**Marie Perrot-Dockès[1] / Céline Lévy-Leduc[1] / Julien Chiquet[1] / Laure Sansonnet[1] / Margaux Brégère[1] / Marie-Pierre Étienne[1] / Stéphane Robin[1] / Grégory Genta-Jouve[2]**

# A variable selection approach in the multivariate linear model: an application to LC-MS metabolomics data

[1] UMR MIA-Paris, AgroParisTech, INRA – Université Paris-Saclay, 75005 Paris, France, E-mail: marie.perrot-dockes@agroparistech.fr
[2] UMR CNRS 8638 Comète – Université Paris-Descartes, CNRS, 75006 Paris, France

**Abstract:**
Omic data are characterized by the presence of strong dependence structures that result either from data acquisition or from some underlying biological processes. Applying statistical procedures that do not adjust the variable selection step to the dependence pattern may result in a loss of power and the selection of spurious variables. The goal of this paper is to propose a variable selection procedure within the multivariate linear model framework that accounts for the dependence between the multiple responses. We shall focus on a specific type of dependence which consists in assuming that the responses of a given individual can be modelled as a time series. We propose a novel Lasso-based approach within the framework of the multivariate linear model taking into account the dependence structure by using different types of stationary processes covariance structures for the random error matrix. Our numerical experiments show that including the estimation of the covariance matrix of the random error matrix in the Lasso criterion dramatically improves the variable selection performance. Our approach is successfully applied to an untargeted LC-MS (Liquid Chromatography-Mass Spectrometry) data set made of African copals samples. Our methodology is implemented in the R package `MultiVarSel` which is available from the Comprehensive R Archive Network (CRAN).

**Keywords:** metabolomics, multivariate linear model, time series, variable selection
**DOI:** 10.1515/sagmb-2017-0077

## 1 Introduction

Metabolomics aims to provide a global snapshot (quantitative or qualitative) of the metabolism at a given time and by extension phenotypic information (see Nicholson, Lindon & Holmes , 1999). It studies the concentration of small molecules called metabolites that are the end products of the enzymatic machinery of the cell. Indeed, minor variations in gene or protein expression levels that are not observable via high throughput experiments may have an influence on the metabolites and hence on the phenotype of interest. Thus, metabolomics is a promising approach that can advantageously complement usual transcriptomic and proteomic analyses. For further details on metabolomics, we refer the reader to Smith, Mathis, and Prince (2014). The analysis of the metabolomic biological samples is often performed using High Resolution Mass Spectrometry (HRMS), Nuclear Magnetic Resonance (NMR) or Liquid Chromatography-Mass Spectrometry (LC-MS) and produces a large number of features (hundreds or thousands) that can explain a difference between two or more populations (see Zhang et al., 2012). It is well-known in the untargeted LC-MS data analysis that the identification of metabolites discriminating these populations remains a major bottleneck and therefore the selection of relevant features (metabolites) is a crucial step, as explained in Verdegem et al. (2016). Our goal is to tackle the task of feature selection by taking advantage of the specificities of the LC-MS spectra.

We consider a typical untargeted metabolomic experiment where LC-MS spectra (intensity vs. $m/z$) are obtained from $n$ samples, resulting in an $n \times q$ data matrix where the $q$ columns are ordered according to their $m/z$ ratio. Note that the abbreviation $m/z$ represents the quantity formed by dividing the ratio of the mass of an ion to the unified atomic mass unit, by its charge number (regardless of sign). Figure 1 displays an example of such a spectrum. It has to be noticed that the data were first pre-processed using the methodology described in Section 4.1. We further assume that the $n$ samples are collected under $C$ conditions and denote $n_c$ the number of samples from Condition $c$, hence $\sum_c n_c = n$. Multivariate ANOVA (MANOVA, see e.g. Mardia, Kent & Bibby

**Marie Perrot-Dockès** is the corresponding author.

, 1979; Muller & Stewart, 2006) provides a natural framework to analyze such a data set. Denoting $\boldsymbol{Y}_{c,r}$ the $q$-dimensional vector corresponding to the spectrum from the $r$th replicate in Condition $c$, the MANOVA model assumes that

$$\boldsymbol{Y}_{c,r} = \boldsymbol{\mu}_c + \boldsymbol{E}_{c,r}, \tag{1}$$

where $\boldsymbol{\mu}_c$ is the theoretical mean spectrum in Condition $c$ and $\boldsymbol{E}_{c,r}$ is a random $q$-dimensional error vector. Each metabolite corresponds to a given component of these three vectors. A "relevant" feature is then defined as the $j$th $m/z$ value, the theoretical concentration $\mu_c^{(j)}$ of which significantly varies between conditions. Stacking the row vectors $\boldsymbol{Y}_{c,r}$ and $\boldsymbol{E}_{c,r}$, the MANOVA model can be rephrased as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}, \tag{2}$$

where $\boldsymbol{Y} = (Y_{i,j})_{1\leq i\leq n,\ 1\leq j\leq q}$ is the $n \times q$ observation matrix, $\boldsymbol{X}$ is the $n \times C$ design matrix of a one-way ANOVA model, $\boldsymbol{B} = (\mu_c^{(j)})_{1\leq c\leq C,\ 1\leq j\leq q}$ is the $C \times q$ coefficient matrix and $\boldsymbol{E} = (E_{i,j})_{1\leq i\leq n,\ 1\leq j\leq q}$ is the $n \times q$ random error matrix. Observe that $C$ corresponds to the number of covariates. For notational simplicity, the samples indexed with $(c, r)$ are now identified with a single index $i \in \{1, \ldots, n\}$, starting with the $n_1$ samples from Condition $c = 1$, then the $n_2$ samples from Condition $c = 2$, etc. In this framework, assuming that the mean spectrum $\overline{\boldsymbol{\mu}} = n^{-1}\sum_n n_c\boldsymbol{\mu}_c$ is set to zero, the problem of determining which metabolites are relevant boils down to finding the non null coefficients in the matrix $\boldsymbol{B}$ and hence can be seen as a variable selection problem in the multivariate linear model. Several approaches can be considered for solving this task: either *a posteriori* methods such as classical statistical tests in ANOVA models (see Mardia, Kent & Bibby , 1979; Faraway, 2004) or methods embedding the variable selection such as Lasso-type methodologies (Tibshirani, 1996). However, a naive application of such approaches does not take into account the potential dependence between the different columns of $\boldsymbol{Y}$, which may affect the identification of the relevant features. This drawback will be illustrated in Section 3.

