


Combining strong sparsity and competitive predictive power with the L-sOPLS approach for biomarker discovery in metabolomics

Baptiste Féraud^{1,2}  · Carine Munaut³ · Manon Martin¹ · Michel Verleysen^{2,4} · Bernadette Govaerts¹

Received: 7 July 2017 / Accepted: 21 September 2017 / Published online: 27 September 2017
© Springer Science+Business Media, LLC 2017

Abstract

Introduction In the context of metabolomics analyses, partial least squares (PLS) represents the standard tool to perform regression and classification. OPLS, the Orthogonal extension of PLS which has proved to be very useful when interpretation is the main issue, is a more recent way to decompose the PLS solution into predictive components correlated to the target Y and components pertaining to the data X but uncorrelated to Y . This predominance of (O)PLS can raise the question of the awareness of alternative multivariate regression and/or classification tools able to find biomarkers. Actually, the search for biomarkers remains a key issue in metabolomics as it is crucial to very accurately target discriminating features.

Objective Most of the time, (O)PLS methods perform well but a drawback often occurs: too many variables can be selected as potential biomarkers even using adapted statistical significance tests. However, for final users (in medical studies for instance), it can be advantageous to deal with only a small number of easily interpretable biomarkers.

Methods This drawback is approached in this paper via the use of sparse methods. The sparse-PLS (sPLS), an extension

of PLS which promotes an inner variable/feature selection, is an interesting existing solution. But a new intuitive algorithm is proposed in this paper to combine sparsity and the advantages of an orthogonalization step: the “Light-sparse-OPLS” (L-sOPLS). L-sOPLS promotes sparsity on a previously optimized deflated matrix which implies the removal of the Y -orthogonal components.

Results A discussion around the compromise between sparsity and predictive modelling performances is provided and it is shown that L-sOPLS produces convincing results, illustrated principally on the basis of ¹H-NMR spectral data but also on genomic RT-qPCR data.

Conclusion The L-sOPLS algorithm allows to reach better predictive performances than (O)PLS and sPLS while taking into account only a very small number of relevant descriptors.

Keywords Biomarker discovery · (O)PLS models · Feature selection · Sparse models · L-sOPLS · ¹H-NMR data · RT-qPCR data

1 Introduction

In a large variety of current metabolomics studies, as for the whole family of -omics, the research of accurate biomarkers is a key issue whether it is to diagnose a disease or to measure the degree of progress of a disease, to estimate the effects of a pharmaceutical treatment, to control the quality of consumer goods, etc. Biomarkers are then a way to explain and to anticipate an event, event which can be of a critical importance for example in case of a medical decision to operate or the choice of a heavy-duty or long-term medical treatment.

✉ Baptiste Féraud
baptiste.feraud@uclouvain.be

¹ Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), Université Catholique de Louvain (UCL), Louvain-la-Neuve, Belgium

² Machine Learning Group, Université Catholique de Louvain (UCL), Voie du Roman Pays 20, bte L1.04.01, B-1348, Louvain-la-Neuve, Belgium

³ Laboratory of Tumor and Development Biology (GIGA-Cancer), Université de Liège (ULg), Liège, Belgique

⁴ SAMM, Université Paris I, Panthéon, Sorbonne, France

In practice, the statistical detection of these biomarkers is carried out by many researchers using partial least squares (PLS) analyses when the response matrix of interest Y is continuous; or PLS-DA (Discriminant Analysis) if Y is categorical or coded as a binary vector y when only two levels are of interest (for instance, $y = 1$ for patients with a disease and $y = 0$ for healthy people). The popularity of PLS regression methods in metabolomics dates from the early 2000s, mainly based on the works of Wold et al. (2001, 2002), and the parallel development of the SIMCA software (see for instance Bylesjo et al. 2006). Since then, this popularity has never stopped growing and the vast majority of the past and current biomarkers' researches have depended on the PLS(-DA) principles.

A bit more recently, the OPLS methodology was proposed (Gabrielsson et al. 2006; Stenlund et al. 2008) and also gained a huge popularity among the metabolomics community. OPLS(-DA) is a more recent way to decompose the PLS solution into components correlated (predictive) to the target Y to predict and components unique in the data table X and uncorrelated (orthogonal) to Y . The common OPLS(-DA) methodology very often leads to nearly the same results than PLS(-DA); the predictions are identical in both cases, but the indisputable advantage of OPLS comes from a better capacity of interpretation of the results, which may facilitate the work of final users.

PLS and OPLS are then massively used for classification, searching for biomarkers, and are obviously efficient to do that. But an issue can rapidly occur in most cases: too many variables (spectral zones in NMR) can be considered and proposed as candidate biomarkers. However, for final users (in medical studies for instance), it can be advantageous to deal with only a small number of -very- significant and ideally easily interpretable biomarkers. In these situations, a critical objective would then be to build the lightest and most effective possible model. In recent years, this challenge is approached and filled via the use of sparse methods, with the objective to reinforce the most significant biomarkers' coefficients and to force the less significant ones to be equal to zero (according to some LASSO-like penalties and to the well known LARS algorithm Efron et al. 2004).

In this paper, the notion of sparsity will be explored in the context of the biomarker discovery issue in metabolomics. An overview of the already known, but not yet enough used, sparse-PLS (sPLS) algorithm will be provided. sPLS can be viewed as an extension of the PLS regression which includes an additional and simultaneous variable/feature selection.

Then, a new methodology, called "Light-sparse-OPLS" (L-sOPLS), both innovative and quite intuitive, will be proposed and detailed. It can be viewed as an extension of OPLS adapted with sparse penalties. The idea is to take advantage of an orthogonalization step by applying sparse algorithms (such as the sPLS, Elastic Net,...) on a previously

optimized decorrelated matrix (called X_d : a deflated matrix where Y -orthogonal components have been optimally removed from the initial data matrix X of spectra).

The goal of this paper is then to demonstrate, on two real data sets, that the L-sOPLS alternative performs well as it can highlight the most relevant biomarkers (and furthermore a very small number of them). The sparse models' predictive performances are also carefully taken into account via the automatic use of cross-validated RMSEP criteria. Often, a trade-off between a strong level of sparsity and a competitive predictive power must be considered. It is shown in this paper that it is the case for sPLS but not for L-sOPLS, which allows to reach better predictive performances than (O)PLS and sPLS while taking into account only a very small number of final accurate descriptors.

The paper is organized as follows. Section 2 provides a detailed description of the two real data sets used to perform the algorithms: a $^1\text{H-NMR}$ one obtained from spiked urine of rats and a genomic RT-qPCR one linked with the endometriosis disease. All the above-mentioned regression models or algorithms (PLS, OPLS, sPLS and L-sOPLS) are detailed in the methodological Sect. 3. Results on both data sets are then shown and discussed in Sect. 4. Finally, a general conclusion and description of further works are given in Sect. 5.

2 Materials: data sets and experimental protocols

In this section, the two selected data sets used to train the classical and sparse algorithms are presented. For both of them, a description and motivational explanations are provided, as well as the main acquisition parameters.

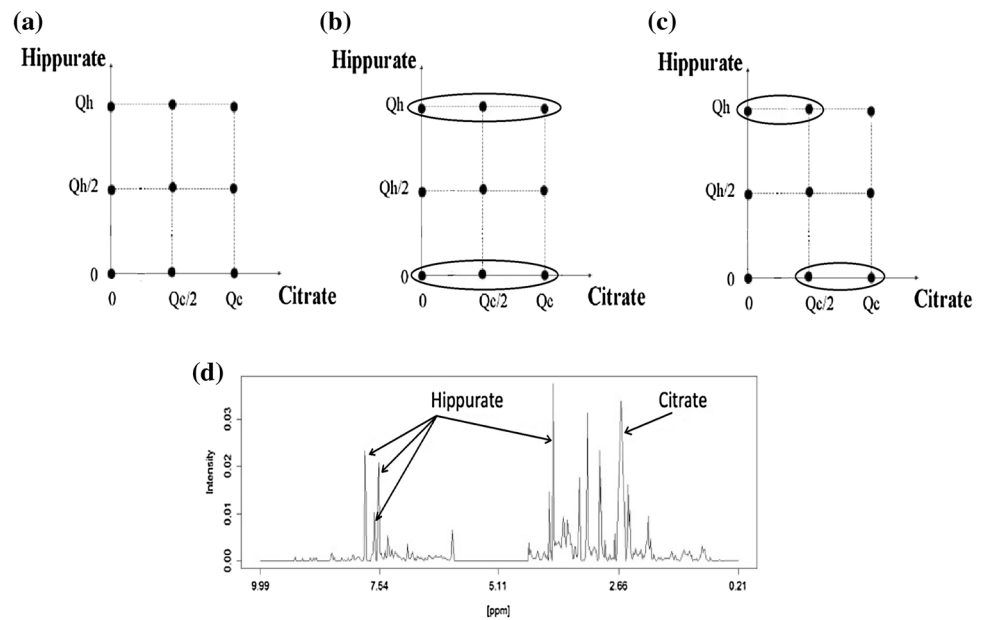
2.1 Spectral $^1\text{H-NMR}$ data set based on urine

The first experimental data set is a one dimensional Proton Nuclear Magnetic Resonance ($^1\text{H-NMR}$) spectral set based on spiked urine samples from rats.

