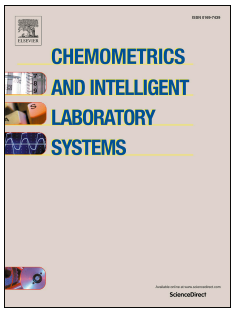
Hervás-Marín, D.; Prats-Montalbán, JM.; Garcia-Cañaveras, J.; Lahoz Rodríguez, AG.; Ferrer, A. (2019). Sparse N-way Partial Least Squares by L1-penalization. Chemometrics and Intelligent Laboratory Systems. 185:85-91. https://doi.org/10.1016/j.chemolab.2019.01.004

Accepted Manuscript



Sparse N-way Partial Least Squares by L1-penalization

D. Hervás, J.M. Prats-Montalbán, J.C. García-Cañaveras, A. Lahoz, A. Ferrer

PII: S0169-7439(18)30461-1

DOI: https://doi.org/10.1016/j.chemolab.2019.01.004

Reference: CHEMOM 3737

To appear in: *Chemometrics and Intelligent Laboratory Systems*

Received Date: 14 August 2018

Revised Date: 2 January 2019

Accepted Date: 10 January 2019

Please cite this article as: D. Hervás, J.M. Prats-Montalbán, J.C. García-Cañaveras, A. Lahoz, A. Ferrer, Sparse N-way Partial Least Squares by L1-penalization, *Chemometrics and Intelligent Laboratory Systems* (2019), doi: 10.1016/j.chemolab.2019.01.004

Sparse N-way Partial Least Squares by L1-penalization

Hervás, D.[a]; Prats-Montalbán, J.M.[b, *], García-Cañaveras, J.C.[c], Lahoz A.[c, *] and Ferrer, A.[b]

a) Biostatistics Unit, Health Research Institute La Fe, Valencia, Spain

b) Multivariate Statistical Engineering Group, Universitat Politècnica de València, Valencia, Spain

c) Biomarkers and Precision Medicine Unit, Analytical Unit Health Research Institute La Fe, Valencia, Spain

* Corresponding author: José M. Prats-Montalbán. Departamento de Estadística e IO Aplicadas y Calidad. Universidad Politécnica de Valencia. Cno. De Vera s/n, Edificio 7A, 46022, Valencia, Spain. Tlf: +34.96.387.70.07 ext. 74949, Fax: +34.96.387.74.99. E-mail: jopramon@eio.upv.es

* Corresponding author: Agustín Lahoz. Biomarkers and Precision Medicine Unit, Analytical Unit (Metabolomics). Health Research Institute La Fe, Torre A-6-19. Avda. Fernando Abril Martorell, 106. 46026, Valencia, Spain. Tlf +34.96.124.66.52, Fax +34.96.124.66.20. E-mail: agustin.lahoz@uv.es

Abstract

N-PLS, as the natural extension of PLS to *N*-way structures, tries to maximize the covariance between an **X** and a **Y** *N*-way data arrays. It provides a useful framework for fitting prediction models to *N*-way data. However, N-PLS by itself does not perform variable selection, which indeed can facilitate interpretation in different situations (e.g. the so-called "–omics" data). In this work, we propose a method for variable selection within *N*-PLS by introducing sparsity in the weights matrices $\mathbf{W}^J$ and $\mathbf{W}^K$ by means of L1-penalization. The sparse version of *N*-PLS is able to provide lower prediction errors by filtering all the noise variables and to further improve interpretability and usability of the *N*-PLS results. To test Sparse *N*-PLS performance two different simulated data sets

were used, whereas to show its utility in a biological context a real time course metabolomics data set was used.

Keywords: *N*-PLS, LASSO, Variable selection, Multiway models

1. INTRODUCTION

In the last decades, advances in technology have enabled the gathering of an increasingly amount of data in the field of biology and biomedicine [1]. The so called "-omics" technologies such as genomics, epigenomics or transcriptomics, metabolomics, among others, produce hundreds, thousands or even millions of variables per dataset. Economic and logistic restrictions often lead to small sample sizes paired to these wide datasets, thus producing the recurring problem of I samples << J variables [2]. A wide variety of methods exist for dealing with these matrix-type data [3-5]. But sometimes, these IxJ datasets can be expanded by taking, for example, repeated measurements at different K time points for each individual, thus having IxJxK datasets that raise more methodological complications to the analyses. These datasets are called three-way data [6]. One useful tool for analysing three-way data, when some **Y** data structure is to be predicted, is *N*-PLS [7]. *N*-PLS reduces the inclusion of noise in the models and obtains more robust parameters (by reducing the number of parameters to be estimated in the model) when compared to PLS while, at the same time, producing easy-to-understand plots.

Related to the problem of I << J datasets, comes the issue of variable selection. Variable selection is essential for facilitating e.g. biological interpretation of the results when analyzing "-omic" data sets [8]. It is often the case that the aim of these analyses is to find a new biomarker or a specific set of biomarkers, also called signature, to diagnose or predict the prognosis of a disease. The *N*-PLS algorithm does not provide (in general) inner variable selection, i.e. the variable selection procedure is not implemented within the algorithm, although some methods have been developed to perform it [9].

In this work, we propose the introduction of L1-penalization in the *N*-PLS algorithm to allow for variable selection within the model-fitting step. This penalization imposes a constrain to the weights matrices, shrinking the coefficients of the model, causing some of them to be exactly zero and thus performing variable selection at the same time. This approach should not only facilitate, e.g. biological, interpretation by producing a reduced model including fewer variables, but should also reduce prediction error by completely eliminating noise features [10] instead of just downweighting them as *N*-PLS does. The method also allows to smoothly adjust its bias-variance trade-off by changing the amount of L1 penalization imposed on the model (Figure 1).

[INSERT FIGURE 1 ABOUT HERE]

In Section 2, the methodological background related to the methods and how to combine them into an embedded version of the *N*-PLS model are presented. In Section 3, the different datasets analyzed are introduced, both for the simulated and the real cases. Section 4 presents the results, and finally Section 5 the conclusions.

2. METHODOLOGICAL BACKGROUND

In the following paragraphs, a brief explanation of the methods *N*-PLS and Lasso is given before introducing our Sparse *N*-PLS algorithm:

2.1. N-PLS

N-PLS [11] studies relationships between some three-way (or *N*-way) $\underline{\mathbf{X}}$ (e.g. $I \times J \times K$) data structure and any $\underline{\mathbf{Y}}$ (e.g. $I \times L \times M$) data structure. It is the natural extension of PLS to *N*-way structures, which tries to maximize the covariance between $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$ data arrays. Considering $\mathbf{X}$ (IxJK) the unfolded version of $\underline{\mathbf{X}}$, *N*-PLS tries to find latent spaces $\mathbf{W}^J$ and $\mathbf{W}^K$ that maximize the covariance between $\mathbf{X}$ and $\mathbf{Y}$, so it can be expressed as:

$$\mathbf{X} = \mathbf{T}(\mathbf{W}^K | \otimes | \mathbf{W}^J)^\mathbf{T} + \mathbf{R_X} \quad \textbf{(1)}$$

afterwards decomposing $\underline{\mathbf{X}}$ from **X** using the improved *N*-PLS version expression [12], in order to obtain residuals with better statistical properties:

$$\mathbf{X} = \mathbf{TG}(\mathbf{W}^{K} \otimes \mathbf{W}^{J})^{\mathbf{T}} + \mathbf{R_X}' \quad \textbf{(2)}$$

In this case, $\mathbf{W}^{J}$ and $\mathbf{W}^{K}$ refer to the weights of the second and third modes, respectively; whereas **T** matrix gathers the scores of the samples at each component extracted, in the 1st mode. $|\otimes|$ is the Khatri-Rao product and $\otimes$ the Kroncecker product, which forbid or allow (respectively) to take interactions between the different modes components into account. **G** is the core array (unfolded) of a Tucker3 decomposition when using **T**, $\mathbf{W}^{K}$ and $\mathbf{W}^{J}$ as loadings, in order to obtain a better (or at least not worse) approximation of the $\underline{\mathbf{X}}$ array [13]. Finally, $\mathbf{R_X}$' incorporates the residuals.