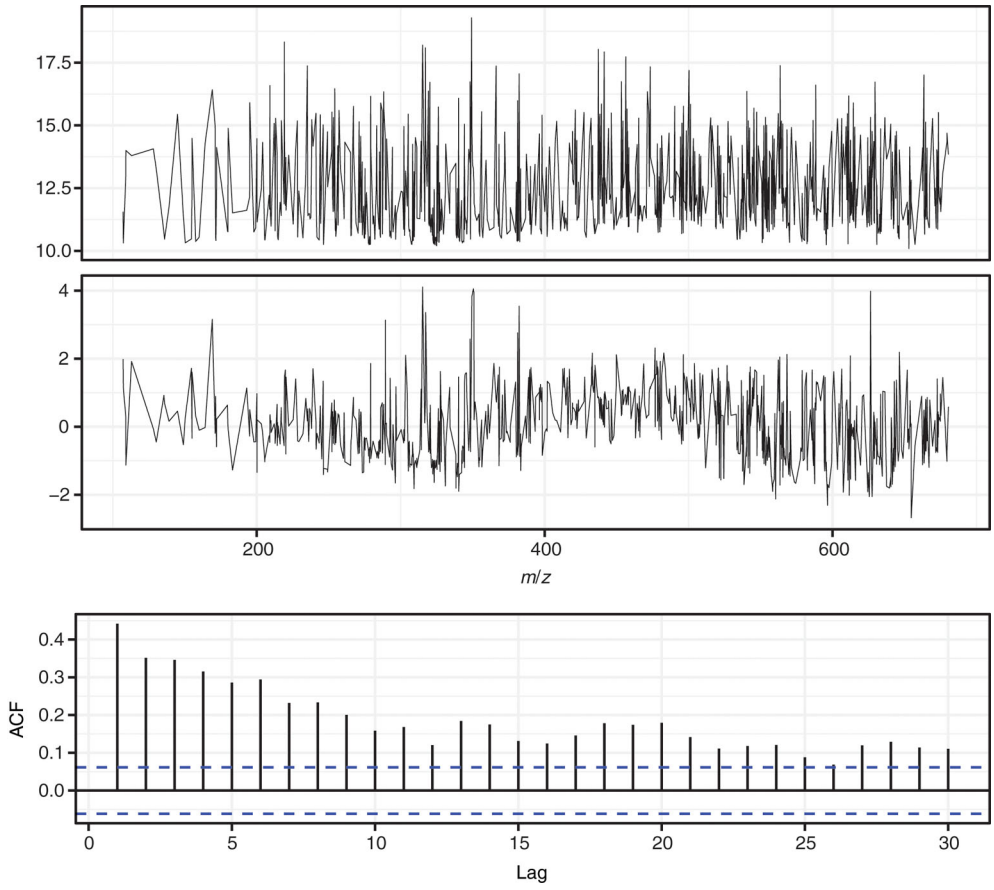


**Figure 1:** An example of a LC-MS spectrum (an instance of $\boldsymbol{Y}_{c,r}$) (top), the same spectrum centered and normalized (middle) and its empirical autocorrelation function (bottom).

Different supervised machine learning approaches have been used to analyze "omics" data during the last few years (see Saccenti et al., 2013; Ren et al., 2015; Boccard & Rudaz, 2016; Zhang et al., 2017). Among them, in

metabolomics, the most popular is the partial least squares-discriminant analysis (PLS-DA) which has recently been extended to sPLS-DA (sparse partial least squares-discriminant analysis) by Lê Cao, Boitard, and Besse (2011) to include a variable selection step.

The originality of our approach lies in the modeling of the dependence that exists among the columns of $\boldsymbol{Y}$ which comes from the fact that usually biomarkers share biosynthetic pathways (Audoin et al. 2014). To account for this dependence, we assume that the samples are all independent, namely, all the rows of $\boldsymbol{E}$ are independent and for each sample $i$, the noise vector $\boldsymbol{E}_i$ has a multivariate Gaussian distribution:

$$\boldsymbol{E}_i = (E_{i,1}, \dots, E_{i,q}) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_q), \tag{3}$$

where $\boldsymbol{\Sigma}_q$ denotes the covariance matrix. The simplest assumption regarding the covariance matrix is $\boldsymbol{\Sigma}_q = \sigma^2 \boldsymbol{I}_q$, where $\boldsymbol{I}_q$ denotes the $q \times q$ identity matrix. In this case the different columns of $\boldsymbol{Y}$ are assumed to be independent. The other extreme assumption consists in letting $\boldsymbol{\Sigma}_q$ free, assuming no specific form for this dependence. However, in such a situation, $q(q+1)/2$ parameters should be estimated which is not possible when $n < q$, which is the most standard case. Our approach lies in between, assuming that some dependence exists but that it has a specific structure. The form we consider is motivated by the nature of LC-MS spectra, which can be seen as random functions of the $m/z$ ratio. This suggests to consider each random vector $\boldsymbol{E}_i$ as a time-series and to borrow classical dependence structure from time-series analysis to model $\boldsymbol{\Sigma}_q$. This approach is consistent with the fact that the empirical autocorrelation function of LC-MS spectra (see Figure 1 for an example) displays the typical characteristics of most time-series such as vanishing autocorrelation when the lag increases.

On top of accounting for the dependence between the columns of $\boldsymbol{Y}$, our methodology can deal with a potentially high number of features (columns of $\boldsymbol{Y}$) thanks to the underlying Lasso-based feature selection and the modeling of the dependence which produces sparse estimates of $\boldsymbol{\Sigma}_q^{-1}$. We also couple the whole procedure to a stability selection step to ensure robustness of the selected features. This methodology is implemented in the R package `MultiVarSel` which is available from the Comprehensive R Archive Network (CRAN).

The rest of the paper is organized as follows. Our method is described in Section 2. Some numerical experiments on synthetic data are provided in Section 3. Finally, an application to a metabolomic data set made of African copals samples is given in Section 4.

## 2 Statistical inference

The strategy that we propose can be summarized as follows.

- First step: Fitting a one-way ANOVA to each column of the matrix $\boldsymbol{Y}$ in order to have access to an estimation $\hat{\boldsymbol{E}}$ of the error matrix $\boldsymbol{E}$.
- Second step: Estimating the matrix $\boldsymbol{\Sigma}_q$ by using the methods described in Sections 2.1.1 and 2.1.2. Then, choosing the most convenient estimator $\hat{\boldsymbol{\Sigma}}_q$ thanks to a statistical test described in Section 2.1.3.
- Third step: Thanks to $\hat{\boldsymbol{\Sigma}}_q$, transforming the data in order to remove the dependence between the columns of $\boldsymbol{Y}$. Such a transformation will be called "whitening" hereafter.
- Fourth step: Applying to the transformed observations the Lasso approach described in Section 2.2.

The first step provides a first estimate $\widetilde{\boldsymbol{B}}$ of $\boldsymbol{B}$. An estimate of $\boldsymbol{E}$ is then defined as

$$\hat{\boldsymbol{E}} = \boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{B}}. \tag{4}$$

In the following, we shall focus on the three other steps.

### 2.1 Estimation of the dependence structure of $\boldsymbol{E}$

We propose hereafter to model each row of $\boldsymbol{E}$ as a realization of a stationary process and hence we shall use time-series models in order to describe the dependence structure of $\boldsymbol{E}$. We refer the reader to Brockwell and Davis (1991) for further details on time series modeling.

We shall consider a large variety of models ranging from the simplest parametric to the most general non-parametric dependence modeling. In each case we focus on the estimation of $\boldsymbol{\Sigma}_q^{-1/2}$ since the use of the following transformation:

$$\boldsymbol{Y}\,\boldsymbol{\Sigma}_q^{-1/2} = \boldsymbol{X}\boldsymbol{B}\,\boldsymbol{\Sigma}_q^{-1/2} + \boldsymbol{E}\,\boldsymbol{\Sigma}_q^{-1/2} \tag{5}$$

removes the dependence between the columns of $\boldsymbol{Y}$. Indeed the covariance matrix of each row of $\boldsymbol{E}\boldsymbol{\Sigma}_q^{-1/2}$ is now equal to the identity matrix. Such a procedure will be called "whitening" hereafter.

### 2.1.1 Parametric dependence

The simplest model among the parametric models is the autoregressive process of order 1 denoted AR(1). More precisely, for each $i$ in $\{1, \ldots, n\}$, $E_{i,t}$ satisfies the following equation:

$$E_{i,t} - \phi_1 E_{i,t-1} = W_{i,t}, \text{ with } W_{i,t} \sim WN(0, \sigma^2), \tag{6}$$

where $|\phi_1| < 1$ and $WN(0, \sigma^2)$ denotes a zero-mean white noise process of variance $\sigma^2$, defined as follows,

$$Z_t \sim WN(0, \sigma^2) \text{ if } \begin{cases} \mathbb{E}(Z_t) = 0, \\ \mathbb{E}(Z_t Z_{t'}) = 0 \text{ if } t \neq t', \\ \mathbb{E}(Z_t^2) = \sigma^2. \end{cases} \tag{7}$$

Note that the closer to one the parameter $\phi_1$ the stronger the dependence between the $E_{i,t}$'s.