2.1.1 Description and motivations

This database was experimentally created according to a design aimed at studying the ability of statistical models to find, as biomarkers, the descriptors of the spectra for which a variability was carefully controlled. This property allows to evaluate the performances of statistical analysis for potentially a large panel of methods. In this experiment, homogenized medium urine samples were spiked with two products (citrate and hippurate) at three levels of concentration and analyzed by spectroscopy. These concentrations of citrate and hippurate are aimed to mimic the variability focused in a biomarker discovery study. The basic design is presented in Fig. 1a and a typical urine spectrum with spiked citrate

Fig. 1 **a** The whole urine experimental design; **b** the hippurate balanced case; **c** the hippurate unbalanced case; **d** A typical urine spectrum with spiked citrate and hippurate



and hippurate is provided in Fig. 1d. For each point of the design, six samples were prepared and analyzed over three days (two replicates per day).

To challenge the biomarker discovery issue, several balanced and unbalanced sub-data sets have been extracted from the full database. In this paper, two of them are used to illustrate the ability of classification methods to recover hippurate peaks as group biomarkers. As shown in Fig. 1b, c, the combinations $[(0, 0), (Qc/2, 0), (Qc, 0)]$ and $[(0, Qh), (Qc/2, Qh), (Qc, Qh)]$ are conserved for the balanced study; and the combinations $[(Qc/2, 0), (Qc, 0)]$ and $[(0, Qh), (Qc/2, Qh)]$ are conserved for the unbalanced study. For each of these subcases, a target binary vector y was created accordingly to determine the groups to discriminate: for instance, for the hippurate balanced case, $y = 1$ if the observation is concerned by one of the $[(0, 0), (Qc/2, 0), (Qc, 0)]$ combinations and $y = 2$ if the observation is concerned by another $[(0, Qh), (Qc/2, Qh), (Qc, Qh)]$ combination.

The goal is to discriminate groups of spectra in these two subcases. Intuitively, in the balanced case, all other things being equal, the search for relevant discriminating biomarkers would primarily lead to the concerned product ppm spectral zone, i.e. hippurate only. In the unbalanced one, it would lead to a mixture of both products ppm spectral zones, i.e. citrate and hippurate, as the changes of these two factors are confounded. The further use of unsparsely and sparse models, in subsequent sections of this paper, has to confirm this intuition.

The whole input data sets are X numeric matrices of dimensions (36×600) for the balanced case, and (24×600) for the unbalanced one. The lines in X , the individual spectra, are identified by the way of the citrate and hippurate levels of concentration.

2.1.2 Acquisition

This database was designed with spectroscopists from Eli Lilly and from the University of Liège (ULg). All the samples preparation and acquisition parameters are already explained in specific details in Rousseau (2011) (in part 2.3).

Some pre-treatments have been incorporated before the statistical analyses as such: the part of the spectrum between 0.2 and 10 ppm has been reduced to 600 descriptors, the ppm values corresponding to the large non-informative urea and to the water zone (4.5–6.0 ppm) were set to zero and the data were normalized via a classical constant sum ($CS = 1$) normalization. It is also possible to integrate into one peak the spectral region around the citrate resonances (2.56–2.72 ppm) to suppress the high shifts of the citrate peaks, but this decision was not taken in the context of this work.

2.2 RT-qPCR genomic data set

In this subsection, a description of a second data set is provided. It consists in Reverse Transcription-quantitative Polymerase Chain Reaction (RT-qPCR) results from miRNAs expression analysis in groups of patients suffering or not from endometriosis (vector y). This leads to the search for discriminant differences of miR expression between the two groups and the potential discovery of endometriosis biomarkers.

2.2.1 Description and motivations

Endometriosis is characterized by the presence and growth of functional endometrial-like tissues outside the uterine cavity. It is a common and benign gynecological disorder

(Giudice and Kao 2004) but women with endometriosis commonly experience a diagnostic delay of 6–12 years, often associated with increased severity of the disease. The current standard to detect the pathology is laparoscopy, a very invasive, expensive and associated to high surgical risks method. Even though extensive studies have been performed there are to date no specific reliable biomarker (Nisenblat et al. 2016).

An emerging strategy to detect pathologies is miRNAs quantification by RT-qPCR in bodyfluids. MicroRNAs (miRNAs) are small non-coding RNAs (20–24 nucleotides) that regulate gene expression through post-transcriptional repression or degradation of messenger RNA (mRNA) (Bartel 2009; Lai 2002). In this study, 19 previously described miRNAs are evaluated to be associated with endometriosis (log_miR_x1 to log_miR_x19, which are log-transformed and for now anonymized). The data set contains information from 120 individuals, 63 suffering from endometriosis and 57 others for control.

2.2.2 Acquisition

Total RNA, including miRNA, was extracted from 200µl of serum using the miRNeasy Serum/Plasma Kit (Qiagen) according to the manufacturers recommendations, and it was then eluted in 30µl of nuclease-free water. A synthetic spike-in control miRNA (*C. elegans* miR-39 mimic, Qiagen) was added for subsequent normalization.

Total miRNA (2µl) from each sample was reverse-transcribed with the miScript II TR kit (Qiagen) according to the manufacturers instructions. The miRNAs were subsequently quantified using the miScript SYBR Green PCR Kit (Qiagen) and specific forward primers. The reaction mixture included 2.5µl of cDNA, 12.5µl of Quantitect SYBR Green PCR Master Mix, 2.5µl of forward primer, 2.5µl of miScript Universal Primer and 5µl of RNase-free water (for a final reaction volume of 25µl). The thermal cycling consisted of an initial denaturation at 95 °C for 15 min, followed by 45 cycles at 95 °C for 15 s, 55 °C for 30 s, and 70 °C for 30 s. All reactions were run in duplicate.

3 Methods: (O)PLS and corresponding sparse solutions: sPLS and L-sOPLS

In this methodologic section, all the tested algorithms are presented, from classical PLS to advanced sparse solutions. First, some reminders about (O)PLS are provided. Then, the sparsing issue is approached with an existing tool, the sparse-PLS (sPLS). Finally, an innovative alternative is presented in details: the “Light-sparse-OPLS” (L-sOPLS), aimed at combining the advantages of the orthogonality and of the use of sparse modelling penalties.

3.1 The PLS(-DA) regression model: some reminders

PLS is a broad spread method for modelling relations between dependant and independant variables. It is very popular and extensively used in chemometrics, and therefore in metabolomics where it has proved its usefulness and efficiency to underline metabolic changes in biofluids (due for example to toxicity or disease process). The PLS regression (Geladi and Kowalski 1986; Wold et al. 2001) is primarily constructed in order to optimize the quality of prediction but is also able to extract factors, called latent variables, to best resume the available information in both regressors and response(s). Thus, the popularity of PLS is principally based on this two levels capacity: to explain and represent the information in X and Y (Y can be univariate or multivariate), and to model the link between X and Y in order to allow further predictions. In practice with omics data, PLS is also often used as a variable selection tool for searching for biomarkers.

The fundamental objective of PLS implies a computation of X and Y scores, in order to capture a high amount of variance in X and Y matrices and also a high amount of “correlations” between X and Y . In most PLS algorithms, the uncorrelated components of the model are extracted sequentially and every iteration deflates from the data the variation that is associated with the last estimated component.

More formally, consider a data matrix X with n observations and m variables and, to simplify, a vector y for the dependent variable of size $(n \times 1)$. To specify the latent component matrix T such that $T = XW$ (linear combinations), PLS requires finding the columns of $W = (w_1, \dots, w_q)$ from successive optimization problems. For the iterative computation of each component $k = 1, \dots, q$, the PLS goal can be written as follow:

$$\max_{\omega} \{ \text{corr}^2(y, Xw_k) \text{var}(Xw_k) \} \equiv \max_{\omega} \{ w_k' X_k' y y' X_k w_k \} \quad (1)$$

subject to $w_k' w_k = 1$, where ω_k is the PLS X -weights vector for dimension k .

Alternatively, this criteria can be written:

$$\min_w \{ -w_k' M w_k \} \\ \text{subject to } w_k' w_k = 1, \text{ where } M = X' y y' X. \quad (2)$$

The outer relation for X is built as follow: the X -scores $T_{(n \times q)}$ are obtained from linear combinations of the original data X with the matrix of weights $W_{(m \times q)}$, such that $T_q = XW_q$. One multiplies T_q with the X -loadings matrix $P'_{(q \times m)}$, as $P_q = X' T_q$. The product is then a good “summary” of X if one obtains small residuals $E_{(n \times m)}$ in $X = T_q P_q' + E$.