For $\underline{\mathbf{Y}}$, similar results can be achieved when unfolding $\underline{\mathbf{Y}}$ into Y:

$$\mathbf{Y} = \mathbf{U}(\mathbf{Q}^{M} \otimes \mathbf{Q}^{L})^{\mathbf{T}} + \mathbf{R_Y} \quad \textbf{(3)}$$

$\underline{\mathbf{Y}}$ scores vectors are called **U**, and weights vectors $\mathbf{Q}^{L}$and $\mathbf{Q}^{M}$, in the case of a three-way array. As for $\underline{\mathbf{X}}$, $\mathbf{R_Y}$ incorporates the residuals. The model is estimated in such a way that the covariance between **T** and **U** is maximized [13]. Finally, the prediction model between **X** and **Y** can be expressed using an inner relationship between **T** and **U**:

$$\mathbf{U} = \mathbf{TB} + \mathbf{R_U} \quad \textbf{(4)}$$

When $\underline{\mathbf{Y}}$ variables have to be predicted on a new sample, score matrix **T** is worked out from Eq. (2), and by using Eq. (4) the scores **U** in $\underline{\mathbf{Y}}$ space are calculated. Finally, the prediction for $\underline{\mathbf{Y}}$ is obtained from Eq. (3).

2.2. Lasso

Lasso is a regression analysis method based on L1 penalization. It was first developed for linear models, and consists in minimizing the usual sum of squared errors, with a

bound on the sum of the absolute values of the coefficients [14]. This bound forces the coefficients of the model to shrink, potentially setting some of them to 0. Since its introduction to linear models, Lasso has been expanded to many other techniques such as generalized linear models, survival analysis and principal component analysis among others [15, 16], or even PLS [17]. The original LASSO for least squares is as follows:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{I} (y_i - \beta_0 - \sum_{j=1}^{J} x_{ij}\beta_j)^2 \tag{5}$$

Subject to the restriction:

$$\sum_{j=1}^{J} |\beta_j| \leq s$$

Reducing *s* increases the penalization and forces shrinkage of the coefficients, producing a simpler model by setting some of them exactly to zero (Figure 2). Thus, assuming data are standardized, Lasso selects the most relevant features and discards the others.

[INSERT FIGURE 2 ABOUT HERE]

The solutions of equation (5) are easily shown to be

$$\hat{\beta}_j^{lasso} = sgn(\hat{\beta}_j^{LS})(|\hat{\beta}_j^{LS}| - \lambda)^+ \tag{6}$$

where $\hat{\beta}_j^{lasso}$ is the lasso estimated regression coefficient for variable *j*, $\hat{\beta}_j^{LS}$ is the least squares estimate of regression coefficient for variable *j*, and λ is the penalization coefficient determined by the condition $\sum_{j=1}^{J}|\beta_j| \leq s$ (see appendix for details).

2.3. Sparse *N*-PLS

To introduce the L1-penalization in the *N*-PLS algorithm, we follow the approach of Lê Cao *et al.* [17]. The main difference between Le Cao et al. algorithm and that proposed

in this paper is that, besides dealing with N-way structures instead 2-way matrices, Le Cao et al. apply soft-thresholding on the loadings vectors for the **X** and **Y** spaces, performing variable selection on **X** and **Y**; whereas Sparse N-PLS applies soft-thresholding to achieve sparse versions of the weights matrices for the second and third mode, thus performing selection in the different modes of **X** and not on **Y** Sparse N-PLS. To achieve these sparse versions of weights $\mathbf{w}^J$ and $\mathbf{w}^K$ for each latent variable, we introduce the soft-thresholding penalty function defined in equation (6), $\hat{\beta}_j^{lasso} = sgn(\hat{\beta}_j^{LS})(|\hat{\beta}_j^{LS}| - \lambda)^+$, in the *N*-PLS algorithm right after the SVD at the $\mathbf{w}^J$ and $\mathbf{w}^K$ determination. The complete algorithm is as follows:

Center **X** and **Y**, and unfold **X** (and **Y** when necessary) into a two-way matrix.

Let **u** be some column of **Y**, and set f (number of components)=1

1. $\mathbf{w}^T=\mathbf{u}^T\mathbf{X}/\mathbf{u}^T\mathbf{u}$
2. Build **Z** by refolding **w** according to the modes dimensions
3. Determine $\mathbf{w}^J$ y $\mathbf{w}^K$ by SVD
4. L1-penalization inclusion
    a. Apply soft-thresholding on $\mathbf{w}^J$: $\widehat{w}_i^{j*} = sgn(\widehat{w}_i^j)(|\widehat{w}_i^j| - \lambda_j)^+$
    b. Apply soft-thresholding on $\mathbf{w}^K$: $\widehat{w}_i^{k*} = sgn(\widehat{w}_i^k)(|\widehat{w}_i^k| - \lambda_k)^+$
    c. Input the new **w** as kronecker($\mathbf{w}^K$, $\mathbf{w}^J$)
5. $\mathbf{t}=\mathbf{X}\mathbf{w}/\mathbf{w}^T\mathbf{w}$
6. $\mathbf{q}=\mathbf{Y}^T\mathbf{t}/\text{norm}(\mathbf{Y}^T\mathbf{t})$
7. $\mathbf{u}=\mathbf{Y}\mathbf{q}$
8. Check for convergence. If it is achieved, continue; otherwise, go to 1
9. $\mathbf{b} = (\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{u}$; where $\mathbf{T}=[t_1\ t_2 \ldots\ t_f]$
10. Deflate both **X** and **Y**: $\mathbf{X} = \mathbf{X}-\mathbf{t}\mathbf{w}^T$ and $\mathbf{Y} = \mathbf{Y}-\mathbf{t}\mathbf{b}\mathbf{q}^T$
11. *f = f+1*. Continue from step 1 until a good description of **Y**

In this work, we perform both the standard regression (continuous response) and the discriminant version of the *N*-PLS model, i.e. *N*-PLS-DA. In the case of *N*-PLS-DA, **Y** is a **y** vector formed by ones and zeros, each of the two values related to one of the two classes to be segregated.

3. MATERIAL AND METHODS

In order to show the performance of the Sparse *N*-PLS model proposed, three different analyses have been performed: two of them using simulated data sets, and one using a real data set from metabolomics data. A description of each one of them is given below.

3.1 Synthetic data 1

First, we tested our implementation of the L1 penalized *N*-PLS using data simulation. In total, fourteen different scenarios with different signal-to-noise ratios were tested. Tuning of the models was performed by 20 repetitions of 10-fold cross-validation.

Our simulations consisted on three-way $\underline{\mathbf{X}}$ arrays with *I*=50 samples, *J*=50 variables and *K*=3 times, where variables were simulated randomly from different kinds of distributions (Poisson, Normal and Uniform) with varying parameters.