In this case, the inverse of the square root of the covariance matrix $\boldsymbol{\Sigma}_q$ of $(E_{i,1}, \ldots, E_{i,q})$ has a simple closed-form expression given by

$$\boldsymbol{\Sigma}_q^{-1/2} = \begin{pmatrix} \sqrt{1-\phi_1^2} & -\phi_1 & 0 & \cdots & 0 \\ 0 & 1 & -\phi_1 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & -\phi_1 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}. \tag{8}$$

Hence, to obtain the expression of $\hat{\boldsymbol{\Sigma}}_q^{-1/2}$, it is enough to have an estimation of the parameter $\phi_1$ and to replace it in (8). For this, we use the estimator $\hat{\boldsymbol{E}}$ defined in (4) and obtained by fitting a standard ANOVA model to the observations, which corresponds to the first step of our method. Then $\phi_1$ is estimated by $\widehat{\phi}_1$ defined by

$$\widehat{\phi}_1 = \frac{1}{n}\sum_{i=1}^{n} \widehat{\phi}_{1,i},$$

where $\widehat{\phi}_{1,i}$ denotes the estimator of $\phi_1$ obtained by the classical Yule-Walker equations from $(\hat{E}_{i,1}, \ldots, \hat{E}_{i,q})$, see Brockwell and Davis (1991) for more details.

More generally, it is also possible to have access to $\boldsymbol{\Sigma}_q^{-1/2}$ for more complex processes such as the ARMA($p$, $q$) process defined as follows: For each $i$ in $\{1, \ldots, n\}$,

$$E_{i,t} - \phi_1 E_{i,t-1} - \ldots - \phi_p E_{i,t-p} = W_{i,t} + \theta_1 W_{i,t-1} + \ldots \theta_q W_{i,t-q}, \tag{9}$$

where $W_{i,t} \sim WN(0, \sigma^2)$, the $\phi_i$'s and the $\theta_i$'s are real parameters.

### 2.1.2 Nonparametric dependence case

In the situation where a parametric modeling is not relevant for $\boldsymbol{\Sigma}_q$, it can be estimated by

$$\hat{\Sigma}_q = \begin{pmatrix} \widehat{\gamma}(0) & \widehat{\gamma}(1) & \cdots & \widehat{\gamma}(q-1) \\ \widehat{\gamma}(1) & \widehat{\gamma}(0) & \cdots & \widehat{\gamma}(q-2) \\ \vdots & & & \\ \widehat{\gamma}(q-1) & \widehat{\gamma}(q-2) & \cdots & \widehat{\gamma}(0) \end{pmatrix}, \tag{10}$$

with

$$\widehat{\gamma}(h) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\gamma}_i(h),$$

where $\widehat{\gamma}_i(h)$ is the standard autocovariance estimator of $\gamma_i(h) = \mathbb{E}(E_{i,t}E_{i,t+h})$, for all $t$. Usually, $\widehat{\gamma}_i(h)$ is referred to as the empirical autocovariance of the $\hat{E}_{i,t}$'s at lag $h$ (*i.e.* the empirical covariance between $(\hat{E}_{i,1}, \dots, \hat{E}_{i,n-h})$ and $(\hat{E}_{i,h+1}, \dots, \hat{E}_{i,n})$). For a definition of the standard autocovariance estimator we refer the reader to Chapter 7 of Brockwell and Davis (1991). The matrix $\hat{\boldsymbol{\Sigma}}_q^{-1/2}$ is then obtained by inverting the Cholesky factor of $\hat{\boldsymbol{\Sigma}}_q$.

### 2.1.3 Choice of the whitening modeling

In order to decide which dependence modeling better fits the data at hand we propose hereafter a statistical test. If the whitening modeling used is well chosen then each row of $\widetilde{\boldsymbol{E}} = \hat{\boldsymbol{E}}\hat{\boldsymbol{\Sigma}}_q^{-1/2}$ should be a white noise as defined in (7), where $\hat{\boldsymbol{E}}$ is defined in (4).

One of the most popular approaches for testing whether a random process is a white noise or not, is the Portmanteau test which is based on the Bartlett theorem (for further details we refer the reader to Brockwell & Davis, 1991, Theorem 7.2.2). By this theorem, we get that under the null hypothesis ($H_0$): "For each $i$ in $\{1, \dots, n\}$, $(\widetilde{E}_{i,1}, \dots, \widetilde{E}_{i,q})$ is a white noise",

$$q\sum_{h=1}^{H} \hat{\rho}_i(h)^2 \approx \chi^2(H), \text{ as } q \to \infty, \tag{11}$$

for each $i$ in $\{1, \dots, n\}$, where $\hat{\rho}_i(h)$ denotes the empirical autocorrelation of $(\widetilde{E}_{i,1}, \dots, \widetilde{E}_{i,q})$ at lag $h$ and $\chi^2(H)$ denotes the chi-squared distribution with $H$ degrees of freedom. Thus, by (11), we have at our disposal a $p$-value for each $i$ in $\{1, \dots, n\}$ that we denote by $\text{Pval}_i$. In order to have a single $p$-value instead of $n$, we shall consider

$$q\sum_{i=1}^{n}\sum_{h=1}^{H} \hat{\rho}_i(h)^2 \approx \chi^2(nH), \text{ as } q \to \infty, \tag{12}$$

where the approximation comes from the fact that the rows of $\widetilde{\boldsymbol{E}}$ are assumed to be independent. Equation (12) thus provides a $p$-value: Pval. Hence, if Pval $< \alpha$, the null hypothesis ($H_0$) is rejected at the level $\alpha$, where $\alpha$ is usually equal to 5and a large value of Pval indicates that the modeling for the dependence structure of $\boldsymbol{E}$ is well chosen.

## 2.2 Estimation of $B$

### 2.2.1 Lasso based approach

Let us first explain briefly the usual framework in which the Lasso approach is used. We consider a high-dimensional linear model of the following form

$$\mathcal{Y} = \mathcal{X}\mathcal{B} + \mathcal{E}, \tag{13}$$

where $\mathcal{Y}$, $\mathcal{B}$ and $\mathcal{E}$ are vectors. Note that, in high-dimensional linear models, the matrix $\mathcal{X}$ has usually more columns than rows which means that the number of variables is larger than the number of observations but $\mathcal{B}$ is usually a sparse vector, namely it contains a lot of null components.

In such models a very popular approach initially proposed by Tibshirani (1996) is the Least Absolute Shrinkage and Selection Operator (LASSO), which is defined as follows for a positive $\lambda$:

$$\widehat{\mathscr{B}}(\lambda) = \operatorname{argmin}_{\mathscr{B}} \left\{ \|\mathscr{Y} - \mathscr{X}\mathscr{B}\|_2^2 + \lambda \|\mathscr{B}\|_1 \right\}, \tag{14}$$

where, for $u = (u_1, \dots, u_n)$, $\|u\|_2^2 = \sum_{i=1}^n u_i^2$ and $\|u\|_1 = \sum_{i=1}^n |u_i|$, i.e. the $\ell_1$-norm of the vector $u$. Observe that the first term of (14) is the classical least-squares criterion and that $\lambda\|\mathscr{B}\|_1$ can be seen as a penalty term. The interest of such a criterion is the sparsity enforcing property of the $\ell_1$-norm ensuring that the number of non-zero components of the estimator $\widehat{\mathscr{B}}$ of $\mathscr{B}$ is small for large enough values of $\lambda$.