Similarly, for the other part of the bilinear decomposition, the outer relation for y is defined as the addition of some summarized information and some error terms:

$y = U_q c'_q + g$, where U_q is the $(n \times q)$ matrix of y -scores, c'_q is the $(q \times 1)$ vector of y -weights and $g_{(n \times 1)}$ is the residual vector of y . Since T_q is a good predictor of y , the following inner relation occurs: $y = T_q c'_q + g^*$, where the y -residuals $g^*_{(n \times 1)}$ are equal to the difference between the modelled and observed responses. Consequently, the goal is to obtain $\|g^*\|$ as low as possible and also to obtain the best possible relation between X and y .

The first component can be retrieved from a singular value decomposition (SVD) of a matrix that combines both information from X and y : $R = X'y$ and gives the singular vector w_1^* known as the first weights column (Hoskuldsson 1988). For each subsequent iteration, the X matrix is deflated by the variation associated with the estimated component. Eventually, the process is maintained until one decides to stop after a certain number of latent variables or until all the components have been calculated. The relevant choice of q is therefore essential to avoid overfitting of the model, giving a poor prediction power especially when the regressors are numerous and/or correlated (Abdi 2010). Usually, a cross-validation criterion is added to decide on the optimal number of latent variables to keep.

Computationally, several PLS algorithms are available according to some specificities: classical PLS, [canonical powered partial least squares (CPPLS) (Indahl et al. 2009)], [straightforward implementation of a modification of PLS (SIMPLS) (De Jong 1993)], [nonlinear iterative partial least squares (NIPALS) (Wold 1975)], etc. In this paper, the SIMPLS algorithm will be used for each PLS routine because of its higher computational speed. With the descriptor matrix X and a the target vector y , this algorithm consists in:

- Centering of X by columns
- Initialization: $r_1 = X'y$
- Iterations from $k = 1$ to q :
 1. Weights normalization: $w_k = \frac{r_k}{\sqrt{r'_k r_k}}$
 2. Scores calculation: $t_k = X w_k$
 3. Scores normalization: $t_k = \frac{t_k}{\sqrt{t'_k t_k}}$
 4. Loadings calculation: $p_k = X' t_k$
 5. Deflation step: $r_{k+1} = r_k - p_k (p'_k p_k)^{-1} p'_k r_k$.
- Choice of q , the optimal number of components of the PLS model using an adequate validation criterion. The [root mean square error of prediction (RMSEP), Mevik and Cederkvist (2004)] is a traditional criterion. It must be minimized and it can be calculated for each size of model by k -fold or Leave-One-Out (LOO) cross-validation. Note that working with an external and independ-

ent validation/test set is very often too expensive and not possible in most metabolomics studies.

- Matrices $T = XW$ and $P = X'T$ are obtained, based on the optimal q .
- Final coefficients of the PLS model are calculated as $b_{PLS} = W(T'T)^{-1}T'y$ and allow to make subsequent predictions on new observations.

From these results, the descriptors with highest (absolute) coefficients can be considered as primary candidate biomarkers.

PLS discriminant analysis (PLS-DA) (Barker and Rayens 2003) is a particular case of PLS regression aiming at predicting one (or several) binary responses y still from a matrix X of descriptors. PLS-DA is specifically suited to deal with problems where the number of predictors is large (compared to the number of observations) and collinear, two major challenges frequently encountered with, for instance, ¹H-NMR data. For spectral biomarker discovery, PLS-DA provides regression parameters that can be used as biomarker scores and the descriptors with the highest coefficients are naturally linked with discriminating zones.

3.2 The OPLS(-DA) algorithm

Projection methods such as PLS remain strongly affected by the potential occurrence of systematic variation in the data source that is not relevant for response prediction. That's why Orthogonal Projection to Latent Structures (OPLS) has emerged in chemometrics as a filtering method enabling the removal of variation from X that is not correlated to the dependant variable Y (or y). It leads to a different model since some of the extracted components are identified as corresponding to systematic or specific orthogonal causes of variation. A generalisation of its use in metabolomics studies [for example, among many others, in Wiklund et al. (2008), Jung et al. (2010) and Weljie et al. (2011)] is explained by the relative simplicity of the algorithm, derived from the NIPALS-PLS one, and the ability to manipulate and interpret the resulting predictive and orthogonal matrices. Note that NIPALS-PLS (Wold 1975) predictions are equal to SIMPLS predictions, described in Sect. 3.1, when considering an univariate target variable y (Wehrens 2011).

As for PLS regressions, the OPLS extracts sequentially each component. Several orthogonal components can be estimated and the X matrix is, at each step, deflated with respect to orthogonal variations. Considering a two-class classification problem, the OPLS(-DA) approach only considers one predictive component along with potentially several orthogonal components. Finally, the goal is to find the projection that classify and discriminate in the best way the y levels, or groups.

The whole statistical methodology behind the OPLS(-DA) approach is available in details in Trygg and Wold (2002). The algorithm differs slightly from the classical NIPALS-PLS one and works as follow for an univariate y :

- Centering of X by columns, as for PLS
- The NIPALS algorithm is sequentially applied to find the predictive component by removing a bilinear structure from the centered X that is t_{ortho}, p'_{ortho} (Wold et al. 2002). Remember that, generally, T_{ortho} is the orthogonal X -scores ($n \times (q - 1)$) matrix for the OPLS(-DA) model with $(q - 1)$ orthogonal components, and P'_{ortho} is the orthogonal X -loadings matrix for the same model.

1. Initialization: $X_1 = X$, starting from the initial mean-centered data matrix X to be deflated.

Iterative orthogonalization from $k = 1$ to $q - 1$:

2. $\tilde{w}_k = \frac{X'_k y}{y' y}$
3. X -weights normalization to $\|\tilde{w}_k\| = 1$: $\tilde{w}_k = \frac{\tilde{w}_k}{\sqrt{\tilde{w}'_k \tilde{w}_k}}$
4. X -scores computation: $t_k = X_k \tilde{w}_k$
5. X -loadings computation: $p_k = \frac{X'_k t_k}{t'_k t_k}$
6. Calculations of the orthogonal components' weights: $\tilde{w}_{ok} = p_k - \frac{\tilde{w}'_k p_k}{\tilde{w}'_k \tilde{w}_k} \tilde{w}_k$, then normalized to $\tilde{w}_{ok} = \frac{\tilde{w}_{ok}}{\sqrt{\tilde{w}'_{ok} \tilde{w}_{ok}}}$
7. Computation of orthogonal scores: $t_{ok} = X_k \tilde{w}_{ok}$
8. And computation of orthogonal loadings: $p_{ok} = \frac{X'_k t_{ok}}{t'_{ok} t_{ok}}$
9. Deflation step on X : $X_{k+1} = X_k - t_{ok} p'_{ok}$
10. A whole deflated matrix can be extracted at the end of the iterations such that:

$$X_d = X - T_o P'_o = X - X_o \tag{3}$$

where $T_o = (t_{o1}, \dots, t_{oq-1})$ and $P_o = (p_{o1}, \dots, p_{oq-1})$.

- The PLS algorithm, using one predictive component, can now be performed on X_d according to the same pattern:

1. $\tilde{w}_d = \frac{X'_d y}{y' y}$, then $\tilde{w}_d = \frac{\tilde{w}_d}{\sqrt{\tilde{w}'_d \tilde{w}_d}}$
2. $t_d = X_d \tilde{w}_d$
3. $p_d = \frac{X'_d t_d}{t'_d t_d}$

- OPLS coefficients are obtained by:

$$b_{OPLS} = \tilde{w}_d (p'_d \tilde{w}_d)^{-1} \frac{y' t_d}{t'_d t_d} \tag{4}$$

- Since the b_{OPLS} are linked with X_d and not directly with the initial data matrix X , note that these coefficients, although very informative, are not strictly interpretable in the same way as b_{PLS} . So, b_{OPLS} can be transformed to match with X as follow:

$$b_{OPLScorr} = (I_m - (\tilde{W}_o (P'_o \tilde{W}_o)^{-1} P'_o)) b_{OPLS} \tag{5}$$

where I_m is the identity matrix involving m columns and $\tilde{W}_o = (\tilde{w}_{o1}, \dots, \tilde{w}_{oq-1})$. The OPLS predictions are then:

$$\hat{y}_{OPLS} = X_d . b_{OPLS} = X . b_{OPLScorr} \tag{6}$$

It is important here to note that the corrected OPLS coefficients ($b_{OPLScorr}$) obtained with $(q - 1)$ orthogonal components and one predictive component will be equal to the corresponding PLS coefficients obtained with q components (Tapp and Kemsley 2009). Therefore, the subsequent predictions will be the same for both models:

$$\hat{y}_{OPLS(q-1)} = \hat{y}_{PLS(q)} \tag{7}$$

As for PLS models, to avoid overfitting, a cross-validation step can be added to optimize a criterion (RMSEP for instance) in order to select the best number of orthogonal components ($q - 1$). From (Eq. 7), it follows that if a solution implying q components was optimally selected during a previous PLS model for a specific study, the solution with $(q - 1)$ orthogonal dimensions is also optimal when using OPLS for this same study.