If Normal: $\mathcal{X} \sim \mathcal{N}(\mu, \sigma)$ where $\mu \sim \mathcal{N}(10,\ 10)$ and $\sigma \sim \Gamma(5,\ 1)$

If Poisson: $\mathcal{X} \sim \mathcal{P}(\lambda)$ where $\lambda \sim \mathcal{N}(10,\ 2.5)$

If Uniform: $\mathcal{X} \sim \mathcal{U}(a, b)$ where $a \sim \mathcal{N}(10,\ 10)$ and $b \sim \mathcal{N}(100,\ 10)$

Only 5 out of the 50 variables were used to construct the response $\underline{\mathbf{Y}}$. They were chosen randomly from the $\underline{\mathbf{X}}$ array and assigned randomly the following coefficients: 0.4, 0.5, 0.6, 0.7 and 0.9. In the first run, only one of the three times of the third mode was involved in the creation of, in this case, vector **y**. In the second run, the three times were involved, but with different coefficients for each variable. In all simulations, random Normal and Poisson errors were added in different amounts to **y**. For each combination of type and amount of random error, simulations were repeated 100 times. Ability to select the real variables involved in **y** generation as well as median and 1st and 3rd quartiles of the mean squared error are provided for each simulation run.

3.2 Synthetic data 2

To further test the performance of the method in selecting highly correlated variables, simulated data resembling the ones of a toxicogenomics data set by Heijne *et al.* [18] was analysed. In this work, the effect of the hepatotoxicant bromobenzene in rats was studied. Groups of rats were treated with different doses of this toxic compound

dissolved in corn oil for a 48 hours period. At three time points from the start of the treatment, rats were sacrificed. Liver samples were used to extract mRNA for microarray profiling and blood and urine was used for metabolite profiling. Additionally, 21 physiological parameters were recorded: Glucose, A/G ratio, GSH, Body Weight, Creatin, GGT, Urea, Kidneys, Kidney/BW, Triglycerides, Liver, Albumin, Total Protein, ALP, Liver/BW, Bilirubin, LDH, Phospholipids, Cholesterol, ASAT, ALAT.

In this work, simulated profiles from 14 of these 21 physiological parameters were used to discriminate between two of the different groups (High and Low doses evolution). 25 samples for high doses and 25 for low doses were simulated, adding to the pattern random normal noise, with standard deviation 0.1.

Figure 3 represents time course levels of the seven patterns corresponding to the 14 physiological parameters evaluated for high and low dose treatment groups. These variables can be grouped attending to their common patterns in the following groups; i) ALAT, ASAT, LDH and GSH; ii) Creatin and Albumin; iii) Kidney and Cholesterol; iv) Liver, Phospholipids and Triglycerides; v) Glucose; vi) A/G Ratio and vii) Urea.

[INSERT FIGURE 3 ABOUT HERE]

3.3 Real dataset study

Rat serum samples preparation

Six-week-old male Oncins France Strain A (OFA) rats (200–240 g) were purchased from Charles River (Barcelona, Spain) and acclimatized to laboratory conditions for at least 7 days. Animals were housed (12-h light-dark cycle, 21–25°C, 30–70% humidity, woodchip bedding) and fed ad libitum with a standard chow diet (Scientific Animal Food and Engineering, Augy, France). Rats were anesthetized with sodium thiobarbital (0.1 g/kg), and blood was collected by cardiac puncture. After coagulation and centrifugation (1,000 g for 10 min at 4°C), serum samples were aliquoted and stored at -80°C until the analysis. All the experimental protocols were approved by the Institutional Animal Ethics Committee. 40 µL of serum sample were mixed with 120 µL

of methanol. After vortexing, samples were kept at -20 ºC for 20 min. Samples were centrifuged (14000 g, 4 ºC, 15 min) and the supernatants transferred to clean tubes and evaporated to dryness. Samples were resuspended in 80 µL of water, centrifuged (14000 g, 4 ºC, 5 min), and the clean supernatants transferred to HPLC vials for their LC-MS analysis. Rats serum samples were separated in two groups of sizes 8 and 6 and subsequently fortified with a set of metabolites to generate the patterns showed in Figure 4. Metabolites and final concentrations used are summarized in Table 1.

LC-MS analysis

Liquid chromatography–mass spectrometry (LC-MS) analyses of rat serum samples were performed in an Agilent 1290 Infinity LC system coupled to an Agilent 6550 Q-TOF mass spectrometer equipped with an ESI source (Agilent Technologies, Santa Clara, CA, USA). LC-MS grade solvents (i.e. water, acetonitrile and methanol) were acquired from Fisher Scientific (Loughborough, UK). All the LC-MS additives and standards were acquired from Sigma–Aldrich/Fluka (Madrid, Spain). Metabolites were separated on an Zorbax SB-Aq column (100 x 2.1 mm; 1.8 µm) (Agilent Technologies, Santa Clara, CA, USA). Mobile phases consisted of (A) 1mM ammonium fluoride and (B) acetonitrile. The separation was conducted under the following gradient at a flow of 0.3 mL/min: 0 min 3 % (B); 0–2 min 40 % (B); 2-5 min 7 % (B); 5-7 min 50 % (B); 7-12 min 100% (B); 12-16min 100% (B); 16-16.5 min 3% (B); 16.5-18 min 3% (B). Sample and column temperatures were maintained at 4 ºC and 40ºC, respectively. The injection volume was 5 µL. The instrument was tuned in the 50-1700 *m/z* range using an Agilent tune mix in 2GHz extended dynamic range mode (mass resolving power 25,000 FWHM). Detection was performed in ESI (-) mode in the 50-1000 m/z range. A reference solution (m/z 119.0360 and m/z 980.0164) was used to correct small mass drifts during acquisition. The following conditions were employed: capillary voltage, 3.5 kV; nozzle voltage -1.0 kV; fragmentor voltage, 175 V; gas temperature, 200 ºC; drying gas (nitrogen), 14 L/min; nebulizer gas (nitrogen), 35 psi; sheath gas temperature, 350 ºC; and sheath gas flow (nitrogen), 11 L/min. The acquisition rate was set at 4 spectra/s in all cases. Data preprocessing was performed using ProgenesisQI software (Nonlinear Dynamics, UK). The MS-data was arranged in an array structure with I x J x K dimensions, where “I” denotes the number of rats used in the assay (i.e., objects); “J” denotes the total number of features detected in the LC-MS analysis (i.e.,

variables) and “K” denotes he time points selected for rats` serum collection (i.e., three time points).

[INSERT FIGURE 4 ABOUT HERE]

[INSERT TABLE 1 ABOUT HERE]

3.3 Software

Both *N*-PLS and Sparse *N*-PLS analyses were performed using the authors’ Sparse *N*-PLS package [19] available at CRAN (version 0.3.31). A comprehensive description of this software can be found elsewhere [20].

4. RESULTS AND DISCUSION

4.1 Synthetic data 1

Results of the different simulations carried out on the first synthetic dataset are provided in Table 2. They show that Sparse *N*-PLS outperforms *N*-PLS regarding mean squared prediction error. When performing Sparse *N*-PLS, the true variables were almost always included in the selected model. Median number of true variables selected in each model was 5 (100%) in most of the simulations (9 out of 14). In the other 5 simulations, the median number of true variables selected was 4 (80%). These simulations were the ones consisting in the more complex and noisy models, with the three times of the third mode affecting **Y** and lower signal-to-noise ratios. Also, a varying amount of other noise variables were erroneously included in the models (false positives). The amount of noise variables that were included in the Sparse *N*-PLS models increased as the signal-to-noise ratio of the data decreased, ranging from a median of 2 (4.4%) to a median of 7 (15.6%) noise variables included in the worst-case simulation.

[INSERT TABLE 2 ABOUT HERE]

4.2. Synthetic data 2

In this case, the goal was to test the ability of Sparse *N*-PLS for gathering those relevant variables within the Lasso selection procedure even in the case that these variables show correlation. Table 3 summarizes the model coefficients obtained in the data set analysis. Overall, results of the analysis showed good agreement with the structure of the data. All the variables following the pattern i (i.e., ALAT, ASAT, LDH and GSH) were selected by the method and similar coefficients were assigned Additionally, Creatinine and Albumin (pattern ii) were also selected with similar coefficients. Interestingly, the A/G ratio (pattern vi), a variable uncorrelated to all the others was also selected. However, the rest of the patterns were not considered/selected by the model, probably because selection was also performed on the third mode and only the third element of the third mode was selected (Table 3).

