Two data pre-processing workflows to facilitate the discovery of biomarkers by 2D NMR metabolomics

Baptiste Féraud, Justine Leenders, Estelle Martineau, Patrick Giraudeau, Bernadette Govaerts, Pascal de Tullio

INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(ISBA)

DISCUSSION PAPER

2018/16

Two data pre-processing workflows to facilitate the discovery of biomarkers by 2D NMR metabolomics

Féraud, B., Leenders, J., Martineau, E., Giraudeau, P., Govaerts, B. and P. de Tullio

Metabolomics manuscript No.
(will be inserted by the editor)

Two data pre-processing workflows to facilitate the discovery of biomarkers by 2D NMR metabolomics

Baptiste Féraud · Justine Leenders · Estelle Martineau · Patrick Giraudeau · Bernadette Govaerts · Pascal de Tullio

Received: date / Accepted: date

Abstract *Introduction* The pre-processing of analytical data in metabolomics must be considered as a whole to allow the construction of a global and unique object for any further simultaneous data analysis or multivariate statistical modelling. For 1D ^{1}H-NMR metabolomics experiments, best practices for data pre-processing are well defined, but not yet for 2D experiments (for instance COSY in this paper).

Baptiste Féraud
Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), Université Catholique de Louvain (UCL), Belgium
Machine Learning Group, UCL, Belgium
Address: Voie du Roman Pays 20, bte L1.04.01, B-1348 Louvain-la-Neuve, Belgium
Tel.: +32 10473053
E-mail: baptiste.feraud@uclouvain.be

Justine Leenders
Université de Liège (ULg), Center for Interdisciplinary Research on Medicines (CIRM), Metabolomics group, Liège, Belgique

Estelle Martineau
Université de Nantes, EBSI Team, Chimie et Interdisciplinarité, Synthèse, Analyse, Modélisation (CEISAM), CNRS, UMR 6230, Nantes, France
Spectromaîtrise, CAPACITES SAS, 26 Bd Vincent Gâche, 44200 Nantes, France

Patrick Giraudeau
Université de Nantes, EBSI Team, Chimie et Interdisciplinarité, Synthèse, Analyse, Modélisation (CEISAM), CNRS, UMR 6230, Nantes, France
Institut Universitaire de France, 1 rue Descartes, 75005 Paris cedex 5, France

Bernadette Govaerts
Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA), Université Catholique de Louvain (UCL), Belgium

Pascal de Tullio
Université de Liège (ULg), Center for Interdisciplinary Research on Medicines (CIRM), Metabolomics group, Liège, Belgique

Objective By considering the added value of a second dimension, the objective is to propose two workflows dedicated to 2D NMR data handling and preparation (the Global Peak List and Vectorization approaches) and to compare them (with respect to each other and with 1D standards). This will allow to detect which methodology is the best in terms of amount of metabolomic content and to advantageously illustrate this selected workflows, as proof of concept, in order to finally be able to find relevant biomarkers.

Methods To select the more informative data source, MIC (Metabolomic Informative Content) indexes are used, based on clustering and inertia measures of quality. Then, to highlight biomarkers or critical spectral zones, the PLS-DA model is used, along with more advanced sparse algorithms (sPLS and L-sOPLS).

Results Results are discussed according to two different experimental designs (one which is unsupervised and based on human urine samples, and the other which is controlled and based on spiked serum media). MIC indexes are shown, leading to the choice of the more relevant workflow to use thereafter. Finally, biomarkers are provided for each case and the predictive power of each candidate model is assessed with cross-validated measures of RMSEP.

Conclusion In conclusion, it is shown that no solution can be universally the best in every case, but that 2D experiments allow to clearly find relevant biomarkers even with a poor initial separability between groups. The MIC measures linked with the candidate workflows (2D GPL, 2D vectorization, 1D, and with specific parameters) lead to visualize which data set must be used as a priority to more easily find biomarkers. The diversity of data sources may often lead to complementary or confirmatory results.

Keywords 2D NMR · ^{1}H-NMR · COSY spectra · Pre-prossessing workflows · Metabolomic Informative Content (MIC) · Biomarker discovery · PLS · sPLS · L-sOPLS

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 its degree of progress, to estimate the effects of a pharmaceutical treatment, to control the quality of consumer goods, etc. Biomarkers are then a way to explain and/or to anticipate an 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.

In practice, the statistical detection of such 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, but note that one can generalize to a multilevel response variable). The popularity of PLS and OPLS (Orthogonal PLS) regression methods in metabolomics dates from the early 2000s, mainly based on the works of Svante Wold and Johan Trygg [Wold, Trygg et al., 2001] [Wold et al., 2001], and the parallel development of the SIMCA software (see for instance [Bylesjo et al., 2006]). Since then, this popu-

larity has never stopped growing and the vast majority of the past and current biomarkers' researches are linked to the PLS(-DA) principles.

Because it can be advantageous to deal with only a small number of -very- significant and ideally easily interpretable biomarkers, several studies involving sparse solutions and methods have been proposed these last years, with the objective to reinforce the most significant biomarkers' coefficients and to force the less significant ones to be equal to zero (according for instance 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 issues in metabolomics. Sparse PLS (sPLS) [Chun and Keles, 2007] and Light sparse Orthogonal PLS (L-sOPLS) [Feraud et al., 2017] are tested (but other sparse techniques are of course possible).

But the main purpose of this paper concerns the input data chosen to feed these algorithms. Best practices for 1D spectral data, and in particular for proton ^{1}H-NMR are now well-known and commonly applied ([Ravanbakhsh et al., 2014] [Craig et al., 2006] [Martin et al., 2017] for example). For two-dimensional data sources, it is not fully the case yet (as for COSY spectra, for COrrelation SpectroscopY, considered here). When the spectral results of an experience or an experimental design have to be considered as a whole, it is critical to create a global and unique data object for further simultaneous statistical analysis or multivariate modelling (as biomarker discovery).

The design of appropriate data pre-processing workflows for 2D NMR data appears timely, since 2D NMR has recently emerged as a promising alternative to 1D NMR in metabolomics studies [Marchand et al., 2017]. 2D spectroscopy offers a broad variety of experiments that can significantly increase the dispersion of NMR signals, and this is particularly useful in the case of complex biological samples whose 1D NMR spectra are severely hampered by peak overlaps. In 2D NMR, peak volumes remain proportional to metabolite concentrations -although the analytical response is peak-specific, contrary to quantitative 1D NMR [Giraudeau, 2014]. The inter-comparison of peak volumes between samples remains possible, as well as the use of 2D spectra for targeted quantification if the peak response factor is calibrated through an external calibration or with standard additions [Giraudeau et al., 2014]. 2D NMR experiments suffer from long experiment times, but this duration can be significantly shortened if needed by relying on fast acquisition methods [Rouger et al., 2017]. Several recent studies have shown the potential of 2D NMR in metabolomics, either for untargeted analysis ([Le Guennec et al., 2014] [Marchand et al., 2018]) or for the targeted quantification of metabolites ([Le Guennec et al., 2012] [Martineau et al., 2012] [Jezequel et al., 2015]). However, all these studies were focused on the improvement of data acquisition methods, but suffered from a lack of standardized approach for the pre-processing of 2D data.

