Statistical contributions to the analysis of 2D NMR spectra in metabolomics studies :

From pre-processing workflows to 2D biomarker discovery

A thesis submitted in partial fulfillment of the requirements for the degree of Docteur en Sciences, Université catholique de Louvain, January 2019

Baptiste Feraud

Thesis committee:

Prof. Bernadette Govaerts, Supervisor (LIDAM, UCLouvain)
Prof. Michel Verleysen, Supervisor (ICTM, UCLouvain)
Prof. Johan Segers, President (LIDAM, UCLouvain)
Prof. Rainer von Sachs, Secretary (LIDAM, UCLouvain)
Prof. Pascal de Tullio (Université de Liège)
Dr. Philippe Bastien (L'Oréal, France)

“Les statistiques sont vraies quant à la maladie et fausses quant au malade ; elles sont vraies quant aux populations et fausses quant à l’individu.”

Léon Schwartzenberg

Acknowledgements

Avant tout, je souhaite évidemment remercier Bernadette. Pour beaucoup de choses, mais peut-être en premier pour ta patience. Ce fut un plaisir d'être ton assistant pour les cours de psycho tout au long de ces années, avec ces moments de satisfaction ou de stress et ces autres moments de surprise (les étudiant(e)s nous surprendront toujours). Ce fut également un plaisir de travailler avec toi pour cette thèse et les papiers qui la composent, même si les passages difficiles ou de découragement ont été bien là. Je te remercie aussi pour l'autonomie que tu as su me donner et pour tes innombrables relectures et discussions. MERCI à toi.

Je remercie Michel pour son soutien et ses conseils toujours avisés. Au début de ce processus de thèse, bien souvent, quelques-unes de tes phrases pouvaient débloquer bien des situations et des raisonnements. Ce fut un réel plaisir de travailler et d'échanger avec toi. Je remercie Rainer pour son soutien également, et sa relecture cruciale à un moment où je pataugeais quelque peu. Je n'oublierai pas.

Merci aussi à Pascal et à Justine pour nous avoir fourni inlassablement la plupart des données utilisées dans cette thèse. Vous avez été notre patient pourvoyeur de matières premières !

Merci à Patrick pour nos échanges spectroscopiques lors de nos rencontres lors de conférences et pour les données envoyées. Merci à Philippe pour d'autres échanges, autour de la notion de sparsité cette fois-ci, et pour avoir accepté

d'être dans mon jury.

Je remercie Johan. J'ai apprécié d'avoir été ton assistant pour le cours d'analyse des données pendant toutes ces années.

Merci aussi à tous les assistants et chercheurs que j'ai croisé ou côtoyé depuis le début de mon contrat à l'institut (la liste est trop longue pour tous vous citer). Avec une mention très spéciale aux footeux qui se reconnaîtront :) Mention aussi à l'équipe des psychologues, aux jeunes mamans et aux compagnons de galères et de journée de 8h de TP en économétrie :)

Je remercie également les membres administratifs et IT pour leur travail et leurs réponses aux questions parfois les plus basiques.

Une pensée pour Rudolf.

Enfin, et par dessus tout, je remercie Cécile, avec moi depuis longtemps déjà. Pour son amour. Merci à mes parents, qui sont loin, mais qui sont là en toutes circonstances. Je leur dois évidemment beaucoup beaucoup. Et merci à mes loulous, mes merveilleux petits garçons, Louis et Roman, qui sont juste mes deux soleils lumineux.

À Gé, pour ton rôle infini dans la dernière ligne droite ...

Contents

INTRODUCTION **13**

1 THE ANALYSIS OF 2D NMR SPECTRA IN METABOLOMICS STUDIES: from pre-processing workflows to 2D biomarker discovery **21**

1.1 NMR spectroscopy: the 1D vs. 2D spectral data sources 21

1.1.1 The 1D ^{1}H-NMR spectroscopy: basics 22

1.1.2 The 1D ^{1}H-NMR limitations for metabolomics studies . 27

1.1.3 The 2D NMR alternatives and the COSY experiment . 30

1.2 Handling of 1D and 2D spectral data and pre-processing workflows for 2D COSY spectra 34

1.2.1 Typical pre-processing steps for ^{1}H-NMR 36

1.2.2 Pre-processing workflows for 2D spectral data 40

1.2.2.1 The Global Peak List (GPL) workflow 41

1.2.2.2 The Vectorization workflow 45

1.2.2.3 Necessary pre-processing for 2D COSY spectra 46

1.3 Evaluation of the quality of the data sources 47

1.3.1 Outlier detection 49

1.3.2 The Metabolomic Informative Content (MIC) 52

1.3.2.1 The choice of a between samples distance measure 53

1.3.2.2 The choice of a clustering algorithm 53

1.3.2.3 The clustering quality measures gathered in the MIC concept 54
1.3.3 The potential use of ASCA and APCA 57
1.3.4 The potential use of ICA 59
1.4 The relevant biomarker discovery issue in metabolomics 61
1.4.1 The use of PCA 62
1.4.2 The PLS(-DA) approach 63
1.4.2.1 PLS modelling 64
1.4.2.2 Highlighting important descriptors as biomarkers with PLS 67
1.4.3 The OPLS approach 69
1.4.4 Introducing the notion of sparsity: the sparse-PLS (sPLS) approach 71
1.4.4.1 The use of penalty terms to obtain sparse solutions: a general overview 72
1.4.4.2 The sPLS algorithm 75
1.4.5 The L-sOPLS algorithm for combining OPLS and sparse results 77
1.4.6 Other sparse methods 80
1.4.6.1 jForest: a Random Forest approach involving importance measures for feature selection . . . 80
1.4.6.2 The Xgboost alternative (as tree ensemble boosting) 83
1.5 Overview of the used datasets 86
1.6 Results and discussions 87
1.7 Appendix of Chapter 1: comparisons of MIC quality measures when considering different dimensionalities for the descriptors . 106

2 PAPER I : Statistical treatment of 2D NMR COSY spectra in metabolomics 111
Abstract 112
2.1 Introduction 113
2.2 Materials: COSY spectra, experimental designs and acquisition parameters 115
2.2.1 Experimental designs 115
2.2.1.1 First design: cell culture media 115
2.2.1.2 Second design: human serum 116

2.2.2 Spectra acquisition 117
2.3 Methods: global peak list, pre-processing steps and clustering analysis 118
2.3.1 Global Peak List matrix 118
2.3.2 Pre-processing steps 119
2.3.3 Control of data resolution 120
2.3.4 Clustering algorithms and distance measures 121
2.3.5 Numerical indexes to evaluate the quality of the clustering results 123
2.4 Results and discussion 126
2.4.1 Results for COSY spectra 126
2.4.2 Comparisons with corresponding ^{1}H-NMR spectra . . . 129
2.4.2.1 Upgrade of ^{1}H-NMR spectra 129
2.4.2.2 Comparisons 130
2.5 Conclusion and further works 130

Local bibliography 132

3 PAPER II : Combining strong sparsity and competitive predictive power with the L-sOPLS approach for biomarker discovery in metabolomics 137
Abstract 138
3.1 Introduction 139
3.2 Materials: data sets and experimental protocols 140
3.2.1 Spectral ^{1}H-NMR data set based on urine 141
3.2.1.1 Description and motivations 141
3.2.1.2 Acquisition 142
3.2.2 RT-qPCR genomic data set 143
3.2.2.1 Description and motivations 143
3.2.2.2 Acquisition 143
3.3 Methods: (O)PLS and corresponding sparse solutions: sPLS and L-sOPLS 144
3.3.1 The PLS(-DA) regression model: some reminders 144
3.3.2 The OPLS(-DA) algorithm 147
3.3.3 The sparsity issue and the Sparse-PLS (sPLS) solution . 149
3.3.4 Sparsity with OPLS: the new "Light-Sparse-OPLS" (L-sOPLS) solution 152

3.4 Results and discussion 153
3.4.1 Results for the ^{1}H-NMR urine spectral data set 154
3.4.2 Results for genomic RT-qPCR data 158
3.5 Conclusion and further works 161

Local bibliography **163**

4 PAPER III : Two data pre-processing workflows to facilitate the discovery of biomarkers by 2D NMR metabolomics **167**
Abstract 167
4.1 Introduction 168
4.2 Materials: data sets and experimental protocols 171
4.2.1 First experimental design: urine donors 171
4.2.1.1 Description and motivations 171
4.2.1.2 Urine collection 171
4.2.2 Second serum based experimental design: spiked products 172
4.2.2.1 Description and motivations 172
4.2.2.2 Blood collection and spiking 172
4.2.3 NMR spectroscopy 173
4.3 Methods 174
4.3.1 The Global Peak List (GPL) approach 174
4.3.2 The vectorization approach 175
4.3.3 The assessment of the Metabolomic Informative Content 177
4.3.4 Biomarker identification 178
4.4 Results and discussion 179
4.4.1 Metabolomic Informative Content results 180
4.4.2 Biomarker discovery results 183
4.5 Conclusions 188

Local bibliography **189**

5 Contributions of faster 2D COSY acquisition experiments to metabolomics: the Non-Uniform Sampling (NUS) and the single scan COSY (NS1) **193**
5.1 Motivations 193
5.2 Introduction to the Non-Uniform Sampling (NUS) method . . . 195
5.3 Data acquisition parameters for the COSY NS1 and COSY NUS50 data sources 197

5.4 Methods: application of the 2D workflows (GPL and Vectorization) and of the MIC on the faster COSY data sources 198
5.5 Results and discussion 199
5.5.1 Descriptive PCA results 200
5.5.2 MIC results 202
5.6 Conclusions 204

Local bibliography **205**

CONCLUSIONS **207**

GENERAL BIBLIOGRAPHY **213**

INTRODUCTION

Metabolomics ([Fiehn, 2002], [Weckwerth, 2003], [Krastanov, 2010], [Patti et al., 2012], [Sevin et al., 2015] among others) mainly refers to studies of endogenous metabolites observations, characterizations or changes in various biological states or media. This quite new science offers very promising tools in many healthcare fields as it provides knowledge about metabolites fluctuations usable to detect and understand biological reactions of the organism in a fast, reasonable cost and generally low invasive way. For instance, physicians need to be able to discover if an individual has developed a pathological reaction; the pharmaceutical industry needs to evaluate if toxicological reactions are developed after the administration of a new drug candidate. Until now, such biological reactions of an organism are usually examined through the occurrence of external endpoints.

For example, the development of a disease can be observed through the examination of fever. But these external biological endpoints are already there and usually follow earlier changes in metabolites. Actually, any complex organism is formed by several interlinked levels of biomolecular organization (genes, proteins, metabolites). Throughout life, when an organism meets stressors, the equilibrium of the different biomolecular levels of this organism is affected with different time maturities (i.e. with more or less fast damages). If these disturbances are of sufficient magnitude, they cannot be controlled. This established molecular disequilibrium can lead to perturbations of the health working of the whole organism, only then translated by external endpoints.

A metabolomic study aims to identify altered metabolites when biological reactions occur in the organism. Based on this knowledge, the examination of biological reactions can be linked with measurements of changes of endogenous metabolites, what may represent a precious gain of time for patients or sick people. Besides medical or pharmaceutical issues, metabolomics can also be of critical interest for quality control or risk measures (food, agro-food industry, toxicology, newborn screening, clinical chemistry, food and medicinal plants quality control, environmental studies, etc...).

Among all the recent technical developments, metabolomics has also emerged as a very challenging approach for drug discovery and personalized medicine. Particularly, personalized medicine is expanding rapidly. For instance, the links between metabolic changes, patient phenotype, physiological and/or pathological status, and treatment are now well established and have thus opened a new area for an extended application of metabolomics in the whole drug discovery process and, consequently, in personalized medicine ([Frederich et al., 2016], [Baraldi et al., 2009], [Hunter, 2009] among others).

More generally, metabolomics belongs to a new family of biological fields, the "Omics", which studies the biological events through different biomolecular levels of an organism. The Omics sciences (metabolomics, genomics, proteomics, transcriptomics...) recently appeared consequently to progress in analytical techniques realized since the 1990's and aim to deeply understand biological systems. Omics are formed by several technological platforms using multiparametric biochemical information derived from the different levels of biomolecular organization (genes, proteins or metabolites).

All of these Omics sciences rely on analytical multi-technology acquisition methods (e.g. spectroscopy and spectrometry in metabolomics), generating large and complex multivariate datasets which require the deployment of statistical, chemometrical and bioinformatical tools ([Spicer et al., 2017], [Barupal et al., 2018]) to be handled and interpretated. A very general view of the Omics philosophy is shown in Fig.1.

Consequently, their high complexity has boosted new methodological developments in related scientific disciplines such as in chemometrics. In fact, most Omics experiments provide complex multivariate datasets typically characterized by a **high dimensionality** (in this thesis, the term of "high dimensionality" always implies a number of correlated variables m much greater than the number of available samples n).

Moreover, to study the effect of several sources of variation on biological systems, omics studies often need and follow multifactorial experimental designs.

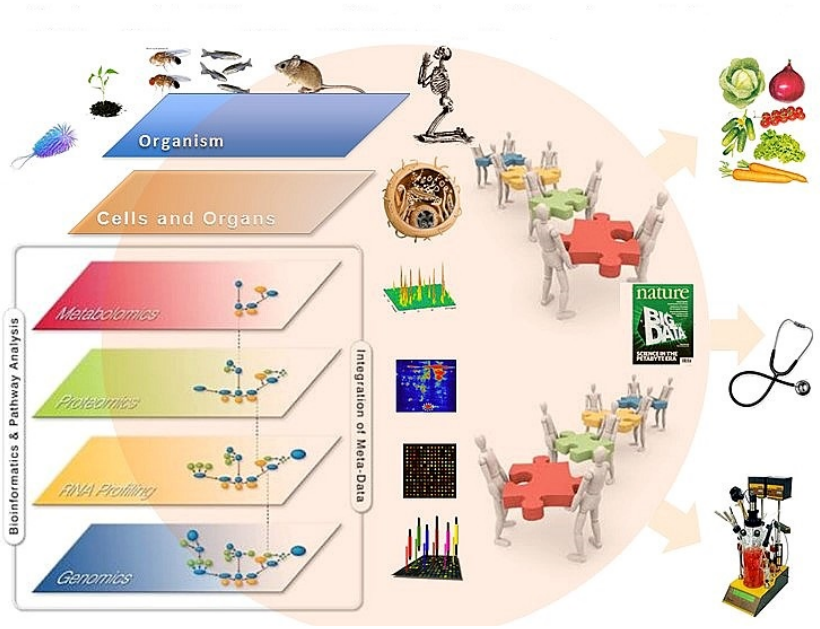


Figure 1: The Omics perspectives (from `https://www.slideshare.net/SN-CM/big-data-14647516`).

In this context, the first crucial experimental step is to define a controlled objective to understand and to explain. Practically, this thesis will deal with datasets which contain **different groups of subjects** (Fig.2). For instance, if a disease is of interest, it is then crucial to work with samples of ill and safe subjects, or suffering from different levels of the disease (the cases involving more than two groups can also be taken into account). Basically, these groups will refer to the main signal, or the valuable information, in the experiments presented in this thesis.

This notion of groups is a key notion as all the subsequent work will focus on the discriminant information contained within each of the available data sources (via the study of the level of the initial informative content of these sources and clustering-based diagnostic criteria) and on the reasons of these discriminations (thus leading to feature selection, biomarker discovery ([San Cristobal et al., 2013]) and prediction purposes).

In this work, **"data sources"** refers to all kinds of data, or data sets, potentially in use during a particular metabolomics study, considering different

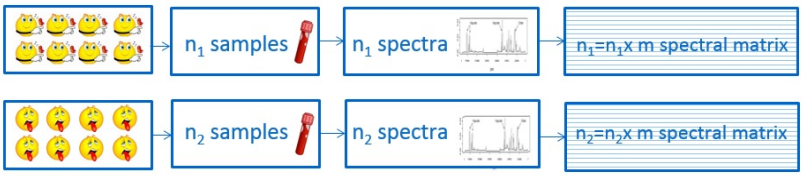


Figure 2: Illustration of a metabolomics study with two groups (ill and healthy people).

acquisition techniques, different acquisition parameters and/or the use of different pre-processing tools and workflows.

Concerning the initial data acquisition, two main techniques, linked with different scientific domains, exist to feed the metabolomics studies and experiments specifically: **NMR** (Nuclear Magnetic Resonance spectroscopy, [Beckonert et al., 2007], [Leenders et al., 2015]) and MS (Mass Spectrometry, [Dettmer et al., 2007], [Courant et al., 2014]).

Mass Spectrometry is supposed to be more sensitive than 1D ^{1}H-NMR (proton NMR spectroscopy) but requires a pre-separation of the metabolic components using either gas chromatography (GC) after chemical derivatisation, liquid chromatography (LC), or methods of ultra-high-pressure LC (UPLC). ^{1}H-NMR has advantages in terms of minimal sampling processing, quantitative calibration and minimal invasiveness [Lindon et al., 2003]. Note that (^{1})H-NMR is initially more devoted to untargeted metabolomics, since it allows in theory to detect all the molecules present in the sample of interest (several hundred and quantification of several dozens), whereas MS is initially more devoted to targeted metabolomics.

Identifying and dosing metabolites is very important in metabolomics, that is why NMR has clearly an important role to play, even if the future is doubtless in the integration of the data.

The NMR and MS data acquisition techniques can be numerous. Within each of these techniques, the variants can be also very numerous and the spectrometers can be more or less powerful or recent. Parameters of acquisition can be tuned, external conditions may differ. All this leads to a very important quantity of potential information (i.e. data sources) to treat.

For NMR experiments, biological samples (e.g. blood, urine, serum, tissue, ...) are typically collected from subjects presenting different states of a reaction of interest. For example, samples can be collected from people or animals with and

without the contact with a pathological agent. Only the first ones are assumed to have biologically reacted by the disease. The concentration of metabolites in biofluid samples is supposed to be altered according to the presence or the degree of the biological reaction of the subject at the sample time. Analytical technology is then used to reveal the composition of metabolites in the samples. ^{1}H-NMR techniques generate spectral profiles describing the structure and the concentrations of metabolites contained in collected biofluid samples.

NMR is the name given to a technique that exploits the magnetic properties of certain nuclei: in a magnetic field, NMR active nuclei absorb energy at a characteristic frequency of the isotope and radiate this energy back out. ^{1}H-NMR spectroscopy is the application of nuclear magnetic resonance with respect to Proton, in order to determine from the irradiated energy the structure and concentration of its molecules [Silverstein et al., 2014]. Note that Carbon NMR, or ^{13}C-NMR, is also well known but generally less sensitive and not sensitive enough for metabolomics studies.

Fig.3 presents a general scheme of a metabolomics study leading to the discovery of endogenous modified metabolites linked with a specific biological reaction.

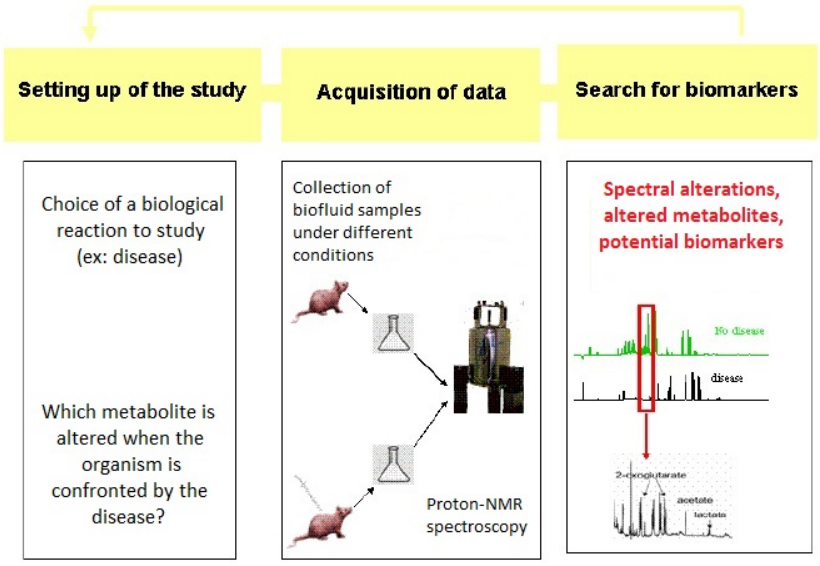


Figure 3: The general scheme of a metabolomics study (from [Rousseau, 2011]).