This methodology cannot be directly applied to our model since we have to deal with matrices and not with vectors. Nevertheless, as explained in Appendix A, Model (2) can be rewritten as in (13) where $\mathscr{Y}$, $\mathscr{B}$ and $\mathscr{E}$ are vectors of size $nq$, $pq$ and $nq$, respectively. Hence, retrieving the positions of the non null components in $\mathscr{B}$ is a first approach for finding relevant variables. However, this approach does not take into account the dependence between the columns of $\boldsymbol{Y}$. Hence, we propose hereafter a modified version of the standard Lasso criterion (14) taking into account this potential dependence.

As explained previously, our contribution consists first in "whitening" the observations, namely removing the dependence that may exist within the observations matrix, by multiplying (2) on the right by $\hat{\boldsymbol{\Sigma}}_q^{-1/2}$, see (5) where $\boldsymbol{\Sigma}_q^{-1/2}$ is replaced by $\hat{\boldsymbol{\Sigma}}_q^{-1/2}$. By using the same vectorization trick that allows us to transform Model (2) into Model (13), the Lasso criterion can be applied to the vectorized version of Model (5) where $\boldsymbol{\Sigma}_q^{-1/2}$ is replaced by $\hat{\boldsymbol{\Sigma}}_q^{-1/2}$. The specific expressions of $\mathscr{Y}$, $\mathscr{X}$, $\mathscr{B}$ and $\mathscr{E}$ are given in Appendix B.

Note that this idea of "whitening" the observations has also been proposed by Rothman, Levina, and Zhu (2010) where the estimation of $\boldsymbol{\Sigma}_q$ and $\boldsymbol{B}$ is performed simultaneously. An implementation is available in the R package `MRCE`. In our approach, $\boldsymbol{\Sigma}_q$ is estimated first and then its estimator is used in (5) instead of $\boldsymbol{\Sigma}_q$ before applying the Lasso criterion. Hence, our method can be seen as a variant of the MRCE method in which $\boldsymbol{\Sigma}_q$ is estimated beforehand. Moreover, after some numerical experiments, we observed that for the values of $n$ and $q$ that we aim at using, the computational burden of the approach designed by Rothman, Levina, and Zhu (2010) is too high for addressing our datasets for fixed regularization parameters, contrary to ours. In addition, in practical situations, the regularization parameters of the MRCE approach have to be tuned. As a consequence, we have not been able to use the MRCE approach for the purpose we consider here.

#### 2.2.2 Model selection issue

Estimator (14) depends on a parameter $\lambda$ which tunes the sparsity level in $\widehat{\mathscr{B}}$. We propose to mix two standard approaches to estimate the positions of the non null components in $\mathscr{B}$: the 10-fold cross-validation method and the stability selection approach of Meinshausen and Buhlmann (2010) which guarantees the robustness of the selected variables.

We first divide our samples into ten groups and remove one group from the dataset thus creating 10 training sets: $Y^{\mathscr{D}_1}, \dots, Y^{\mathscr{D}_{10}}$. For each training set $Y^{\mathscr{D}_k}$, we apply the first three steps of our approach and the Lasso criterion with a 10-fold cross-validation procedure to get $\lambda_{CV}^{(k)}$. Then, we randomly select a subsample of size $q \times n_k/2$, where $q \times n_k$ denotes the size of the vector of observations $\mathscr{Y}^{\mathscr{D}_k} = \mathrm{Vec}(Y^{\mathscr{D}_k})$. We then apply the Lasso criterion to this subsample with $\lambda = \lambda_{CV}^{(k)}$ and store the indices $i$ of the non null $\hat{B}_i$. This random selection of a subsample of the training set and the application of the Lasso criterion are repeated $N$ times. At the end, we have access to the number of times $N_i^{(k)}$ where each component $\hat{B}_i$ is non null among the $N$ replications for each group $k$. We only keep in the final set of selected variables the indices $i$ such that $(\sum_{k=1}^{10}(N_i^{(k)}/N))/10$ is larger than a given threshold. The influence of $N$ and the choice of the threshold will be investigated in Section 3.

For some theoretical results supporting our approach we refer the reader to Perrot-Dockès et al. (2018).

## 3 Simulation study

The goal of this section is to assess the statistical performance of our methodology implemented in the R package `MultiVarSel`. In order to emphasize the benefits of using a whitening approach from the variable selection point of view, we shall first compare our approach to standard methodologies. Then, we shall analyze the performance of our statistical test for choosing the best dependence modeling. Finally, we shall investigate the performance of our model selection criterion.

To assess the performance of the different methodologies, we generate observations $\boldsymbol{Y}$ according to Model (2) with $q = 1000$, $p = 3$, $n = 30$ ($n_1 = 9$, $n_2 = 8$ and $n_3 = 13$) and different dependence modelings, namely different matrices $\boldsymbol{\Sigma}_q$ corresponding to the AR(1) model described in (6) with $\sigma = 1$ and $\phi_1 = 0.7$ or 0.9. Note that the values of the parameters $p$, $q$ and $n$ are chosen in order to match the metabolomic data analyzed in Section 4.

We shall also investigate the effect of the sparsity and of the signal to noise ratio (SNR). The sparsity level $s$ corresponds to the proportion of non null elements in $\mathcal{B}$. Different signal to noise ratios are obtained by multiplying $\boldsymbol{B}$ in (2) by a coefficient $\kappa$.

## 3.1 Variable selection performance

The goal of this section is to compare the performance of our different whitening strategies to standard existing methodologies. More precisely, we shall compare our approaches to the classical ANOVA method (denoted ANOVA), the standard Lasso (denoted Lasso), namely the Lasso approach without the whitening step and to sPLSDA (Lê Cao, Boitard & Besse , 2011), implemented in the `mixOmics` R package and also in MetaboAnalyst, which is widely used in the metabolomics field. By ANOVA, we mean the classical one-way ANOVA applied to each column of the observations matrix $\boldsymbol{Y}$ without taking the dependence into account. Our different whitening approaches (described in Sections 2.1.1 and 2.1.2) are denoted by AR1 and Nonparam. These methods are also compared to the Oracle approach where the matrix $\boldsymbol{\Sigma}_q$ is known, which is never the case in practical situations.

We shall use three classical criteria for comparison: ROC curves, AUC (Area Under the ROC Curve) and Precision-Recall (PR) curves. ROC curves display the true positive rates (TPR) as a function of the false positive rates (FPR) and the closer to one the AUC the better the methodology. PR curves display the Precision as a function of the Recall. Since the features selected by sPLSDA are not assigned to a given condition $c$, we shall consider that as soon as a feature is selected it is a true positive, which gives a great advantage to sPLSDA.

We can see from Figure 2, Figure 3 and Table 1 that in the case of an AR(1) dependence, taking into account this dependence provides better results than sPLSDA and than approaches that consider the columns of the matrix $\boldsymbol{E}$ as independent. Moreover, we observe that the performance of the non parametric modeling are on a par with those of the parametric and the oracle ones. We also note that the larger the sparsity level $s$ the smaller the difference of performance between the approaches. As expected, the larger the signal to noise ratio $\kappa$ the better the performance of the different methodologies. We also conducted numerical experiments in a balanced one-way ANOVA framework. Since the conclusions are similar, we did not report the results here but they are available upon request.