Two methodologies, or workflows, to handle 2D COSY spectral data are presented and compared in order to fill this lack. First, the Global Peak List (GPL) workflow [Feraud et al., 2015] implies the construction of individual peak lists linked with each initial individual spectrum and the intelligent merging of these

individual peak lists into a global matrix, called GPL. The GPL workflow also allows to control the resolution (and also the final size of the final object) and basic data treatments. Second, a vectorization workflow is proposed which is not dependant of any external software and which allows to fully and more precisely fix the size of the final matrix. The goal of this work is to consider these two workflows, to see how they can be tuned via different parameters and to compare them to each other and with the 1D ^{1}H-NMR approach in terms of amount of metabolomic content, or useful information, contained in the resulting data matrices. When a specific workflow is declared the best for classifying and separating homogeneous groups, the biomarker discovery step can advantageously be implemented on it.

The paper is organized as follows. Section 2 provides a detailed description of the two data sets used to demonstrate and compare the workflows: a COSY design involving different urine donors and a controlled COSY design involving spiked doses of two products (threonine and glutamate) in serum media. The questions of interest will be respectively the following for these two data sets: how can we blindly retrieve and separate the urine donors? How can we identify the threonine and glutamate molecules' fingerprints and then accurately find these two products as primary biomarkers?

The GPL and vectorization workflows, the assessment of the amount of captured metabolomic content (via the Metabolomic Informative Content, MIC [Feraud et al., 2015]) and the biomarker discovery statistical models (PLS, sPLS, L-sOPLS) are detailed in the methodological Section 3. Results for both data sets are then shown and discussed in Section 4. Finally, a general conclusion is given in Section 5.

2 Materials: data sets and experimental protocols

In this section, the two selected data sets used to train the two workflows and to illustrate the biomarkers selection issue are presented. For both of them, a description and motivational explanations are provided, as well as the main acquisition parameters.

2.1 First experimental design: urine donors

2.1.1 Description and motivations

This first data set was built in the context of a wider inter-laboratory study about repetability of different 1D ^{1}H-NMR and 2D COSY spectral measures and acquisition protocols. The experience involves three factors of variation: three different donors, two urine dilution levels and four different days of acquisition. Eight measures are finally available for each donor.

In this paper, note that the focus is strictly on the group factor (i.e. the donors) as it corresponds to the signal we want to capture and explain in subsequent sections. In this regard, urine dilution and days factors can be considered as additional sources of noise and will not be commented separately in details.

Concretely, the idea is to handle all these data sources with the two presented workflows, to visualize which of them succeeds best in capturing the signal or main information (i.e. the donors) and finally to illustrate the allowed benefits

when finally searching for relevant biomarkers (i.e. metabolites which contribute to classify the data into homogeneous groups).

2.1.2 Urine collection

In order to conduct this experiment and to design the collection of urine samples, the morning urine of three different fasting donors was collected. For each donor, four aliquots of 400 μl and four aliquots of 320 μl were placed at -80°C. Then, on each consecutive day (four days), six aliquots were thawed (3 donors $\times$ 2 quantities) and routinely prepared as follows.

Urine samples of 400 μl were supplemented with 300 μl of deuterated phosphate buffer (DPB, pH 7.4), while 320 μl urine aliquots were supplemented with 380 μl of the same buffer. 10 μl of a 10 mg/ml TMSP solution was then added to all aliquots. The four aliquots of each dilution were put in 5mm NMR tube for NMR acquisition and analyzed. For each day, the order of measurement was held constant across the six sub-samples. A total of 24 1D and 2D collected signals are finally available.

All spectra are internally labelled as $S_i\_Dj\_E_k$, where S corresponds to the donor label ($i = 1, ..., 3$), D is the dilution ($j = 0$: no dilution; $j = 1$: 25% diluted) and E is the day of acquisition ($k = 1, ..., 4$).

The acquisition of 1D and 2D NMR spectra is described below in Section 2.3.

2.2 Second serum based experimental design: spiked products

2.2.1 Description and motivations

The second dataset consists in proton-NMR and 2D COSY spectra acquisition of sera samples spiked with known concentrations of metabolites. In this dataset, glutamate and threonine have been used as spiking references. Indeed, ^{1}H-NMR identification of these two metabolites is usually difficult to perform on clinical samples due to spectral crowding and peak overlapping in their region. Threonine is completely overlapped by lactate at 1.32 ppm, while glutamate is overlapped at 2.12 ppm and 2.34 ppm by proline, lipoproteins and glutamine [Marjanska et al., 2008]. The purpose of this dataset is to assess the separation between spiked and non-spiked samples using the two presented workflows, in order to determine which one is the best at highlighting these spiked metabolites.

2.2.2 Blood collection and spiking

Serum from a single donor was collected and 36 aliquots of 500 μl were prepared and stored at -80°C. On three consecutive days of experiment, 12 aliquots were thawed and spiked under four different conditions as follows: i) 500 μl serum without any product added (internally labelled as "$P1D1\ P2D1$"); ii) 500 μl serum spiked with 7 μl of a 17 mM threonine solution ("$P1D2\ P2D1$"); iii) 500 μl serum spiked with 10 μl of a 8.3 mM glutamate solution ("$P1D1\ P2D2$"); iv) and 500 μl serum spiked with 7 μl of the 17 mM threonine solution and with 10 μl of the 8.3 mM glutamate solution ("$P1D2\ P2D2$"). 30 μl of a 10 mg/ml TMSP solution and 50 μl of a 35 mM maleic acid solution were then added to all samples. Finally, samples were supplemented with different volumes of deuterated

phosphate buffer (DPB, pH 7.4) in order to reach a final volume of exactly 700 μl in each case. Samples were put in 5 mm NMR tube for NMR acquisition and were analyzed. A total of 36 spectral measures are finally available.

In this experiment, spectra are labelled as $J_i\ P1D_j\ P2D_j\ R_k$, where J is the day of measurement ($i = 1, ..., 3$), $P1$ corresponds to threonine and $P2$ to glutamate, D refers to the spiking condition ($j = 1$: no spiking; $j = 2$: spiking of the corresponding metabolite) and R is the repetition of the experiment ($k = 1, ..., 3$ for each day).

The acquisition of 1D and 2D NMR spectra is described below in Section 2.3.

2.3 NMR spectroscopy

All samples were analyzed at 298K on a Bruker Avance spectrometer operating at 500.13 MHz for proton signal acquisition. The instrument was equiped with a 5 mm TCI cryoprobe with a Z-gradient. ^{1}H-NMR spectra were acquired using a 1D NOESY pulse sequence with presaturation for urine samples and a CPMG relaxation-editing sequence with presaturation for serum samples. The NOESY experiment with water signal presaturation used a RD-90°-Tau-90°-Tm-90°-acquire sequence with a relaxation delay of 4 s, a mixing time (Tm) of 10 ms, and a fixed Tau delay of 4 μs. The water suppression pulse was placed during the relaxation delay (RD). The CPMG experiment used a RD-90-(t-180-t)n-sequence with a relaxation delay (RD) of 2 s, a spin echo delay (t) of 400 μs and a number of loops equal to 80. The number of transients was typically 32, and a number of 4 dummy scans was chosen. The acquisition time was fixed to 3.28 s. The data were then processed with the Bruker Topspin 3.2 software with a standard parameter set.

Before Fourier transformation, FIDs were subjected to exponential multiplication resulting in an additional line-broadening of 0.3 Hz. Phase and baseline corrections were performed manually over the entire range of the spectra, and the δ scale was calibrated to 0 ppm, using the internal standard TMSP.

For 2D experiments, gradient enhanced magnitude COSY (pulse sequence cosygpprqf supplied by Bruker) with a pre-saturation during relaxation delay was used for urine and sera samples 2D measurements. Spectra were collected with 4096 points in F2 and 300 points in F1 over a sweep width of 10 ppm, with 4 scans per F1 value. The acquisition times were fixed to 0.256 s in F2 and 0.0187 s in F1. The resulting COSY spectra were processed in Topspin 3.2 using standard methods, with sine-squared apodization in both dimensions and zero filling in F1 to yield a transformed 2D dataset of 2048 by 2048 points. Finally, Fourier transformation, baseline correction and symmetrisation were performed manually on all FIDs.