In this thesis, the emphasis is mainly on two-dimensional NMR techniques, and in particular on 2D COSY (COrrelation SpectroscopY). The well-known and widely used ^{1}H-NMR is also considered throughout this work as a robust way of comparison (see Section 1.1). COSY is called "homonuclear" and is a method which gives data plotted in a space defined by two frequency axes rather than one. Two-dimensional NMR spectra are supposed to provide more information about molecules than one-dimensional NMR spectra and are especially useful in determining the structure of a molecule, particularly for molecules that are too complicated to visualize when using one-dimensional NMR. Specifically, 2D spectra, as COSY, allow to detect more easily multi-peaks molecules, whose peaks can overlap with other molecules' peaks in 1D representations (see ^{1}H-NMR limitations in Section 1.1.2).

The first two-dimensional experiment, COSY, was proposed by Jean Jeener in 1971. This experiment was later implemented in [Aue et al., 1976] and [Martin and Zekter, 1988]. In Section 1.1, a general comparison between ^{1}H-NMR and COSY spectra is provided and the use of 2D data will be motivated. The COSY data acquisition process is also fully detailed in Section 1.1.3.

Once all the available data sources, as defined previously, are obtained for a particular study, different pre-processing methodological steps can be applied or not. The choice to apply pre-processing may increase again the amount of available information. The main pre-processing steps are detailed in Section 1.2. In this section, two innovative workflows about the handling and the statistical treatment of 2D spectral objects are detailed specifically. Actually, the use and the main pre-processing steps applied on 1D spectra in metabolomics studies are well known and well mastered ([Martin et al., 2017], [Xia and Wishart, 2016] for very recent examples), but it is not totally the case for 2D ones, and in particular for COSY spectra (but it is universally the same philosophy for any other 2D data sources). **By introducing the Global Peak List (GPL) concept (Section 1.2.2.1) and the Vectorization approach (Section 1.2.2.2), these two workflows represent the first contribution of this thesis**.

To face this data source diversity, it is then very important and challenging to quantify their quality in order to select the best one for further statistical modelling and predictions. The above mentioned presence of groups of interest in the datasets is all making sense here. These groups correspond to the signal, the main information we want to capture, independently of any external sources of noise. Naturally, unsupervised clustering algorithms are adapted to identify these groups. In Section 1.3, such algorithms are discussed and different quality measures are presented in order to diagnose and evalu-

ate the final amount of captured signal linked with the use of a specific data source. **All these clustering quality measures are grouped together in the so-called Metabolomic Informative Content (MIC, Section 1.3.2) concept which represents the second main contribution of this thesis in the field of metabolomics**.

When considering a whole experimental design, the possible use and added values of other techniques like ASCA (ANOVA-Simultaneous Component Analysis), APCA (ANOVA-Principal Component Analysis) and ICA (Independant Component Analysis) will be also mentioned in Sections 1.3.3 and 1.3.4.

In average, MIC indexes or ASCA/APCA/ICA studies ideally lead to the emphasis of a "best" data source, which allows to capture the signal (i.e. to discrimine the groups) in the best possible way. This data source would then be chosen to advantageously perform further statistical studies to discover relevant biomarkers and to make predictions. In Section 1.4, the biomarker discovery issue and challenges in metabolomics are deeply discussed. From PCA (Principal Component Analysis), PLS (Partial Least Squares) and OPLS (Orthogonalized PLS) regressions to more advanced sparse algorithms, an overview of different existing methods is presented. **An innovative algorithm is then proposed: the L-sOPLS (Light-sparse OPLS, Section 1.4.5), which represents the third key contribution of this thesis**.

Apart from PLS-based algorithms, some additional works about other advanced sparse techniques are mentioned in Section 1.4.6.

The notion of sparsity is crucial in such methods. Most of the time, (O)PLS methods lead to a high number of biomarkers considered of interest for further investigations in metabolomics experiments, which may quickly induce difficulties of interpretation. If the decision rule is based on 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 quite subjective. To make this decision more objective, a solution is to maintain 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 some penalties. 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.

After a brief summary of the different datasets used throughout this thesis in Section 1.5, the main results and lessons will be discussed in Section 1.6.

The idea is to visualize a whole process leading to best practices, implying successive questions as follows:

- Which are the potential data sources? 1D or 2D? How to cleverly handle them?
- Which are the best and more suitable pre-processing steps to apply?
- Which data sources provide the best Metabolomic Informative Content (MIC) performances in average?
- Which is the best modelling algorithm in order to find a reasonable pool of relevant biomarkers?

Globally, the main goal of this thesis is then to try to answer in the best way these successive questions and to propose some novel elements for each step (GPL and Vectorization workflows for 2D COSY spectra, MIC, L-sOPLS).

This work is a thesis based on published or submitted original scientific papers. Sections 1.1 to 1.7 will then refer constantly to papers referenced as Paper I to Paper III. These three papers are integrally available in their original versions in Chapter 2, 3 and 4 respectively. Only the page layout was harmonized. Note that some notations can very slightly differ from a paper to another.

In Chapter 5, a draft for a fourth paper is presented and concerns the use of faster COSY acquisition experiments in order to solve the prohibitive 2D acquisition times drawback.

CHAPTER 1

THE ANALYSIS OF 2D NMR SPECTRA IN METABOLOMICS STUDIES: from pre-processing workflows to 2D biomarker discovery

In Chapter 1, an overview of the whole thesis is deeply presented, with constant references to scientific Papers I to III (fully available in subsequent chapters). This chapter is built with the aim of being as far as possible self-content.

1.1 NMR spectroscopy: the 1D vs. 2D spectral data sources

In this section, an introductory description is provided for both ^{1}H-NMR and 2D COSY spectral data sources. Note that this work aims to use and analyse the final spectral data. Therefore, the stages preceding the acquisition of the raw spectra (directly usable for pre-processing and statistical purposes) will not be described in great details. The potential limitations specific to 1D NMR spectroscopy are exposed (Section 1.1.2) and the benefits of 2D COSY are highlighted accordingly (Section 1.1.3).

1.1.1 The 1D ^{1}H-NMR spectroscopy: basics

As already mentioned in the introduction and based on different potential media (urine, blood, serum, ...), ^{1}H-NMR spectroscopy represents a widely and routinely used and well-known analytical technique employed for metabolomics studies, along with the family of Mass spectrometry tools. Now, ^{1}H-NMR spectroscopy finds applications in several areas of sciences as the structural determination of organic compounds. The use of ^{1}H-NMR for metabolic studies was described as early as 1977 when it was shown that proton signals could be observed from a range of compounds in a suspension of red blood cells, including lactate, pyruvate, alanine and creatine [Brown et al., 1977].

^{1}H-NMR uses the magnetic properties of hydrogen-1 nuclei (^{1}H or proton): in a magnetic field, these nuclei absorb energy from applied pulse of an adequate frequency and radiate this energy back out after the pulse. ^{1}H-NMR spectroscopy uses this later energy in order to evaluate the physical, chemical and biological properties of matter.

As the selected nuclei, proton is characterized by a particular magnetic moment and by nuclear spins having two possible basic orientations, the Alpha and Beta levels. The Alpha which aligns spins with the field (north of the nucleus to south of the field and vice versa) is the low energy orientation. The Beta orientation has the north poles and south poles facing. This, as expected, is the high energy orientation. NMR basically consists in transferring spins from the Alpha level to the Beta level via Larmor frequencies ([Dixon, 1984], see Fig.1.1).

In this context, the first state of a NMR experiment is the magnetization (noted B_0 in Fig.1.1, inside an NMR instrument the sample is exposed to a very strong magnetic field). Then, in modern NMR, radiofrequency (Rf) is applied in the form of pulses. A pulse can be defined as a Rf irradiation sequence with pre-established amplitude and period. It is the excitation step. The energy from the radiofrequency pulse is enough to flip the nuclei from its Alpha position to the Beta.

After the excitation step, a relaxation time in needed so that the nuclei relax back to its Alpha position. The excitation/relaxation sequence is repeated several times (thus implying several scans) and successive measurements are recorded between the Rf pulse and the relaxation time.

The information recording and acquisition leads to **Free Induction Decays (FIDs)**. A FID is an interferogram which reflects the absorption and relaxation curve after the disturbance engendered by the pulse. A FID contains all the

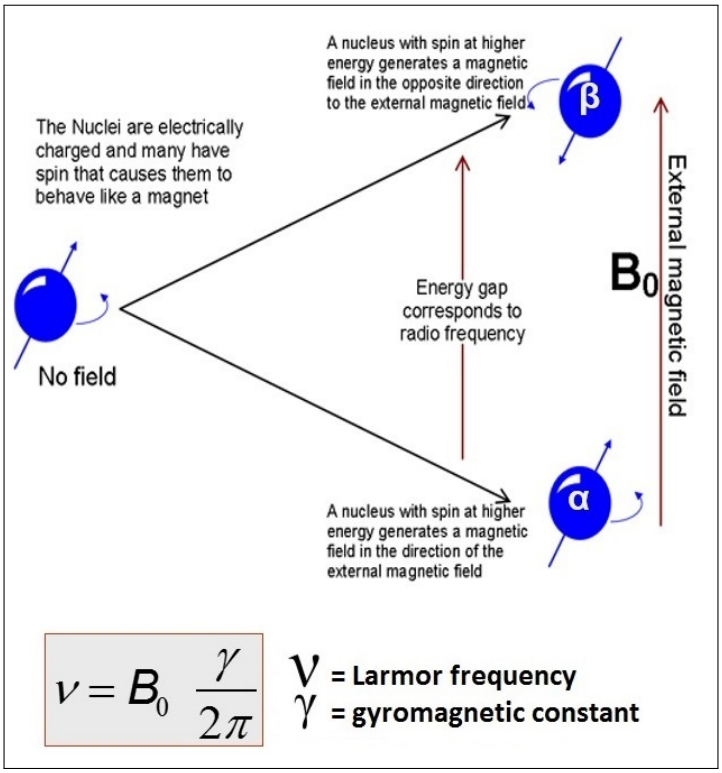


Figure 1.1: From Alpha to Beta: magnetic basis for NMR phenomenon (figure provided by Pascal de Tullio, `http://chem.ch.huji.ac.il/nmr/whatisnmr/whatisnmr.html`).

sample signals, all the signals linked with an internal reference and background noises.

Basically, the FID of a single component is characterized by three elements: its frequency υ (in Hertz), an amplitude s_0 and a decay $T2$ (see the upper part of Fig.1.2). The frequency υ corresponds to the Larmor frequency ([Dixon, 1984]) of the ^{1}H nuclei emitting the FID component and attests of the presence of a given chemical group in the sample. The concentration of the ^{1}H nuclei involved in this chemical group is proportional to the amplitude s_0. During the ^{1}H-NMR experiment, several kinds of ^{1}H nuclei with different Larmor frequencies emit $s(t)$ with specific values of υ, s_0 and $T2$. Being the sum of all these components, the recorded FID has a quite complex form (see the bottom part of Fig.1.2)

which can be simply described as a function of time with real and imaginary parts:

$$S(t) = S_x(t) + i.S_y(t). \tag{1.1}$$

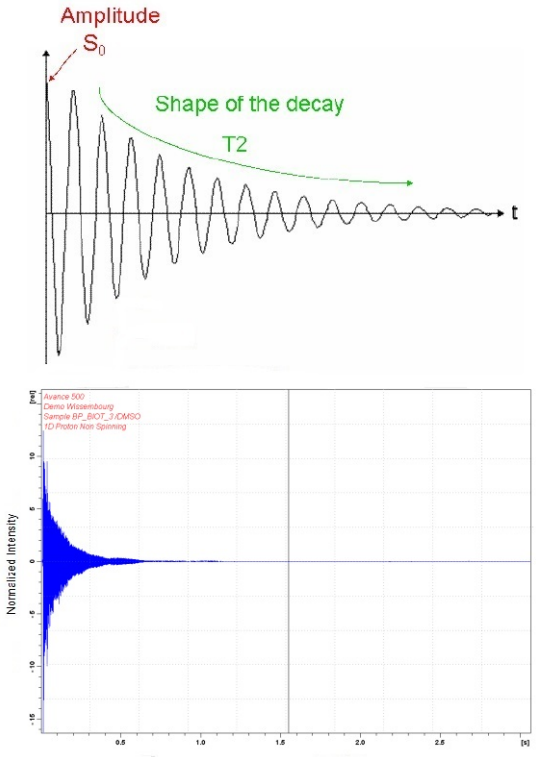


Figure 1.2: The characteristics of the FID component and a proper FID.

Upstream pre-processing steps, including a Fourier transformation, transform the individual FID in a frequency spectrum with the aim to render subsequent analysis easier, robust and accurate. These methods tend to reduce sources of variation, external noise and potential artefacts which are not of interest or which might interfere with further statistical data analyses. The Fourier Transform (FT) is used to recover from the FID the different molecular components involved in its formation. The FT represents each recovered FID component by a peak in a frequency domain spectrum. The final aspect of a spectral peak is determined by different parameters (again the initial frequency, the amplitude

and the shape of the decay [Claridge, 2016]). The resulting spectrum is the translation in a frequency domain of the FID time domain signal.

The position of a peak expresses the frequency of emission and attests the presence in the sample of a proton with specific electron surrounding. Note that small interactions (couplings) between different protons involved in a same molecule can lead to the appearance of several peaks for a same group of proton (the so-called "multiplicity of peaks", see Fig.1.3 which represents the ethanol ^{1}H-NMR spectrum). Based on established couplings of peaks, the molecules contained in the sample are deduced from the observation of several peaks in some specific positions. The area under any peak is proportional to the concentration in the sample of the kind of proton emitting at this frequency [Tofts and Wray, 1988]. Consequently, the concentration of a molecule in the sample may be obtained from the areas under its corresponding spectral peaks. Peak areas estimation is most of the time realized by integrals.

It must be noted that the scale in Hertz is finally calibrated and converted into a chemical shift vector expressed in "parts per million" (ppm), a dimensionless scale that is unrelated to the magnetic field strength of the spectrometer ([Keeler, 2010], [Martin et al., 2017]). By the way, the frequencies expressed in ppm are independent of the used spectrometer.

Besides the Fourier Transform step, allowing to work on the basis of a frequency domain, other supplementary pre-treatments have to be performed before the spectra become fully and statistically exploitable. Indeed, due to instrumental ^{1}H-NMR parameters and limitations, the original data inevitably contain noise or artefacts. Considering the biological character of the samples, the spectra are also altered under the influence of pH, diuretic fluctuations, waiting times, bacterial contaminations among others.

At this level, three main inherent problems can occur:

- Phase shift (or peak disturbances). For instrumental reasons, the axis along which the signal appears cannot be predicted, so in any practical situation there is an unknown phase shift. In general, this leads to a situation in which the real part of the spectrum (which is normally the displayed part) does not show a pure absorption lineshape. For phase correction, frequency independent (0-order) or frequency dependent (first order) solutions are generally used. In practice, to phase the spectrum correctly usually requires some iteration of the zero- and first-order phase corrections ([Keeler, 2010]).
- Baseline artefacts (see Section 1.2.1 for solutions).

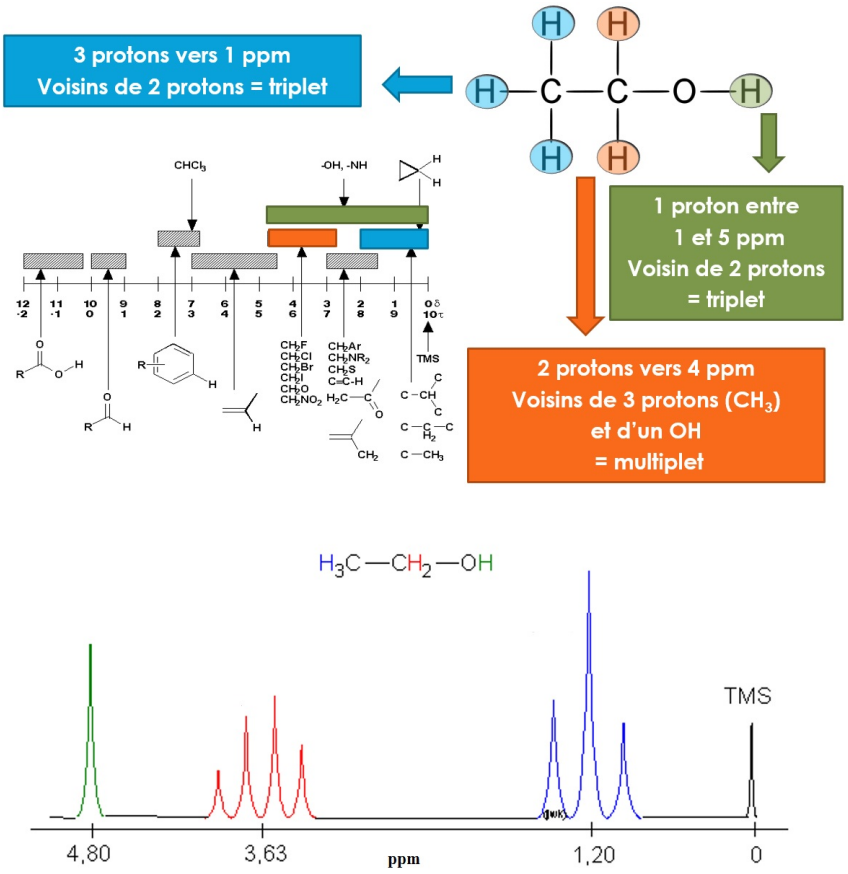


Figure 1.3: The multiplicity of peaks issue: the example of the ethanol ^{1}H-NMR spectrum (figure provided by Pascal de Tullio).

- Calibration. The zero on the ppm axis is generally adjusted on a reference signal, usually the TMSP (trimethylsilylpropionate) signal. Two principal methods may be used for metabolite quantification in ^{1}H-NMR: the absolute method with reference to an internal standard compound, and the calibration approach with reference to some calibration model.

All these problems modify the quality of the spectral description of metabolites and/or create spectral variations that are non-informative for statistical purposes and for the biomarker discovery goal. Data pre-processing aims at removing these undesired variabilities. A data pre-processing scheme consists in a

combination of operations based on statistical and/or mathematical principles, performed on time or frequency ^{1}H-NMR data.

For 1D ^{1}H-NMR, these operations are discussed in Section 1.2.1 and are fully explained in [Rousseau, 2011] and [Martin et al., 2017].

Finally, a typical obtained ^{1}H-NMR spectrum of urine sample contains thousands of sharp lines from predominantly low molecular weight metabolites (see the upper part of Fig.1.4, [Duarte et al., 2014]). Blood contains low and high molecular weight components, and this gives a wide range of signal line widths (see the lower part of Fig.1.4, [Duarte et al., 2014]). Broad bands from protein and lipoprotein signals contribute strongly to these ^{1}H-NMR spectra, with sharp peaks from small molecules superimposed on them. Spectral databases respectively exist to identify molecules in urine and blood spectra according to the positions of spectral peaks.

1.1.2 The 1D ^{1}H-NMR limitations for metabolomics studies

As already mentioned, the ^{1}H-NMR data source is a standard in the field of metabolomics and very powerful pre-processing and analytical tools adapted to these data still demonstrate their potential. Nevertheless, when the studied medium is very rich in different metabolites and/or when the design is by construction uncontrolled (for instance without spiked metabolites or products during the experiment), it is natural to observe a very large number of peaks. Based on complex biofluids, a resulting ^{1}H-NMR spectrum can be interpreted as a superimposing of individual spectra, each stemming from a particular pure metabolite.