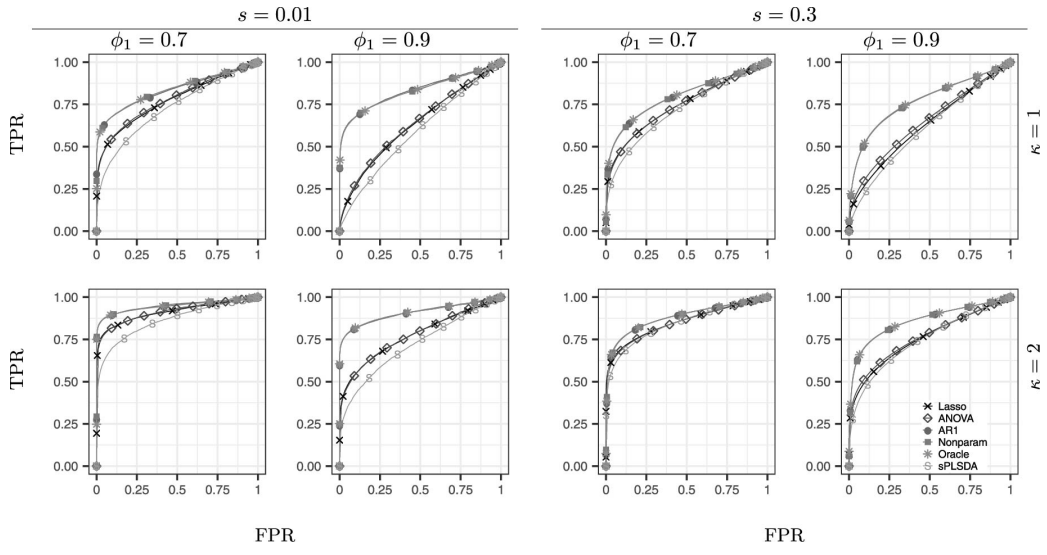


**Figure 2:** Means of the ROC curves obtained from 200 replications for the different methodologies in the AR(1) dependence modeling; $\kappa$ is linked to the signal to noise ratio (first row: $\kappa = 1$, second row $\kappa = 2$); $\phi_1$ is the correlation level in the AR(1) and $s$ the sparsity level (*i.e.* the fraction of nonzero elements in $\boldsymbol{B}$).

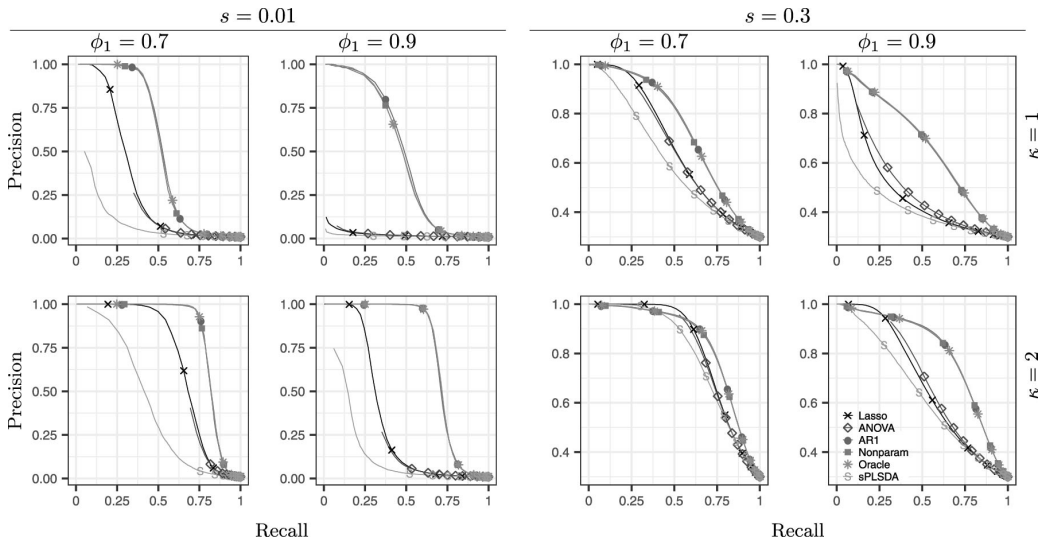


**Figure 3:** Means of the precision-recall curves obtained from 200 replications for the different methodologies in the AR(1) dependence modeling; $\kappa$ is linked to the signal to noise ratio (first row: $\kappa = 1$, second row $\kappa = 2$); $\phi_1$ is the correlation level in the AR(1) and $s$ the sparsity level (*i.e.* the fraction of nonzero elements in $B$).

**Table 1:** AUC of the different methods corresponding to Figure 2.

| SNR | $\phi_1$ | $s$ | Lasso | ANOVA | AR1 | Nonpar | Oracle | sPLSDA |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.7 | 0.01 | 0.78 | 0.78 | 0.83 | 0.84 | 0.84 | 0.73 |
| 1 | 0.7 | 0.3 | 0.74 | 0.74 | 0.80 | 0.80 | 0.80 | 0.72 |
| 1 | 0.9 | 0.01 | 0.63 | 0.64 | 0.83 | 0.83 | 0.83 | 0.58 |
| 1 | 0.9 | 0.3 | 0.63 | 0.64 | 0.77 | 0.77 | 0.77 | 0.61 |
| 2 | 0.7 | 0.01 | 0.91 | 0.91 | 0.95 | 0.95 | 0.95 | 0.86 |
| 2 | 0.7 | 0.3 | 0.85 | 0.85 | 0.88 | 0.88 | 0.88 | 0.84 |
| 2 | 0.9 | 0.01 | 0.77 | 0.77 | 0.91 | 0.91 | 0.91 | 0.72 |
| 2 | 0.9 | 0.3 | 0.75 | 0.76 | 0.86 | 0.86 | 0.86 | 0.74 |

## 3.2 Choice of the dependence modeling

The goal of this section is to assess the performance of the whitening test proposed in Section 2.1.3. We generated observations $\boldsymbol{Y}$ as described at the beginning of Section 3, with AR(1) dependence, a sparsity level $s = 0.01$ and SNR such that $\kappa = 1$. The corresponding results are displayed in Figure 4.

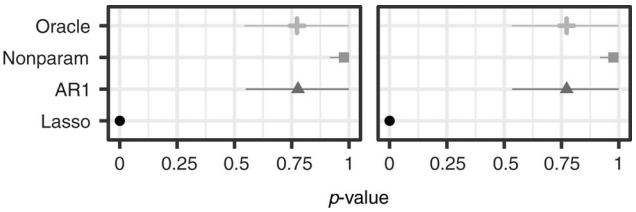


**Figure 4:** Means and standard deviations of the $p$-values of the test described in Section 2.1.3 of the main paper for the different approaches in the AR(1) dependence modeling when $\phi_1 = 0.7$ (left) and $\phi_1 = 0.9$ (right).

We observe that our test behaves properly: it provides $p$-values close to zero in the case where no whitening strategy is used (Lasso) and that when one of the proposed whitening approaches is used the $p$-values are larger than 0.7.

## 3.3 Choice of the model selection criterion

We investigate here the performance of our model selection criterion described in Section 2.2.2. Figure 5 displays the TPR and the FPR for different values $N$ of the sampling replicates and different thresholds. We can see from

this figure than taking $N$ larger than 1000 and a threshold of 0.999 ensures a small false positive rate and a large true positive rate.