3 Methods

This methodological section discusses the data pre-processing steps to apply when the full spectral acquisition has been performed. As previously said, when a whole experimental design is considered, a global and unique object has to be built for further data analysis or multivariate statistics. Two workflows are presented in

this paper: a workflow based on the concept of "Global Peak Lists" (GPL, see [Feraud et al., 2015] and Section 3.1) and another which implies the vectorization of the initial spectral matrices (Section 3.2). In this paper, note that we make the implicit and intuitive assumption that 2D data requires a less advanced level of pre-processing than for 1D data.

When an unique data object has been built, any multivariate statistical tool can then be applied on it: clustering, classification, PCA, PLS regressions, sparse regressions, etc... Different data sources, or matrices, built according to differents acquisition parameters or conditions, can first be compared in order to determine which set contains most information about the signal. Inertia or Metabolomic Informative Content (MIC) measures [Feraud et al., 2015] are presented to answer this question (Section 3.3).

In the context of biomarker identification, this unique and global object can then be considered as input into usual (O)PLS regressions or into more advanced sparse models (Section 3.4).

3.1 The Global Peak List (GPL) approach

This approach is fully detailed in [Feraud et al., 2015] and is structured as illustrated in Fig.1 (left part). First, each of the n initial individual 2D spectra is converted into an individual peak list, i.e. a $(t_i \times 3)$ matrix which contains the two ppm coordinates and the concentration intensity of t_i existing peaks. After some manipulations detailed below, the n peak lists are merged, resulting in a $(T \times M_{GPL})$ Global Peak List (GPL) matrix. This matrix includes the T pairs of coordinates that appear in at least one of the individual spectra and all the corresponding intensities. Note that the number of rows T is not known in advance.

The individual peak lists can be obtained from initial spectra with, for example, the ACD/Labs free software (ACD/NMR processor). It implies the choice of a threshold to determine when an intensity level begins to be relevant and can not be associated with pure noise only. This threshold has to be maintained at a low level, typically between 0.02 and 0.05 in ACD/Labs.

A list of pre-processing steps can be applied on the individual initial peak lists in order to increase the impact of the informative sources and to remove potential unnecessary artefacts. These steps include, for instance, the symmetrisation of homonuclear 2D spectra with respect to the diagonal, or the removal of negative intensities.

With biological samples, one major problem is the strong residual water signal, even with a previous application of some solvent signal suppression techniques. As a result, a water zone deletion is very useful to avoid over-representations (it mainly concerns the water zone, but may also concern urea, maleic acid or lipoproteins). A normalization of the intensities (using Constant Sum = 1) and a further log-transformation can also be of crucial interest and added in order to stabilize distribution variances.

Finally, a dimension reduction step is also applied. In ^{1}H-NMR spectroscopy, bucketing tools are common and widely used to control the spectral resolution and/or to overcome the misalignments problems. Classical and more advanced bucketing methods have already shown their usefulness for ^{1}H-NMR spectra [Rousseau, 2011]

[Sousa et al., 2013]. In this workflow, a soft bucketing step adapted to 2D COSY is used in order to control the resolution level of the two-dimensional spectra. Practically, a variation of the number of decimals of the coordinates is simply proposed. The intensities belonging to a bucket are then aggregated. For example, if the couples of coordinates [3.286; 4.194], [3.281; 4.189] and [3.278; 4.191] provide positive intensities INT1, INT2 and INT3 respectively, the couple [3.28; 4.19] provides an intensity equal to INT1+INT2+INT3 when adjusting the number of allowed decimals from 3 to 2. Using this method, the width of the COSY peaks is adjusted and the size of the resulting database is adjusted simultaneously. Furthermore, intermediate resolutions can be computed in a similar way.

All these steps can be easily implemented using the R programming language (`http://www.R-project.org`).

3.2 The vectorization approach

Unlike the GPL approach, the vectorization one can be directly applied on raw Bruker text files. The choice of the intensity threshold when using ACD/Labs is then not anymore an issue.

The principle is quite straightforward: each individual initial $(m \times p)$ 2D spectral matrix is first bucketed twice (by rows and by columns) and summarized into a $(M_V \times M_V)$ object. The high dimensionality of the data and small residual peak shifts can indeed impede the future multivariate data analyses [Liland, 2011] and bucketing reduces such problems by integrating the p original spectral intensities into M_V predefined intervals, or buckets, with $M_V < p$. For convenience with standards, M_V is often chosen equal to 256 or 512 here (or subsequent 2^k multiples according to the initial resolution) in order to control the dimension reduction and to avoid truncating extremities.

In this step, the optimal trade-off lies between keeping the spectral information and removing the peaks shifts as well as decreasing the total number of variables. Among possible binning methods, The R package PepsNMRs ([Martin et al., 2017]) bucketing function proposes two integration options, either trapezoidal or rectangular, with equally sized buckets and is generalized to cut the original axis at any chosen location. For 2D COSY, rectangular bucketing is chosen for its more intuitive aspect linked with a kind of pixelation of a spectral grid.

These bucketed 2D objects are then vectorized, transformed into a $(1 \times {M_V}^2)$ row vector. Finally, these row vectors are stacked to form a global matrix of size $(n \times {M_V}^2)$ containing the whole information from all the initial spectra.

Before this bucketing step, the water zone is set to zero and the normalization (Constant Sum = 1) step is performed. A subsequent log-transformation could also be considered for the same reason as the GPL methodology.

The vectorization workflow is displayed in simplified form in Fig.1 (right part). The different patterns represented in this figure could involve different sources for the initial data, different acquisition techniques or set of parameters, etc. In this figure, all the steps are displayed from the signal acquisition to the final obtained global matrices. Note that the initial FID are transformed and pre-processed via, for example, the Bruker TopSpin software (double Fourier transform, baseline