In this context, different problems can occur. In metabolomics analyses, the presence of a huge amount of water and sometimes large proteins leads to undesired and over-represented signals that could mask the metabolites signals. These effects are generally reduced upstream via the use of devoted sequences (respectively *noesypresat* ([Wright et al., 1995]) for water and the *cpmg* ([Meiboom and Gill, 1958], [Loria et al., 1999]) routine for proteins) which are available for instance in Bruker softwares.

A more major and usually encountered problem involves signals which overlap heavily, with ambiguous or overlapping resonances (due to the presence of numerous metabolites and to the multiplicity of peaks). In other words, a peak associated with a high intensity in these cases is not necessarily and directly linked with a single metabolite but may be linked with several aggregated ones,

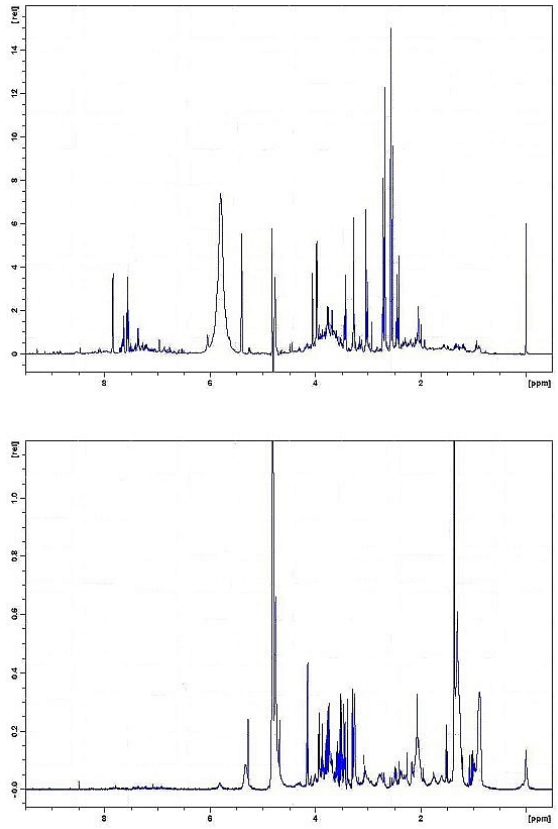


Figure 1.4: High resolution ^{1}H-NMR spectrum of urine (top) and of blood serum (bottom) on a ppm scale.

with very similar chemical shifts. This leads to a significative loss of information about the masked metabolites and to difficulties of identification.

Generally, two cases often occur: for a same metabolite, all its peaks can overlap (as in Fig.1.5); and peaks associated with different nearby metabolites can overlap too (as in Fig.1.6).

Fig.1.5 shows an illustration based on the octane hydrocarbon, where the theoretical number of signals may not equal the observed number because of overlap.

In a more quantitative regard, 1D NMR is a recognized method for quantitative analysis ([Malz and Jancke, 2005] [Pauli, 2001]) with applications in various fields such as metabolism studies (ex: [Billault et al., 2004]) or the au-

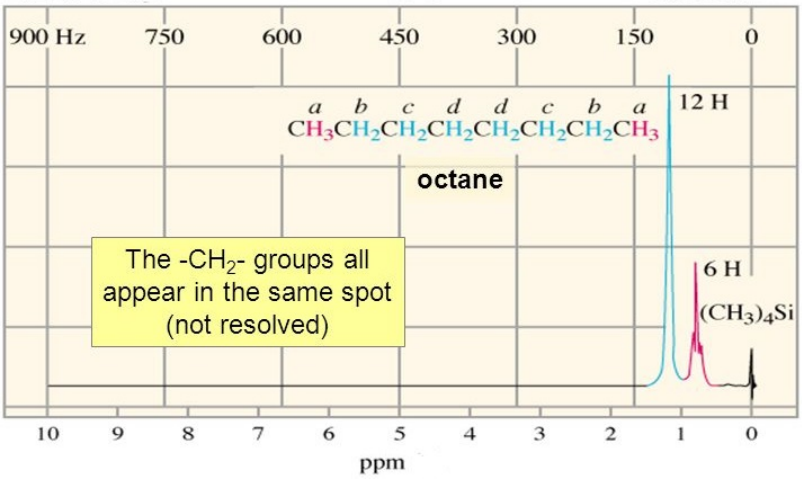


Figure 1.5: NMR overlapping proton signals: protons b, c and d are roughly in the same environment, and their chemical shifts are also about the same (from [Giraudeau, 2010]).

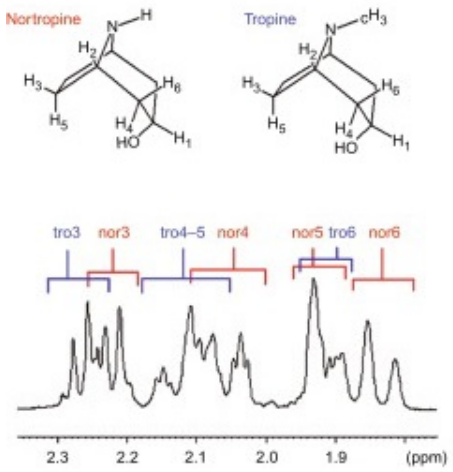


Figure 1.6: Nortropine and tropine molecules, and 1D spectrum of an equimolar mixture of these molecules showing the overlap of some of the peaks (from [Giraudeau, 2010]).

thentication of bio-based products (ex: [Tenailleau et al., 2004]). This method has however some limitations, in particular when the above-mentioned overlap

between 1D peaks is such that it is impossible to determine the area under these peaks with a good correctness and a good precision. This is the case when the 1D spectrum is concerned by numerous couplings, or when we want to measure mixtures of similar compounds. Indeed, when two molecules have nearby structures, their 1D spectra are very close and the 1D spectrum of a mixture of these two molecules does not allow to quantify them separately ([Giraudeau, 2010] and see an example in Fig.1.6).

Finally, metabolite chemical shifts are sensitive to the chemical environment and subtle matrix effects, such as differences in pH and ionic strength, generally lead to inter-sample peak position variations ([Weljie et al., 2006] [Tredwell et al., 2016]).

Unlike the water and proteins issues, there is no obvious answer to solve the 1D overlapping issue. In Section 1.1.3, the use of 2D data sources instead of 1D is deeply promoted and argumented in this way.

1.1.3 The 2D NMR alternatives and the COSY experiment

To address the potential problems of multiplicity and overlapping encountered in 1D NMR, some 2D spectroscopy techniques have been considered these recent years for metabolomics purposes. Depending of the complexity of the molecules, 2D NMR gives a different kind of additional informations than ^{1}H-NMR. Interactions between different nuclei can be displayed more concretely.

Technically, these interactions can be seen in the ^{1}H-NMR spectrum through homonuclear coupling constants (which measure the interaction between a pair of protons usually in frequency units, Hz). But for compound with a certain level of complexity, full peaks overlapping or very similar coupling constants will likely occur, which is where 2D NMR comes into play.

In two- (or multi-) dimensional NMR spectroscopy, multipulse sequences are employed to provide additional information not obtainable (or not easily obtainable) from one-dimensional spectra. The most useful and commonly used forms of 2D NMR spectroscopy provide correlations between proton (or other NMR-active nuclei) signals based on some interaction between them, most frequently J-coupling (H-H, H-C (carbon) or even C-C), but it could also be dipolar coupling (the NOE interaction), rates of chemical exchange or relative diffusion rates.

Among many others, some common types of 2D experiments are of interest for metabolomics analysis and are shortly listed below:

- COSY (COrrelation SpectroscopY). As a homonuclear correlated spectroscopy tool, COSY is based on the correlation between protons that are coupled to each other ([Bax and Freeman, 1981]). Many modified modern versions of the basic COSY experiment are tested in recent literature.

- TOCSY (TOtal Correlation SpectroscopY, also referred to as the HOHAHA experiment, [Dalvit and Bovermann, 1995]). TOCSY uses a spin-lock for coherence transfer. During the spin-lock all protons of a coupled system become "strongly coupled", leading to cross peaks between all resonances of a coupled system. It has some advantages over COSY as long-distance couplings can be detected (in COSY small couplings can give very weak cross peaks because the peaks are antiphase).

- HSQC (Heteronuclear Single Quantum Coherence). HSQC is a heteronuclear correlation experiment, usually between ^{1}H and ^{13}C resonances, which uses proton detection of the ^{13}C signals using an INEPT sequence. It shows higher resolution in the C-dimension than does related experiments like HMQC (Proton detected heteronuclear multiquantum coherence, [Bax and Summers, 1986]).

- HMBC (Heteronuclear Multi-Bond Connectivity). This experiment is a lot like HMQC, but it is optimized to detect proton-carbon correlation over two and/or three bonds. Although the sensitivity is better than direct detection methods, it is still not optimal, since the magnetization has to be transferred twice, first from proton to carbon and then back to proton via weak long-range interactions, which require relatively long delay times and thus much opportunity for loss of signal intensity by times of relaxation. For distinguishing two from three-bonds coupling, see [Kock et al., 1996].

Note that the COSY and TOCSY experiments provide symmetric final 2D spectra, because of their homonuclear basis. Being heteronuclear, it is not the case for HSQC and HMBC. HSQC and HMBC are also less sensitive for deep metabolomics studies and generally require longer analysis times.

In this thesis, the **COSY experiment** is used as a matter of priority because it is probably the most useful and the most popular of all 2D NMR experiments ([Xi et al., 2006], [Marchand et al., 2017]), and it provides easily accessible information about coupling pathways. COSY provides a high level of sensitivity and structural information. The COSY pulse sequence in its simplest form is shown in Fig.1.7 (note that there are many variants with additional pulses for achieving higher quality correlations and removing artefacts).

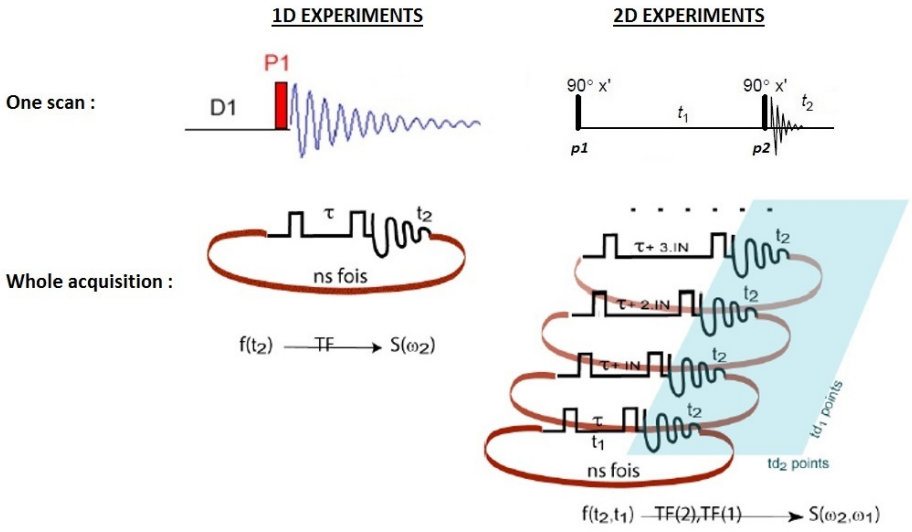


Figure 1.7: The 2D COSY pulse sequence (right) compared with the 1D pulse sequence (left) (adapted from `http://thcgerm.free.fr/IMG/pdf/RMNMutidimensionnelle_PBerthault.pdf`).

COSY is used to identify spins which are coupled to each other. It consists in a single radiofrequency pulse (p_1) followed by the specific evolution time (t_1) followed by a second pulse (p_2) followed by a detection and measurement period (t_2). Practically, during the t_1 period, each of the signals is labeled by the chemical shifts and couplings. After the second 90^o pulse, the component of the magnetization which used to be in the y' direction is now in the z-direction, and becomes undetectable. Only the component in the x' direction evolves and is detected during collection of the FID. What cannot be shown in vector diagrams is that during the t_2 period the magnetization in the $x-y$ plane contains components from both the nucleus observed as well from the frequency of its coupling partner, mediated by the coupling between them. When the spectra are Fourier transformed in the t_1 dimension, the normal coherence is detected, such signals appear along the diagonal of the resulting 2D spectrum. In addition, the coherence from the frequency of the coupled partner is also detected and results in cross peaks at that position ([Bax and Freeman, 1981]).

In other words, the variation of the evolution time t_1 in a collection of experiments provides an identification of the spins where the magnetizations originate ([Gu et al., 2009]). The result is a collection of FID measurements (Fig.1.8 left),

whose Fourier resulting spectrum shows signals connecting the frequencies of neighboring nuclei (Fig.1.8 right).

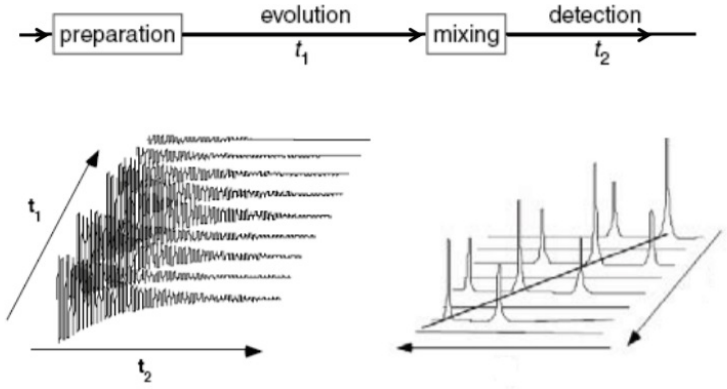


Figure 1.8: A schematic description of a 2D experiment: FID measurements as a function of t_1 and t_2 (left) and Fourier resulting spectrum of the signal (right) (from [Gu et al., 2009]).

Thus, final COSY spectra show two types of peaks. **Diagonal peaks** have the same frequency coordinate on each axis and appear along the diagonal of the plot, while **cross peaks** have different values for each frequency coordinate and appear off the diagonal. Diagonal peaks correspond to the peaks in a 1D-NMR experiment, while the cross peaks indicate couplings between pairs of nuclei (much as multiplet splitting indicates couplings in 1D-NMR).

Cross peaks result from a phenomenon called magnetization transfer, and their presence indicates that two nuclei are coupled (they have the two different chemical shifts that make up the cross peak's coordinates). Each coupling gives two symmetric cross peaks above and below the diagonal. That is, a cross peak occurs when there is a correlation between the signals of the spectrum along each of the two axes at these values. One can thus determine which metabolite peaks are connected to another (within a small number of chemical bonds) by looking for cross peaks between various signals. An easy visual way to determine which couplings a cross peak represents is to find the diagonal peak which is directly above or below the cross peak, and the other diagonal peak which is directly to the left or right of the cross peak. The nuclei represented by those two diagonal peaks are coupled [Keeler, 2010].

Fig.1.9 shows a pedagogical example which highlights the informative gain provided by a 2D COSY spectrum. Let us imagine several individuals speaking

each a single language. When representing them on a single dimension, it can become quickly unreadable if the number of individuals is high (middle part of Fig.1.9). This can refer to the multiplicity and to the overlapping issues encountered in 1D spectra when several metabolites are present in a particular biofluid. In a 2D represention inspired by the COSY experiment (bottom part of Fig.1.9), the individuals and their relationships ("correlations") are much more clear and interpretable. A diagonal peak can be seen as an individual who understands himself, a cross peak can be interpreted as two individuals who understand each other.

Consequently, by the addition of the cross peaks outside the main diagonal, COSY spectra naturally contain more information than corresponding 1D spectra. One of the challenges of this thesis consists in demonstrating that this extra information is not linked only with potential sources of noise, but helps to capture the main signal (the existence of groups of samples, as already mentioned before) and helps to discover further relevant biomarkers during metabolomics studies. It will be shown that results in Section 1.6 and in Papers I to III are consistent with this challenge.

Note that further technical details on the COSY technique (and secondarily on other homonuclear and heteronuclear 2D techniques mentioned above) are available in [Giraudeau, 2010]. It concerns for instance the double Fourier transform step, the signal amplitude and phase modulations and the ways to reduce noise.

Note also that faster 2D COSY NMR acquisition techniques, also detailed in [Giraudeau, 2010] and in [Giraudeau and Frydman, 2014], are mentioned and taken into consideration in Chapter 5 (Non-Uniform Sampling and single-scan COSY). These faster NMR studies intend to obtain two-dimensional NMR spectra in a shortest time by optimizing the whole pulse sequences protocol steps.

1.2 Handling of 1D and 2D spectral data and pre-processing workflows for 2D COSY spectra

In this section, the main pre-processing tools for ^{1}H-NMR are first discussed as an established best practice in Section 1.2.1. In this thesis, **a very important starting intuition is made: we assume that 2D spectra are expected to require a lower level of pre-processing to express at least as much**

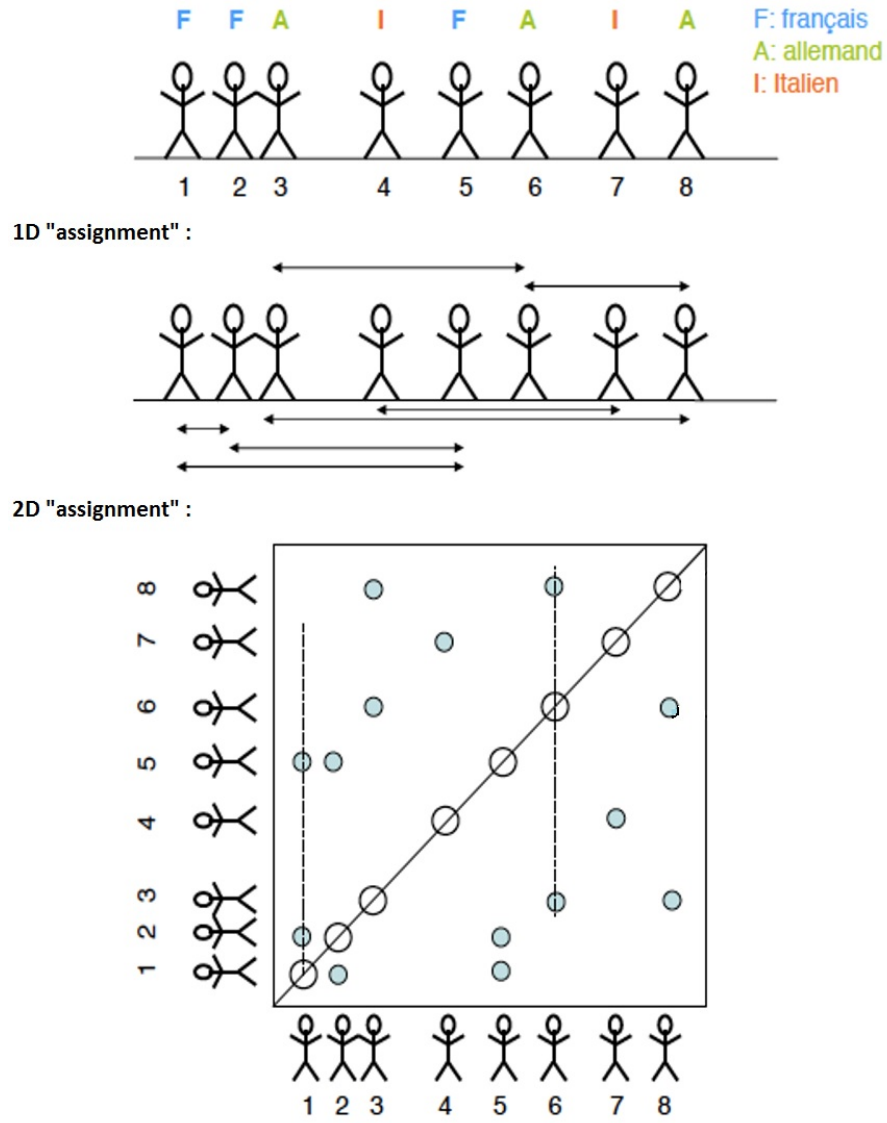


Figure 1.9: A pedagogical view of the 2D COSY additional information compared to 1D representations (adapted from `https://docplayer.fr/6372253-La-resonance-magnetique-nucleaire-rmn-dr-julien-furrer-1.html`).

information as corresponding 1D ones. In other words, this assumption

places trust in the fundamental relevance and quality of the additional informative content inherent in the 2D spectral data sources (see Section 1.1.3).