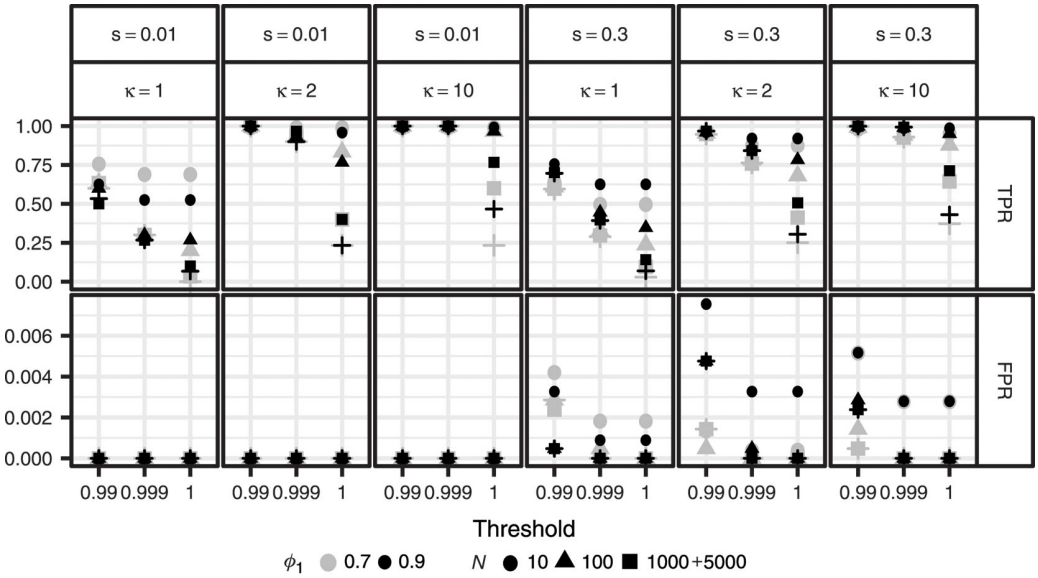


**Figure 5:** Influence of the number of replications $N$ and of the threshold.

Bullets ('•') in Figure 6 show the positions of the variables selected by our four-step approach for two possible thresholds (0.999 and 1) from $N = 1000$ replications. The positions of the non null coefficients in $\boldsymbol{B}$ are displayed with '+'. Here $\boldsymbol{Y}$ is generated with the parameters described at the beginning of Section 3 in the case of an AR(1) dependence with $\phi_1 = 0.9$ and $\kappa = 10$. We observe from this figure that the positions of the non null coefficients are recovered for both thresholds. However, the performance are slightly better when the threshold is equal to 0.999.

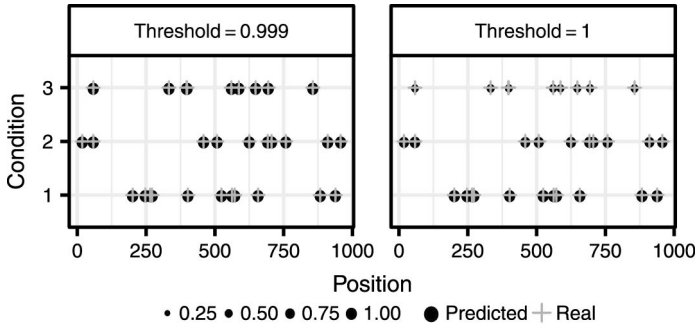


**Figure 6:** Positions of the variables selected by our approach ('•') when $\kappa = 10$.
Values on the $y$-axis correspond to the 3 conditions. The results obtained when the threshold is equal to 0.999 (resp. 1) are on the left (resp. on the right). The size of the bullets is all the more large that the selection frequency is high.

## 3.4 Numerical performance

In order to investigate the computational burden of our approach, we generated matrices $\boldsymbol{Y}$ satisfying Model (2) with $n = 30$ and $q \in \{100, 1000, 2000, \ldots, 5000\}$. Here, the rows of the matrix $\boldsymbol{E}$ are generated as realizations of an AR(1) process and the level of sparsity $s$ of $\boldsymbol{B}$ is equal to 0.01. Figure 7 displays the computational times of `MultiVarSel`, including the model selection step described in Section 2.2.2, for different number of replications in the stability selection stage. Timings were obtained on a workstation with 16 GB of RAM and Intel Core i7 (3.66GHz) CPU, using 8 cores for parallel computing. Our implementation uses the R language (R Core Team, 2017) and relies on the `glmnet` and `Matrix` packages (Friedman, Hastie & Tibshirani , 2010; Bates & Maechler, 2017). We can see from this figure that the computational burden of `MultiVarSel` is very low and that it takes only a few minutes to analyze matrices having 5000 columns.

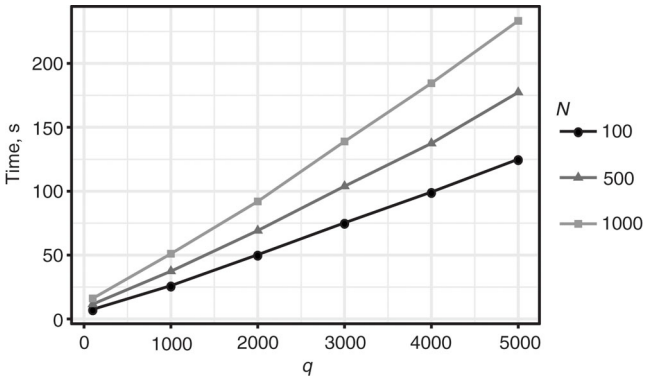


**Figure 7:** Computational times (in seconds) of `MultiVarSel`.
The number of replications corresponds to the number $N$ of subsamplings in the stability selection step.

## 4 Application to the analysis of a LC-MS data set

In this section, `MultiVarSel` is applied to a LC-MS (Liquid Chromatography-Mass Spectrometry) data set made of African copals samples. The samples correspond to ethanolic extracts of copals produced by trees belonging to two genera *Copaifera* (C) and *Trachylobium* (T) with a second level of classification coming from the geographical provenance of the *Copaifera* samples (West (W) or East (E) Africa). Since all the *Trachylobium* samples come from East Africa, we can use the modeling proposed in Equations (1) and (2) with $C = 3$ conditions: CE, CW and TE such that $n_{\text{CE}} = 9$, $n_{\text{CW}} = 8$ and $n_{\text{TE}} = 13$. Our goal is to identify the most important features (the $m/z$ values) for distinguishing the different conditions. In this section, we also compare the performance of our method with those of other techniques which are widely used in metabolomics.

### 4.1 Data pre-processing

LC-MS chromatograms were aligned using the `R` package `XCMS` proposed by Smith et al. (2006) with the following parameters: a signal to noise ratio threshold of 10:1 for peak selection, a step size of 0.2 min and a minimum difference in $m/z$ for peaks with overlapping retention times of 0.05 amu. Sample filtering was also performed: To be considered as informative, as suggested by Kirwan et al. (2013), a peak was required to be present in at least 80% of the samples. Missing values imputation was realized using the KNN algorithm described in Hrydziuszko and Viant (2012). Subsequently, the spectra were normalized to equalize signal intensities to the median profile in order to reduce any variance arising from differing dilutions of the biological extracts and probabilistic quotient normalization (PQN) was used, see Dieterle et al. (2006) for further details. In order to reduce the size of the data matrix which contains 6327 metabolites, selection of the adducts of interest $[M+H]^+$ was then performed using the CAMERA package of Kuhl et al. (2012). A $n \times q$ matrix $\boldsymbol{Y}$ was then obtained with $q = 1019$ and submitted to the statistical analyses.

### 4.2 Application of our four-step approach

The observations matrix $\boldsymbol{Y}$ is first centered and scaled.

- First step: A one-way ANOVA is fitted to each column of the observation matrix $\boldsymbol{Y}$ in order to have access to an estimation $\hat{\boldsymbol{E}}$ of the matrix $\boldsymbol{E}$. Then, the test proposed in Section 2.1.3 is applied to $\hat{\boldsymbol{E}}$ that is without "whitening" the observations. We found a $p$-value equal to zero which indicates that the columns of $\hat{\boldsymbol{E}}$ cannot be considered as independent and hence that applying the whitening strategy should improve the results.
- Second step: The different whitening strategies described in Section 2.1 were applied and the highest $p$-value for the test described in Section 2.1.3 is obtained for the nonparametric whitening. More precisely, the $p$-values obtained for the AR(1) and the nonparametric dependence modeling are equal to 0 and 0.664, respectively. Hence, in the following we shall use the nonparametric modeling.
- Third step: Observations were whitened with $\hat{\boldsymbol{\Sigma}}_q$ obtained by using the nonparametric modeling.