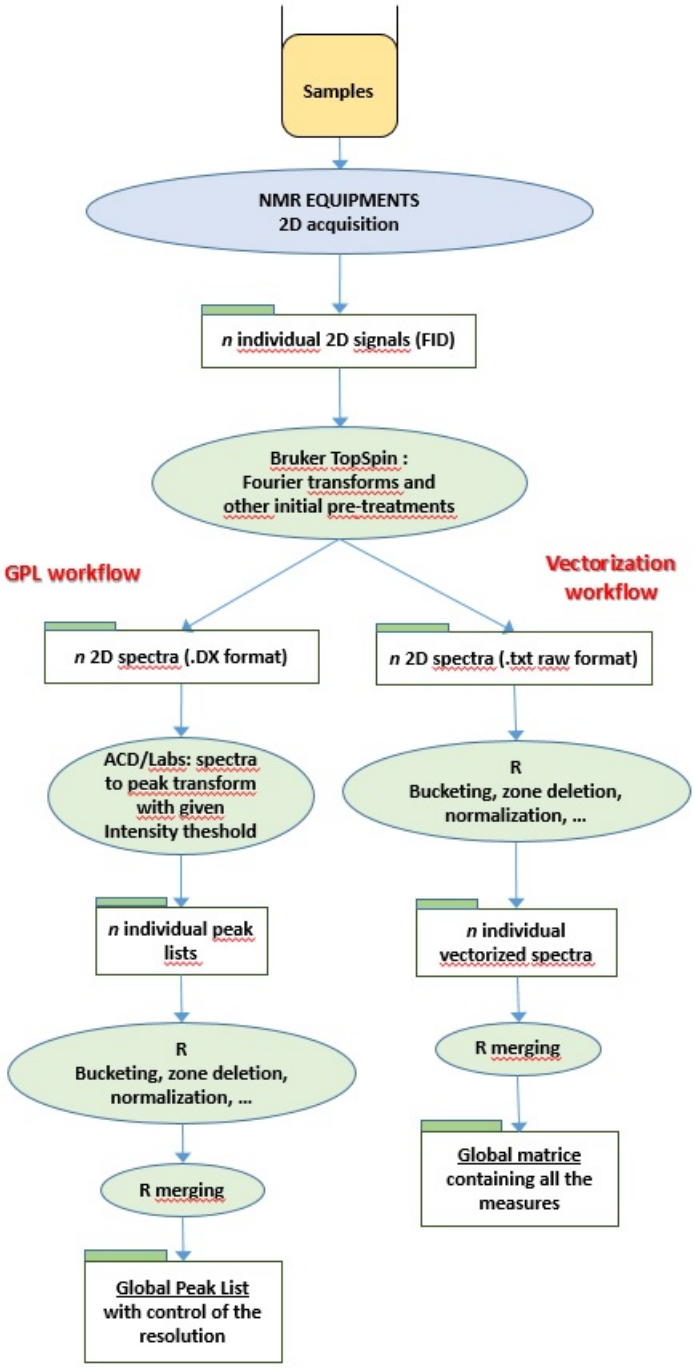

Fig. 1 The GPL and vectorization approaches workflows for 2D experiments: steps and resulting data files.

correction, etc.). Recommended data file formats are also mentioned. For instance, the .dx format is more adapted to the ACD/Labs environment.

3.3 The assessment of the Metabolomic Informative Content

Once a set of global objects is obtained for a given experimental study, Global Peak Lists or Vectorized spectra, stemming from various sources (urine, blood, serum, ...), various experiences (COSY, TOCSY, ...) or from the variation of some parameters of acquisition at any step of the workflows, a natural progress involves determining which one is the "best" to consider further more elaborate statistical studies. "Best" means here the capacity to distinguish and highlight the useful information, the signal, by opposition to the experimental or environmental sources of variability.

Of course, repetitions of the sample measures have to be ideally planned during the data acquisition when these signal/noise studies are of major interest. Moreover, the presence of groups to recover into the data is very important and helpful by taking the role of the above-mentioned signal.

When such data are available, quality criteria to compare distinct pre-processing strategies can be derived from unsupervised as well as supervised chemometric tools. In this paper, the selected criteria are gathered under the Metabolomic Informative Content (MIC) concept, developed in [Feraud et al., 2015] and include complementary inertia measures, clustering analysis quality measures and PCA or PLS-DA related criteria.

First, the inertia analysis decomposes the total variance into two complementary parts: the variance between the groups (maximized in a good partition) and the variance within the groups of observations (minimized in a good partition). Second, the clustering results obtained via the Ward's algorithm ([Ward, 1963], [Murtagh and Legendre, 2011]) or the K-means one ([McQueen, 1967]) are summarized into some criteria. The (adjusted) Rand indexes measure the true class recovery efficiency and should be maximized, with a maximal value of 1, while the Dunn and Davies-Bouldin indexes measure the clustering homogeneity and they have to be respectively maximized and minimized (formula details can be found in [Feraud et al., 2015]). Finally, PCA and PLS-DA can allow to graphically recover the groups from the spectral data.

A priori, the spectral pool which provides the best MIC performances on average would be the pool that allows in the best way to capture the relevant information and to discern the useful signal relative to the noise.

3.4 Biomarker identification

After data preparation and data quality analysis, the next step consists in applying multivariate statistical tools to model the information and ideally discover relevant biomarkers.

In order to apply and illustrate the two workflows, as proof of concept, and the quality measurements, the idea is now to discriminate the urine donors and to blindly identify discriminant biomarkers or spectral zones when using the first

dataset; and to discriminate the groups linked with different doses of threonine and glutamate when using the second dataset. In the later controlled case, the biomarkers to be found should be directly linked with threonine and/or glutamate.

The conventional PLS ([Wold et al., 2001] [Trygg and Wold, 2002]) and the already well established sparse PLS (sPLS) ([Chun and Keles, 2007] [Chung et al., 2012] [Feraud et al., 2017]) methods are applied here, along with the innovative sparse L-sOPLS solution fully detailed in [Feraud et al., 2017]. L-sOPLS remains quite intuitive. It requires to start with the OPLS algorithm, to follow the process until the construction of a data matrix X_d (deflated by the y-orthogonal components) and then 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 ensure the best possible predictive ability (i.e. for each step, minimization of the Root Mean Square Error of Prediction (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, and is based on a number of predictive components (q) and a penalty term (λ_1) also both chosen optimally.

1D ^{1}H-NMR and 2D COSY spectra will be used and results will be compared in both cases.

4 Results and discussion

In this section, all the obtained results are discussed for the two data sets, but only some of them are shown for convenience and readability. Section 4.1 provides MIC indexes and Section 4.2 shows biomarker discovery results. In Section 4.2, optimal parameters will be also provided for each PLS, sPLS and L-sOPLS model.

During the pre-processing stage, the water zone was set to zero (typically deletion of a square area between 4.5 and 5.5 ppm) and a classical outlier detection was performed. In the second experimental design based on serum, it was also found better to delete lipoprotein zones (1.26-1.33 ppm and 0.91-0.945 ppm) as they are tending to vary according to the time which the sample spent at room temperature before NMR measurements. A log transformation was applied on all GPL matrices. The 1D ^{1}H-NMR spectra were bucketed at 0.02 ppm and were submitted to constant sum normalization (using PepsNMR [Martin et al., 2017]).

4.1 Metabolomic Informative Content results

For both experimental designs, the MIC indexes mentioned in Section 3.3 and described in [Feraud et al., 2015] were calculated on the basis of different degrees of pre-processing of the initial COSY spectra. More precisely, different log-transformed GPL were generated by tuning the ACD/Lab significance threshold

and/or the number of retained decimals of the coordinates in the 2D bucketing step. The vectorization approach (Section 3.2) was also taken into account in the process, along with ^{1}H-NMR spectra. The main results are described in Table 1.

As a reminder, the goal is to recover the three different donors in the case of the first design and the four doses combinations of threonine and glutamate in the case of the second one.

FIRST DESIGN - 3 urine donors									
Data	ACD/Lab threshold	GPL decimals	MIC indexes (Ward / K-means) Dunn	Davies-Bouldin	Rand	Adj-Rand	Inertia Btw (%)	Wth (%)	Columns in X
GPL	0.02	1	0,7659 / 0,7659	1,5571 / 1,5571	**1,00 / 1,00**	**1,00 / 1,00**	38,72	61,28	1183
GPL	0.02	2	0,7285 / 0,7285	1,8991 / 1,8991	**1,00 / 1,00**	**1,00 / 1,00**	20,52	79,48	4420
GPL	0.02	3	**0,9095 / 0,9095**	0,8618 / 0,8618	0,3732 / 0,3732	0,008 / 0,008	9,29	90,71	6886
GPL	0.05	1	**0,9146 / 0,9146**	1,3067 / 1,3067	**1,00 / 1,00**	**1,00 / 1,00**	**44,95**	**55,05**	619
GPL	0.05	2	0,8138 / 0,8138	1,704 / 1,704	**1,00 / 1,00**	**1,00 / 1,00**	23,28	76,72	2367
GPL	0.05	3	0,939 / 0,8655	0,9028 / 1,5901	0,3732 / 0,5107	0,008 / 0,0565	9,54	90,46	4080
COSY vectorization			0,7797 / 0,7627	0,571 / 0,7514	**1,00 / 1,00**	**1,00 / 1,00**	**79,15**	**20,85**	65536
Pre-processed ^{1}H-NMR			0,7088 / 0,7088	0,7518 / 0,7518	**1,00 / 1,00**	**1,00 / 1,00**	**85,22**	**14,78**	500