3.3 The sparsity issue and the sparse-PLS (sPLS) solution

Both PLS(-DA) and OPLS(-DA) are efficient and massively used in -omics studies for biomarker discovery because of the $m \gg n$ characteristic of the X matrix. In these models, descriptors with highest absolute coefficients can be selected as candidate biomarkers. Other tools can also be used as, for instance, the [variable importance in the projection criterion (VIP) (Afanador et al. 2013; Lu et al. 2014)] or the popular S-plots for OPLS. Most of the time, these methods lead to a high number of biomarkers considered as of interest for further investigations in metabolomics experiments, which quickly induces difficulties of interpretation.

If decision rule is based on the highest absolute coefficients, the choice of a threshold to determine what is of interest and what is not is inevitably needed. But what threshold to use? This question is obviously subjective. To objectify this decision, a solution is to reinforce the more significant biomarkers' coefficients and to force the other ones to be equal to zero. This represents the key basis of the sparsity (Hastie et al. 2015) and leads to the additional application of penalties (LASSO-like, Ridge, Elastic Net (Zou and Hastie 2005,...)). Moreover, a lighter and more interpretable model, not less efficient or relevant, should be ideally provided to

final users for medical decisions, treatment adjustments, etc. Of course, the quality of prediction of the sparse alternatives must be assessed and, sometimes, a compromise between predictive performances and sparsity needs to be found.

In this context, the use of the Sparse-PLS (sPLS) method is increasing in metabolomics studies. sPLS keeps the spirit of PLS models but provides a restricted number of final selected biomarkers. Different algorithms exist in the literature, the two main of them being respectively implemented in the *mixOmics* R package [detailed in Lê Cao et al. (2008)] and in the *spls* R package [detailed in Chun and Keles (2007) and Chung et al. (2012)].

On the basis of the PLS objective [see Eq. (2)], the sPLS algorithm of Chun and Keles, used in this paper, provides an efficient implementation based on the LARS objective function:

$$\min_{w,c} \{(-\beta w'_k M w_k) + (1 - \beta)(c - w_k)'M(c - w_k) + \lambda_1 c_1 + \lambda_2 c_2\}$$

subject to $w'_k w_k = 1$ for $k = 1, \dots, q$, where $M = X'yy'X$.

(8)

This formulation promotes exact zero property onto the weights by imposing L_1 penalty (λ_1) onto a surrogate direction vector c (linked with the Y -weights) instead of the original direction w (linked with the X -weights), while keeping w and c close to each other. In other words, this L_1 penalty encourages sparsity on c . The L_2 penalty (λ_2) takes care of the potential singularity of matrix M when solving for c . Finally, the effect of the concave part, and the local solution issue, is reduced by using a small additional parameter β (Chun and Keles 2007).

The generalized regression formulation (Eq. 8) is then solved by alternatively iterating between solving for w for fixed c , and solving for c after fixing w . For the problem of solving w for fixed c , the objective function becomes:

$$\min_w \{(-\beta w'_k M w_k) + (1 - \beta)(c - w_k)'M(c - w_k)\}$$

subject to $w'_k w_k = 1$ for $k = 1, \dots, q$.

(9)

This constrained least squares problem can be solved via the method of Lagrange multipliers (Chun and Keles 2007). When solving for c for fixed w , it becomes:

$$\min_c \{(R'c - R'w_k)'(R'c - R'w_k) + \lambda_1 \|c\|_1 + \lambda_2 \|c\|_2\}$$

(10)

with $R = X'y$. This second problem is equivalent to the naive elastic net (EN) problem of Zou and Hastie (Zou and Hastie 2005) when Y in the naive EN is replaced with $R'w$ and can be solved efficiently via the least angle regression spline algorithm (LARS) (Efron et al. 2004). sPLS often requires a very large λ_2 -value to solve Eq. (10) because R' is a matrix with usually a small number of lines, i.e. one line when $Y = y$ is univariate. As a remedy, an EN formulation with $\lambda_2 = \infty$ is used.

In order to promote sparse solutions in a restricted X -space, sPLS searches for relevant variables, the so-called active variables. Specifically, at each step of either the NIPALS or the SIMPLS algorithm, Eq. (8) is optimized and all direction vectors are updated to form a Krylov subsequence (Chapman and Saad 1997) on the subspace of the active variables. This is achieved by conducting PLS regression by using the selected features. Let Π be an index set for active variables, q the number of components and X_Π the submatrix of X whose column indices are contained in Π . The sPLS-SIMPLS algorithm works as follow (Chun and Keles 2007):

- Initialization step: $b_{PLS_1} = \{.\}$; $\Pi_1 = \{.\}$; and $X_1 = X$
 - While $1 \leq k \leq q, 1$. Find \hat{w}_k by solving the objective (8) with $M = X'_{\Pi_k} y y' X_{\Pi_k}$,
 - 2. Update $\Pi = \Pi_k \cup \{i : \hat{w}_{ki} \neq 0\} \cup \{i : b_{PLS_{ki}} \neq 0\}$,
 - 3. Fit then a PLS (SIMPLS) model with X_{Π_k} as explanatory matrix, using k number of latent components,
 - 4. Compute active weights W_{Π_k} and the loading matrix P_{Π_k} such that $P_{\Pi_k} = X'_{\Pi_k} X_{\Pi_k} W_{\Pi_k} (W'_{\Pi_k} X'_{\Pi_k} X_{\Pi_k} W_{\Pi_k})^{-1}$,
 - 5. Define and update the SIMPLS regression parameters b_{PLS_k} by the new estimates of the direction vectors,
 - 6. Update X_k by deflation of the subset Π_k of active columns through $X_{\Pi_{(k+1)}} \leftarrow X_{\Pi_k} (I - P_{\Pi_k} (P'_{\Pi_k} P_{\Pi_k})^{-1} P'_{\Pi_k})$.

Note that the sPLS-SIMPLS algorithm, used in this paper, has similar attributes to the sPLS-NIPALS one. It selects more than one variable at each step and handles multivariate responses as well.

An additional cross-validation step (for minimizing RMSEP for instance) is greatly useful to determine the optimal couple formed by the number of predictive components q to keep, as for PLS, and the λ_1 penalty term. The final descriptors with non-zero coefficients (i.e. in X_{Π_q}) can finally be considered as primary candidate biomarkers. In terms of interpretation, final biomarkers are directly available through their non-zero coefficients and the number of biomarkers is obtained as a result of the objective optimization of (q, λ_1) . Using the VIP or any variable ranking criterion at the end of a (O)PLS regression, this number would be inevitably subjective.

3.4 Sparsity with OPLS: the new “light-sparse-OPLS” (L-sOPLS) solution

If one want to keep the OPLS interpretability advantages which consist of a withdrawal of the Y (or y)-orthogonal effects relative to X , and if one also want to propose sparse

results, a natural way is to explore the possible combinations of OPLS and sPLS. Although some interesting work already exist about this topic (van Gerwen and Heskes 2010; Munoz-Romero et al. 2015), the theory is not always very clear and no intuitive and user-friendly solution is proposed yet to the -omics community.

The innovative proposal of this paper consists in the development of the “Light-Sparse-OPLS” (L-sOPLS) model. The idea is quite intuitive and requires to start from the OPLS algorithm described in details in Sect. 3.2, to follow the process until the construction and the recovery of the data matrix X_d [deflated by the y -orthogonal components, see Eq. (3)] and to apply a sparsing model on this filtered matrix specifically, instead of on the initial data matrix X . The sparsing technique may be freely chosen between sPLS, Lasso logistic regression, Elastic Net, etc. In this paper, L-sOPLS combines the OPLS orthogonalization step with sPLS.

For interpretability and intuitiveness concerns, the L-sOPLS algorithm implies two optimization steps in order to maintain the best possible predictive ability (i.e. for each step, minimization of the RMSEP via an adapted cross-validation technique). The first step is aimed at optimally selecting the number of orthogonal components, as in a classic OPLS process. The deflated resulting matrix X_d is then built according to this number (q_{ortho}). The second step builds the sPLS sparse model, using X_d as input, whose number of predictive components (q_{pred}) and penalty term λ_1 are chosen optimally.