Handling a whole set of 2D spectra is not an obvious and usual task in metabolomics studies. Section 1.2.2 presents two novel devoted workflows: the Global Peak Least (GPL) one and the Vectorization one. **These workflows represent the first contribution of this thesis**.

1.2.1 Typical pre-processing steps for ^{1}H-NMR

As the ^{1}H-NMR data source is widely used in metabolomics studies, the key pre-processing steps to apply on it are known and increasingly well-controlled. These main steps are briefly discussed here. Interesting further details can be found in [Martin et al., 2017] with the use of the PepsNMR R package.

As already mentioned, ^{1}H-NMR spectra can be altered by overlaps, artefacts, external noise factors or systematic peak misalignments. In the analysis of biological samples, control over experimental design and data acquisition procedures alone cannot ensure well-conditioned ^{1}H-NMR spectra (with a maximal signal recovery for instance). But another major point can affect the accuracy and robustness of final metabolomics analyses: the data pre-processing itself.
The usual approach is to use proprietary software provided by the analytical instruments' manufacturers to conduct the entire pre-processing strategy. This widespread practice has a number of advantages such as a user-friendly interface with graphical facilities, but it involves non-negligible drawbacks: a lack of methodological information and automation, a dependency of subjective human choices, the limitation to standard processing possibilities and a lack of objective quality criteria to evaluate the whole pre-processing quality.

To solve these issues and to provide clean inputs for further statistical models, a list of tools are generally performed on the raw spectra. The main steps are presented in the green part[1] of Fig.1.10 and are the following ones (the PepsNMR workflow presented in [Martin et al., 2017] is taken as reference here):

- *Zero order phase correction.* For technical reasons, a spectrum can exhibit a zero order phase shift error of a certain angle ϕ_0. This phase shift

[1] It is very important to distinguish what is within the competence of initial treatment (in blue in Fig.1.10) at the FID level, and therefore before the acquisition of raw spectra, and what concerns the "pre-processing" itself (in green in Fig.1.10). Relevant statistics cannot be made in the blue part, what motivates the choice not to develop and focus on the steps included in the blue part in this thesis. Thus, all the statistical contributions contained in this thesis have to be considered once the raw spectra are acquired and available.

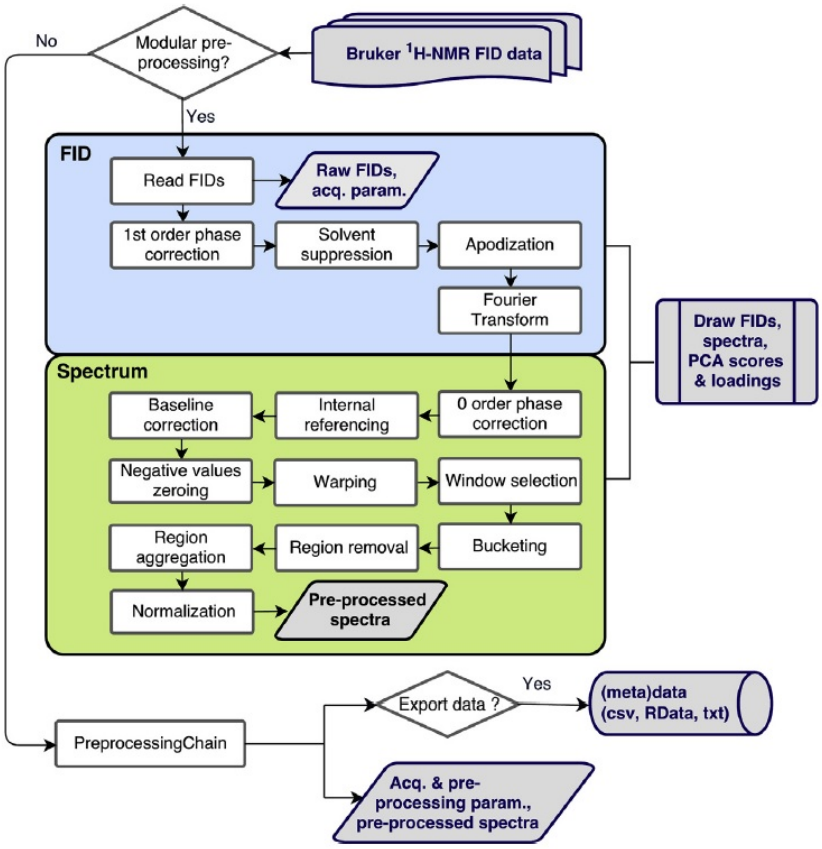


Figure 1.10: Complete flowchart for ^{1}H-NMR data pre-processing (from [Martin et al., 2017])

is independent of the spectral frequencies. To solve this problem, an optimal angle ϕ_0 has to be found by maximizing an adequate criterion of the positiveness of the real part of the spectrum.

- *Referencing with an internal standard.* In order to be more reliable, with a more accurate alignment between spectra, the scale can be referenced to a known standard, i.e. an internal reference compound whose chemical shifts are minimally influenced by external factors (temperature or concentration) and ideally located outside the spectral region to be clearly identifiable. The reference compound used in NMR is usually a silane derivative such as Trimethylsilylpropionate (TMSP). The chemical shift of these standard peaks is referenced to 0 ppm by convention.

- *Baseline correction.* Ideal spectra are expected to have a flat baseline but, even after phase correction, baseline artefacts can still appear. These artefacts result from multiple sources, such as the presence of macromolecules or calibration errors from the initial pulse [Euceda et al., 2015]. Baseline distortions in 1D NMR spectra are also mainly caused by the corruption of the first few data points in FID (free induction decay). These corrupted data points add low frequency modulations in the Fourier-transformed spectrum, and thus formed the distorted baseline. Correction of these distortions is a necessary step in NMR spectra data processing because they offset the intensity values and result in inaccuracy in peak assignment and quantification. These errors could be critical in the study of metabolomics, which involves many small but statistically significant peaks that are sensitive to baseline distortions. Incorrect quantification of these peaks may result in failures in detection of important metabolites or identifying potential biomarkers ([Xi and Rocke, 2008]). It is critical to solve this issue with an appropriate methodology in order to enhance the warping efficiency and to perform accurate data analysis on meaningful and unaltered peak shapes. Several algorithms are suggested and mainly involve iterative penalized least squares methods, robust baseline estimation, iterative polynomial fitting, local means or wavelets transforms ([Liland et al., 2010], [Marion, 2016], [Eilers and Boelens, 2005] for instance). Fig.1.11 shows a simple illustration of baseline correction for a simple 1D ^{1}H-NMR reference spectrum of DSS (2,2-Dimethyl-2-silapentane-5-sulfonic acid) with 65536 data points.

- *Negative value zeroing.* Despite the application of baseline and phase corrections, spectra may still have negative intensities at specific frequency values. This step simply sets all these negative values to zero since they cannot be properly interpreted.

- *Warping.* In metabolomics applications, due to the nature of biological samples and variations in experimental conditions (e.g. pH, temperature or concentration), peak shifts, or misalignment between identical features from different spectra, are observed. In the literature, peak alignment solutions commonly applied to NMR spectra are warping and bucketing. Warping will indeed enhance the similarity between profiles by the way of shifts, stretches and/or compressions along their horizontal axis. Recent reviews by [Bloemberg et al., 2013], [Vu and Laukens, 2013] and [Eilers, 2004] report the advantages and drawbacks of the most important warping procedures for signal alignment, such as parametric time warping (PTW), semi-parametric time warping (SPTW) or dynamic time warping (DTW), among others.

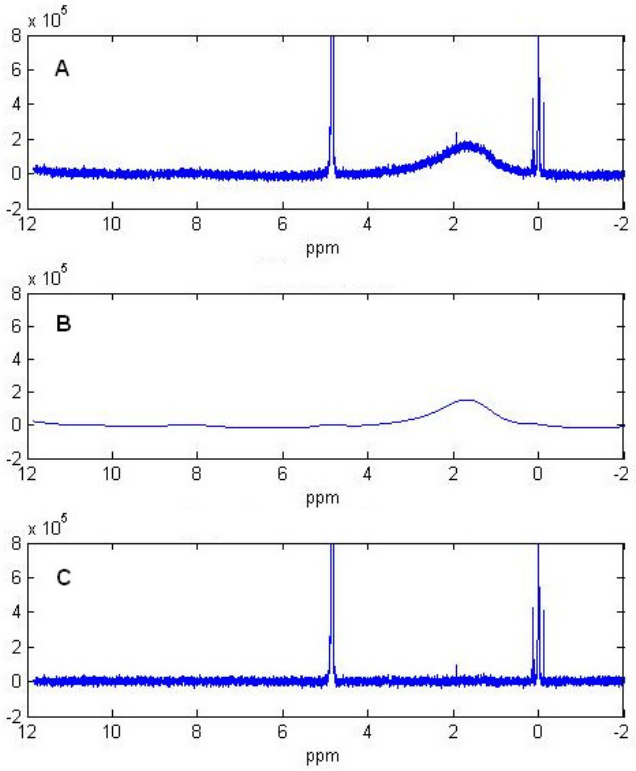


Figure 1.11: Baseline correction by penalized parametric smoothing method. (A) Original 1D proton NMR spectrum of DSS reference with distorted baseline. (B) Detected baseline curve by penalized parametric smoothing method. (C) Corrected spectrum after baseline subtraction (from [Xi and Rocke, 2008].

- *Window selection.* In order to focus the data analysis on particularly informative spectral areas, one can trim the spectral window between right and left ppm chemical shift bounds, thus decreasing the number of descriptors m in the spectrum.

- *Bucketing.* The high dimensionality of the data and small residual peak shifts can impede future multivariate data analyses [Liland, 2011]. Bucketing integrates the original spectral intensities into predefined intervals. Since the warping already corrects for chemical shift misalignments, a soft bucketing technique is sufficient to smooth and correct for small remaining deviations, provided that shifts of a same peak are small enough to be included in the same interval. Also called data reduction, bucketing

therefore reduces the number of potential descriptors. In this step, there is a trade-off between keeping all the spectral information and removing peaks shifts as well as decreasing the total number of descriptors.

- *Region removal or aggregation.* Resonances without interest can eventually alter data analysis introducing unwanted variation. With biological samples, one major problem is the presence of water resonance residuals and positive values created by the baseline correction step, even with a previous application of pre-saturation and solvent signal suppression techniques [Smolinska et al., 2012].

- *Normalization.* Normalization is defined as an operation where each spectrum is multiplied by a constant term. It mostly addresses the dilution factor problem: variations in absolute concentrations can be observed between biofluid samples whereas data analysis focuses only on relative peak differences between them. Typically, the constant sum normalization (CSN) technique is very popular and advised for urine, sera, cell media or biopsy samples where large differences between groups of patients are not expected. Guidance on the choice of a normalization approach based on the type of sample of interest can be found in [Wu and Li, 2016].

Once every individual spectrum has been submitted to these pre-processing transformations, all the resulting individual intensities are summarized into a $(1 \times m)$ row vector, with m being here the final number of ppm signals (descriptors) which can be principally tuned via the bucketing step. A whole experimental design, implying n measures or samples, is then fully represented into a $(n \times m)$ matrix by stacking all the individual spectral vectors.

These steps, along with the other steps applied upstream on the raw FIDs (in blue in Fig.1.10), were put into practice, in diverse degrees, every time 1D data were used within the framework of this thesis.
By "diverse degrees", it must be noted that some steps can be optional (according to the nature of the used medium and to the chosen level of pre-processing) and are submitted to tuning parameters.

1.2.2 Pre-processing workflows for 2D spectral data

As discussed in Section 1.2.1, good practices are well established and known for pre-processing ^{1}H-NMR spectral data before statistical modelling. However, it is nowadays not the case when using 2D data sources. A few years back, decisions were often taken only on the basis of graphical comparisons between individual 2D spectrum (visualizing differences between peak positions from a

spectrum to another, or between peak intensities, for example in [Schroeder et al., 2007]). These practices were obviously subjective and inadequate when trying to highlight small variations. More recently, image processing or graphical tools contributed to semi-automate the process ([Xia et al., 2008]), but automation is not generalized yet for 2D spectra handling and pre-processing. In this context, the use of 2D NMR data sources is very far from competing with the standard use of ^{1}H-NMR.

In this thesis, two novel dedicated workflows are presented to fully handle 2D NMR COSY spectra, from each raw individual spectrum involving in an experimental design to a global object containing all the information for further statistical analyses. The Global Peak List (GPL) one is presented in Section 1.2.2.1. The Vectorization approach is presented in Section 1.2.2.2. These two workflows are displayed in Fig.1.12 and are fully detailed in Paper III.

It is crucial to repeat that our initial intuition is that 2D spectral data, as COSY, require a less advanced pre-processing level than corresponding 1D spectra. A priori, if the additional information contained in the 2D data is sufficiently relevant, fewer efforts may be needed for their pre-processing.

It is also very important to note that the GPL and Vectorization approaches may be generalized and applied on any 2D spectral sources (Section 1.1.3) implying similar structures, issued from NMR, Mass Spectrometry or others, and are not specifically dedicated to COSY spectra.

1.2.2.1 The Global Peak List (GPL) workflow

Any initial point of a 2D spectrum (for instance COSY) can be summarized with only three items: its two ppm coordinates and its intensity (which corresponds to the integral under the point).

For this reason, it is relatively natural to store a whole individual 2D spectrum into a corresponding 3-columns matrix. Because a 2D spectrum contains a lot of empty areas, with intensities very close to zero, this matrix can be reduced in order to only reveal existing detectable peaks. It is then called "individual peak list" and there are as many individual peak lists as individual initial measures, or samples (n), in the experimental design.

Since every individual peak list contains the information about detectable peaks in its corresponding spectrum, the size of each peak list is unique (with a number of rows t_i) and is not known in advance.

This notion of **detectability** is directly linked with the choice of a threshold to determine when an intensity level begins to be relevant and can not

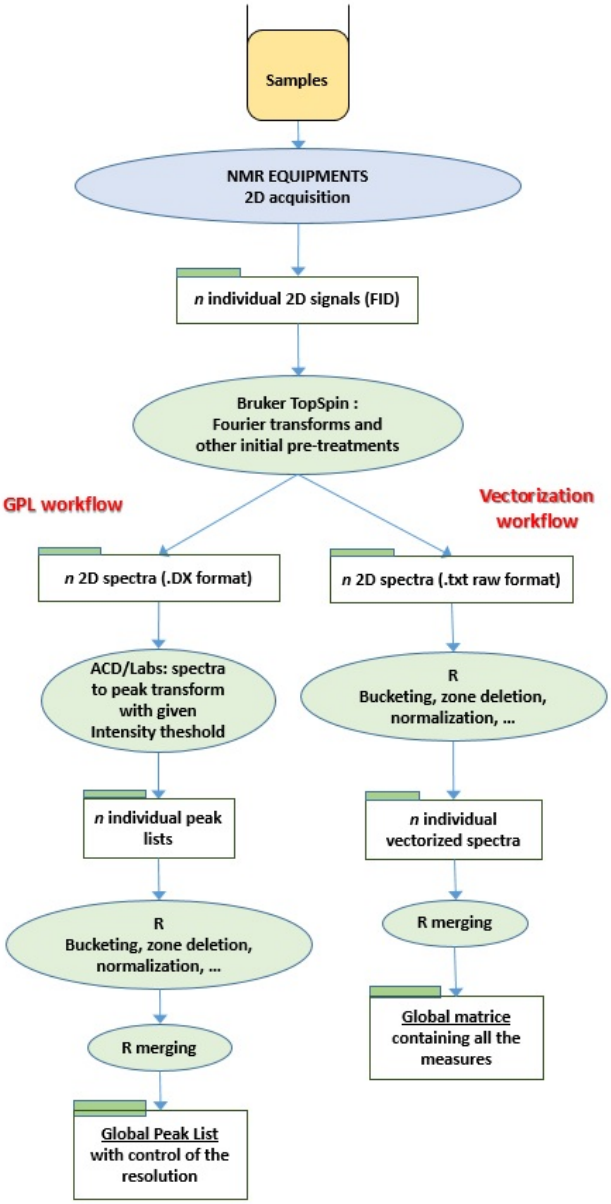


Figure 1.12: The GPL and vectorization approaches workflows for 2D COSY experiments: steps and resulting data files.

be associated with pure noise only or background noise. This minimal level threshold has to be maintained at a low value, empirically between 0.02 and 0.05 when using a software like ACD/Labs (ADC/NMR processor) to extract signal from a 2D spectrum. Initially, in ACD/Labs, the minimal level threshold is set according to a noise level calculated using an autodetect tool after Fourier Transform. However, empirically, with threshold values lower than 0.02, or equal to zero in the extreme case, all the background noisy parts of the spectrum are kept; with values higher than 0.05, some potentially interesting peaks can be removed and lost. So, such threshold values help to avoid negative or very small noisy intensities and don't affect the peaks shape pattern. An illustration is provided in Fig.1.13 where the use of a minimal threshold lower than 0.02 involves a dramatic increase in the number of detected peaks from a 2D COSY spectrum.

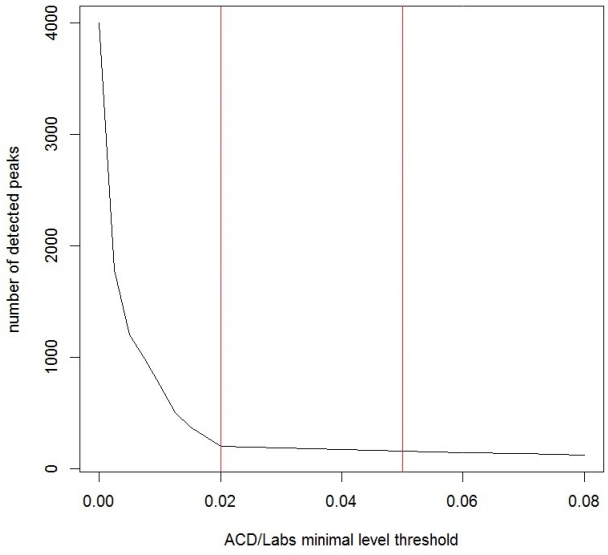


Figure 1.13: Illustration of the evolution of the number of detected peaks according to the chosen minimal ACD/Labs detectability threshold.

It is very important here to highlight that the notion of detectability has for consequence a **dimension reduction** when using the GPL workflow.

In practice, note that the use of ACD/Labs to obtain the individual peak lists from the raw spectra means that every initial spectrum, extracted in a standard .dx format, is transformed into a 3-columns peak list, basically in a text format.

As explained in [Davies and Lampen, 1993], the .dx format is a protocol for the exchange of NMR spectral data without any loss of information and in a format that is compatible with all storage media and computer systems.

As already mentioned, in scientific research, experimental design numerical results have to be prepared and considered as a whole to allow the construction of a global and unique object before any further simultaneous data analysis or multivariate statistical modelling. To fully use the whole information, it is fundamental to gather and merge the individual peak lists of size $(t_i \times 3)$ in a single and unique global object: the so-called Global Peak List (GPL).

The GPL approach was first presented in Paper I and the proper workflow is fully described in Paper III. For an entire experimental design involving n individual spectra, a GPL will contain information about all existing peaks (i.e. all T peaks which appear at least one time in an individual spectrum). This GPL is a $(T \times M_{GPL})$ intensity matrix. Typically, the column dimension M_{GPL} corresponds to $2 + n$ (i.e. the two ppm coordinates and the corresponding n individual spectra peak intensities). Individual and Global Peak List are summarized in Fig.1.14.