[INSERT TABLE 3 ABOUT HERE]

4.3. Results of the metabolomics dataset

Simulated data sets provide a useful suitable first approach to test the performance of the new Sparse *N*-PLS model. However, to exemplify Sparse *N*-PLS utility in a more complex context, the proposed method was faced to the analysis of a real dataset, which was derived from a metabolomics study. In the metabolomics data sets usually hundreds of variables, with high noise and high correlation are obtained, which dramatically hinders biomarker discovery and variable selection for predictive models building. Thus, real processed rat serum samples were used to artificially generate two different groups by adding a set of standards at different final concentrations that additionally showed different trends along time (Figure 4). In our opinion, this experimental design provides a suitable frame work to assess Sparse NPLS capabilities when facing real -*omics* data sets.

Our cross-validation procedure (20 repetitions of 5-fold cross-validation) selected as the optimum parameter values 30 features of $\mathbf{W}^{J}$, 3 features of $\mathbf{W}^{K}$ and 2 components. Therefore, 60 variables among the initial 1220 obtained from the LC-MS analysis (30 in each component) were selected by our final Sparse *N*-PLS model. Out of the four variable classes which were different between both groups by design (Table 1), our

model included at least one representative variable for each class. The model also included other variables not present among the four controlled variable classes, but many of them showed similar patterns to those included and could be derivatives or adducts of the original metabolites. Overall, the selection provided by the new model showed a quite feasible result, where not only the real assignable variables (added metabolites) but also those interfering ones can be selected. A list of all the selected variables is presented in Table 4. The first column lists those variables selected by Sparse *N*-PLS, while the second column indicates on which component these variables were selected. The third column shows whether these variables belong or not to one of the classes described in Table 1. Finally, column four shows whether those variables that do not belong to any of the assayed classes follows or not a pattern similar to those variables included (Table 1). Interestingly, variables of the classes 1, 2 and 3 and its derivatives or analogues were all exclusively selected in the first component and variables of the class 4 and its derivatives or analogues were all exclusively selected in the second component. Variables with different patterns to those of the four experimentally generated classes were included in both components, but were more prominent in the second one (13 in the first component versus 20 in the second).

[INSERT TABLE 4 ABOUT HERE]

Finally, the performance of our Sparse *N*-PLS model was compared with the (standard) *N*-PLS model. To this end, the metabolomics data set was analysis using both approaches. The Sparse *N*-PLS model clearly discriminated between the two rat groups (Figure 5A). However, similar groups' separation was also obtained by using *N*-PLS (Figure 5B). The differences between both models appear when comparing Figs 5B vs 5G, and 5C vs 5H, related to $\mathbf{W}^J$ and $\mathbf{W}^K$, respectively; or Figs 5D vs 5I, and 5E vs 5J respectively, which are alternative $\mathbf{W}^J$ and $\mathbf{W}^K$ representations. For interpretation purposes, it seems better to compare Figs 5B vs 5G for $\mathbf{W}^J$, and Figs. 5E vs 5J for $\mathbf{W}^K$. For $\mathbf{W}^J$, it seems quite clear that the selection made from sparse *N*-PLS allows a clear interpretation of the metabolites responsible for the separation between the two groups. In the first component, those metabolites belonging to the classes 1, 2 and 3 (see Table 4) are represented. While, the second component is related to completely independent metabolites (with respect to component one), which could be related to the separation of

rats 6 and 13 from the rest. Many of these metabolites are from class 4 (Table 4), although some of them are not apparently related to any of the designed variable groups.

These interpretations are much more difficult to do when using *N*-PLS due to the high number of variables to deal with in the second (metabolites) mode, so from this perspective the proposed approach seems to improve *N*-PLS when trying to directly select the variables of interest (metabolites in this case). However, it should be highlighted that variable selection is out of the *N*-PLS scope. These results show that when variable selection is of prior relevance for interpretation or validation purposes Sparse *N*-PLS comes up as a valid alternative.

For the interpretation of $\mathbf{W}^K$, Figs. 5 E and J have been selected. These plots show, for the first component, a similar pattern, although a slight shift downwards is observed for *N*-PLS. The similar trend observed for both methods strengths the use of the Sparse *N*-PLS results, as it provides extra information as discussed above. However, regarding the second component, they did not provide the same result, which could be related to the clear separation of rats 6 and 13 observed in Sparse *N*-PLS (Fig. 5A).

[INSERT FIGURE 5 ABOUT HERE]

5. CONCLUSIONS

Overall, the results presented here show that Sparse *N*-PLS provides a straightforward method for variable selection in both synthetic and real experimental data sets. Sparse *N*-PLS reduces mean squared error compared to *N*-PLS in our synthetic simulations (although this might not always be the case). Furthermore, when the model was challenged to analyze a real metabolomic data set, it was able to identify all the discriminating metabolite classes between the two defined groups. Significantly from a biological point of view, the model is able to retrieve correlated variables when they are related to the response. In summary, the new Sparse *N*-PLS method enables variable selection and simplifies data interpretation, which is of utmost importance in the data analysis of untargeted approaches focus on the discovery of new biomarkers in biomedicine. Still, further work is required to compare this intra variable selection

method with other variable selection procedures (e.g. selectivity ratio, permutation tests, etc.) and assess for the best one, globally or at each problem at hand.

Appendix

Derivation of the soft-thresholding operator as a solution of the Lasso lagrangian form:

1. Assuming **X** (matrizied version of **$\underline{\mathbf{X}}$**) is composed of orthogonal columns, the least-squares solution is

$$\hat{\beta}^{LS} = (X^T X)^{-1} X^T y = X^T y \quad \text{(A. 1)}$$

2. Using the Lagrangian form, an equivalent problem to that considered would be

$$\min_\beta \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \quad \text{(A. 2)}$$

3. Expansion of the first term gives

$$\frac{1}{2}y^T y - y^T X\beta + \frac{1}{2}\beta^T\beta \quad \text{(A. 3)}$$

Since $y^T y$ does not contain any of the variables of interest, it can be discarded, and we can consider the following equivalent problem

$$\min_\beta \left(-y^T X\beta + \frac{1}{2}\|\beta\|_2\right) + \lambda\|\beta\|_1 \quad \text{(A. 4)}$$

Which can be rewritten as

$$\min_\beta \sum_{j=1}^{p} -\hat{\beta}_j^{LS}\beta_j + \frac{1}{2}\beta_j^2 + \lambda|\beta_j| \quad \text{(A. 5)}$$

So, we have a sum of objectives as the objective function. Since each of them corresponds to a separate β_j, this means that each variable may be solved individually

4. For a certain *j*, we want to minimize

$$\mathcal{L}_j = -\hat{\beta}_j^{LS}\beta_j + \frac{1}{2}\beta_j^2 + \lambda|\beta_j| \quad \text{(A. 6)}$$

If $\hat{\beta}_j^{LS} > 0$, then $\beta_j \geq 0$, otherwise we could just change its sign and get a lower value for the objective function. Correspondingly, if $\hat{\beta}_j^{LS} < 0$, then $\beta_j \leq 0$

5. In the first case, if $\hat{\beta}_j^{LS} > 0$ and $\beta_j \geq 0$, then

$$\mathcal{L}_j = -\hat{\beta}_j^{LS}\beta_j + \frac{1}{2}\beta_j^2 + \lambda\beta_j \quad \text{(A. 7)}$$

After differentiating respect to β_j and setting equal to zero, we get $\beta_j = \hat{\beta}_j^{LS} - \lambda$. Since $\beta_j \geq 0$, the right-hand side must be nonnegative, so the solution would be

$$\hat{\beta}_j^{lasso} = (\hat{\beta}_j^{LS} - \lambda)^+ = sgn(\hat{\beta}_j^{LS})(|\hat{\beta}_j^{LS}| - \lambda)^+ \quad \text{(A. 8)}$$

Which is the soft-thresholding operator.

