Journal of Chemometrics*, vol. 11, pp. 73–85, 1997.

[17] F. Arteaga and A. Ferrer, "Dealing with missing data in mspc: several methods, different interpretations, some examples," *Journal of Chemometrics*, vol. 16, pp. 408–418, 2002.

[18] F. Arteaga and A. Ferrer, "Framework for regression-based missing data imputation methods in on-line mspc," *Journal of Chemometrics*, vol. 19, pp. 439–447, 2005.

[19] F. Arteaga, "A note on "missing-data theory in the context of exploratory data analysis"," *Technical Report. MEDA Toolbox*, 2011.

[20] E. Saccenti, A. K. Smilde, J. A. Westerhuis, and M. M. W. B. Hendriks, "Tracy-widom statistic for the largest eigenvalue of autoscaled real matrices," *Journal of Chemometrics*, vol. 25, no. 12, pp. 644–652, 2011.

[21] A. Lorber, L. E. Wangen, and B. R. Kowalski, "A theoretical foundation for the pls algorithm," *Journal of Chemometrics*, vol. 1, no. 1, pp. 19–31, 1987.

[22] S. Wold, L. Eriksson, J. Trygg, and N. Kettaneh, "The pls method–partial least squares projections to latent structures–and its applications in industrial rdp (research, development, and production)," *Unea University*, 2004.

[23] H. Shen and J. Z. Huang, "Sparse principal component analysis via regularized low rank matrix approximation," *Journal of multivariate analysis*, vol. 99, no. 6, pp. 1015–1034, 2008.

[24] E. Szymańska, E. Saccenti, A. K. Smilde, and J. A. Westerhuis, "Double-check: validation of diagnostic statistics for pls-da models in metabolomics studies," *Metabolomics*, vol. 8, no. 1, pp. 3–16, 2012.

[25] P. Bernini, I. Bertini, C. Luchinat, L. Tenori, and A. Tognaccini, "The cardiovascular risk of healthy individuals studied by nmr metabonomics of plasma samples," *Journal of proteome research*, vol. 10, no. 11, pp. 4983–4992, 2011.

[26] E. Saccenti, M. Suarez-Diez, C. Luchinat, C. Santucci, and L. Tenori, "Probabilistic networks of blood metabolites in healthy subjects as indicators of latent cardiovascular risk," *Journal of proteome research*, vol. 14, no. 2, pp. 1101–1111, 2014.

[27] B. Wise, N. Gallagher, R. Bro, J. Shaver, W. Windig, and R. Koch, *PLSToolbox 3.5 for use with Matlab*. Eigenvector Research Inc., 2005.

[28] E. Szymańska, G. H. Tinnevelt, E. Brodrick, M. Williams, A. N. Davies, H. J. van Manen, and L. M. C. Buydens, "Increasing conclusiveness of clinical breath analysis by improved baseline correction of multi capillary column - ion mobility spectrometry (MCC-IMS) data," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 127, pp. 170–175, 2015.