The L-sOPLS algorithm can be summarized as follow:

1. Application of an OPLS model (see Sect. 3.2) on the initial centered spectral matrix X .
Minimization, via k -fold or LOO cross-validation, of the predictive quality error criterion (i.e. RMSEP, MSPE, ...) in order to select the optimal number of orthogonal components q_{ortho} .
2. Computation of the q_{ortho} -deflated matrix $X_d = X - X_o$.
3. Application of a SIMPLS-sPLS model (see Sect. 3.3) using X_d as input. This implies the inner identifications of $M_d = X_d' y y' X_d$ and $R_d = X_d' y$ instead of M and R in Eqs. (9) and (10). Note again that the process is similar for a multivariate Y .
Minimization, via cross-validation, of the predictive quality error criterion (i.e. RMSEP, MSPE,...) in order to select the optimal number of predictive components q_{pred} and the optimal penalty criterion λ_1 .
4. The final non-zero coefficient vector b^* is obtained by conducting PLS regression on the selected variables only. The length of b^* is obviously smaller or equal than the length of b_{PLS} or b_{OPLS} .
5. The descriptors (features) associated with b^* , which strictly contains non-zero coefficients, at the end of the

sPLS process are finally considered as primary candidate biomarkers (see Sect. 3.3).

Note that, as opposed to OPLS with a binary y , L-sOPLS allows the potential use of several predictive dimensions q_{pred} , combined with potentially several orthogonal dimensions q_{ortho} . The optimal number of orthogonal dimension(s) q_{ortho} is obtained after a first cross-validation step within an OPLS framework, and the optimal number of predictive dimension(s) q_{pred} is independently obtained after a second cross-validation step this time within a sPLS framework. This makes the modelling even more flexible and focused on the prediction of the target Y (or y). Indeed, since the sPLS step is performed on a deflated object, which is supposed to be very strongly linked with Y by construction, the predictive goal is probably strengthened and prioritized at the expense of the descriptive goal when using L-sOPLS. This intuition is confirmed in the next results section.

It can't seem so obvious that the potential use of more than one predictive dimension in L-sOPLS can really lead to parsimonious models (compared to OPLS). But the fact that q_{pred} is chosen by cross-validation simultaneously along with the λ_1 term avoids high risks of final non-parsimonious or very huge models (see further results in Table 1).

Note also that, according to some subsequent trials, L-sOPLS seems to be robust to overfitting toward the major class (i.e. if $P(y = 0) \neq P(y = 1)$).

This new method was implemented in R version 3.3.2 and a function is available here: <https://github.com/ManonMartin/MBXUCL>. A subsequent “LSOPLS” R package is under construction.

4 Results and discussion

In this section, all the obtained results are discussed, only some of them are shown for convenience and readability. Parameters will be provided for each PLS, OPLS, sPLS and L-sOPLS optimal model and for the two data cases.

4.1 Results for the ¹H-NMR urine spectral data set

Starting from the ¹H-NMR urine of rat design, two main sub-data sets are considered, with balanced and unbalanced hippurate doses (see Sect. 2.1.1). 600 descriptors, all potential spectral biomarkers in X (without metadata or medical extra information), are available to explain the membership in a group, i.e. the corresponding binary target vector y . For the balanced case, the discovery of only one product spectral zone(s) is principally expected to classify the observations into the two groups. For the unbalanced case, the spectral zones of both hippurate and citrate would be of interest to

Table 1 (O)PLS and sparse results for the ¹H-NMR citrate and hippurate data: the balanced and unbalanced hippurate cases (a “-” means that the peak is not among the first fifteen coefficients in absolute value for PLS or OPLS)

The hippurate balanced case (see Fig. 1d)					
Biomarkers (in ppm)	Zone	PLS coefficients $q = 5$	OPLS b_{OPLS} $q = 4 + 1$	sPLS selection $q_{pred} = 5, \lambda_1 = 0.6$	L-sOPLS selection $q_{ortho} = 4, q_{pred} = 3, \lambda_1 = 0.72$
		LOO-RMSEP = 0.0197	LOO-RMSEP = 0.0197	LOO-RMSEP = 0.0206	LOO-RMSEP = 0.0162
2.478		- 7.636	-	No	No
2.593	Citrate	-	2.906	No	No
2.609	Citrate	-	3.331	No	No
2.626	Citrate	-	2.906	No	No
2.642	Citrate	-	2.505	No	No
3.017		- 10.677	4.728	Yes (- 9.329)	No
3.050		- 5.730	3.331	Yes (- 5.381)	No
3.279		- 13.839	5.646	Yes (- 20.556)	No
3.295		- 7.153	-	Yes (- 13.606)	No
3.442		7.573	-	Yes (- 13.651)	No
3.801		- 7.260	-	No	No
3.981	Hippurate	17.115	28.699	Yes (16.829)	Yes (22.737)
3.997	Hippurate	19.408	6.744	Yes (19.907)	Yes (27.409)
6.055		- 9.734	-	No	No
7.558	Hippurate	11.335	16.204	Yes (11.777)	Yes (5.543)
7.574	Hippurate	23.517	13.906	Yes (24.738)	Yes (16.282)
7.639	Hippurate	-	5.368	Yes (6.128)	No
7.656	Hippurate	7.135	7.889	Yes (6.113)	Yes (- 7.126)
7.836	Hippurate	14.159	14.768	Yes (16.952)	Yes (39.541)
7.852	Hippurate	24.122	18.855	Yes (23.682)	Yes (26.641)
The hippurate balanced case (see Fig. 1e)					
Biomarkers (in ppm)	Zone	PLS coefficients $q = 5$	OPLS b_{OPLS} $q = 4 + 1$	sPLS selection $q_{pred} = 5, \lambda_1 = 0.6$	L-sOPLS selection $q_{ortho} = 4, q_{pred} = 3, \lambda_1 = 0.73$
		LOO-RMSEP = 0.0187	LOO-RMSEP = 0.0187	LOO-RMSEP = 0.0184	LOO-RMSEP = 0.0114
2.478		- 7.822	-	No	No
2.560	Citrate	-	- 4.385	Yes (0.370)	No
2.576	Citrate	-	- 5.847	Yes (0.492)	No
2.593	Citrate	-	- 7.309	Yes (0.615)	Yes (- 17.008)
2.609	Citrate	-	- 8.771	Yes (0.738)	Yes (- 20.409)
2.625	Citrate	-	- 7.309	Yes (0.615)	Yes (- 17.008)
2.642	Citrate	-	- 5.847	Yes (0.492)	No
2.658	Citrate	-	- 4.385	Yes (0.370)	No
3.017		- 9.581	-	Yes (13.995)	No
3.279		- 13.519	-	Yes (- 44.365)	No
3.295		- 6.299	-	No	No
3.442		7.802	4.953	Yes (- 44.365)	No
3.801		- 6.120	-	No	No
3.981	Hippurate	17.592	22.496	Yes (11.672)	Yes (- 5.564)

Table 1 (continued)

The hippurate balanced case (see Fig. 1e)					
Biomarkers (in ppm)	Zone	PLS coefficients $q = 5$	OPLS b_{OPLS} $q = 4 + 1$	sPLS selection $q_{pred} = 5, \lambda_1 = 0.6$	L-sOPLS selection $q_{ortho} = 4, q_{pred} = 3, \lambda_1 = 0.73$
		LOO-RMSEP = 0.0187	LOO-RMSEP = 0.0187	LOO-RMSEP = 0.0184	LOO-RMSEP = 0.0114
3.997	Hippurate	18.548	–	Yes (18.164)	No
6.055		– 7.750	–	No	No
7.558	Hippurate	11.061	12.686	Yes (9.313)	Yes (15.340)
7.574	Hippurate	23.622	10.917	Yes (29.902)	Yes (14.196)
7.639	Hippurate	5.664	4.251	Yes (10.180)	No
7.656	Hippurate	7.294	6.172	Yes (1.588)	No
7.836	Hippurate	12.166	11.554	Yes (12.701)	Yes (27.561)
7.852	Hippurate	25.717	14.693	Yes (30.735)	Yes (40.921)

The selected sparse biomarkers are highlighted in bold

explain this discrimination. The design also involves three different days of measurements.

On each set, PLS (using the SIMPLS algorithm), OPLS, sPLS (using the *spls* R package) and L-sOPLS algorithms were applied, along with LOO cross-validation steps in order to choose optimal parameters for prediction purpose. For each optimal model, the objectives are focused on the relevance and the final number of the selected biomarkers on one hand (and their graphical interpretability), and on the analysis of the predictive abilities on the other hand. For the sparse models, the penalty parameter λ_1 was in all cases considered between 0.6 and 0.99 in order to provide sparse enough results.

The choices of the optimal numbers of components for PLS and OPLS were made by minimizing the RMSEP criterion via LOO cross-validation. Finally, five components are used for PLS and OPLS (four orthogonal dimensions and one predictive dimension for OPLS), providing a minimal LOO-RMSEP equal to 0.0197 for the balanced case. For the unbalanced one, the same solution is found for a minimal LOO-RMSEP equal to 0.0187.

For the sparse methods, the optimal parameters (the number of predictive components q_{pred} and the penalty term λ_1) were also chosen to minimize the LOO cross-validated RMSEP.

For sPLS and for the balanced hippurate case, the optimal parameter combination is $q_{pred} = 5$ and $\lambda_1 = 0.6$ (not very severe) and leads to a minimal LOO-RMSEP equal to 0.0206. For L-sOPLS: $q_{pred} = 3$ and $\lambda_1 = 0.72$ after a deflation of matrix X through four orthogonal components (i.e. $q_{ortho} = 4$). For this last model, LOO-RMSEP is equal to 0.0162, which is lower than the minimal value obtained with the initial (O)PLS.