6. In the other case, if $\hat{\beta}_j^{LS} < 0$ and $\beta_j \leq 0$, then

$$\mathcal{L}_j = -\hat{\beta}_j^{LS}\beta_j + \frac{1}{2}\beta_j^2 - \lambda\beta_j \quad \text{(A. 9)}$$

After differentiating respect to β_j and setting equal to zero, we get $\beta_j = \hat{\beta}_j^{LS} + \lambda$. Since we need $\beta_j \leq 0$ the solution is

$$\hat{\beta}_j^{lasso} = sgn(\hat{\beta}_j^{LS})(|\hat{\beta}_j^{LS}| - \lambda)^+ \quad \text{(A. 10)}$$

Which, again, gives the soft-thresholding operator.

REFERENCES

[1] Martin-Sanchez F, Verspoor K. Big Data in Medicine is Driving Big Changes. Yearbook of Medical Informatics. 9(1) (2014), 14-20.

[2] Diao G. and Vidyashankar AN. Assessing genome-wide statistical significance for large p small n problems. *Genetics* 194 (2013), 781–783

[3] Breiman L. Random Forests. *Mach Learn* 45(5) (2001)

[4] Vigneau E, Devaux MF, Qannari EM and Robert P. Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration. *J Chemometr* 11(3) (1997), 239-249.

[5] Barker M, Rayens W. Partial least squares for discrimination. *J Chemometr* 17(3) (2003), 166-73.

[6] Smilde AK. Three-way analyses problems and prospects. *Chemom. Intell. Lab. Syst*. 15(2-3) (1992), 143-57.

[7] Andersson CA, Bro R. The N-way Toolbox for MATLAB *Chemom. Intell. Lab. Syst*. 52 (1) (2000), 1-4.

[8] Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, Ballestar E, Bongcam-Rudloff E, Conesa A, Tegnér J. Data integration in the era of omics: current and future challenges. *BMC Syst Biol*. 8(2) (2014), I1.

[9] Favilla S, Durante C, Vigni ML, Cocchi M. Assessing feature relevance in NPLS models by VIP. *Chemom. Intell. Lab. Syst*. 129 (2013), 76-86.

[10] Donoho D. and Elad M. Optimally Sparse Representation in General (Nonorthogonal) Dictionaries via L1-Norm Minimizations. Proceedings of the National Academy of Science USA 1005 (2002), 2197–2202

[11] Bro R. Multilinear Calibration. Multilinear PLS, *J Chemometr* 10 (1996), 31–45.

[12] Bro R, Smilde AK, de Jong S. On the Difference Between Low-rank and Subspace Approximation: Improved Model for Multi-linear PLS regression. *Chemom. Intell. Lab. Syst*. 58(1) (2001), 3-13.

[13] Smilde A, Bro R, Geladi P. *Multi-way Analysis: Applications in the Chemical Sciences* (2005). John Wiley & Sons.

[14] Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *J. Roy. Stat. Soc. B Met*. 58(1) (1996). 267-288

[15] Zou H, Hastie T, Tibshirani R. Sparse principal component analysis. *J Comput Graph Stat*. 15(2) (2006), 265-86.

[16] Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw*. 39(5) (2011), 1.

[17] Lê Cao KA, Rossouw D, Robert-Granié C, Besse P. A sparse PLS for variable selection when integrating omics data. Statistical applications in genetics and molecular biology. 7(1) (2008), 35-2.[18] Heijne WH, Slitt AL, van Bladeren PJ, Groten JP, Klaassen CD, Stierum RH, van Ommen B. Bromobenzene-induced hepatotoxicity at the transcriptome level, Toxicological Science 79 (2004) 411–422.

[19] Hervas D. sNPLS: NPLS Regression with L1 Penalization. R package version 0.2.0. 2017 Jun. https://CRAN.R-project.org/package=sNPLS

[20] Hervás D, Prats-Montalbán JM, Lahoz A and Ferrer A. Sparse N-way Partial Least Squares with R Package sNPLS. *Chemom. Intell. Lab. Syst*. 179 (2018), 54-63.

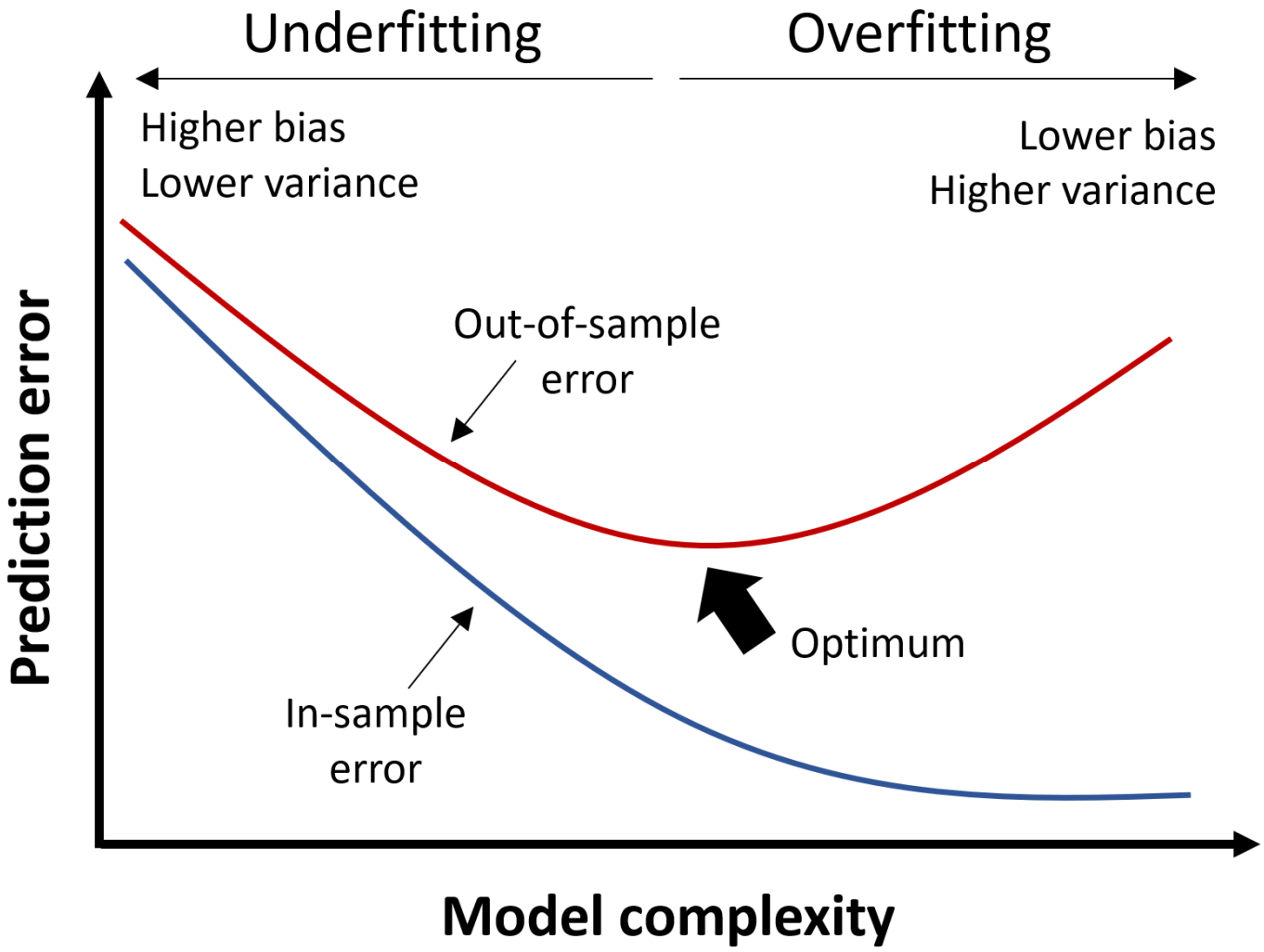


Figure 1: Bias-variance trade-off as a function of model complexity. In-sample error (error on the same data used to fit the model) and out-of sample error (error on new data) are represented along different values of model complexity. Increasing the amount of L1-penalization reduces model complexity, thus producing an increase in bias at the same time it produces a decrease in variance. This results in less overfitting.