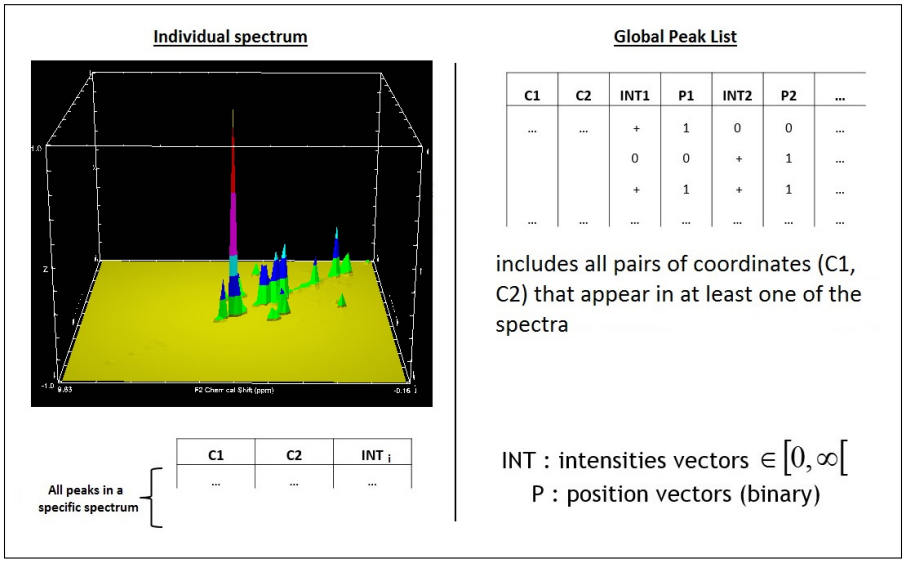


Figure 1.14: From individual to GPL: an overview.

In Paper I, the positions of peaks were also taken into account for subsequent analyses, involving the potential addition of n other binary column vectors

(with a value of 1 if the peak appears in the corresponding spectrum; and 0 otherwise). Working on the signals' positions or, in other words, on the simple existence of signals is motivated by a biological justification: a signal, or a particular metabolite, can be observed or not for a particular donor or in a particular medium. If a signal is present in detectable quantity, this presence or absence is supposed to be stable, whereas intensities are variable from one measure to another according to potential uncontrolled factors (from Paper I).

The number of rows (T) of a GPL can not be known in advance. It depends of the number of existing peaks in the experiment and can vary greatly according to the medium of interest, the conditions of acquisition, etc... To handle this issue, the GPL workflow proposes a way to partially control this number and impose a selected resolution level.

In this workflow, a **soft "bucketing" step** adapted to 2D COSY (but extensible to any 2D data) 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 accordingly. 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.

Finally, by tuning the ACD/Labs detectability threshold and/or the number of decimals in the GPL "bucketing" step, different resulting GPL matrices, having different dimensions, may be generated and used to represent a same experimental design.

These GPL matrices will be later tested and compared in terms of Metabolomic Informative Content and of biomarker discovery efficiency (see Paper I, Paper III and Section 1.6 for numerical results and conclusions).

1.2.2.2 The Vectorization workflow

Along with the GPL workflow, a second approach is proposed based on the systematic vectorization of 2D COSY spectra. Unlike the GPL approach, the Vectorization one can be directly applied on raw Bruker text files because peak lists are not required at all. The choice of the intensity threshold when using the ACD/Labs software is then not anymore an issue.

The straightforward principle of this approach is fully detailed in Paper III (see Section 4.3.2). It starts by an integration bucketing step on a full individual 2D grid similar to the 1D one but applied twice, by rows and by columns, to obtain a first square matrix of size $(M_V \times M_V)$. For convenience with standards in metabolomics studies, M_V is often chosen equal to 256, 512 or 1024 in order to control the resolution and to avoid truncating extremities.

The vectorization of each individual square matrix naturally leads to a $(1 \times {M_V}^2)$ row vector. Finally, all the 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 involved in a specific design. Note that the number ${M_V}^2$ is fully known in advance via the bucketing step, while it is not the case for T in the GPL approach.

In this resulting intensity matrix, the n spectra are represented in rows and the ${M_V}^2$ couples of ppm coordinates in columns.

In summary, the GPL workflow provides a global matrix for which the dimensionality is partially controlled and allows dimension reduction by considering detectable peaks only (according to a devoted threshold). The Vectorization workflow provides a global matrix with a fully controlled dimensionality and takes into account all the initial information (because, like in 1D, there is no peak selection).

1.2.2.3 Necessary pre-processing for 2D COSY spectra

As already mentioned, a very important starting intuition throughout this thesis is that 2D spectral data need a lesser level of pre-processing than ^{1}H-NMR before statistical analyses because of the additional information they contain. It was shown in Section 1.2.1 that pre-processing steps for ^{1}H-NMR are crucial to make the data exploitable analytically, are numerous and may be very difficult to understand and to perform in an efficient manner. Following our assumption, some of these pre-processing steps (such as baseline correction or warping) are no longer critical when using 2D data sources[2]. But some others are still crucial.

In the previous Sections 1.2.2.1 and 1.2.2.2, the bucketing step has been highlighted once again for controlling the dimensionality and to fix the resolution of the final global matrices (obtained via GPL or Vectorization).

Along with bucketing, negative value zeroing (mainly via the ACD/Labs detectability threshold) and normalization (for instance Constant Sum) of the

[2]For instance, the baseline correction issue is naturally solved by using the positive ACD/Labs detectability threshold in the 2D GPL workflow.

intensities are still necessary for 2D data. Moreover, COSY correlation peaks have to be symmetric around the diagonal. So small non-symmetric artefacts have to be removed[3].

Also, even if a presaturation sequence is applied previously as for 1D (Section 1.1.2), the water zone, as an important source of unwanted and over-represented variations, must be assigned to zero with a typical deletion of a square area between 4.5 and 5.5 ppm. Some other zones have to be deleted too, according to the nature of the used medium ([Martin et al., 2017]) and to the user's objective.

Finally, a log transformation of the intensities may be considered in order to stabilize distribution variances, especially for the GPL objects.

1.3 Evaluation of the quality of the data sources

At this point of the work, it is important to understand that numerous potential data sources can be generated within the framework of a single metabolomics study, by taking into consideration:

- different possible data acquisition techniques (1D NMR, 2D NMR, MS, ...),
- if ^{1}H-NMR is chosen, different scenarii during the pre-processing steps shown in Section 1.2.1,
- if 2D NMR is chosen, different possible experiments (COSY, TOCSY, HSQC, ...),
- if COSY is chosen, different upstream spectroscopic parameters,
- the choice between GPL or Vectorization for any 2D experiment,
- if COSY and GPL are chosen (for instance), different values for the ACD/Labs threshold and/or for the number of decimal(s) during the soft bucketing step,
- etc...

In other words, the amount and the variety of potentially available data can be more and more massive under the effect of multiple acquisition methods, multiple parameters of acquisition and/or multiple levels of pre-processing.

[3]This is the case for COSY and for any homonuclear 2D data sources pre-processing. This experiments involve symmetric spectra by construction. By opposition, heteronuclear experiments, such as HSQC, HMBC and HMQC, do not provide symmetric spectra.

Of course, it is rarely interesting or even possible to compare a high number of different data sources. In NMR-based metabolomics, as already said, ^{1}H-NMR is widely used and 2D (COSY or other) is still not yet widespread. So this thesis is particularly focused on the comparison between ^{1}H-NMR and COSY experiments (and between different COSY experiments), with the basic idea to demonstrate the contribution and the informative relevance of the COSY's second dimension.

Once one have a set of global objects for a same experimental study, Global Peak Lists or Vectorized spectra, stemming from various media (urine, blood, serum, ...), possible various experiments (here inspired by COSY) or from the variation of some parameters of acquisition at any step of the workflows, a natural process is to determine which one is the "best" to consider for 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 noisy 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, and as already highlighted, the presence of groups to recover in the data is crucial and helpful by taking the role of the above-mentioned signal.

On the basis of these groups (different blood or urine donors, different culture cells media, ...), a natural way to measure the quality of data sources is to visualize to what extent we can blindly retrieve these groups via clustering analyses.

Independently of any external source of noise and with the use of a same clustering scheme, if a data source allows to discriminate the groups in a better way than other data sources, this source will be considered to have a better *Informative Content*.

Having approached briefly the question of the detection of outliers in Section 1.3.1, Section 1.3.2 is devoted to the definition of the **Metabolomic Informative Content (MIC)**. To objectivize the data sources quality and to propose a hierarchy of these data sources, clustering numerical quality indexes are needed and are regrouped into the MIC principle.

Sections 1.3.3 and 1.3.4 show the potential additional contributions of the use of ASCA (ANOVA-Simultaneous Component Analysis), APCA (ANOVA-Principal Component Analysis) and ICA (Independent Component Analysis) to measure the quality of data sources.

1.3.1 Outlier detection

After pre-processing (and normalization in particular) and before any statistical analysis, it is important to detect outliers in the initial pool of spectral measures. For instance, some spectra can contain unexpected extreme signal values due to exceptional external factors or instrumental concerns. These outliers should be identified before data analysis and removed because they could be too influential.

PCA can offer a first graphical indicator when a sample is projected far away from the others in a first two components' score plots (see Fig.1.15). In parallel, Euclidean and Mahalanobis distances can be calculated between raw spectra and/or between group centered spectra in order to detect anomalous measurements.

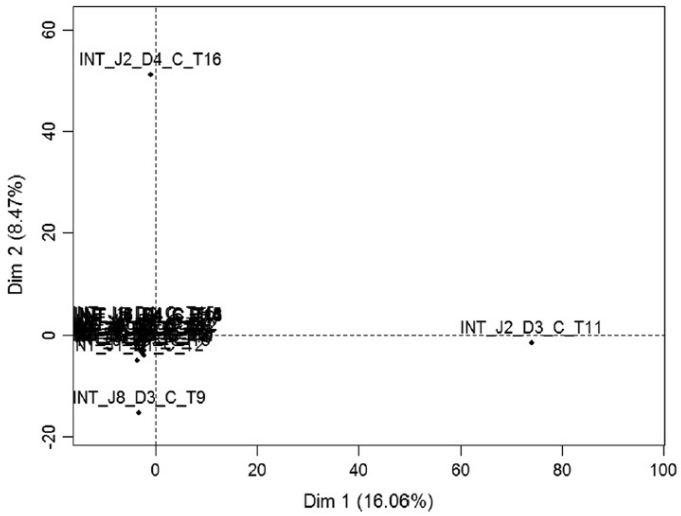


Figure 1.15: Graphical identification of outliers using PCA's score plot (from Paper I).

A secondary work at the beginning of this thesis was to try to go beyond PCA analyses with the use of clustering based anomaly detection (AD) algorithms. Basic anomaly detection assumes that outliers are very rare compared to normal data and that outliers are "different" with respect to their feature values. In this context, differences can be considered between local and global anomalies (see an illustration with Fig.1.16).

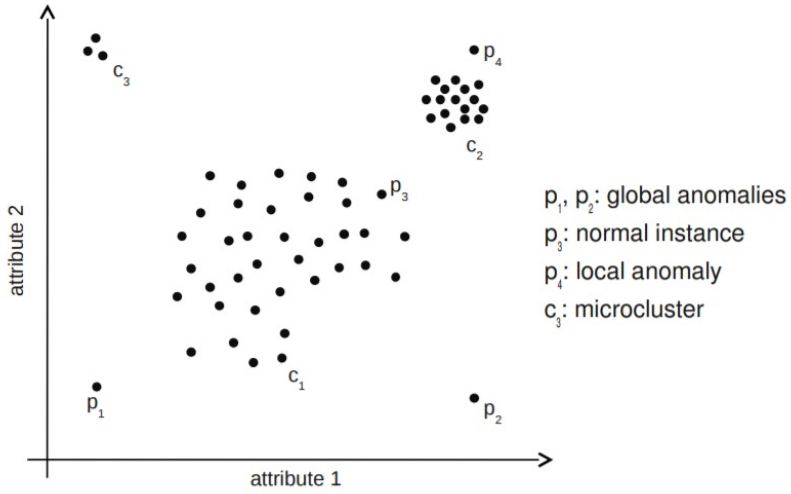


Figure 1.16: Local vs. global anomalies (from [Breunig et al., 2000]).

Based on the k-Nearest Neighbors (k-NN) principle, the Local Outlier Factor (LOF) algorithm is the most used AD algorithm [Breunig et al., 2000]. LOF is based on a concept of a local density, where locality is given by k-NN (whose euclidean distance is used to estimate the density). By comparing the local density of an object to the local densities of its neighbors, one can identify regions of similar density and points that have a substantially lower density than their neighbors. These are considered to be outliers (see a simple illustration, with $k = 3$, in Fig.1.17).

The local density is estimated by the typical distance at which a point can be "reached" from its neighbors. The definition of "reachability distance" used in LOF is an additional measure to produce more stable results within clusters.

Formally, let $kdist(A)$ be the distance of the object A to the k-th nearest neighbor. Note that the set of the k nearest neighbors includes all objects at this distance, which can in the case of a tie be more than k objects. This set is denoted as $N_k(A)$. This distance is used to define what is called "reachability distance":

$$reachdist_k(A, B) = \max\{kdist(B), d(A, B)\} \tag{1.2}$$

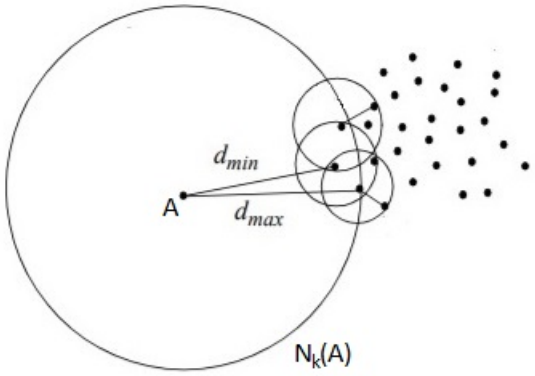


Figure 1.17: Illustration of the LOF principle ($k = 3$) (from [Breunig et al., 2000]).

In words, the reachability distance of an object A from B is the true distance between the two objects, but at least $kdist(B)$ to get more stable results. The local reachability density of an object A is then defined by:

$$lrd(A) = \frac{1}{\left(\frac{\sum_{B \in N_k(A)} reachdist_k(A,B)}{|N_k(A)|}\right)} \tag{1.3}$$

which is the inverse of the average reachability distance of the object A from its neighbors. The local reachability densities are then compared with those of the neighbors using:

$$LOF_k(A) = \frac{\sum_{B \in N_k(A)} \frac{lrd(B)}{lrd(A)}}{|N_k(A)|} \tag{1.4}$$

which is the average local reachability density of the neighbors divided by the object's own local reachability density. A value of approximately 1 indicates that the object is comparable to its neighbors (and thus not an outlier). A value below 1 indicates a denser region, while values significantly larger than 1 indicate outliers.

Based on LOF, other algorithms exist such as Local Outlier Probability (LoOP, [Kriegel et al., 2009]) and Local Correlation Integral (LOCI, [Papadimitriou et al., 2003]), among others. LoOP uses normal distribution for density estima-

tion while LOCI can grow the considered neighborhood from k to a maximum (thus considering local and global possible outliers together).

LOF, LoOP and LOCI have been tested on 1D data and on 2D GPLs in the context of Paper I studies, and have largely confirmed the graphical indications obtained via PCA. For LOF, LoOP and LOCI, the ELKI open source data mining software was used (`https://elki-project.github.io/`).

1.3.2 The Metabolomic Informative Content (MIC)

The principle of the Metabolomic Informative Content (MIC, published in Paper I) is to propose a pool of objective measures to resume the pre-processed data quality when these data contain groups of interest. This principle is related to the notion of signal/noise ratio: we want to measure in what extent a specific data source allows to retrieve and to discriminate the signal (the presence of different groups) in spite of the existence of numerous potential sources of noise (experimental conditions, replicates along time, freezing/defrosting of samples, bacterial degradation of samples, etc...).

Naturally, the presence of groups in the data leads to the use of **unsupervised multivariate clustering methods** (unsupervised because the idea is to try to rediscover the initial groups of interest into blindly formed homogeneous clusters).
These methods are then applied on 1D data matrices, on 2D COSY GPL matrices and/or on matrices resulting from the Vectorization workflow. Practically, the inputs are as follows:

- $(n \times m)$ pre-processed intensity data matrices for 1D NMR data sources;
- $(n \times {M_V}^2)$ intensity matrices resulting from the 2D Vectorization workflow (${M_V}^2$ being equal to 65536 in practice);
- for the GPL workflow, a proper GPL matrix is of size $(T \times M_{GPL})$, with M_{GPL} being equal to $2+n$ or $2+2n$ according to if binary positions are take into account or not (see Section 1.2.2.1). From this GPL matrix, an intensity (or position) matrix can be obtained by keeping the n intensity (or position) columns, transposing the matrix and allowing the couple of ppm coordinates as new column names. Final input GPL objets are then of size $(n \times T)$ and the intensities are pre-processed (negative value zeroing, water zone removal, normalization).

As for any clustering analysis, the choices of a distance measure between individuals and of a clustering algorithm are needed. These two points are

shortly discussed in Sections 1.3.2.1 and 1.3.2.2. Then, Section 1.3.2.3 details the selected pool of quality indexes which are grouped together in the MIC concept.

1.3.2.1 The choice of a between samples distance measure

When an experimental design involves numerous measures (e.g. numerous spectra), with replicates and different groups, it is crucial to choose the data source which captures the useful information, the signal, in the best possible way. To perform that, clustering algorithms firstly need the choice of an appropriate similarity measure, or distance, to quantify the neighborhood relation between two objects.

As mentioned in Paper I, when working on continuous normalized intensities, log-transformed or not, the euclidean distance is systematically used because it is widely known and recognized ([Taylor et al., 2002]).

When working on binary data, like the presence/absence of a peak at a given ppm position as it was also tested in Paper I, the choice of an adequate distance measure is not as obvious. In this case, the devoted Jaccard and Ochiai similarity measures were chosen (see Paper I, Section 2.3.4, for references and mathematical details).
Obviously, two spectra with many common peaks are supposed to provide high Jaccard and Ochiai measures. The Jaccard measure can be interpreted as the size of the intersection divided by the size of the union of the sample sets; and the Ochiai measure can be interpreted as a geometric mean of the probabilities that if one object has the attribute, the other object has it too. Because product increases weaker than sum when only one of the terms grows, Ochiai will be really high only if both of the two proportions (probabilities) are high. It implies that the objects must share a great part of their attributes to be considered similar by Ochiai (Paper I).

1.3.2.2 The choice of a clustering algorithm

Among the very impressive literature on this subject, only two well-known clustering algorithms were finally considered: the hierarchical Ward algorithm and the K-Means algorithm (for continuous intensities only in this second case[4]). Details and references are fully provided in Paper I (Section 2.3.4).

[4]The K-Means algorithm does not use a proper distance matrix, does not work on pairwise distances but only needs the deviation of a point from a center. The core idea of K-Means is minimizing variance (which is the same as minimizing squared euclidean distances). If a Jaccard or Ochiai distance matrix is plugged into K-Means, it will often yield a somewhat useable result, but quite incorrect. Rather than comparing points by Jaccard or Ochiai, it

This choice to consider only very popular and quite intuitive algorithms is explained by a will to be easily understood by final metabolomics practitioners and users for further implementation and decision making (illustrations available in [Steuer et al., 2007] and in [Liland, 2011]).

Note that a preliminary denoising step might be possible by using sparse clustering techniques such as sparse hierarchical clustering and sparse K-Means ([Witten and Tibshirani, 2010]) before MIC indexes calculations. Nevertheless, the central notion of sparsity will be introduced later, in Section 1.4.4, in the context of final biomarker discovery.

1.3.2.3 The clustering quality measures gathered in the MIC concept

The MIC concept proposes a methodology to diagnose, quantify and compare with adequate indexes the useful information (in a sense of advantageous signal capture compared to noise) contained in different sets of spectra. Thus, **the MIC can be considered as a development and validation tool for any available spectral acquisition methods. The MIC concept represents the second contribution of this thesis**.