SECOND DESIGN - 4 doses combinations									
Data	ACD/Lab threshold	GPL decimals	MIC indexes (Ward / K-means) Dunn	Davies-Bouldin	Rand	Adj-Rand	Inertia Btw (%)	Wth (%)	Columns in X
GPL	0.02	1	0,6947 / 0,6947	1,5996 / 1,6095	0,5698 / 0,5762	0,0139 / 0,0085	22,72	77,28	330
GPL	0.02	2	0,7636 / 0,7906	1,7776 / 1,7797	0,6016 / 0,6032	0,0228 / 0,0371	19,75	80,25	1461
GPL	0.02	3	0,8825 / 0,8607	1,8657 / 1,9144	0,5857 / 0,6206	0,0153 / 0,0821	18,51	81,49	3824
GPL	0.05	1	0,6345 / 0,6345	1,7023 / 1,6758	**0,6809 / 0,6762**	0,1148 / 0,1037	24,58	75,42	196
GPL	0.05	2	0,7179 / 0,6073	1,6296 / 1,8345	0,5921 / 0,6524	0,0975 / 0,0646	20,1	79,9	769
GPL	0.05	3	0,9116 / 0,8999	1,8189 / 1,8518	0,5857 / 0,6159	0,2333 / 0,1195	18,69	81,31	2362
COSY vectorization			0,4478 / 0,337	1,7401 / 2,0394	0,5635 / 0,6365	0,0086 / 0,0241	23,38	76,62	65536
Pre-processed ^{1}H-NMR			0,3001 / 0,332	1,2367 / 1,3039	**0,7476 / 0,727**	**0,3664 / 0,3444**	24,35	75,65	476

Table 1 MIC results for the two designs, according to the choice of different pre-processing workflows (GPL and vectorization for 2D COSY or ^{1}H-NMR)

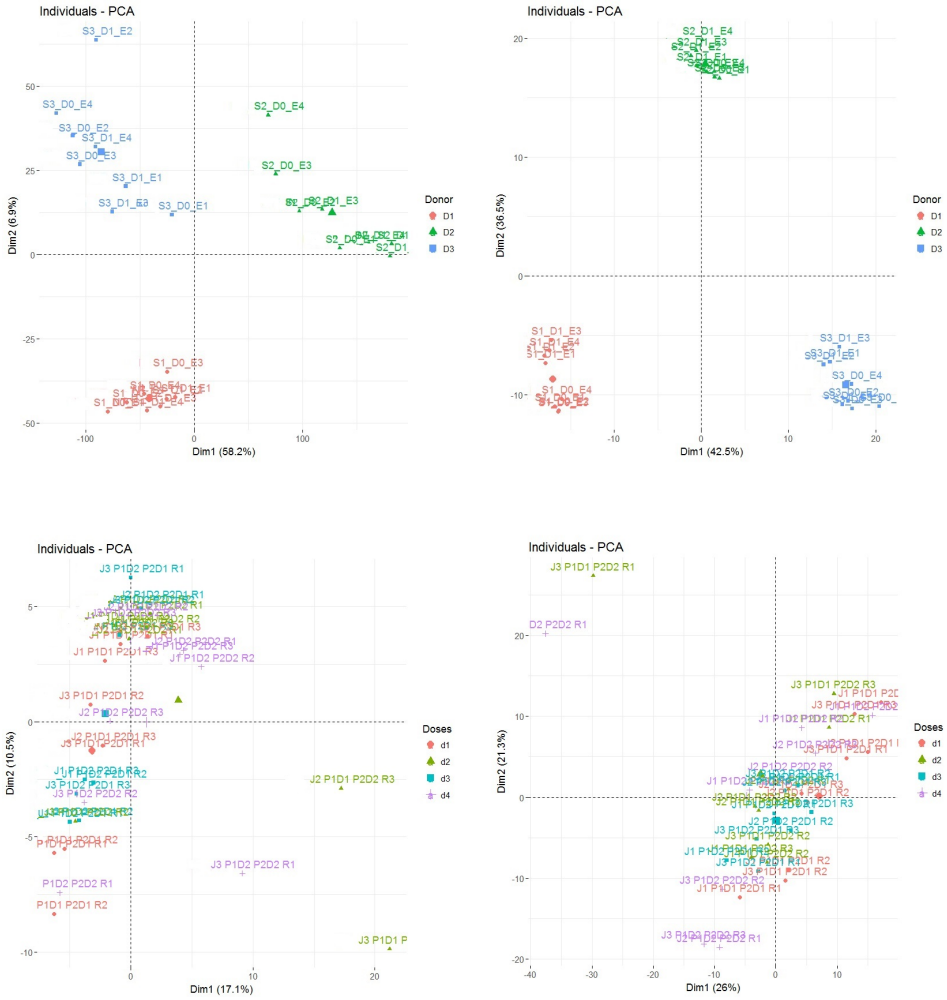

Fig. 2 First two PCA factors for the best GPL in terms of MIC results (left) and for the corresponding pre-processed ^{1}H-NMR data (right). The upper individuals PCA graphs are related to the first urine design; the lower ones to the second threonine/glutamate design. Note again that each individual in these score plots are internally defined by S=donor, D=dilution and E=day of acquisition for the first design; and by J=day, P=product, D=spiking condition and R=repetition for the second design.

For the first experimental design, the indexes tend to show that the best informative content and separability are reached by log-transformed GPL approaches involving small matrices (with one or two decimals for the coordinates). In these cases, the initial groups can be blindly retrieved without any error (Rand indexes = 1, good between inertia, etc.). It is also the case when using pre-processed ^{1}H-NMR spectra, but with lower Dunn index values. Note that inertia measures are somehow far better with 1D processed data.

In the upper part of Fig.2, the first two PCA factors are displayed for the GPL with threshold = 0.05 and number of decimals = 1, and for the pre-processed ^{1}H-NMR data. In both cases, perfect separability is reached between the urine donors, as expected by the perfect Rand indexes.

Note that the urine dilution factor does not interfere in the creation of the donors clusters.

For the second design, the results in Table 1 are far more homogeneous between the various techniques and separability between groups is harder to obtain (lower Rand and between inertia indexes for instance). However, the use of the 1D NMR seems to supply the best performances to separate the doses levels. The vectorization approach provides similar results than GPL ones (except the GPL with threshold = 0.05 and number of decimals = 1, which is again better). Remember that the benefits of vectorization allow to directly select the dimensions of the matrix on which the user wants to work and allow to bypass the use of ACD/Labs to arbitrarily select the initial significance threshold.

In the lower part of Fig.2, the first two PCA factors are displayed for this second design, illustrating a quite low discriminating power between the doses. Note that the percentage of explained variance by the first two principal components is better when using 1D data. This later PCA does not seem very informative about the groups separability but Section 4.2 demonstrates that this is not necessarily bad news in terms of relevant biomarker discovery.

These poor PCA results can be explained by two reasons. First, PCA remains a descriptive and unsupervised technique which, by definition, takes into account all the factors of the experiment (in other words, PCA is not appropriate to detect one factor precisely, here the dose levels, among other factors). PLS and relative models are on the contrary supervised and focused on a target to explain.