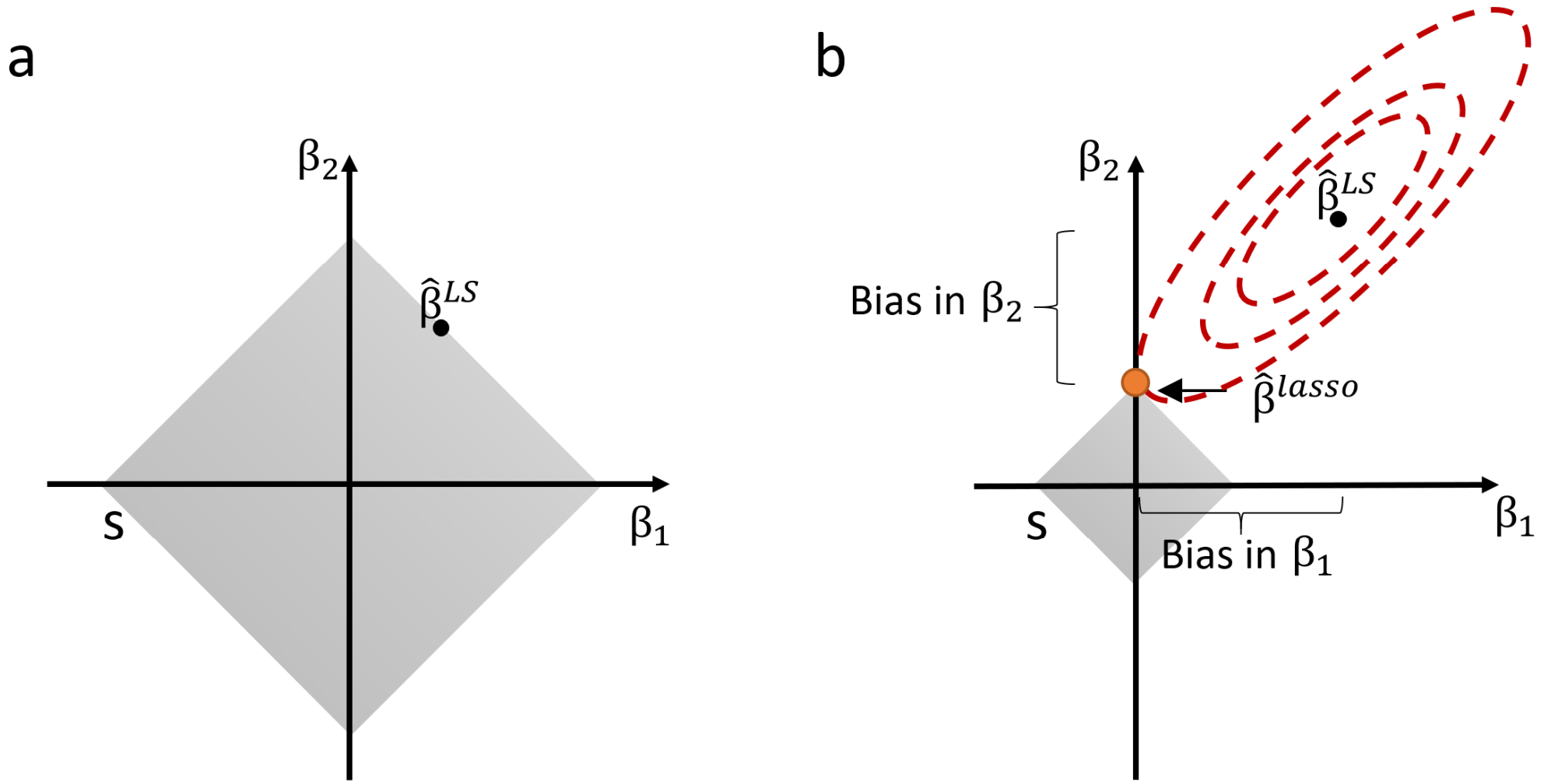


Figure 2: : Effect of the parameter *s*, imposing the L1-penalization restriction, in the estimation of the model coefficients. When *s* is large, the estimated least square coefficients are not modified, since they lie inside the defined restriction space (grayed area). When *s* decreases, coefficients are shrunken towards zero in order to satisfy the imposed restriction.