Among many possibilities, it includes PCA analyses, inertia measures and the calculation of four complementary quality criteria: the Dunn, Davies-Bouldin, Rand and Adjusted Rand indexes. The two first evaluate the homogeneity of the resulting clusters regardless of the initial groups content; the two last measure the correctness of the unsupervised clustering process according to the true initial groups (as would perform classic misclassification rates).

The use and the value of PCA descriptive analyses were already mentioned by graphically visualizing the individual objects, and potentially the groups, in the plane formed by the first two principal components. Beyond the search for outliers (see Section 1.3.1), PCA provides initial reports about the facility in separating the groups.

Inertia analyses then propose a global view on separabily and homogeneity of final obtained clusters. Let P_K be a partition of the set $\Omega = \{1, ..., i, ...n\}$ of n individuals (measures) in a set $(C_1, ..., C_k, ..., C_K)$ of non empty and non-overlapping K clusters, via an initial data matrix X. Using:

would cluster them by squared euclidean of their distance vectors. That is why K-Means expects continuous numerical input data for clustering, and does not support arbitrary distance functions ([Raykov et al., 2016]).

- weights w_i associated with each individual i, usually $w_i = \frac{1}{n}$ or simply $w_i = 1$,
- the cluster's weight

$$g_k = \sum_{i \in C_k} w_i \tag{1.5}$$

- the gravity center of the cluster

$$\mu_k = \frac{1}{g_k} \sum_{i \in C_k} w_i X_i \tag{1.6}$$

- and the squared euclidean distance between two lines of X, X_i and $X_{i'}$

$$d^2(X_i, X_{i'}) = \sum_{j=1}^{m} (x_{ij} - x_{i'j})^2, \tag{1.7}$$

the total inertia T measures the dispersion of the cloud of n individuals as (where $\boldsymbol{\mu}$ is the gravity center of all the individuals):

$$T = \sum_{i=1}^{n} w_i d^2(X_i, \boldsymbol{\mu}) \tag{1.8}$$

This total inertia is independant of the partition P_K. The between-clusters inertia B of P_K is the inertia of the gravity centers of the clusters weighted by g_k and measures then the separation between the clusters:

$$B = \sum_{k=1}^{K} g_k d^2(\mu_k, \boldsymbol{\mu}) \tag{1.9}$$

The within-cluster inertia W of the partition P_K is the sum of the inertia of the clusters and measures then the heterogeneity within the clusters:

$$W = \sum_{k=1}^{K} \sum_{i \in C_k} w_i d^2(X_i, \mu_k) \tag{1.10}$$

Ideally, a good final partition has a large between-cluster inertia and a small within-cluster inertia. Note that the inertia (total, between and within) calculated with the standardized data are equal to the inertia calculated with all the p corresponding principal components of X (standardized).

Having $T = W + B$, minimizing the within-cluster inertia (the heterogeneity) is equivalent to maximizing the between-clusters inertia (the separability).

Finally, the MIC concept provides a pool of four critera (Dunn, Davies-Bouldin, Rand and Adjusted Rand) to diagnose the quality of a partition. In particular, the choice of these four criteria was motivated by the will to consider and evaluate at the same time the homogeneity of the clusters and the clusters' accuracy. MIC's users may then choose to focus on one of these two objectives, or to consider them in average.

As mentioned, the Dunn and Davies-Bouldin indexes, respectively referenced as DI and DBI, evaluate the homogeneity of clusters regardless of the initial groups content. Consequently, they can be related to the within-cluster inertia measure W.

The Rand and adjusted-Rand indexes, referenced as RI and ARI respectively, quantify the correctness of the unsupervised clustering process according to the true groups. A perfect partition, with no misclassification, would lead to RI=ARI=1.

Note that DI, RI and ARI have to be as high as possible to represent good clustering partitions. DBI has to be as small as possible. Finally, all the candidate data sources may then be ranked according to these MIC performances.

For full mathematical definitions and references, and to avoid redundancy, please see Paper I (Section 2.3.5).

Generally, it is important to understand that MIC analyses can be fully performed on any spectral data sources, issued from NMR (1D or 2D), Mass Spectrometry, among others... In the context of this work, the MIC concept is principally applied in order to compare 1D proton-NMR and 2D COSY spectral data (using different scenarii and pre-processing workflows as seen before). Papers I and III contain real studies based on MIC results. Note that MIC also became a usually used tool in the Institute of Statistics for metabolomics purposes.

Because MIC indexes are built to assess and compare different spectral matrices having different dimensionalities (i.e. different numbers of descriptors m), an important question may arise which is linked with the curse of dimensionality issue. Indeed, higher dimensionality may induce a reduction in relevance for some indexes and the question is about the real capacity to compare measures from a dimensionality to another.

This question is briefly discussed in Appendix (Section 1.7) on the basis of some experiments. They show that the comparison of MIC clustering quality measures when different numbers of features/descriptors (m) are considered has to be done with caution, more especially for homogeneity indexes and

inertia (accuracy indexes are more stable) and when a significant amount of noise exists in the data.

The MIC procedure is fully implemented in R via the *ClustMIC*, *binClustMIC* and *Inertia* functions available in `https://github.com/ManonMartin/MBXUCL`.

1.3.3 The potential use of ASCA and APCA

In metabolomics studies, high dimensionality, with high-dimensional multivariate databases where the number of variables (descriptors) tends to be much larger than the number of experimental units, is a common issue. The clustering approaches of Section 1.3.2 are adapted to this problem and are able to measure the quality of the data sources' informative content.

In addition, some other approaches, based on other algorithms or concepts, can be useful to answer this same question. For instance ASCA (ANOVA-Simultaneous Component Analysis) and APCA (ANOVA-Principal Component Analysis), and their ASCA+ and APCA+ extensions for unbalanced experimental designs, can offer an other perspective. These methods are presented in details and applied in [Thiel et al., 2017] (paper which I am co-author).

Basically, ANOVA is a well-known statistical method to study the behaviour of a continuous variable depending on one or more independent categorical variables called "factors", whose possible values are called "levels". In this context, the use of ANOVA I, or of multifactorial ANOVA if other potential effects have to be considered, can be applied here if group membership is considered as a proper factor in the initial experimental design. The continuous response variable would then be the intensities of spectral peaks, stemming from 1D spectra, 2D GPL matrices or global matrices resulting from the Vectorization workflow.

ASCA and APCA are extensions of ANOVA to multivariate cases. In an ANOVA II (considering two factors), the y vector can be decomposed into vectors of parameters as:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk} \tag{1.11}$$

for $(i = 1, ..., a; j = 1, ..., b; k = 1, ..., n)$ and where y_{ijk} is the observed response for replicate k of level i of factor A, and level j of factor B. Note a the number of levels of factor A, b the number of levels of factor B and n the number of replicates of each treatment. Let us also define N as the total number of experiments, with $N = a \times b \times n$ in crossed ANOVA II ([Thiel et al., 2017]).

In ASCA and APCA, a matrix Y of dimension $(N \times m)$ is decomposed in matrices of effects that have the same dimensions $(N \times m)$, leading to:

$$Y = M_0 + M_A + M_B + M_{AB} + E \tag{1.12}$$

where each effect matrix regroups the estimated ANOVA II model parameters for each of the m response columns of Y.

In ASCA, a PCA decomposition is then applied on these different effect matrices such as:

$$Y = M_0 + T_A P'_A + T_B P'_B + T_{AB} P'_{AB} + E \tag{1.13}$$

where the T and P matrices correspond to scores and loadings for each effect.

In APCA, the PCA decomposition is not directly performed on effect matrices but on the addition of the effect matrices and the error matrix (for instance on $M_A + E$ instead of only M_A). Score plots obtained in APCA show the groups of observations for the different levels of the variables, whereas ASCA shows only means of these levels.

Having briefly introduced ANOVA, ASCA and APCA, the interest here is then to quantify a percentage of variation explained by each potential effect of the design (if in a multivariate case), and very particularly the group effect. [Vis et al., 2007] proposes to calculate the sums of squares of the matrices of effects with the square of the Frobenius norm defined as follows for the Y matrix:

$$\|Y\|^2 = \sum_{i=1}^{N} \sum_{j=1}^{m} |y_{ij}|^2 \tag{1.14}$$

In balanced situation, the percentage of variation of Y explained by a proper group effect would then be obtained by:

$$\%Var_{group} = \frac{\|M_{group}\|^2}{\|Y\|^2 - \|M_0\|^2} \times 100 \tag{1.15}$$

It must be noted that the results obtained by the ASCA case involving only one factor (i.e. the group factor only here) are equivalent to inertia calculations. The real interest here is the potential simultaneous study of other effects linked with the data (with factors like the laboratories, the operators or other categorical metadata considerations) in addition to the group factor.

In unbalanced situation, the ASCA+ and APCA+ extensions proposed in [Thiel et al., 2017] also allow to calculate such percentages of variation corresponding to each effect.

Finally, if different data sources (involving different acquisition techniques or different pre-processing parameters or workflows as already mentioned) are compared, it would be very interesting to at least calculate $\%Var_{group}$ and $\%Var_{error}$ in each case and to rank the data sources accordingly. Other factors are welcome when using ASCA or APCA as they can be considered simultaneously.

1.3.4 The potential use of ICA

Another possible approach to measure the informational quality of data sources is the Independant Component Analysis (ICA) [Hyvarinen and al., 2004]. In signal processing, ICA is a computational method for separating a multivariate signal into additive subcomponents. This is done by assuming that these subcomponents are non-Gaussian signals and that they are statistically independent from each other. ICA is then a tool for blind source separation. The question is whether it is possible to separate these contributing sources from the observed total signal.

The ICA model is presented in [Feraud et al., 2017]. Note that this paper is mainly Rejane Rousseau's work and was published later thanks to some personal enhancements. In this paper, and in the context of ^{1}H-NMR metabolomics data, the analyzed biofluid (e.g. serum, urine) can be seen as a mixture of individual metabolites; NMR spectra may then be interpreted as weighted sums of NMR spectra of these single metabolites. If the data matrix X is rich enough, the application of ICA on these data should then ideally recover components included in the mixture, interpretable as spectra of pure or complex metabolites. Consequently, each ICA source is finally connected with a particular descriptor or metabolite.

Nevertheless, the use of ICA in [Feraud et al., 2017] was strictly focused on biomarkers identification. But ICA can also be used as a pre-treatment step to turn the spectral data sets X into matrices of independant sources' weights. Then, these latter matrices can be submitted to MIC indexes calculations (Section 1.3.2) in order to evaluate their ability to detect a main signal and to be ordered accordingly.

Practically, let X be an initial $(n \times m)$ spectral data matrix, for instance acquired via ^{1}H-NMR experiments. ICA assumes that X' can be modelled

as linear combination of p ($p \leq m$) independent sources vectors, grouped in $S = (\mathbf{s}_1, ..., \mathbf{s}_p)$, with some matrix of unknown coefficients A, named mixing matrix, such that:

$$X' = SA' \tag{1.16}$$

Both S and A are unknown. The "unmixing" problem considered by ICA is typically to estimate the independant sources S. Solving this ICA problem aims then at finding a theoretical matrix that maximizes the non-gaussianity of the estimated sources, under the constraint that their variances are constant. The non-gaussianity may be estimated by the negentropy, as in the frequently used FastICA algorithm [Hyvarinen, 1999].

An adapted ICA analysis with FastICA can be applied on metabolomics data as follows:

- Pre-processing step 1: transpose the X spectral matrix and center it by columns.
- Pre-processing step 2 ("Whitening"): reduce by PCA the centered matrix X' to a $(m \times p)$ matrix of scores T ($p \leq \min(n, m)$): $X' = T^\star P^\star = TP + E$. The column vectors of the PCA score matrix $T^\star$ are centered, uncorrelated and their variances are equal to one. $P^\star$ is a $(n \times n)$ matrix defined according to the eigenvectors of the covariance matrix $((X')'X')/n$. Note that this PCA differs from the usual PCA in metabolomics as the created components are linear combinations of observations (spectra) and not of variables (spectral descriptors), and as centering is done by spectra and not by descriptors. The number of sources p to be estimated can be fixed to less than $\min(m, n)$. This is performed by selecting the p first scores vectors (columns) of $T^\star$. A matrix T of dimension $(m \times p)$ can then be built. P is finally defined as the p first lines of P^* and E is the error matrix ($= 0$ if $p = \min(m, n)$).
- Application of ICA to T and calculation of S and A'. The FastICA algorithm, with parallel extraction of components, proceeds in the following steps:
  - compute a $(p \times p)$ unmixing matrix W such that $TW = S$ where S is the $(m \times p)$ matrix which contains the independent sources. W is chosen to maximize the negentropy of the columns of S. As the variance of TW must be constrained to unity for the whitened data, this is equivalent to constrain the norm of W to be unity.
  - define the mixing matrix A as:

$$A' = W^{-1}P \tag{1.17}$$

in order to obtain the required following ICA decomposition: $X' = SA' + E$.

The $(m \times p)$ matrix S contains p estimated independent components (IC). Each IC has a zero mean and a unit variance and at least $(p-1)$ sources are non gaussian distributed. The A mixing matrix, the transposed matrix of $A' = W^{-1}P$, is of dimension $(n \times p)$. Each column of A, a_j is then a $(n \times 1)$ vector containing the weights or contributions of the corresponding source s_j in the construction of the n observed spectra.

Finally, to evaluate different data sources, different $(n \times p)$ mixing matrices A can be obtained and considered as new inputs in hierarchical clustering analyses using Ward's method or K-means. MIC quality indexes can then be calculated strictly in the same way as previously to evaluate the clustering quality and the informative content associated with any available data source.

1.4 The relevant biomarker discovery issue in metabolomics

At this stage of this work, spectral data issued from different 1D or 2D experiments have been handled, pre-processed and transformed into global objects (Section 1.2). Then, these objects are diagnosed and compared via the MIC quality criteria in order to highlight their separability power and to measure their Metabolomic Informative Content when capturing the main signal (Section 1.3).

Consequently, it is now possible to determine which data source(s) is(are) the most suitable for further statistical modelling.

Indeed, the intuitive next step of most clustering analyses consists in understanding why the clusters have been determined in this way and if there exists at least a relevant variable (a descriptor, a feature) which can explain the different assignments of individuals between these clusters. Here, the term "variable" or "feature" refers to spectral zones, to spectral peaks, which are ideally and finally interpretable in terms of metabolites. All these considerations lead to the biomarkers discovery issue.

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, 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 this section, a chronological review of the techniques used in metabolomics is proposed, from basic PCA (Section 1.4.1) to advanced sparse algorithms (Light-sparse-OPLS, in Section 1.4.5), including PLS(-DA) (Section 1.4.2), OPLS (Section 1.4.3) and sparse-PLS (Section 1.4.4).

Note here that the decision is not to go away from the Partial Least Squares modelling scheme, because PLS is by far the most used and encountered technique in biomarker discovery studies in metabolomics. In Section 1.4.6, some other very different approaches will however be discussed.

1.4.1 The use of PCA

Since the emergence of the metabolomics science, Principal Component Analyses were used not only for descriptive statistics and/or to detect outliers and/or for groups identification purposes, but also for biomarker identification. PCA is a multivariate exploratory tool for data analysis and remains essential in metabolomics and chemometrics. Its principle goes back to [Pearson, 1901] and [Hotelling, 1933].
In the same way as the closenesses between individuals are highlighted, the individuals/variables duality in PCA allows to quantify the correlations between the measured variables of the dataset. Groups of variables having identical trends are so identified. It is then very informative to visualize the initial variables in the vectorial space formed by the main two principal components (PCs, which are hierarchical and orthogonal). The projection distance of a variable on these PCs corresponds to a "loading", or factorial contribution, which gives information about its weight during the PCs' computation.

As a PC is a linear combination of the initial variables, the loadings directly correspond to the parameters p in the following equations (where max denotes the number of columns, or descriptors, in the input matrix, which can be a 1D grid, a 2D GPL, a matrix issued from the 2D Vectorization workflow, etc...) :

$$\begin{aligned} PC1 &= p_{1,1}X_1 + p_{1,2}X_2 + ... + p_{1,max}X_{max} \\ PC2 &= p_{2,1}X_1 + p_{2,2}X_2 + ... + p_{2,max}X_{max} \\ &... \end{aligned} \tag{1.18}$$

Thus, the analysis of these loadings, and in particular those which have the largest absolute values in the first PCs expressions, can help to classify the variables by their importance to explain the whole variability into X ([Brindle et al., 2003]).

Another commonly used index to evaluate variable importance, directly linked with the loadings, is the communality ([Norris and Lecavalier, 2010]). Communalities indicate the amount of variance in each variable that is accounted for. Initial communalities are estimates of the variance of each variable accounted for by all components. When only considering the first two PCs, the communality for variable j is obtained by:

$$Comm_j = p_{1,j}^2 + p_{2,j}^2 \tag{1.19}$$

A high communality (near 1) for a variable means that this variable is well projected in- the first two PCs' plan and that it plays a relevant role to explain the total variance in X.

Of course, the main drawback here is that classical PCA is by construction unsupervised. The interpretation of high loadings in absolute values or of communalities gives only indications on the importance of variables to explain the total variance in the X data matrix.
The group membership, the main signal we want to explain and to use to discover final biomarkers, contributes to this total variance but is not this total variance. In other words, the group membership must be taken into account specifically and explicitly, and the problem has to become supervised in order to gain relevance.

This had led to the extensive use of (O)PLS regression models instead of PCA ones in metabolomics (Section 1.4.2).

Note that sparse principal components solutions, interested in obtaining a decomposition made up of sparse vectors, might be possible at this point by applying some Penalized Matrix Decomposition (PMD, [Witten et al., 2009], [Jolliffe et al., 2003]). Nevertheless, the central notion of sparsity will be introduced later, in Section 1.4.4, in the context of final biomarker discovery using supervised and oriented models.

1.4.2 The PLS(-DA) approach

In metabolomics and chemometrics, as already mentioned previously, high dimensionality is a common issue as the initial number of variables is much

greater than the number of individuals. This is why the classical linear regression model, which implies the well known $\hat{\beta} = (X'X)^{-1}XY$ least squares solution, becomes impossible to solve in this case, especially since the variables tend to be highly correlated. The PLS (Partial Least Squares) regression allows to tackle this issue and rapidly became the most widespread and used regression method in the field of metabolomics.

1.4.2.1 PLS modelling

Instead of PCA which strictly decomposes X in loadings and scores vectors, the PLS methodology allows to use the information of at least one dependent variable Y to build the decomposition. PLS is based on a simultaneous variability modelling into X and Y, with the calculation of latent variables which maximise the extracted variance of both matrices at the same time as their correlation.

Practically, PLS consists in decomposing X and Y under the constraint that the factorial coordinates T, extracted from X, are as correlated as possible with the factorial coordinates U, extracted from Y (see Fig.1.18).

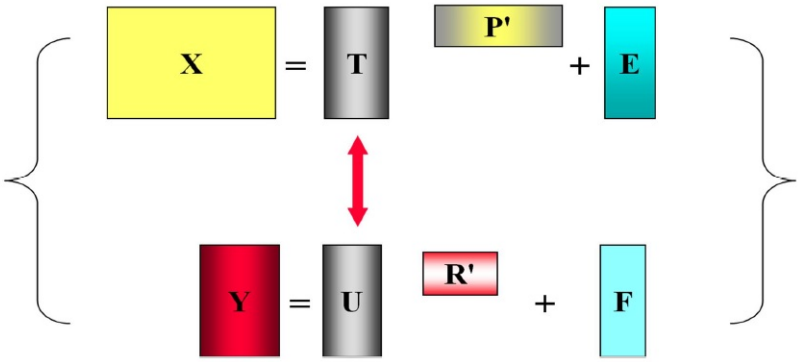


Figure 1.18: Calculation of X and Y factorial coordinates.