Secondly, this threonine/glutamate experimental design proposes conditions of low variabilities between samples, which is doubtless close to real conditions. Indeed, the spiking process described in Section 2.2.2 implied only a 1 to 1.4% change between the samples in terms of volume (7 or 10 μl among a total of 700 μl).

4.2 Biomarker discovery results

When the source of data containing most information is identified via the MIC measures, the user can have an idea on which workflow to use in order to separate individuals and to highlight biomarkers or discriminating spectral zones.

In this section, PLS, sPLS and L-sOPLS are used to detect relevant biomarkers from the best 2D COSY solution (i.e. GPL with threshold = 0.05 and number of decimal = 1) and from corresponding ^{1}H-NMR spectra. Each model is optimized in order to minimize the RMSEP via LOO cross-validation. For PLS, the number of predictive component(s) q is optimized by this way. For sPLS, the number of predictive component(s) q and the penalty term λ_1 are optimized. And, finally, for L-sOPLS, the number of orthogonal component(s) q_{ortho} is first optimized and, in a second step, the number of predictive component(s) q and the penalty term λ_1 are optimized too.

For convenience, and because the expected biomarkers to found are known, results for the second experimental design are primarily discussed in details.

In 1D, threonine appears in these following zones: 1.33 ppm, 3.59 ppm and 4.29 ppm (submitted to 0.02-0.03 ppm variations according to the pH). Glutamate appears in these following zones: 2.05 ppm (large signal), 2.34 ppm and 3.76 ppm (submitted to 0.02-0.03 ppm variations according to the pH). In 2D COSY, correlation peaks have to appear outside the diagonal. For threonine: 1.356 - 4.288 ppm (and 4.288 - 1.356 ppm) and 3.611 - 4.288 ppm (and 4.288 - 3.611 ppm). For glutamate: 2.165 - 2.487 ppm (and 2.487 - 2.165 ppm) and 2.165 - 3.803 ppm (and 3.803 - 2.165 ppm). These correlation peaks are shown in Fig.3.

Ideally, the objective is to already detect the 1D peaks using the pre-processed ^{1}H-NMR spectra (used first because of better MIC performances for this design, see Table 1).

The PLS, sPLS and L-sOPLS results are shown in Table 2 (as the associated optimal parameters). For PLS, the twenty first selected features, associated with the twenty highest coefficients in absolute values, are arbitrarily displayed. For optimized sPLS and L-sOPLS, the sparse decisions concerning these biomarkers are also shown (Yes/No to indicate if the biomarker is selected, and the corresponding final sparse coefficient if yes).

		PLS coefficients $q = 4$	sPLS selection $q = 5$, $\lambda_1 = 0.72$	L-sOPLS selection $q_{ortho} = 3$, $q = 5$, $\lambda_1 = 0.72$
Biomarkers (in ppm)	Zone	LOO-RMSEP =0.1755	LOO-RMSEP =0.1514	LOO-RMSEP =0.0759
1.349	Threonine	345.88	**Yes (352.92)**	**Yes (196.14)**
2.369	Glutamate	340.53	**Yes (647.35)**	**Yes (158.22)**
2.389	Glutamate	313.39	**Yes (578.24)**	**Yes (145.98)**
3.609	Threonine	261.96	**Yes (362.60)**	**Yes (366.93)**
3.789	Glutamate	200.84	No	No
1.329	Threonine	-180.44	No	No
4.289	Threonine	170.24	**Yes (223.17)**	**Yes (274.41)**
3.269		-168.86	Yes (-272.94)	Yes (-310.98)
1.37	Threonine	158.61	**Yes (87.14)**	**Yes (348.69)**
3.25		-156.90	No	No
2.149		140.05	No	No
3.909		-125.45	Yes (-197.68)	Yes (-348.36)
1.389		-115.01	Yes (-169.57)	Yes (-229.57)
4.69		-112.42	No	No
3.449		110.32	Yes (182.42)	Yes (127.44)
0.929		-107.69	No	No
3.289		-92.20	No	No
3.49		-91.15	Yes (-64.97)	No
2.089	Glutamate	89.46	No	No
0.949		-89.13	No	No

Table 2 PLS, sPLS and L-sOPLS biomarker selection for ^{1}H-NMR data (threonine-glutamate design).

One can see that the threonine and glutamate 1D peaks are well retrieved by the PLS and the sparse algorithms (in bold for these last ones) even if the MIC and PCA results don't provide a smart separation between the groups of spectra. Furthermore, the use of sparse techniques, and in particular L-sOPLS, leads to a

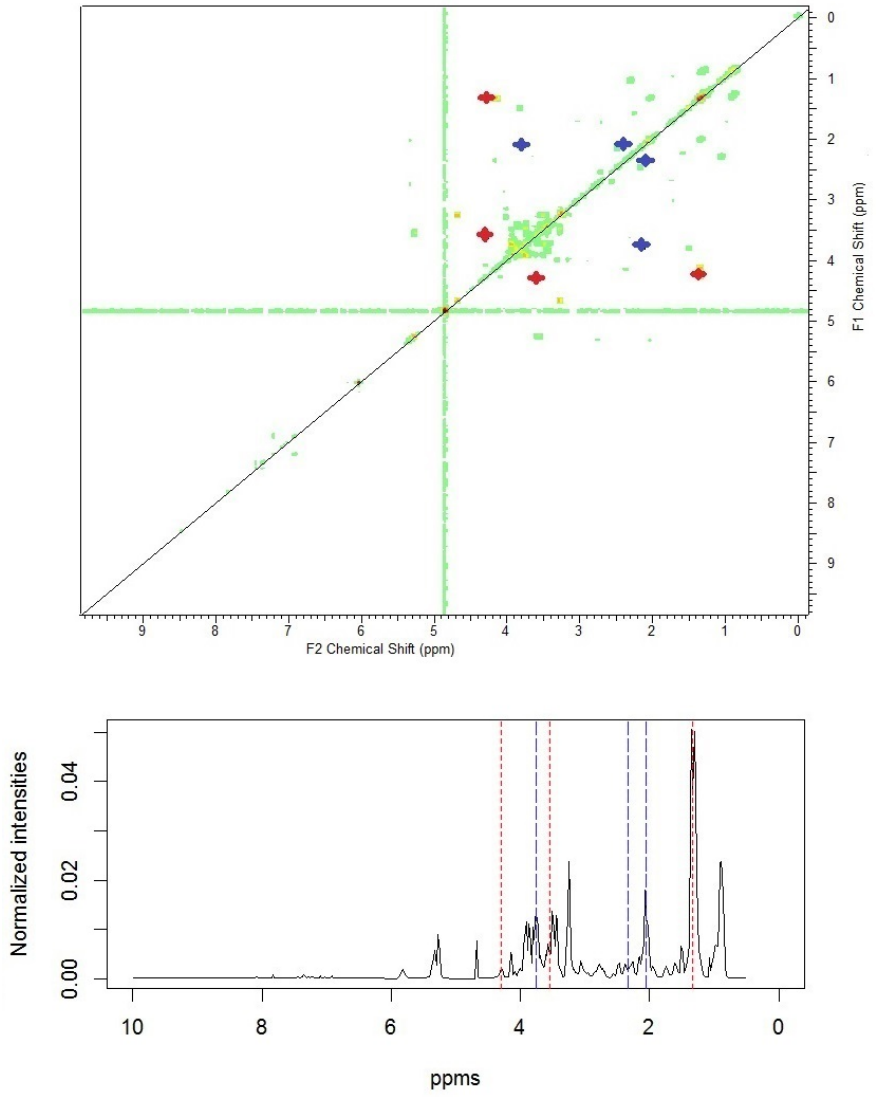

Fig. 3 Threonine (in red) and glutamate (in blue) 2D correlation peaks in a raw COSY spectrum and corresponding peaks in a 1D NMR spectrum (red dashes for threonine and blue long dashes for glutamate).

lower value of the RMSEP of the model, thus significantly improving the predictive power (from 0.1755 to 0.0759).