For the unbalanced case, the optimal parameter combination for sPLS is $q_{pred} = 5$ and $\lambda_1 = 0.6$ and leads to a minimal LOO-RMSEP equal to 0.0184. For L-sOPLS: $q_{pred} = 3$ and $\lambda_1 = 0.73$ after a deflation of matrix X through four orthogonal components. This last model has a LOO-RMSEP equal to 0.0114 and contains only eight final descriptors with non-zero coefficients (see the last column of the second part of Table 1). Thus, L-sOPLS provides (very) sparse models with a better predictive power than (O)PLS.

In Table 1, the first two columns lead to the main biomarkers selected by PLS and OPLS (fifteen for PLS and fifteen for OPLS, most of them being concerned by both) on the basis of their higher coefficients in absolute values. The number fifteen was arbitrarily chosen. The hippurate spectral zones are mainly represented in these highest coefficients for both PLS and OPLS; the citrate zone appears with lower coefficients in the OPLS results only, with positive low values in the balanced case and negative values in the unbalanced one (as expected according to the design). The sparse decisions concerning these biomarkers are also shown in Table 1 for optimized sPLS and L-sOPLS methods (Yes/No to indicate if the biomarker is selected, and the corresponding final sparse coefficient if yes).

For the balanced case, the optimal sPLS model finally selects more than twenty biomarkers from the 600 initial input variables. Among them, all the hippurate peaks are selected but not those of the citrate spectral zone. The optimal L-sOPLS model leads to the selection of only seven final biomarkers. It is very important to note that all of them are connected to the hippurate spectral peaks only.

For the unbalanced case, the optimal sPLS model selects twenty biomarkers (see also Table 2) from the 600 initial input variables. All the hippurate and citrate peaks are

Table 2 Evolution of the LOO-RMSEP criterion for both sPLS and L-sOPLS models according to different values for the q_{pred} and λ_1 parameters

The hippurate unbalanced case										
λ_1	q_{pred} for sPLS					q_{pred} for L-sOPLS (with $q_{ortho} = 4$)				
	1	2	3	4	5	1	2	3	4	5
0.6	0.1164 (2)	0.0317 (6)	0.0284 (12)	0.0235 (17)	0.0184 (20)	0.0171 (2)	0.0121 (6)	0.0123 (12)	0.0139 (17)	0.0131 (20)
0.65	0.1706 (2)	0.0409 (6)	0.0286 (11)	0.0263 (17)	0.0205 (19)	0.0171 (2)	0.0121 (6)	0.0123 (11)	0.0141 (17)	0.0141 (19)
0.7	0.1759 (1)	0.0557 (5)	0.0284 (9)	0.0266 (15)	0.0269 (19)	0.0222 (1)	0.0123 (5)	0.0121 (9)	0.0141 (15)	0.0140 (19)
0.75	0.1759 (1)	0.0559 (4)	0.0279 (7)	0.0263 (12)	0.0279 (16)	0.0222 (1)	0.0115 (4)	0.0121 (7)	0.0133 (12)	0.0135 (16)
0.8	0.1759 (1)	0.0549 (3)	0.0587 (5)	0.0257 (8)	0.0292 (12)	0.0222 (1)	0.0127 (3)	0.0122 (5)	0.0133 (8)	0.0135 (12)
0.85	0.1759 (1)	0.0552 (3)	0.0681 (5)	0.0258 (6)	0.0326 (9)	0.0222 (1)	0.0127 (3)	0.0122 (5)	0.0136 (6)	0.0135 (9)
0.9	0.1759 (1)	0.0548 (2)	0.0611 (4)	0.0609 (5)	0.0289 (6)	0.0222 (1)	0.0126 (2)	0.0120 (4)	0.0131 (5)	0.0131 (6)
0.95	0.1759 (1)	0.0548 (2)	0.0566 (3)	0.0593 (4)	0.0462 (5)	0.0222 (1)	0.0126 (2)	0.0134 (3)	0.0119 (4)	0.0124 (5)

The corresponding number of final selected biomarkers is in brackets. The combinations leading to better LOO-RMSEP performances than initial (O)PLS models (RMSEP = 0.0187) are highlighted in bold

selected among them. The optimal L-sOPLS model only selects eight final biomarkers. It is very important to note that all of these later biomarkers are either hippurate or citrate descriptors only, which coincides with the corresponding design (Fig. 1c). Furthermore, the selected citrate descriptors are well associated with negative sparse coefficients and localized at the center of the citrate peak.

Figure 2 shows the final coefficients and the first two loadings of each model for the unbalanced hippurate case. The main hippurate and citrate peaks are highlighted by vertical lines. One can see that the orthogonal models offer better performances to find the negative effect of the citrate dose. The optimal L-sOPLS model only involves eight non-zero final coefficients strictly linked with the hippurate or citrate spectral peaks only (positive coefficients for the majority of hippurate peaks and negative ones for citrate, as expected). The first L-sOPLS loadings reveal these effects very quickly in a quite intuitive and iterative way.

Figure 3 shows the obtained scores for PLS, OPLS, sPLS and L-sOPLS. For all the models, the discrimination between the two groups of interest is very clear, as already seen via very low LOO-RMSEP values, and one can see the interest of the orthogonalization which perfectly aligns the groups. The four subgroups linked with the different concentrations of hippurate and citrate may be also well separated (see symbols in Fig. 3). Finally, the day effect does not seem to be very relevant for the group classification (the numbers are not always ordered in Fig. 3).

For the unbalanced hippurate case’s sPLS and L-sOPLS, Table 2 illustrates the LOO-RMSEP variations over a grid of (q_{pred}, λ_1) different values. Remember that the (O)PLS reference value of LOO-RMSEP is 0.0187 in that case. One can see that only one (q_{pred}, λ_1) combination leads to a better, i.e. lower, LOO-RMSEP value when using sPLS. What is already very positive. But, in other words, for all other

combinations, a compromise has to be made between sparsity (degree or severity of feature selection) and predictive power.

For the L-sOPLS solution, almost all the (q_{pred}, λ_1) combinations lead to lower LOO-RMSEP values. Sparsity, even very severe, goes hand in hand with better predictive ability when using L-sOPLS instead of non-sparse PLS or OPLS. Furthermore, since the vast majority of possible models have a similar better predictive power, one can imagine that final users can be free to “manually” select some parameters and, consequently, to choose their L-sOPLS model according to the final number of biomarkers which they want to explore (beyond the automatic approach of the algorithm).

The evolution of the final number of selected biomarkers (displayed in brackets in Table 2) shows, as expected, that high values of λ_1 lead to a restricted number of selected descriptors because of a greater degree of severity. Moreover, when q_{pred} increases into the model, the number of selected biomarkers also increases.

Note that all these results and conclusions are very similar when considering the citrate case sub-data sets: the citrate peak regions are primarily selected as final biomarkers for the balanced case, both hippurate and citrate ones for the unbalanced case. And similar conclusions can be highlighted about the models’ predictive power.

4.2 Results for genomic RT-qPCR data

In the RT-qPCR data set described in Sect. 2.2, $m = 19$ anonymized, log-transformed and standardized miRNAs are considered as explanatory variables of the model, and $n = 120$ observations are available. These data aim at showing how the algorithms behave when the number of variable is drastically limited, just like the level of information. The y