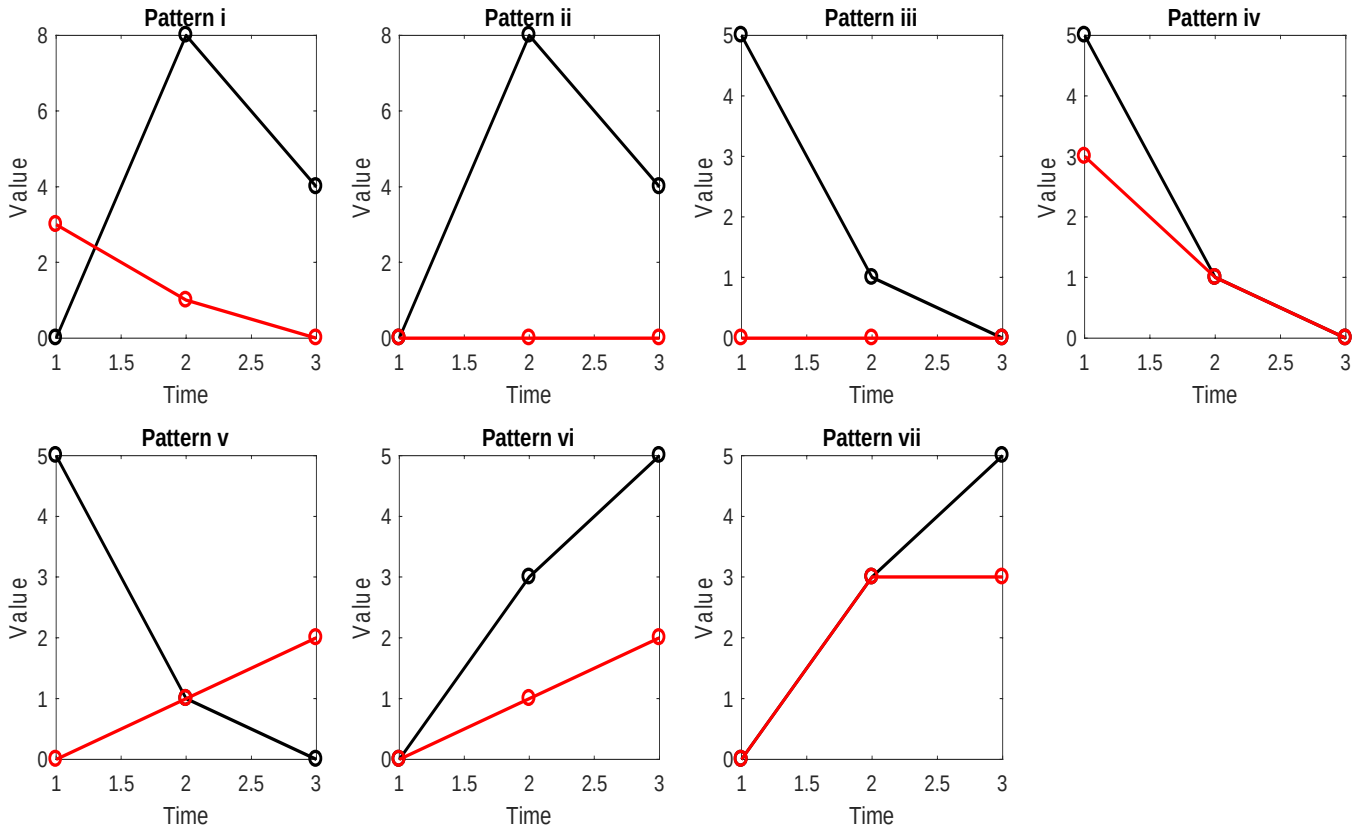


Figure 3: Patterns for the 14 different variables in both groups: High doses (red lines) and low doses (black lines).

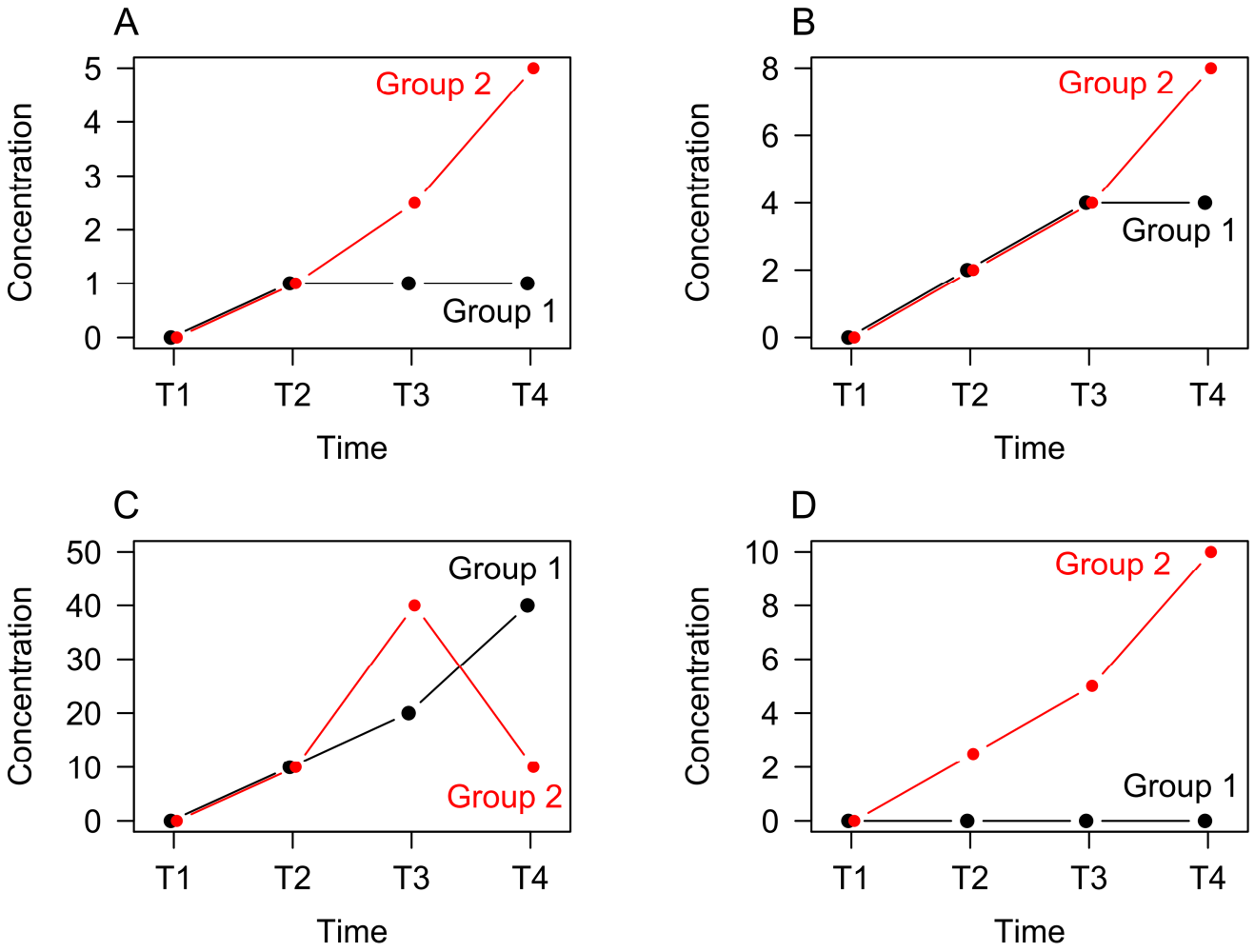


Figure 4: Expected patterns for the four different metabolite classes in group 1 (n=8) and group 2 (n=6).