But one can also see that some artefacts or other ppm zones are selected, often with high coefficients (and importance). For example, the peaks in 3.909 and 1.389 ppm are selected and can be easily confused with some nearby threonine and/or glutamate peaks. It can refer to a contiguous peak zones problem which may require the addition of a deeper warping step. This confirms the interest to

apply PLS, sPLS and L-sOPLS on 2D data in order to ideally highlight relevant correlation peaks outside the diagonal. These peaks would then confirm and reinforce the first 1D results. The search for non-diagonal biomarkers based on the best 2D COSY solution (i.e. GPL with threshold = 0.05 and number of decimal = 1) is summarized in Table 3. Even if all the information was taken into account, biomarkers located on the diagonal are not displayed, being redundant with the biomarkers discovered in 1D. For PLS, the non-diagonal peaks whose coefficients are among the thirty higher global coefficients in absolute values are displayed.

		PLS coefficients $q = 5$	sPLS selection $q = 4$, $\lambda_1 = 0.78$	L-sOPLS selection $q_{ortho} = 4$, $q = 5$, $\lambda_1 = 0.60$
Biomarkers (in ppm)	Zone	LOO-RMSEP =0.6393	LOO-RMSEP =0.4872	LOO-RMSEP =0.1912
4.3 - 3.6	Threonine	0.146	**Yes (0.2089)**	**Yes (0.0994)**
3.6 - 4.3	Threonine	0.142	**Yes (0.1884)**	**Yes (0.0942)**
2.1 - 2.4	Glutamate	0.0582	**Yes (0.0933)**	**Yes (0.0289)**
4.3 - 1.4	Threonine	0.0544	**Yes (0.0884)**	**Yes (0.0555)**
3.8 - 2.1	Glutamate	0.0495	**Yes (0.0567)**	**Yes (0.0308)**
4.3 - 1.3	Threonine	0.0492	**Yes (0.0692)**	**Yes (0.0514)**
2.4 - 2.1	Glutamate	0.0466	**Yes (0.0367)**	**Yes (0.0324)**
3.8 - 2.2	Glutamate	-0.0288	No	No
1.4 - 4.2	Threonine	0.0269	No	**Yes (0.0119)**
3.3 - 4.0		-0.0252	No	No
4.4 - 4.9		-0.0244	No	No
3.0 - 1.7		-0.0198	No	No
3.8 - 4.9		-0.0182	No	No

Table 3 PLS, sPLS and L-sOPLS non-diagonal biomarker selection applied on the best GPL solution in terms of MIC measures (threonine-glutamate design).

One can see in Table 3 that the PLS-DA technique can already detect the threonine and glutamate correlation peaks in the 2D COSY spectral data, with higher coefficients in absolute values. But the corresponding predictive power, based on LOO-RMSEP criteria, is very poor. It is very important to observe that sparse sPLS and L-sOPLS only detect threonine and glutamate correlation peaks (and nothing else outside the diagonal). This result is very comforting, especially as the increase in predictive quality is very significant (from 0.6393 with PLS to 0.1912 with L-sOPLS).

Of course, note that these latter values of LOO-RMSEP are not better than the previous 1D ones, which is consistent with the initial MIC measures (better separability can be reached when using pre-processed 1D data here).

Concerning the first data design, the process would be a little bit different, and more or less inverted. As MIC indexes tend to promote the use of some 2D GPL as a priority (see Table 1), PLS, sPLS and L-sOPLS have to be applied on these data source first. So, relevant 2D correlation peaks can be quickly found in a blind and non-supervised way as biomarkers.

The discovery of such spectral correlation peaks, which perfectly discriminate the three initial urine donors, can be directly put in connection with recent studies. For instance in [Thevenot et al., 2015] and [Rist et al., 2017], the effects of age, gender or Body Mass Index (BMI) as sources of variation on the

human metabolome are proved. Moreover, some particular peaks are highlighted as biomarker zones, in 1D and 2D (onto the diagonal), and contribute to discrimine the three donors: principally Trimethylamine N-oxide (TMAO), urea and acetaminophen (paracetamol).

Finally, the application of sparse algorithms on ^{1}H-NMR data would only serve as confirmation studies.

Note that the same conclusions are observed in terms of predictive power: L-sOPLS always allows better results.

5 Conclusions

In this article, two pre-processing workflows are presented to handle 2D COSY spectral data issued from an experimental design and to summarize them into a single and global object. The Global Peak List (GPL) workflow, especially when considering low resolutions and a log transformation, performs very well in terms of metabolomic content and tends to allow a high separability power between existing groups in the data. The vectorization approach also has advantages: it may seem more intuitive for users, with no external significance threshold to chose and final resulting matrices whose dimensions are strictly known in advance.

On the basis of two different data designs (masked group donors and controlled threonine/glutamate doses), GPL matrices, vectorization objects and preprocessed 1D ^{1}H-NMR data were compared in terms of MIC (Metabolomic Informative Content), involving unsupervised clustering and inertia quality measures, in order to visualize which data source is allowing the best separability. For the first design, some GPL matrices obtained the best performances. And for the second design, the 1D data source was slightly the best.

In each case, once a data source seemed better than the others, this source was chosen as priority input for the biomarkers discovery algorithms. In this paper, PLS-DA, sPLS and L-sOPLS were then applied. Because there is no methodology or workflow that can be the best every time, the biomarkers discovery step should ideally be applied on both 1D and 2D data sources to highlight significant 1D peaks or 2D correlations. These two subsequent data sources can then be used for confirmation or reinforcing purposes (1D peaks which can confirm the previous highlighting of 2D correlation peaks; or 2D correlation peaks which can reinforce the previous highlighting of 1D peaks and corresponding metabolites).

The conclusion is then definitively focused on the complementarity between the different data sources that can be available during an experience. The use of different worflows or data sources, coupled with the use of different algorithms, can lead to complementary and/or confirmatory results.

It is also very important to enhance that the second example, with the threonine-glutamate experimental design, demonstrated that 2D COSY spectra allow to discover the relevant molecules' fingerprints without any equivocation via the non-diagonal biomarkers (Table 3), even if the initial MIC and PCA results were not very optimistic (Table 1 and Fig.2).

For the biomarker research part of this article, it is demonstrated that sparse derivatives of the PLS model provide very good performances in terms of biomarker identification and predictive power. By selecting a (very) small number of (very) relevant features as biomarkers, by providing, consequently, lighter and more interpretable cross-validated optimal models to practitionners, and by offering very low RMSEP values, the L-sOPLS seems to be a very interesting and promising tool. The conclusions observed in [Feraud et al., 2017] are now confirmed on 2D COSY data here.

Acknowledgements Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is gratefully acknowledged. Support from the CORSAIRE metabolomics platform (Biogenouest network) is also acknowledged. Pascal de Tullio is Research Director of the Fonds de la Recherche Scientifique (FNRS).

Author Contribution Statement BF, BG, PG and PT conceived and designed research. EM, JL and PT collected and supplied the data. BF analyzed data and wrote the manuscript. All authors read and approved the manuscript.

Conflict of Interest All authors declare that they have no conflict of interest.

Compliance with ethical requirements This study analyzes collected data which involved human participants.

Softwares availability statement The raw data were processed with the Bruker Topspin 3.5 software. Peak lists were extracted using ACD/Labs 12.00 (ACD/NMR processor). The R software (`http://www.R-project.org`) environment was exclusively used for statistical purpose, via existing packages (*pls*, *spls*, *ropls*), or coded ad hoc (PepsNMR package; MIC indexes, L-sOPLS, functions which are available here: `https://github.com/ManonMartin/MBXUCL`.).