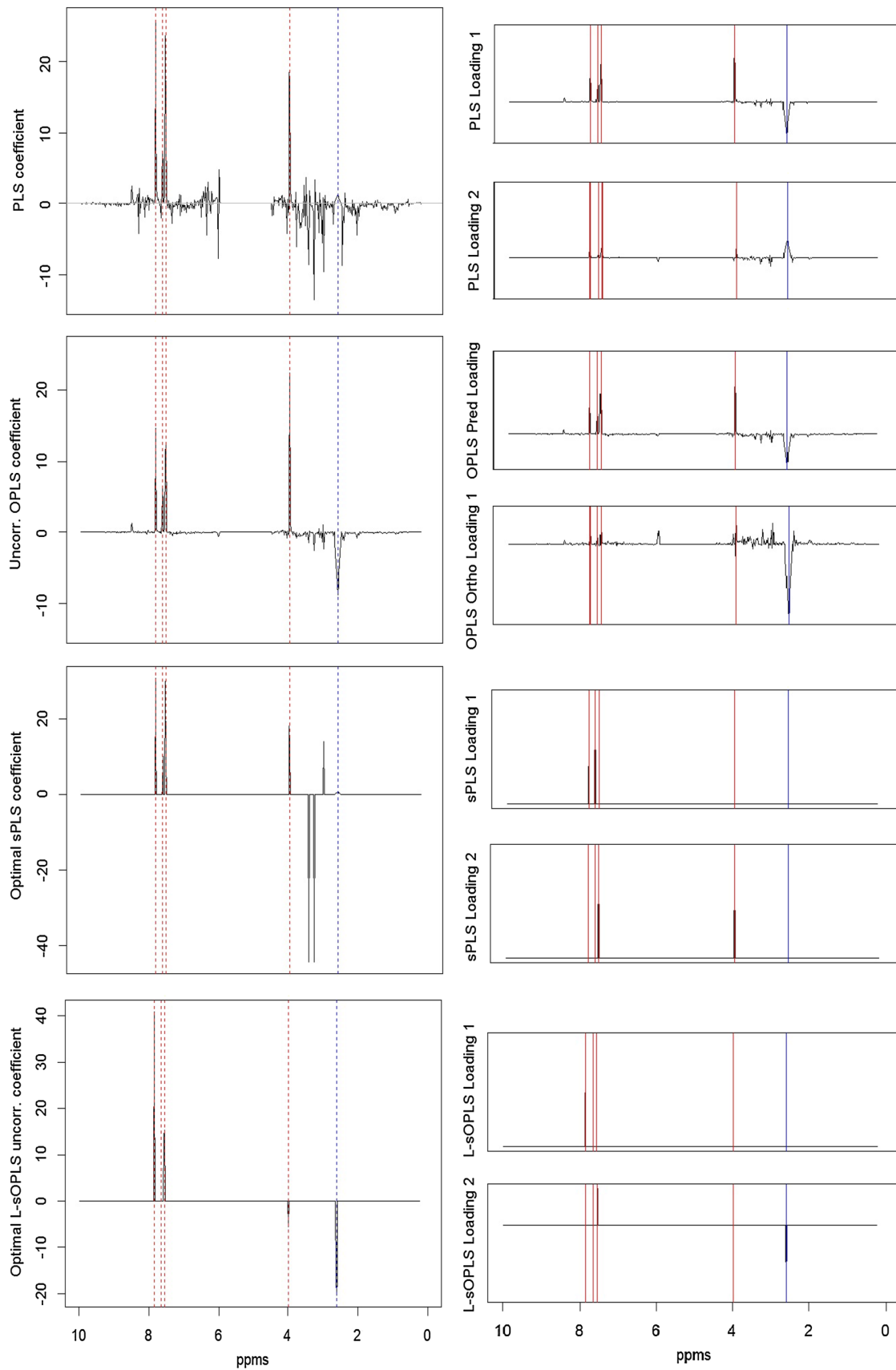
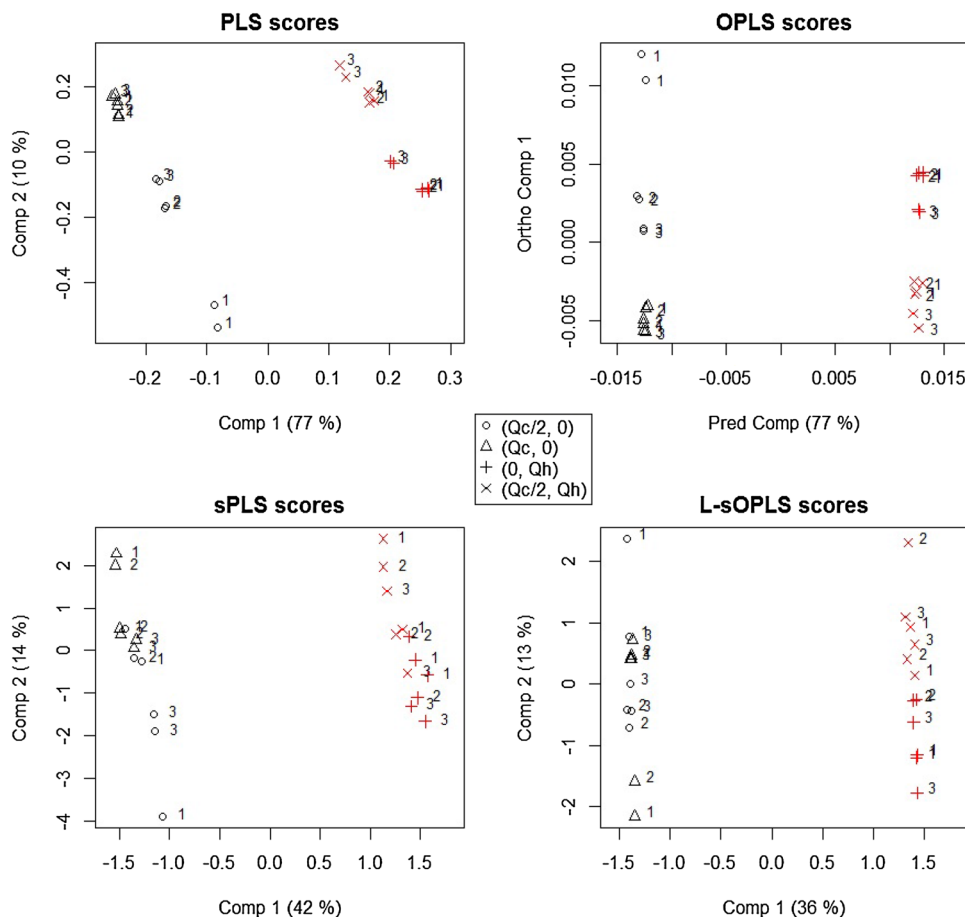


Fig. 2 (O)PLS, sPLS and L-sOPLS coefficients (left) and two first loadings (right) for the unbalanced hippurate case

Fig. 3 (O)PLS, sPLS and L-sOPLS scores for the unbalanced hippurate case. Colours discriminate Group 1 and Group 2, symbols refer to the citrate and hippurate concentrations and numbers refer to the day of measurement



binary vector to predict is the presence/absence of the endometriosis pathology.

In the same way as for the previous data, cross-validated PLS and OPLS coefficients and sparsing results obtained via sPLS and L-sOPLS are displayed in Table 3, along with the optimal parameters and the obtained minimal LOO-RMSEP. An underlined “Yes” in the sparsing selection columns means that the corresponding miRNA variable is always selected as important biomarker for any penalty term, even very severe; when the “Yes” is not underlined, it means that the corresponding variable is often selected when the penalty term is slightly modified.

The optimal PLS regression involves here two predictive components; optimal OPLS involves one orthogonal and one predictive component; optimal sPLS involves two predictive components and $\lambda_1 = 0.68$; and finally, optimal L-sOPLS involves $(q_{ortho}, q_{pred}, \lambda_1) = (1, 1, 0.67)$.

As for the previous data set, the global minimal LOO-RMSEP (= 0.4147) is obtained with a L-sOPLS model providing seven final biomarkers; and the majority of tested L-sOPLS models, for different (q_{pred}, λ_1) combinations, provide better predictive performances than the best (O) PLS reference model. So, again, no compromise must be done between very competitive predictions on one side and

sparsity on the other side (leading to simpler and more interpretable models) when using L-sOPLS.

Two clear primary biomarkers are identified by the four algorithms: \log_{miR_x2} and \log_{miR_x3} (Table 3). These miRNAs are among the higher coefficients found by PLS/OPLS and are also selected by the sparse solutions.

5 Conclusion and further works

In this article, objective sparse solutions are provided in order to solve the problem linked with the subjective selection of biomarkers in -omics studies (how many? What thresholds? What cut-offs?) when using PLS or OPLS. Sparse-PLS (sPLS) is already implemented in computer software and begins to be used in the metabolomics community, but an innovative and quite intuitive approach combining sparsity in the feature selection and the orthogonalization advantages of OPLS is proposed in this paper with the Light-sparse-OPLS (L-sOPLS).