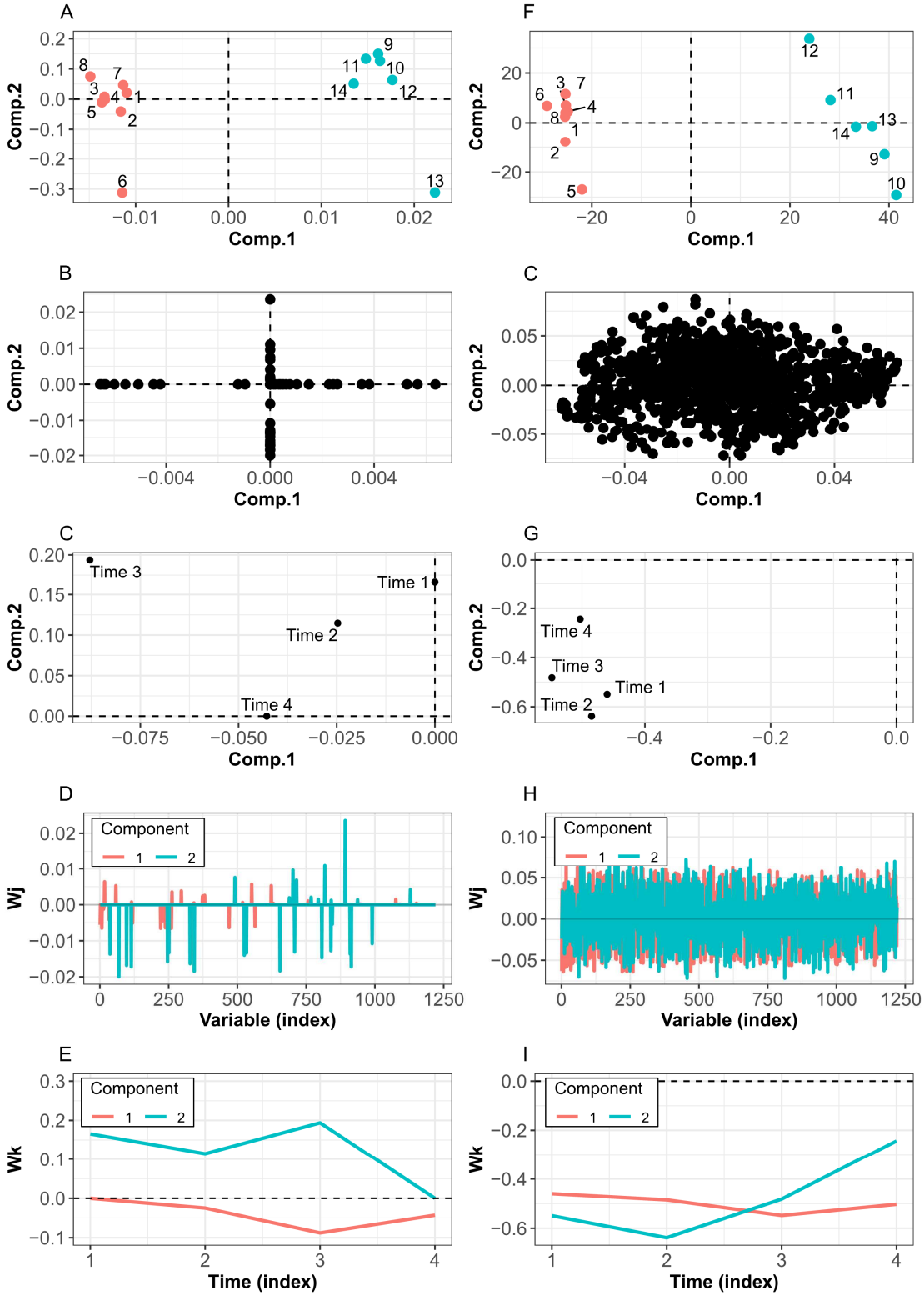


Figure 5. Plots of the Sparse *N*-PLS model (left), and the *N*-PLS model (right). Score plots of the two first components in the **T** matrix (A, F); weighting plots of the $\mathbf{W}^J$ matrix (B, G); weighting plot of the $\mathbf{W}^K$ matrix (C, H); plots of the loadings of the second (D, I) and third (E, J) modes.

Table 1. List of metabolites for each variable class. Metabolites are grouped attending to their physical and chemical properties. Class A, comprises fatty acid; Class B, comprises bile acids; Class C, comprises amino acids; Class D comprises miscellaneous compounds.

Variable Class	Metabolites
A	Capric Acid, Lauric Acid, Myristic Acid, Myristoleic Acid, Palmitic Acid, Palmitoleic Acid, Octadecanoic Acid, Oleic Acid, Linoleic Acid, Linolenic Acid
B	Cholic acid, Glycocholic acid, Taurocholic acid, Chenoeoxycholic acid, Glycochenodeoxycholic acid, Taurochenodeoxycholic acid, Deoxycholic acid, Glycodeoxycholic acid, Taurodeoxycholic acid, Lithocholic aid, Glycolithocholic acid, Taurolithocholic acid
C	Valine, Leucine, Isoleucine, Phenylalanine, Methionine, Cysteine, Proline, Tyrosine, Aspartic acid, Alanine, Glycine, Lysine
D	Ornithine, Glutamate, Glutamine, Citrulline, Arginine, Argininosuccinic Acid, γ-glutamyl-glutamic acid, γ-glutamyl-glutamine, γ-glutamyl-2aminobutyric acid, ophthalmic acid

Table 2. Results of the analyses performed using *N*-PLS and Sparse *N*-PLS on the different simulations. Median (1st, 3rd quartile) of the mean squared error and a 95% confidence interval for the difference in mean squared error between Sparse *N*-PLS and *N*-PLS is also provided. True variables selected column indicates the median of the occasions these are included in the models, as well as the 1st and 3rd quartiles (True positives). Noise variables selected column presents analogous results for the Noise variables (False positives).

	Mean Squared Error			Variable selection	
	N-PLS	sparse *N*-PLS	95% CI for difference	True variables selected	Noise variables selected
One time					
Normal error (sd=1)	85.58 (71.36, 96.24)	66.23 (52.72, 90.54)	[-21.97, -9.72]	5 (5, 5)	3 (1, 6)
Normal error (sd=1.5)	103.16 (85.98, 114.21)	90.71 (70.54, 121.8)	[-16,32, 0.61]	5 (5, 5)	2 (1, 6)
Normal error (sd=2)	116.2 (101.05, 135.47)	99.17 (84.34, 121.62)	[-23.93, -8.02]	5 (4.75, 5)	4 (2, 12)
Normal error (sd=3)	166.33 (134.1, 199.73)	149.57 (117.21, 198.9)	[-26.84, 2.10]	5 (3, 5)	4 (1, 11)
Poisson error (mean=1)	86.46 (70.46, 99.98)	60.03 (45.83, 101.65)	[-26.66, -9.52]	5 (5, 5)	3 (1, 5)
Poisson error (mean=3)	109.21 (92.02, 131.57)	83.08 (67.22, 107.67)	[-32.09, -16.21]	5 (5, 5)	3 (1, 6)
Poisson error (mean=5)	138.29 (108.9, 153.86)	99.55 (78.81, 125.76)	[-40.58, -21.47]	5 (4,5)	5 (2, 11)
Three times					
Normal error (sd=1)	165.1 (137.81, 202.86)	107.97 (91.79, 126.32)	[-68.02, -46.07]	5 (3,5)	5 (2, 11)
Normal error (sd=1.5)	171.95 (147.6, 206.47)	115.75 (98.58, 133.18)	[-69.71, -49.51]	5 (3, 5)	6 (2, 12)
Normal error (sd=2)	195.57 (159.96, 230.1)	122.73 (100.32, 145.4)	[-85.29, -59.27]	4 (3, 5)	7 (3, 19)
Normal error (sd=3)	245.71 (199.71, 292.2)	130.97 (107.48, 169.2)	[-125.2, -92.03]	4 (3, 4.5)	7 (2, 18)
Poisson error (mean=1)	153.67 (135.87, 193.6)	103.92 (85.16, 123.6)	[-65.17, -45.07]	5 (3, 5)	5 (2, 5)
Poisson error (mean=3)	186.22 (154.9, 213.56)	115.77 (91.33, 134.17)	[-79.64, -57.47]	4 (3, 5)	6 (3, 13)
Poisson error (mean=5)	205.18 (169.3, 236.78)	118.58 (94.83, 140.54)	[-98.13, -73.24]	4 (3, 5)	6 (3, 15)

Table 3.Coefficients of the model Coefficients:

	T1	T2	T3
Glucose	0	0	0.000
Phospholipids	0	0	0.000
Kidney	0	0	0.000
Liver	0	0	0.000
Cholesterol	0	0	0.000
Tryglycerids	0	0	0.000
A/G ratio	0	0	0.077
Urea	0	0	0.000
Creatinine	0	0	0.082
Albumin	0	0	0.076
ALAT	0	0	0.221
ASAT	0	0	0.221
LDH	0	0	0.221
GSH	0	0	0.101

Table 4: Variables selected by the final sparse *N*-PLS model and their corresponding assigned variable classes.

Variable	Component	Variable Class	Profile similar to variable class
V8, V16	1	A	-
V27, V28, V32	1	B	-
V54	1	C	-
V58	2	D	-
V187, V466, V853	1	-	A
V470	1	-	B
V388, V405, V422, V660, V661, V672	1	-	C
V112, V151, V179, V434, V449, V587, V608, V612, V967, V990	2	-	D
V95, V180, V527, V955, V1034, V1056, V1165, V1183, V1512, V2041, V2463, V2520, V2683	1	-	-
V897, V1235, V1322, V1354, V1378, V1389, V1535, V1601, V1627, V1647, V1711, V1715, V1729, V1873, V1935, V1945, V2011, V2077, V2180, V2616	2	-	-

- A new version of *N*-PLS for embedding LASSO-based variable selection, Sparse *N*-PLS, is presented.
- Both *N*-PLS and Sparse *N*-PLS are compared in a metabolomics scenario
- Sparse *N*-PLS method enables variable selection and simplifies data interpretation.