– Fourth step: The Lasso approach described in Section 2.2 was then applied to the whitened observations. The stability selection is used with $N = 1000$ replications and a threshold equal to 0.999.

Figure 8 displays the Venn diagram of the features ($m/z$ values) selected for each condition CE, TE and CW. Among the 1019 features, 98 features have been selected by MultiVarSel: 77 have been selected for Condition TE, 28 for Condition CW and 5 for Condition CE. Note that there were no features selected for all the conditions, 10 for both TE and CW and 2 for both CW and CE.

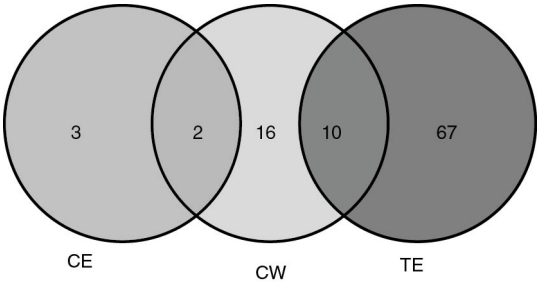


**Figure 8:** Venn diagram of the features selected for each condition by `MultiVarSel`.

## 4.3 Comparison with existing methods

The goal of this section is to compare our approach with the sparse partial least square discriminant analysis (sPLS-DA) which is classically used in metabolomics.

### 4.3.1 Additional simulations

Since in the case of real data, the position of the relevant features is of course unknown, we propose the following additional simulations in order to further compare these two approaches. We start by applying the first step of our approach in order to get $\hat{\boldsymbol{E}}$. Then, we perform $M$ random samplings with replacement among the rows of $\hat{\boldsymbol{E}}$. Let $E^\star$ denote one of them, then we generate a new observation matrix $\boldsymbol{Y}^\star = \boldsymbol{X}^\star\boldsymbol{B} + \boldsymbol{E}^\star$, where $X^\star$ is the same as $\boldsymbol{X}$ except that its rows are permuted in order to ensure a correspondence between the rows of $E^\star$ and $X^\star$. The matrix $\boldsymbol{B}$ is obtained as in Section 3 with $s = 0.01$ and $\kappa = 0.5$ and 1. ROC curves averaged over $M = 50$ random samplings are displayed in Figure 9. We can see from this figure that our approach outperforms the classical ones. Other values of $s$ and $\kappa$ have been tested. The corresponding results are not reported here but available upon request.

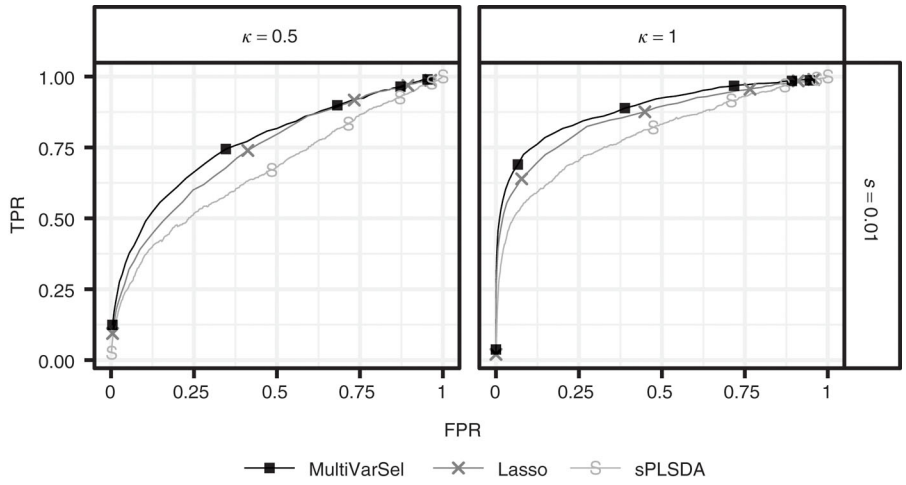


**Figure 9:** Means of the ROC curves obtained by `MultiVarSel`, Lasso and sPLS-DA.

### 4.3.2 Results on the LC-MS data set

As recommended by Lê Cao, Boitard, and Besse (2011), we used two components for sPLS-DA. Moreover, in order to make sPLS-DA comparable with our approach, 49 variables are kept for each component. However, as

explained in Section 3, the main difference between our approach and sPLSDA is that the features selected by sPLSDA are not assigned to a given condition $c$, and thus less interpretable.

Figure 10 displays the location of the features ($m/z$ values) selected by our approach and sPLS-DA. We can see from this figure that the features selected for the condition TE are mainly located between 400 and 500 $m/z$ whereas those selected for the condition CE are around 600 $m/z$. The features selected by the first component of the sPLS-DA are also mainly located between 400 and 500 $m/z$. However, as previously explained, the features selected by sPLSDA are assigned to a component built by the method and not to a condition of the experimental design. Venn diagrams comparing the features selected by both methods are available in Figure 11. We observe from these Venn diagrams that the features selected in each component of sPLS-DA do not characterize the conditions of the MANOVA model contrary to ours.

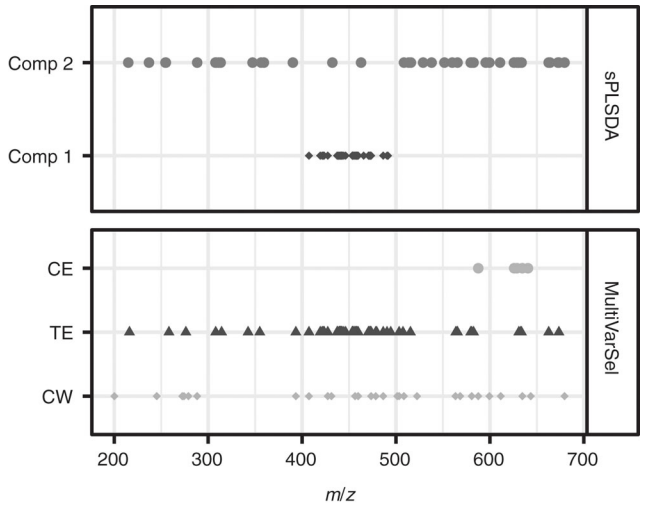


**Figure 10:** Comparison of the features selected by `MultiVarSel` and `sPLS-DA`.

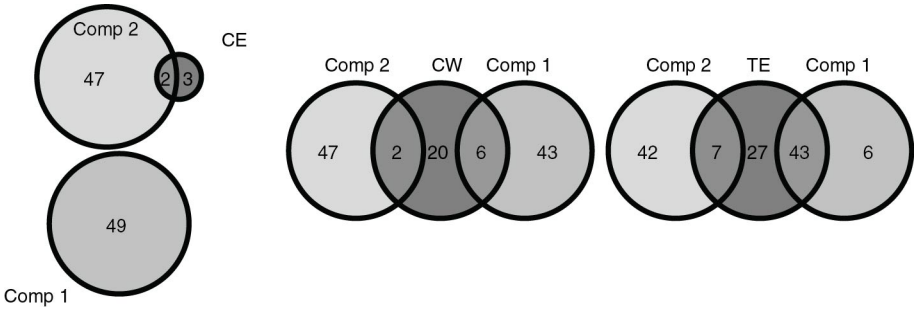


**Figure 11:** Venn diagrams comparing the features selected by `MultiVarSel` in the three conditions with those selected by sPLS-DA in its two components.

## 5 Conclusion

In this paper, we proposed a novel approach for feature selection taking into account the dependence that may exist between the columns of the observations matrix. Our approach is implemented in the R package `MultiVarSel` which is available from The Comprehensive R Archive Network (CRAN). We have shown that our method has two main features. Firstly, it is very efficient for selecting a restricted number of stable features characterizing each condition. Secondly, its very low computational burden makes its use possible on very large LC-MS metabolomics data.