Data availability statement The metabolomics and metadata reported in this paper are available on demand from the Institute of Statistics, Biostatistics and Actuarial Sciences, UCLouvain, Belgium.

References

[Bylesjo et al., 2006] Bylesjo M., Rantalainen M., Cloarec O., Nicholson J., OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification, Journal of Chemometrics, 20(8-10), 341-351 (2006).

[Chun and Keles, 2007] Chun H., Keles S., Sparse Partial Least Squares Regression with an Application to Genome Scale Transcription Factor Analysis. Department of Statistics, University of Wisconsin, Madison (2007).

[Chung et al., 2012] Chung D., Chun H., Keles S, Spls: Sparse partial least squares (SPLS) regression and classification. R package, version, 2, 1-1 (2012).

[Craig et al., 2006] Craig A., Cloarec O., Holmes E., Nicholson J. K. and Lindon J. C., Scaling and normalization effects in NMR spectroscopic metabonomic data sets. Analytical chemistry, 78(7), 2262-2267 (2006).

[Efron et al., 2004] Efron B., Hastie T., Johnstone I., Tibshirani R., Least Angle Regression, Annals of Statistics, 32(2), 407499 (2004).

[Feraud et al., 2015] Feraud B., Govaerts B., Verleysen M., De Tullio P., Statistical treatment of 2D NMR COSY spectra in metabolomics: data preparation, clustering-based evaluation of the Metabolomic Informative Content and comparison with ^{1}H-NMR, Metabolomics, 11(6), 1756-1768 (2015).

[Feraud et al., 2017] Feraud B., Munaut C., Martin M., Verleysen M., Govaerts B., Combining strong sparsity and competitive predictive power with the L-sOPLS approach for biomarker discovery in metabolomics. Metabolomics, 13(11), 130 (2017).

[Friedman et al., 2010] Friedman J., Hastie T., Tibshirani R., A note on the group lasso and a sparse group lasso, arXiv preprint arXiv:1001.0736 (2010).

[Giraudeau, 2014] Giraudeau, P., Quantitative 2D liquid-state NMR. Magnetic Resonance in Chemistry, 52(6), 259-272 (2014).

[Giraudeau et al., 2014] Giraudeau, P., Tea, I., Remaud, G. S., and Akoka, S., Reference and normalization methods: essential tools for the intercomparison of NMR spectra. Journal of pharmaceutical and biomedical analysis, 93, 3-16 (2014).

[Jezequel et al., 2015] Jezequel, T., Deborde, C., Maucourt, M., Zhendre, V., Moing, A., and Giraudeau, P., Absolute quantification of metabolites in tomato fruit extracts by fast 2D NMR, Metabolomics, 11(5), 1231-1242 (2015).

[Le Guennec et al., 2014] Le Guennec, A., Giraudeau, P., and Caldarelli, S., Evaluation of fast 2D NMR for metabolomics. Analytical chemistry, 86(12), 5946-5954 (2014).

[Le Guennec et al., 2012] Le Guennec, A., Tea, I., Antheaume, I., Martineau, E., Charrier, B., Pathan, M., ... and Giraudeau, P., Fast determination of absolute metabolite concentrations by spatially encoded 2D NMR: application to breast cancer cell extracts. Analytical chemistry, 84(24), 10831-10837 (2012).

[Liland, 2011] Liland K. H., Multivariate methods in metabolomics, from pre-processing to dimension reduction and statistical analysis. TrAC Trends in Analytical Chemistry, 30(6), 827841 (2011).

[Marchand et al., 2017] Marchand, J., Martineau, E., Guitton, Y., Dervilly-Pinel, G., and Giraudeau, P., Multidimensional NMR approaches towards highly resolved, sensitive and high-throughput quantitative metabolomics. Current opinion in biotechnology, 43, 49-55 (2017).

[Marchand et al., 2018] Marchand, J., Martineau, E., Guitton, Y., Le Bizec, B., Dervilly-Pinel, G., and Giraudeau, P., A multidimensional ^{1}H-NMR lipidomics workflow to address chemical food safety issues. Metabolomics, 14(5), 60 (2018).

[Marjanska et al., 2008] Marjanska M., Henry P. G., Ugurbil K., and Gruetter R., Editing through multiple bonds: Threonine detection. Magnetic Resonance in Medicine, 59(2), 245-251 (2008).

[Martin et al., 2017] Martin M., Legat B., Leenders J., Vanwinsberghe J., Rousseau R., et. al., PepsNMR for the 1H-NMR metabolomic data pre-processing. ISBA Discussion Paper, 2017/22, http://hdl.handle.net/2078.1/187159 (2017).

[Martineau et al., 2012] Martineau, E., Tea, I., Akoka, S., and Giraudeau, P., Absolute quantification of metabolites in breast cancer cell extracts by quantitative 2D 1H INADEQUATE NMR. NMR in Biomedicine, 25(8), 985-992 (2012).

[McQueen, 1967] MacQueen J. B., Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1, University of California Press, pp. 281-297 (1967).

[Murtagh and Legendre, 2011] Murtagh F., Legendre P., Ward's hierarchical clustering method: clustering criterion and agglomerative algorithm, arXiv preprint arXiv:1111.6285 (2011).

[Ravanbakhsh et al., 2014] Ravanbakhsh S., Liu P., Bjorndahl T., Mandal R., Grant J. R., Wilson M., ... and Greiner R., Accurate, fully-automated NMR spectral profiling for metabolomics. arXiv preprint arXiv:1409.1456 (2014).

[Rist et al., 2017] Rist, M. J., Roth, A., Frommherz, L., Weinert, C. H., Kruger, R., Merz, B., ... and Gorling, B., Metabolite patterns predicting sex and age in participants of the Karlsruhe Metabolomics and Nutrition (KarMeN) study. PloS one, 12(8), e0183228 (2017).

[Rouger et al., 2017] Rouger, L., Gouilleux, B., and Giraudeau, P., Fast n-dimensional data acquisition methods. Encyclopedia of Spectroscopy and Spectrometry, third ed., Academic Press, Oxford, 588-596 (2017).

[Rousseau, 2011] Rousseau R., Statistical contribution to the analysis of metabonomic data in ^{1}H-NMR spectroscopy, PhD Thesis, UCL, http://hdl.handle.net/2078.1/75532 (2011).

[Sousa et al., 2013] Sousa S.A., Magalhaes A., Castro Ferreira M.M., Optimized bucketing for NMR spectra: Three case studies, Chemometrics and Intelligent Laboratory Systems, 122, pp. 93-102 (2013).

[Thevenot et al., 2015] Thevenot, E. A., Roux, A., Xu, Y., Ezan, E., and Junot, C., Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. Journal of proteome research, 14(8), 3322-3335 (2015).

[Trygg and Wold, 2002] Trygg J., Wold S., Orthogonal Projections to Latent Structures (O-PLS), Journal of Chemometrics, 16(3), 119-128 (2002).

[Ward, 1963] Ward J.H., Hierarchical Grouping to optimize an objective function, Journal of American Statistical Association, 58(301), pp.236-244 (1963).

[Wold, Trygg et al., 2001] Wold S., Trygg J., Berglund A., Antti H., Some recent developments in PLS modeling, Chemometrics and Intelligent Laboratory Systems, 58(2), 131-150 (2001).

[Wold et al., 2001] Wold S., Sjostrom M., Eriksson L., PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems, 58(2), 109-130 (2001).