Applied on urine ¹H-NMR spectral data or on genomic q-PCR data, this new sparse algorithm leads to very convincing results, by selecting a (very) small number of (very) relevant features as biomarkers and by providing,

Table 3 (O)PLS and sparse results for the RT-qPCR data

miRNAs	PLS coefficients $q = 2$	OPLS b_{OPLS} $q = 1 + 1$	sPLS selection $q_{pred} = 2, \lambda_1 = 0.68$	L-sOPLS selection $q_{ortho} = 1, q_{pred} = 1,$ $\lambda_1 = 0.67$
	LOO-RMSEP = 0.4356	LOO-RMSEP = 0.4356	LOO-RMSEP = 0.4297	LOO-RMSEP = 0.4147
log_miR_x1	– 0.083	– 0.031	Yes (– 0.115)	No
log_miR_x2	0.066	0.050	Yes (0.073)	Yes (0.072)
log_miR_x3	0.065	0.045	Yes (0.081)	Yes (0.065)
log_miR_x4	– 0.053	– 0.009	No	No
log_miR_x5	0.053	0.034	Yes (0.065)	Yes (0.049)
log_miR_x6	– 0.049	0.014	No	No
log_miR_x7	– 0.048	– 0.019	No	No
log_miR_x8	0.042	0.037	Yes (0.046)	Yes (0.054)
log_miR_x9	– 0.031	– 0.001	No	No
log_miR_x10	0.026	0.040	Yes (0.014)	Yes (0.058)
log_miR_x11	0.024	0.039	Yes (0.009)	Yes (0.056)
log_miR_x12	0.022	0.028	No	No
log_miR_x13	0.022	0.038	Yes (0.009)	Yes (0.055)
log_miR_x14	0.015	0.009	No	No
log_miR_x15	– 0.015	– 0.004	No	No
log_miR_x16	0.007	0.015	No	No
log_miR_x17	0.002	0.032	Yes (– 0.020)	No
log_miR_x18	0.001	0.023	No	No
log_miR_x19	– 0.001	0.030	No	No

The six higher coefficients in absolute value are highlighted in bold for PLS and OPLS

An underlined "Yes" denotes that the corresponding variable is always selected as important biomarker for any sparsity penalty term

consequently, lighter cross-validated optimal models to practitioners which may be much easier to interpret and to deploy. However, often, the sparsity goal can be only reached at the expense of lower predictive performances, when classifying new observations, compared to non-sparse models. It is mainly the case for sPLS models, for which compromises have to be made between drastic feature selection and predictive power. But such compromises do not seem to be necessary for the L-sOPLS, for which sparsity, even at a severe level, goes hand in hand with a better predictive ability than (O)PLS and sPLS models.

The objective of this article remains focussed on biomarker identification. A further interesting work would consist in more deeply investigate the predictive performances of L-sOPLS (binary predictions, false positives/negatives, ROC curves,...) and compare them to standard PLS or OPLS predictions. Then, it would probably emphasize even more a question of compromise between “perfect” predictions and interpretable sparse predictions.

Another further possible work would be to address the correlation or colinearity into the data that very often characterizes metabolomics spectral databases. Several signals, for

instance peaks in $^1\text{H-NMR}$, can overlap, move together and/or be associated to a same molecule. So, an idea would be to apply L-sOPLS on 2D-NMR spectra (COSY for instance Féraud et al. 2015) or to take into account groups of features instead of individual features as inputs of the sparse algorithms. Methods like Group-LASSO, Sparse-Group-LASSO or Overlay-Group-LASSO (Friedman et al. 2010) could then be tested in this context.

5.1 Softwares

As mentioned regularly all along this paper, the R software (<http://www.R-project.org>) environment was exclusively used, via existing packages (*pls*, *spls*, *ropls*), or coded ad hoc (OPLS and L-sOPLS, functions which are available here: <https://github.com/ManonMartin/MBXUCL>).

Acknowledgements The authors thank the GIGA-Cancer laboratory and Eli Lilly and Company for providing the data used in this paper. Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is also gratefully acknowledged.

Compliance with ethical standards

Conflict of interest Authors declare that they have no conflict of interest.

Ethical approval This study analyzes collected data which involved human participants. For the q-PCR data set, the study was approved by our local Ethics Committee (CHR Citadelle, Liège, number B412201215082-1267) and all patients gave their informed consent.

References

- Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (pls regression). *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1), 97–106.
- Afanador, N. L., Tran, T. N., & Buydens, L. (2013). Use of the bootstrap and permutation methods for a more robust variable importance in the projection metric for partial least squares regression. *Analytica Chimica Acta*, 768, 49–56.
- Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, 17(3), 166–173.
- Bartel, D. P. (2009). MicroRNAs: Target recognition and regulatory functions. *Cell*, 136(2), 215–233.
- Bylesjo, M., Rantalainen, M., Cloarec, O., & Nicholson, J. (2006). OPLS discriminant analysis: Combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics*, 20(8–10), 341–351.
- Chapman, A., & Saad, Y. (1997). Deflated and augmented Krylov subspace techniques. *Numerical Linear Algebra with Applications*, 4(1), 43–66.
- Chun, H., & Keles, S. (2007). *Sparse partial least squares regression with an application to genome scale transcription factor analysis*. Madison: Department of Statistics, University of Wisconsin.
- Chung, D., Chun, H., & Keles, S. (2012). Spls: Sparse partial least squares (SPLS) regression and classification. R package, version, 2, 1–1.
- De Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3), 251–263.
- Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32(2), 407499.
- Feraud, B., Govaerts, B., Verleysen, M., & De Tullio, P. (2015). Statistical treatment of 2D NMR COSY spectra in metabolomics: Data preparation, clustering-based evaluation of the metabolomic informative content and comparison with ¹H-NMR. *Metabolomics*, 11(6), 1756–1768.
- Friedman J., Hastie T., & Tibshirani R. (2010). A note on the group lasso and a sparse group lasso. arXiv preprint [arXiv:1001.0736](https://arxiv.org/abs/1001.0736).
- Gabrielsson, J., Jonsson, H., Airiaub, C., & Schmidt, B. (2006). OPLS methodology for analysis of pre-processing effects on spectroscopic data. *Chemometrics and Intelligent Laboratory Systems*, 84(1–2), 153–158.
- Geladi, P., & Kowalski, B. R. (1986). Partial least squares regression: A tutorial. *Analytica Chimica Acta*, 185, 1–17.
- Giudice, L. C., & Kao, L. C. (2004). Endometriosis. *Lancet*, 364, 178999.
- Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. Boca Raton: CRC Press.
- Hoskuldsson, A. (1988). PLS regression methods. *Journal of Chemometrics*, 2(3), 211–228.
- Indahl, U. G., Liland, K. H., & Ns, T. (2009). Canonical partial least squares: A unified PLS approach to classification and regression problems. *Journal of Chemometrics*, 23(9), 495–504.
- Jung, Y., Lee, J., Kwon, J., Lee, K. S., Ryu, D. H., & Hwang, G. S. (2010). Discrimination of the geographical origin of beef by ¹H-NMR-based metabolomics. *Journal of Agricultural and Food Chemistry*, 58(19), 10458–10466.
- Lai, E. C. (2002). Micro RNAs are complementary to 3 UTR sequence motifs that mediate negative post-transcriptional regulation. *Nature Genetics*, 30, 363.
- Lê Cao, K. A., Rossouw, D., Robert-Grani, C., & Besse, P. (2008). A sparse PLS for variable selection when integrating omics data. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 35.
- Lu, B., Castillo, I., Chiang, L., & Edgar, T. F. (2014). Industrial PLS model variable selection using moving window variable importance in projection. *Chemometrics and Intelligent Laboratory Systems*, 135, 90–109.
- Mevik, B. H., & Cederkvist, H. R. (2004). Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). *Journal of Chemometrics*, 18(9), 422–429.
- Munoz-Romero, S., Arenas-Garca, J., & Gmez-Verdejo, V. (2015). Sparse and kernel OPLS feature extraction based on eigenvalue problem solving. *Pattern Recognition*, 48(5), 1797–1811.
- Nisenblat V., Bossuyt P. M., Shaikh R., Farquhar C., Jordan V., Scheffers C. S., ... & Hull M. L. (2016). Blood biomarkers for the non-invasive diagnosis of endometriosis. The Cochrane Library.
- Rousseau, R. (2011). Statistical contribution to the analysis of metabolomic data in ¹H-NMR spectroscopy (Doctoral dissertation, Université Catholique de Louvain, Belgium), permalink: <http://hdl.handle.net/2078.1/75532>.
- Stenlund, H., Gorzdas, A., Persson, P., Sundberg, B., & Trygg, J. (2008). Orthogonal projections to latent structures discriminant analysis modeling on in situ FT-IR spectral imaging of liver tissue for identifying sources of variability. *Analytical Chemistry*, 80(18), 6898–6906.
- Tapp, H. S., & Kemsley, E. K. (2009). Notes on the practical utility of OPLS. *TrAC Trends in Analytical Chemistry*, 28(11), 1322–1327.
- Trygg, J., & Wold, S. (2002). Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, 16(3), 119–128.
- van Gerven, M. A. J., & Heskes, T. (2010). Sparse orthonormalized partial least squares. In Benelux conference on artificial intelligence.
- Wehrens, R. (2011). *Chemometrics with R: Multivariate data analysis in the natural sciences and life sciences* (pp. 155–165). New York: Springer.
- Weljie, A. M., Bondareva, A., Zang, P., & Jirik, F. R. (2011). ¹H-NMR metabolomics identification of markers of hypoxia-induced metabolic shifts in a breast cancer model system. *Journal of Biomolecular NMR*, 49(3–4), 185–193.
- Wiklund, S., Johansson, E., Sjöström, L., Mellerowicz, E., Edlund, U., Shockcor, J. P., et al. (2008). Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Analytical Chemistry*, 80(1), 115–122.
- Wold, H. (1975). *Path models with latent variables: The NIPALS approach* (pp. 307–357). New York: Academic Press.
- Wold, S., Trygg, J., Berglund, A., & Antti, H. (2002). Some recent developments in PLS modeling. *Chemometrics and Intelligent Laboratory Systems*, 58(2), 131–150.
- Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2), 109–130.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301–320.