In other words, the fundamental objective of PLS implies a computation of T and U scores, in order to capture a high amount of variance in the X and Y matrices and also a high amount of "correlations" between T and U. 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.

Note that, in the context of this thesis, Y is simplified as a single categorical vector (and can be noted y) which contains the group membership of the indi-

viduals. With such a categorical target, the PLS regression can be interpreted as a PLS-DA (Discriminant Analysis) problem.

More formally, consider a data matrix X with n observations and m variables and 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 $W = (w_1, ..., w_q)$ from successive optimization problems. For the iterative computation of each component $k = 1, ..., q$, the PLS criterion to optimize can be written as follows ([Wold et al., 2001]) :

$$\max_{w} \left\{ corr^2(y, Xw_k) var(Xw_k) \right\} \equiv \max_{w} \{ w_k' X_k' y y' X_k w_k \} \tag{1.20}$$

subject to $w_k' w_k = 1$, where w_k is the PLS X-weights vector for dimension k. Note that X and y are supposed to be previously centered.

Computationally, two main PLS algorithms, among others, are available and very popular according to some specificities: **NIPALS** (Non-linear Iterative Partial Least Squares [Wold H., 1975]) and **SIMPLS** (Straightforward Implementation of a Modification of PLS [De Jong, 1993]). Unlike NIPALS, SIMPLS is actually derived to solve a specific objective function, i.e. to **maximize covariance** for a multivariate Y. Note that NIPALS predictions are equal to SIMPLS predictions when considering an univariate response y ([Wehrens, 2011]). In this paper, the SIMPLS algorithm is used for each subsequent PLS routine because of its higher computational speed. **Formal and algorithmic details for SIMPLS** are developed and referenced in Paper II (see Section 3.3.1).

PLS discriminant analysis (PLS-DA) is indeed a particular case of PLS regression aiming at predicting a binary responses y 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, ^{1}H-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 in absolute values are naturally linked with discriminating zones.

The goal is to obtain a model which can well explain y and which also avoids overfitting. Mathematical regression models can be built in order to fit closer and closer the data by simply increasing q (see Fig.1.19, a first illustration found in the Chemooc[5] slides), what can be bad in terms of generalization and

[5] `https://chemproject.org/wakka.php?wiki=PresentationChemoocs`

deployment of the model (this is linked with the notion of overfitting when the use of too many latent components implies the inclusion of noise in the final coefficient vector, see Fig.1.20).

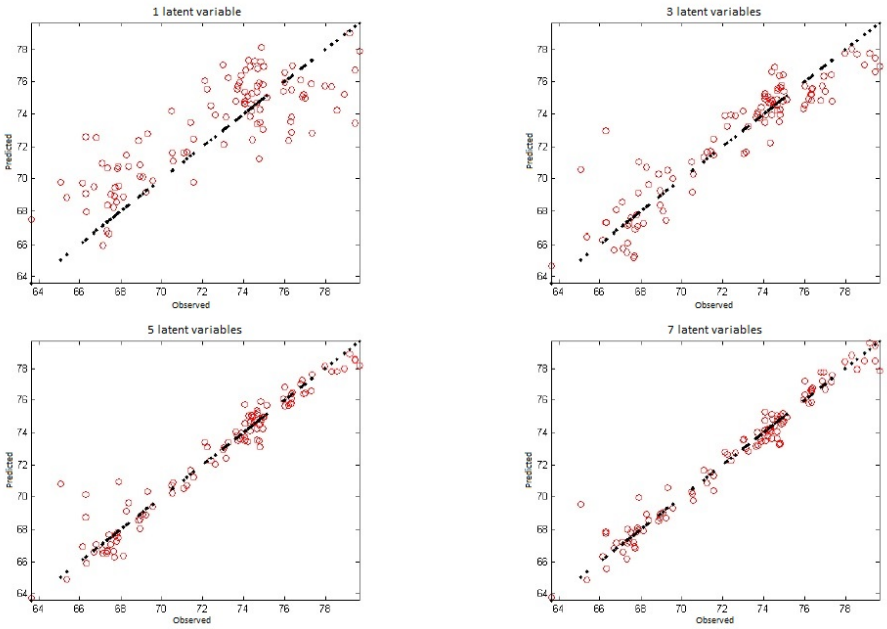


Figure 1.19: Evolution of predicted values according to the number of latent components q in PLS model (from Chemooc slides).

Therefore, the number q of predictive components (ot latent variables) should not be too high to avoid overfitting. To derive a suitable PLS(-DA) model, this number has to be optimally chosen using an adequate validation criterion. The Root Mean Squared Error of Prediction (RMSEP, [Mevik and Cederkvist, 2004]) was chosen as such a validation criterion. It must be minimized and can be calculated for each size of model by k-fold or Leave-One-Out (LOO) cross-validation.

Note that working with a fully external and independent validation/test set is very often too expensive and not possible in most metabolomics studies.

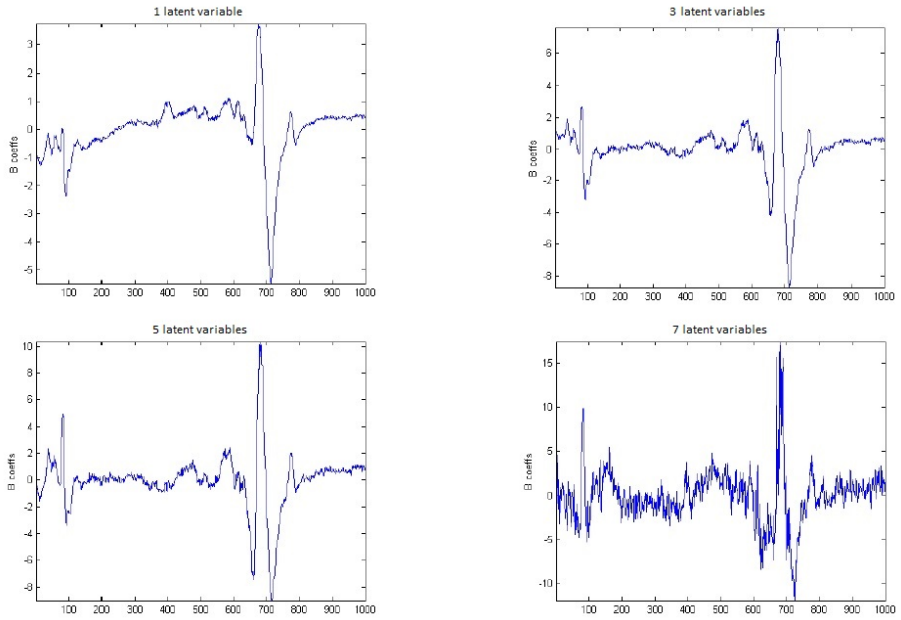


Figure 1.20: Evolution of regression coefficient vectors according to the number of latent components q in PLS model (from Chemooc slides).

1.4.2.2 Highlighting important descriptors as biomarkers with PLS

Finally, the matrices of scores $T = XW$ and of loadings $P = X'T$ are calculated based on the optimal q. Final coefficients of the PLS model are calculated as:

$$b_{PLS} = W(T'T)^{-1}T'y \tag{1.21}$$

and potentially allow to make subsequent predictions on new observations, by simply using:

$$\hat{y} = Xb_{PLS} \tag{1.22}$$

From these results, the descriptors with highest (absolute) coefficients can therefore be considered as primary candidate biomarkers. The decision to consider biomarkers can depend then on such rankings. Note that if X is not standardized or normalized, the final coefficients may change because of the occurence of different scales.

In the PLS literature, two main techniques also exist to highlight important features as biomarkers (independently of scales effects).

The first one is the use of S-plot [Wiklund et al., 2008]. S-plots combine the modelled covariances and modelled correlations between X variables and the scores of each latent component (or with y predicted values) from the (O)PLS-DA model in a scatter plot. Covariances and correlations are respectively calculated as follows for the first scores vector t_1:

$$Cov(t_1, X) = \frac{t_1' \times X}{n-1} \tag{1.23}$$

$$Corr(t_1, X) = \frac{Cov(t_1, X)}{\sigma_{t_1} \sigma_X} \tag{1.24}$$

If the data X has variation in peak intensities, S-plots will effectively look like a S. The covariance axis describes the magnitude of each variable in X. The correlation axis, always between -1 and 1, represents the reliability of each variable in X.

Variables with high covariances and correlations are declared more important for the latent component's contruction (see an illustration with Fig.1.21). Ideal biomarker have high magnitude and high reliability, implying a smaller risk for spurious correlations.

Finally, the second common tool is the Variable Importance in the Prediction (VIP) criterion. VIP allows to estimate the importance of variable X_j ($j = 1, ..., m$) for the prediction of y in a PLS model with q components using the following formula ([Mehmood and al., 2011]):

$$VIP_j = \sqrt{m \times \frac{\sum_{h=1}^{q} R^2(y, t_h).b_{hj}^2}{\sum_{h=1}^{q} R^2(y, t_h)}} \tag{1.25}$$

where $R^2(y, t_h)$ denotes the coefficient of determination between the y vector and the X-scores t_h associated with the latent component h ($h = 1, ..., q$); and b_{hj}^2 is the squared regression coefficient for variable X_j in the latent component h ($h = 1, ..., q$). This relation takes into account the correlation between the latent scores and the dependent variable y, and the contribution of each variable in X to the latent components of the PLS model. Note that volcano plots can be based on VIP calculations to quickly identify changes in large data sets ([Feng et al., 2016]).

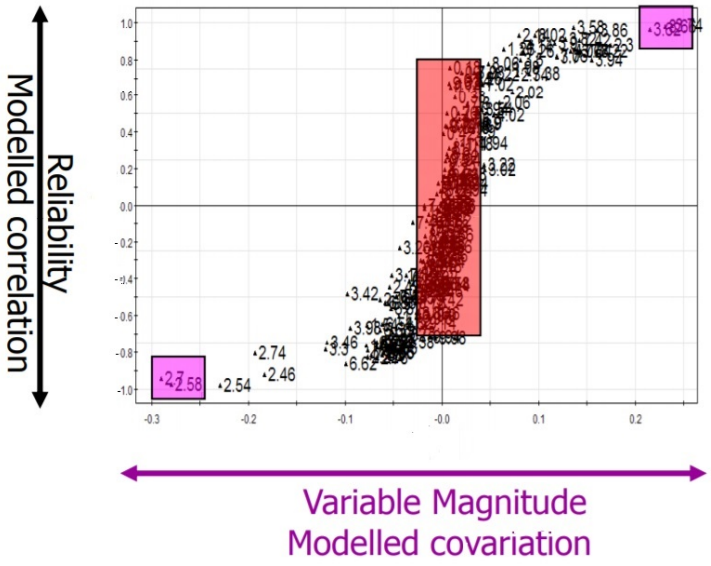


Figure 1.21: Typical S-plot from an (O)PLS-DA modelling.

1.4.3 The OPLS approach

Projection methods such as PLS remain strongly affected by the potential occurrence of systematic variations in the data source that are not relevant for response prediction. This is why Orthogonal Projection to Latent Structures (OPLS, [Trygg and Wold, 2002]) 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 slightly different model since some of the extracted components are identified as corresponding to systematic or specific orthogonal causes of variation. In OPLS language, "predictive variations" are related to correlated variations between X and Y (or y) and "orthogonal variations" to uncorrelated variations between X and Y (or y) (see Fig.1.22).

A generalization of its use in metabolomics studies (several references are available in Paper II) is explained by the relative simplicity of the algorithm, derived from the PLS one, and the ability to manipulate and interpret the resulting predictive and orthogonal matrices. Thus, OPLS provides a significant gain in interpretability for final practitioners compared to standard PLS models.

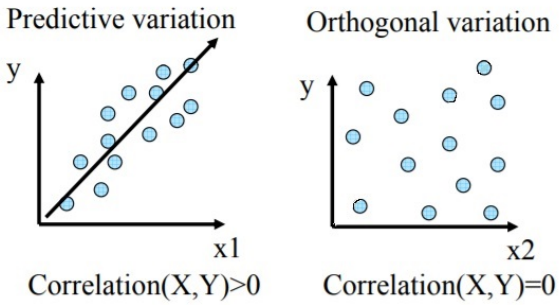


Figure 1.22: Predictive and orthogonal variations in OPLS (from [Wiklund et al., 2008]).

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 classifies and discriminates in the best way the y levels, or groups of interest.

All **technical and algorithmic details for the OPLS methodology**, and the final calculations of OPLS coefficients, are fully available in Paper II (Section 3.3.2). Note an interesting property of OPLS which connects the X-weights ($\tilde{\boldsymbol{w}}$) and the X-loadings ($\boldsymbol{p}$): once all the y-orthogonal part is extracted, thus implying the use of more than $(q-1)$ orthogonal dimensions, one can observe that $\tilde{\boldsymbol{w}} = \boldsymbol{p}$.

When considering a response vector y, it is just important to note that the final corrected OPLS coefficients ($b_{OPLScorr}$, see paper II) obtained with $(q-1)$ orthogonal components and one predictive component are stricly equal to the corresponding PLS coefficients obtained with q latent components [Tapp and Kemsley, 2009]. Therefore, the subsequent predictions will be the same for both models, as:

$$\hat{y}_{OPLS(q-1)} = \hat{y}_{PLS(q)} \tag{1.26}$$

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 optimal

number of orthogonal components $(q-1)$. From (Eq. 1.26), 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.

1.4.4 Introducing the notion of sparsity: the sparse-PLS (sPLS) approach

As explained in Sections 1.4.2 and 1.4.3, the PLS and OPLS solutions strictly depend on the selected number of latent components (predictive or orthogonal). Therefore, this number will have consequences on the calculation of the final (O)PLS coefficients. In (O)PLS, the selection of relevant biomarkers, as mentioned previously, can be based on the higher final coefficients absolute values, on S-plots or on the VIP criterion. These practices are very common in metabolomics studies, but not having concrete thresholds or critical values to determine if a biomarker is really relevant or not remains an important drawback.

For example, if a decision rule is based on highest absolute coefficients, the choice of a threshold to determine what is of interest for further investigations/interpretations and what is not is inevitably needed. But what threshold to use? This question is somehow subjective.
Indeed, most of the time (O)PLS methods lead to a high number of spectral peaks or zones considered as of interest during metabolomics experiments, which quickly induces difficulties of interpretation. Finally, a lighter and probably more interpretable model should be provided to final users for medical decisions, treatment adjustments, etc...

To make this above question more objective, a solution is to keep the more significant biomarkers' coefficients and to force the other ones to be equal to zero. This represents the key basis of the sparsity concept and leads to the additional application of penalties. General aspects of such penalties are discussed in Section 1.4.4.1. In this context, the use of sPLS (sparse-PLS) is discussed in Section 1.4.4.2. Considering the massive use of PLS in metabolomics studies, sPLS is the natural tool for combining the efficience in tackling the $m >> n$ characteristic in the X data matrix and for proposing sparse interpretable solutions.

1.4.4.1 The use of penalty terms to obtain sparse solutions: a general overview

The first immediate reason to consider sparsity refers to the curse of dimensionality issue. High-dimensional spaces are by definition vast and data points are isolated in their immensity. Separating signal from noise becomes in general almost impossible in this context. For example, when the dimensionality (m) increases, the notion of "nearest" points vanishes, the minimal distance between two points increases, all the points are at a similar distance between each other and local methods cannot work anymore ([Keogh and Mueen, 2011]).

Fortunately, high-dimensional data are often much more low dimensional that they seem to be. Usually, these data are not uniformely spread in a R^m space, but rather concentrate around smaller "structures". As an oriented and supervised technique, PLS can smartly address this issue if this structure is a sub-space adapted to linear models. But the lack of an objective threshold criterion when highlighting final biomarkers prevents from fully answering the question to separate the signal of the noise.

Thus, additional regularization is needed to simplify the final obtained models and to obtain more stability for prediction purposes (when the curse of dimensionality issue joins the overfitting issue). In this context, sparsity could allow to avoid overfitting, to reach better repetability and generalization and to provide a gain in interpretation.

Historically, sparsity is born with the use of the Lasso (Least Absolute Shrinkage and Selection Operator [Tibshirani, 1996]), which is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces. Lasso was originally formulated for least squares models and this simple case reveals a substantial amount about the behavior of the estimator, including its relationship to ridge regression ([Hoerl and Kennard, 1970] [Hoerl et al., 1975]) and best subset selection and the connections between lasso coefficient estimates and so-called soft thresholding. It also reveals that (like standard linear regression) the coefficient estimates need not to be unique if covariates are collinear.

For least squares, the general objective of lasso is to solve:

$$\min_{\beta_0,\beta} \left\{ \frac{1}{n} \left\| y - \beta_0 - X\beta \right\|_2^2 \right\} \tag{1.27}$$

subject to $\|\beta\|_1 \leq t$, where t is a pre-specified free parameter that determines the amount of regularization.

Let us simply define the L_1 norm as $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$ and the L_2 norm as $\|\mathbf{x}\|_2^2 = \sum_{i=1}^{n} |x_i|^2$ (which is the squared euclidean distance), for any vector $\mathbf{x} = (x_1, ..., x_n)$.

Lasso can set coefficients to zero, while ridge regression, which appears superficially similar, cannot. This is due to the difference in the shape of the constraint boundaries in the two cases. Both lasso and ridge regression can be interpreted as minimizing the same objective function (Eq. 1.27), but with respect to different constraints: $\|\beta\|_1 \leq t$ for lasso and $\|\beta\|_2^2 \leq t$ for ridge regression.

Fig.1.23 shows that the constraint region defined by the L_1 norm is a rotated square so that its corners lie on the axes, while the region defined by the L_2 norm is a circle, which is rotationally invariant and, therefore, has no corners. As seen in the figure, a convex object that lies tangent to the boundary, such as the shown line, is likely to encounter a corner (or in higher dimensions an edge or higher-dimensional equivalent) of a hypercube, for which some components of β are identically zero, while in the case of an n-sphere, the points on the boundary for which some of the components of β are zero are not distinguished from the others and the convex object is no more likely to contact a point at which some components of β are zero than one for which none of them are.

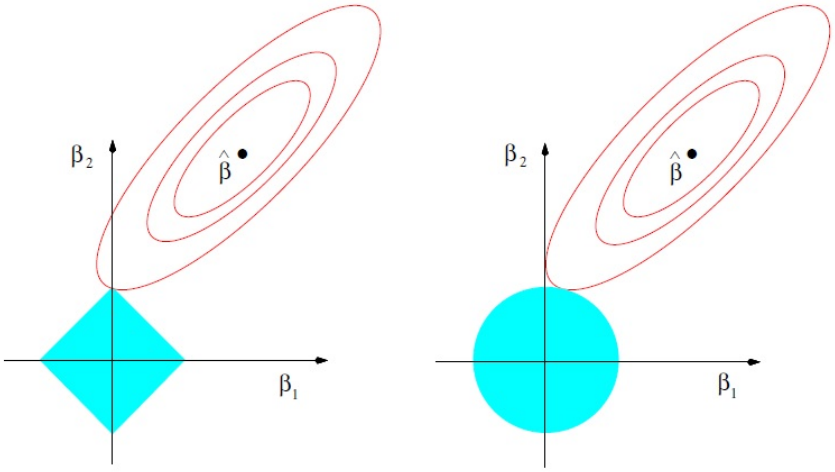


Figure 1.23: Forms of the constraint regions for lasso (left, using the L_1 norm) and ridge regression (right, using the L_2 norm) (from [Tibshirani, 1996]).

A number of lasso variants have been created in order to remedy certain limitations of the original technique and to make the method more useful for particular problems. Almost all of these focus on respecting or using different types of dependencies among the covariates. Among them, the Elastic Net (EN) regularization ([Zou and Hastie, 2005]) adds an additional ridge regression-like penalty which improves performance when the number of predictors is larger than the sample size, allows the method to select strongly correlated variables together and improves overall prediction accuracy. Indeed, when $m > n$ (the number of descriptors is greater than the sample size), lasso can select only n variables (even when more are associated with the outcome) and it tends to select only one covariate from any set of highly correlated covariates.