## Acknowledgement

This project has been funded by La mission pour l'interdisciplinarité du CNRS in the frame of the DEFI ENVIROMICS (project AREA). The authors thank the Musée François Tillequin for providing the samples from the Guibourt Collection.

## Appendix A

Let $vec(\boldsymbol{A})$ denote the vectorization of the matrix $\boldsymbol{A}$ formed by stacking the columns of $\boldsymbol{A}$ into a single column vector. Let us apply the $vec$ operator to Model (2), then

$$vec(\boldsymbol{Y}) = vec(\boldsymbol{XB} + \boldsymbol{E}) = vec(\boldsymbol{XB}) + vec(\boldsymbol{E}).$$

Let $\mathscr{Y} = vec(\boldsymbol{Y})$, $\mathscr{B} = vec(\boldsymbol{B})$ and $\mathscr{E} = vec(\boldsymbol{E})$. Hence,

$$\mathscr{Y} = vec(\boldsymbol{XB}) + \mathscr{E} = (\mathrm{I}_q \otimes \boldsymbol{X})\mathscr{B} + \mathscr{E},$$

where we used that

$$vec(AXB) = (B' \otimes A)vec(X),$$

see (Mardia, Kent & Bibby , 1979, Appendix A.2.5). In this equation, $B'$ denotes the transpose of the matrix $B$. Thus,

$$\mathscr{Y} = \mathscr{X}\mathscr{B} + \mathscr{E},$$

where $\mathscr{X} = \mathrm{I}_q \otimes \boldsymbol{X}$ and $\mathscr{Y}$, $\mathscr{B}$ and $\mathscr{E}$ are vectors of size $nq$, $pq$ and $nq$, respectively.

## Appendix B

Let us apply the $vec$ operator to Model (5) where $\boldsymbol{\Sigma}_q^{-1/2}$ is replaced by $\hat{\boldsymbol{\Sigma}}_q^{-1/2}$, then

$$vec(\boldsymbol{Y}\hat{\boldsymbol{\Sigma}}_q^{-1/2}) = vec(\boldsymbol{XB}\hat{\boldsymbol{\Sigma}}_q^{-1/2}) + vec(\boldsymbol{E}\hat{\boldsymbol{\Sigma}}_q^{-1/2}) = ((\hat{\boldsymbol{\Sigma}}_q^{-1/2})' \otimes \boldsymbol{X})vec(\boldsymbol{B}) + vec(\boldsymbol{E}\hat{\boldsymbol{\Sigma}}_q^{-1/2}).$$

Hence,

$$\mathscr{Y} = \mathscr{X}\mathscr{B} + \mathscr{E},$$

where $\mathscr{Y} = vec(\boldsymbol{Y}\hat{\boldsymbol{\Sigma}}_q^{-1/2})$, $\mathscr{X} = (\hat{\boldsymbol{\Sigma}}_q^{-1/2})' \otimes \boldsymbol{X}$ and $\mathscr{E} = vec(\boldsymbol{E}\hat{\boldsymbol{\Sigma}}_q^{-1/2})$.

## References


Audoin, C., V. Cocandeau, O. Thomas, A. Bruschini, S. Holderith, and G. Genta-Jouve (2014): "Metabolome consistency: additional parazoanthines from the mediterranean zoanthid parazoanthus axinellae," Metabolites, 4, 421–432.

Bates, D. and M. Maechler (2017): Matrix: sparse and dense matrix classes and methods. R package version 1.2-8. https://CRAN.R-project.org/package=Matrix.

Boccard, J. and S. Rudaz (2016): "Exploring omics data from designed experiments using analysis of variance multiblock orthogonal partial least squares," Anal. Chim. Acta, 920, 18–28.

Brockwell, P. and R. Davis (1991): Time series: theory and methods, Springer Series in Statistics, Springer-Verlag, New York.

Dieterle, F., A. Ross, G. Schlotterbeck, and H. Senn (2006): "Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. application in 1h nmr metabonomics," Anal. Chem., 78, 4281–4290.

Faraway, J. J.(2004): Linear models with R, Chapman & Hall/CRC, New York.

Friedman, J., T. Hastie, and R. Tibshirani (2010): "Regularization paths for generalized linear models via coordinate descent," J. Stat. Softw., 33, 1–22.

Hrydziuszko, O. and M. R. Viant (2012): "Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline," Metabolomics, 8, 161–174.

Kirwan, J., D. Broadhurst, R. Davidson, and M. Viant (2013): "Characterising and correcting batch variation in an automated direct infusion mass spectrometry (dims) metabolomics workflow," Anal. Bioanal. Chem., 405, 5147–5157.

Kuhl, C., R. Tautenhahn, C. Boettcher, T. R. Larson, and S. Neumann (2012): "CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets," Anal. Chem., 84, 283–289.

Lê Cao, K.-A., S. Boitard, and P. Besse (2011): "Sparse pls discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems," BMC Bioinformatics, 12, 253.

Mardia, K., J. Kent, and J. Bibby (1979): Multivariate analysis, Probability and mathematical statistics, Academic Press, Londan.

Meinshausen, N. and P. Buhlmann (2010): "Stability selection," J. R. Stat. Soc., 72, 417–473.

Muller, K. E. and P. W. Stewart (2006): Linear model theory: univariate, multivariate, and mixed models, John Wiley & Sons.

Nicholson, J. K., J. C. Lindon, and E. Holmes (1999): "'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data," Xenobiotica, 29, 1181–1189.

Perrot-Dockès, M., C. Lévy-Leduc, L. Sansonnet, and J. Chiquet (2018): "Variable selection in multivariate linear models with high-dimensional covariance matrix estimation," J. Multivar. Anal., 166, 78–97.

R Core Team (2017): R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.

Ren, S., A. A. Hinzman, E. L. Kang, R. D. Szczesniak, and L. J. Lu (2015): "Computational and statistical analysis of metabolomics data," Metabolomics, 11, 1492–1513.

Rothman, A. J., E. Levina, and J. Zhu (2010): "Sparse multivariate regression with covariance estimation," J. Comput. Graph. Stat., 19, 947–962.

Saccenti, E., H. C. J. Hoefsloot, A. K. Smilde, J. A. Westerhuis, and M. M. W. B. Hendriks (2013): "Reflections on univariate and multivariate analysis of metabolomics data," Metabolomics, 10, 361–374.

Smith, C., E. Want, G. O'Maille, R. Abagyan, and G. Siuzdak, (2006): "XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification," Anal. Chem., 78, 779–787.

Smith, R., A. Mathis, and J. Prince (2014): "Proteomics, lipidomics, metabolomics: a mass spectrometry tutorial from a computer scientist's point of view," BMC Bioinformatics, 15, S9.

Tibshirani, R. (1996): "Regression shrinkage and selection via the Lasso," J. R. Stat. Soc. B, 58, 267–288.

Verdegem, D., D. Lambrechts, P. Carmeliet, and B. Ghesquière (2016): "Improved metabolite identification with midas and magma through ms/ms spectral dataset-driven parameter optimization," Metabolomics, 12, 1–16.

Zhang, A., H. Sun, P. Wang, Y. Han, and X. Wang (2012): "Modern analytical techniques in metabolomics analysis," Analyst, 137, 293–300.

Zhang, H., Y. Zheng, G. Yoon, Z. Zhang, T. Gao, B. Joyce, W. Zhang, J. Schwartz, P. Vokonas, E. Colicino, A. Baccarelli, L. Hou, and L. Liu (2017): "Regularized estimation in sparse high-dimensional multivariate regression, with application to a DNA methylation study," Stat. Appl. Genet. Mol. Biol. 16, 159–171.