Concretely, the EN extends lasso by adding an additional L_2 penalty term, giving:

$$\min_{\beta \in \mathbb{R}^m} \left\{ \|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2 \right\} \tag{1.28}$$

which is equivalent to solving:

$$\min_{\beta_0, \beta} \left\{ \|y - \beta_0 - X\beta\|_2^2 \right\} \tag{1.29}$$

subject to $(1-\alpha)\|\beta\|_1 + \alpha\|\beta\|_2^2 \leq t$, where $\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$. So, the results of the EN penalty is a combination of the effects of the lasso and ridge penalties.

Furthermore, highly correlated variables will tend to have similar regression coefficients, with the degree of similarity depending on both $\|\beta\|_1$ and λ_2, which is very different from basic lasso. This phenomenon, in which strongly correlated variables have similar regression coefficients, is referred to as the grouping effect ([Zou and Hastie, 2005], [Zhou, 2013]) and is generally considered desirable since, in many applications, such as identifying genes associated with a disease, one would like to find all the associated features, rather than selecting only one from each set of strongly correlated variables, as lasso generally does. In addition, selecting only a single variable from each group will typically result in increased prediction error, since the model is less robust (this is why ridge regression often outperforms lasso in this particular case).

This presence of correlated variables also leads to the use of other lasso variants such as the Group Lasso ([Yuan and Lin, 2007], [Friedman et al., 2010]), the Fused Lasso ([Tibshirani et al., 2005]) or the Overlay-Group Lasso ([Friedman et al., 2010]) which will not be discussed and detailed in this thesis.

In this whole context, implying the need of sparse and more interpretable final models, a gap had to be filled to put together the use of such penalties and the use of PLS models (the standard models in the field of metabolomics these recent years, Section 1.4.2 and Paper II). It allowed the emergence of the sparse-PLS (sPLS) model ([Chun and Keles, 2007], [Chung et al., 2012]) detailed below.

1.4.4.2 The sPLS algorithm

sPLS, used nowadays to an increasing extent in metabolomics, keeps the spirit of PLS models but provides a restricted number of final selected biomarkers via the use of penalty terms. Different algorithms exist in the literature, the two main of them being respectively implemented in the *mixOmics* R package ([Le Cao et al., 2008]) and in the *spls* R package ([Chun and Keles, 2007], [Chung et al., 2012]). In this thesis, the second one was preferred in order to allow a fast optimization of PLS parameters.

To get a sparse enough solution, the PLS objective function (Eq. 1.20) is reformulated by generalizing a regression formulation. The formulation of the sPLS algorithm of Chun and Keles promotes an exact zero property onto the weights. To perform that, sPLS imposes a L_1 penalty (λ_1) onto a surrogate direction vector c instead of the original direction w (linked with the X-weights), while keeping w and c sufficiently close to each other. The vector c can then be considered as a "sparse version" of w in the closest possible direction as the L_1 penalty encourages sparsity on c.
The addition of a L_2 penalty (λ_2) takes care of the potential singularity of matrix $M = X'yy'X$ when solving for c. Finally, the effect of the concave part, and the local solution issue, is reduced by using a small additional parameter noted θ here ([Chun and Keles, 2007]).

Practically, this algorithm provides an efficient implementation based on the following lasso-like objective function to minimize:

$$\min_{w_k, c_k} \{(-\theta w_k' M w_k) + (1-\theta)(c_k - w_k)' M (c_k - w_k) + \lambda_1 \|c_k\|_1 + \lambda_2 \|c_k\|_2\} \tag{1.30}$$

subject to $w_k' w_k = 1$, for $k = 1, ..., q$ and where $M = X'yy'X$.

The generalized regression formulation (Eq. 1.30) 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:

$$\min_{w_k} \{(-\theta w_k' M w_k) + (1-\theta)(c_k - w_k)' M (c_k - w_k)\} \tag{1.31}$$

subject to $w_k' w_k = 1$ for $k = 1, ..., q$.

This constrained least squares problem can be solved via the method of the Lagrange multipliers ([Chun and Keles, 2007]).

When solving for c for fixed w, it becomes:

$$\min_{c_k} \{(R' c_k - R' w_k)' (R' c_k - R' w_k) + \lambda_1 \|c_k\|_1 + \lambda_2 \|c_k\|_2\} \tag{1.32}$$

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 Selection algorithm (LARS [Efron et al., 2004]). sPLS often requires a very large λ_2-value to solve (Eq. 1.32) 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 and this yields the solution to have the form of an univariate soft thresholded (UST) estimator ([Zou and Hastie, 2005]).

The sPLS goal is to impose sparsity in the dimension reduction step of PLS so that sparsity can play a direct principled role. In order to promote sparse solutions in a restricted X-space, sPLS searches for relevant features, the so-called active variables. Thus, the first step of the **sPLS (using SIMPLS) algorithm**, fully described in Paper II (Section 3.3.3), is to obtain sparse weights by solving (Eq. 1.30). Descriptors associated with non-zero weights are then kept in a submatrix of X (noted X_{Π_k} in Paper II), which can be considered as the subspace of the active variables. In a second step, a PLS model is applied on these selected active descriptors only and PLS coefficients b_{PLS_k} are obtained for iteration $k, 1 \leq k \leq q$.

From an iteration to another, by upgrading the considered number of predictive component(s), all direction vectors are deflated and updated to form a Krylov subsequence ([Kramer, 2007], [Chapman and Saad, 1997]) on the subspace of the active variables. New PLS regression coefficients are updated by these new estimates of the direction vectors and, finally, X_{Π_k} is updated with all descriptors associated with non-zero weights or non-zero updated PLS coefficients.

An additional cross-validation step (for minimizing RMSEP for instance) is again 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 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 directly obtained as a result of the objective optimization of (q, λ_1).

1.4.5 The L-sOPLS algorithm for combining OPLS and sparse results

In this context, a potential major improvement in this field of PLS model-based biomarker discovery consists in combining the orthogonality of OPLS and the sparsity of sPLS. Indeed, if one wants 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 wants to propose sparse and interpretable results, a natural way is to explore the possible combinations of OPLS and sPLS.

Although some interesting work already exist about this topic ([van Gerven 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.

To this end, **the third innovative proposal of this thesis consists in the development of the so-called "Light-Sparse-OPLS" (L-sOPLS) model**. The idea is quite intuitive and requires to start from the OPLS algorithm described in details in Section 1.4.3 and in Paper II (Section 3.3.2), to follow the process until the construction and the recovery of a deflated X_d data matrix (matrix which is deflated by all the y-orthogonal components and effects) such that:

$$X_d = X - T_o P_o' = X - X_o \tag{1.33}$$

where $T_o = (t_{o1}, ..., t_{oq-1})$ contains the calculated orthogonal scores and $P_o = (p_{o1}, ..., p_{oq-1})$ contains the calculated orthogonal loadings from the $(1, ..., q-1)$ orthogonal components. Then, a sparsing model is specifically applied on this filtered matrix X_d, instead of on the initial data matrix X.

The sparsing technique may be freely chosen between lasso logistic regression, Elastic Net, sPLS, etc... Here, L-sOPLS combines the OPLS orthogonalization step with sPLS (see Section 1.4.4.2).

For interpretability and intuitiveness concerns, the L-sOPLS algorithm implies two successive 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 the classic OPLS process. The deflated resulting matrix X_d, which no longer contains sources of variations not connected to y, is then built according to this optimal number (called q_{ortho}). The second step is aimed at building the final sPLS sparse model, using X_d as input, whose number of predictive components (q_{pred}) and penalty term (λ_1) are again chosen optimally.

Note again that the application of the sparsity criterion and the choice of the number of predictive components are performed on a deflated matrix, after the initial and optimal choice of the number of orthogonal dimensions. These two successive steps precisely explain why this novel method is called "Light"-sOPLS, because, in general, it is a method that has to remain intuitive, fast and interpretable for metabolomics (-omics) final practitionners. For them, it seems great to first visualize a concrete deflated object by separating what is of interest to explain y and what is not. And, once this object is optimally found, the sparsity step will be more interpretable and concrete too. Interpretability and intuitiveness were the main concerns when developing L-sOPLS.

The L-sOPLS algorithm can be summarized as follows (from Paper II):

1. Application of an OPLS model 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 $\boldsymbol{X_d = X - X_o}$.

3. Application of a SIMPLS-sPLS model (see Section 1.4.4.2) using X_d as input. This implies the inner matrix products $\boldsymbol{M_d = X_d' y y' X_d}$ and $\boldsymbol{R_d = X_d' y}$ instead of M and R in (Eq. 1.31) and (Eq. 1.32). 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. Direction vectors are updated and deflated during the sPLS process (see Section 1.4.4.2). Finally, the final non-zero coefficient vector b^* is obtained by conducting PLS regression on the active variables only (i.e.

the variables which are associated with non-zero loadings in the predictive components). 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 contain non-zero coefficients at the end of the sPLS process, are finally considered as primary candidate biomarkers.

Considering a binary y, note that, as opposed to OPLS, L-sOPLS allows the potential use of several predictive dimensions q_{pred}, combined with potentially several orthogonal dimensions q_{ortho}. Consequently, L-sOPLS can allow more flexibility.

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 and in Paper II results and conclusions.

It can appear counterintuitive that the possible use of more than one predictive dimension in L-sOPLS can really lead to parsimonious and simple final models (compared to OPLS). But the fact that q_{pred} is chosen by cross-validation simultaneously along with the λ_1 penalty term avoids high risks of final non-parsimonious or very huge models (see further results in Paper II, Tables 3.1 and 3.2).

Note also that, according to some subsequent trials, L-sOPLS seems to be robust to overfitting toward the major class (i.e. if $P(y = 0) \neq P(y = 1)$). Indeed, the final selected biomarkers are stable even if the target $y = 1$ becomes a rare event in the y response vector.

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

1.4.6 Other sparse methods

In this thesis, PLS based models, sparse or not, were privileged for biomarker discovery because of their widespread use in metabolomics for years. But it is obvious that many other methods exist in machine learning, data mining, chemometrics or multivariate statistics in order to highlight and select relevant variables or features.

Among many possibilities, recent methods used in metabolomics can be inspired by clustering of variables (for instance in [Vigneau and Qannari, 2003], [Vigneau et al., 2015] and [Chen M. and Vigneau, 2016]), by Bayesian approaches (for instance the GSPPCA, Globally Sparse Probabilistic PCA, in [Bouveyron et al., 2016]; or in [Song Q. and Liang, 2015]), by the use of Support Vector Machines (for instance in [Henneges at al., 2009] and [Lin et al., 2012]) or by the use of decision trees and tree ensembles methodologies.

To a lesser extent, this later approach was tested during this thesis along with (O)PLS, sPLS and L-sOPLS via the use of the jForest algorithm (see [Paul and Dupont, 2015]) and the Xgboost algorithm (see [Chen T. and Guestrin, 2016] and [Song R. et al., 2016]). These predictive methods can include sparsity techniques to perform feature selection, and are rapidly presented below.

1.4.6.1 jForest: a Random Forest approach involving importance measures for feature selection

In order to propose another approach which can distance itself from the (O)PLS world, jForest, a Random Forests (RF) solution coupled with the use of statistical importance measures, is proposed here, directly inspired by [Paul and Dupont, 2015], and applied to metabolomics data for the first time during this thesis (a different approach of Random Forests for omics data can be seen in [Rinaudo et al., 2016]). In contrast to PLS, RF ([Breiman, 2001]) are well know for its ability to jointly use continuous and categorical (meta)data, which can represents a real additional asset in metabolomics for its interpretation. Therefore, RF allows to handle heterogeneous data along with high dimensionality (small n, big m).

In the Paul and Dupont solution, in the context of personalised medicine, a quite huge number K of decision trees (minimum $K = 1000$) are built to allow convergence for the results and to avoid overfitting of the data (using a single decision tree is in general too sensitive to input data and leads to systematic overfitting). Each individual tree, called k, grows from a "bootstrap" sample bag B_k of the n initial data observations (rows) and a final democratic decision is taken, according to a majority vote. Note that bag B_k corresponds to samples

at the root of k-th tree and out-of-bag $\bar{B}_k$ corresponds to samples left aside. An illustration is shown in Fig.1.24.

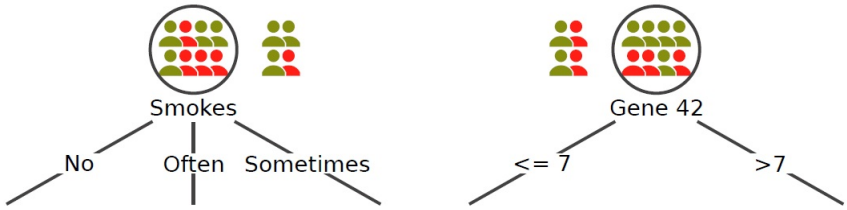


Figure 1.24: Illustration of the bag and out-of-bag notions (from J. Paul thesis defense slides).

In order to compute feature importances, Breiman ([Breiman, 2001]) proposes a permutation test procedure based on classification error. For each variable x_j, there is one permutation test per tree in the forest. For an out-of-bag sample $\bar{B}_k$ corresponding to the k-th tree of the ensemble, one considers the original values of the variable x_j and a random permutation $\tilde{x_j}$ of its values on $\bar{B}_k$. The difference in prediction error using the permuted and original variable is recorded and averaged over all the out-of-bag samples in the forest.

Here, the question is to detect if a variable x_j is useful and relevant to predict y. To perform that, the classification performances of the forest have to be assessed as such and without this variable x_j.

The emphasis is on the additional use of an importance measure from these RF to evaluate the classification performances, linked with the response vector y, in these two cases. In consequence, an observed decrease in RF performances will be proportional to the "importance" of the corresponding variable x_j. This importance of a feature, as candidate biomarker, is defined as the average decrease in accuracy over all trees when randomly permuting values, and can be written as:

$$Imp(x_j) = \frac{1}{K}\sum_{k=1}^{K}(ACC_k(\bar{B}_k) - ACC_k(\bar{B}_k^{\tilde{x_j}})) \tag{1.34}$$

where $ACC_k()$ can be any classical accuracy criterion of k-th tree's predictions, $\bar{B}_k$ is the out-of-bag sample of k-th tree and $\bar{B}_z^{\tilde{x_j}}$ is the out-of-bag sample of k-th tree with x_j permuted. These out-of-bag samples can be considered as validation sets on which class vote distributions are estimated.

A statistical Pearson's χ^2 test is then used to assess whether the permuted versions significantly differ from the original x_j. For interpretability, rejecting the null hypothesis with a low associated p-value means that permuting variable x_j significantly influences the class vote distribution and, therefore, that x_j is important in the current predictive model.

Since the importance of several features is typically evaluated through these tests on the same data, p-values must be corrected for multiple testing. The popular Benjamini-Hochberg correction ([Benjamini and Hochberg, 1995]) is used to control the final false discovery rate (FDR). Let $p_{\chi^2}^{fdr}(x_j)$ be the value of $p_{\chi^2}(x_j)$ after FDR correction, the final importance measure is defined as:

$$J_{\chi^2}(x_j) = p_{\chi^2}^{fdr}(x_j) \tag{1.35}$$

Finally, the variable x_j will be significantly important when $J_{\chi^2}(x_j) < 5\%$. The significance threshold of 5% is generally chosen and corresponds to the probability of falsely identifying a variable as important. The process is represented in Fig.1.25, where the key notions of true and false positives or negatives concretely appear. As for the use of lasso penalties (Sections 1.4.4 and 1.4.5), non-important features/biomarkers (with Pearson's p-values higher than 5%) are then not taken into account, thus implying sparsity.

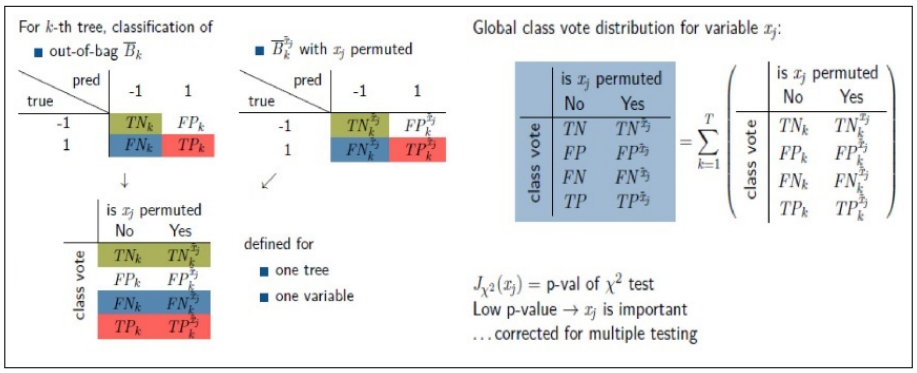


Figure 1.25: The implementation of Pearson's χ^2 tests as statistical importance measures for Random Forests (from J. Paul thesis defense slides).

Note that this index can easily be generalized to multi-class problems. In such cases, the formal contingency table simply has $c^2 \times 2$ entries, where c is the number of classes in the y response vector.

This methodology, implemented in the *jForest* R package (`https://github.com/jeromepaul/jForest`), was also applied in order to identify a small number of relevant biomarkers. Some results are available in Section 1.6.

1.4.6.2 The Xgboost alternative (as tree ensemble boosting)

In this section, Boosting methods are very superficially approached and all mathematical and algorithmic details can be obtained through references.

Boosting is a machine learning ensemble meta-algorithm for primarily reducing bias, and also variance, in supervised learning, and a family of machine learning algorithms that convert weak learners to stronger ones ([Freund and Schapire, 1999]). Boosting is based on the initial question posed by Kearns and Valiant ([Kearns and Valiant, 1989]): "Can a set of weak learners create a single strong learner?". A weak learner is defined to be a classifier that is only slightly correlated with the true classification (it can label examples better than random guessing). In contrast, a strong learner is a classifier that is more strongly and directly correlated with the true classification.

While Boosting is not algorithmically constrained, most boosting algorithms consist of iteratively learning weak classifiers with respect to a distribution and adding them to build a final strong classifier.
When they are added, they are typically weighted in some way that is usually related to the weak learners' accuracy. After a weak learner is added, the data are reweighted: examples that are misclassified gain weight and examples that are classified correctly lose weight. Thus, future weak learners focus more on the examples that previous weak learners misclassified.
There are many Boosting algorithms. The main variation between them is their method of weighting training data points. AdaBoost ([Schapire, 2003], [Hastie, Rosset et al., 2009]), with decision trees as the weak learners, is very popular and perhaps the most significant historically as it was the first algorithm that could adapt to these weak learners. However, there are many more recent algorithms adapted to particular types of data sets.

In this context, sparsity can be included into the Boosting framework in order to obtain few final and relevant descriptors. Among other trees based algorithms, along with an inner importance measures ranking, the Sparse Boosting ([Buhlmann and Yu, 2006], [Wang et al., 2015]) and Xgboost, for eXtreme Gradient Boosting ([Chen T. and Guestrin, 2016], [Song R. et al., 2016]), algorithms can offer another approach to highlight important biomarkers.

Xgboost, which is becoming popular and generally leads to good prediction results, was tested during this work. Xgboost is used for supervised learning problems, where a training data (with multiple features) X is used to predict a target variable y. Xgboost is based on a tree ensembles model which is a set of classification and regression trees (CART). By assigning final scores to individuals, CART is a bit different from decision trees, where the leaf only contains decision values. In CART, a real score is associated with each of the leaves, which provides richer interpretations that may go beyond classification.

Mathematically, a simple predictive model for observation i can be written as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_i), f_k \in F \tag{1.36}$$

where K is the total number of trees, f is a function in the functional space F, and F is the set of all possible CARTs. Therefore, an objective function to be optimized can be written as follows:

$$obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{1.37}$$

where l is typically linked with training loss, and Ω is the regularization term. The training loss measures how predictive the model is on training data. For example, a commonly used training loss is the mean squared error. The regularization term controls the complexity of the model, which helps to avoid overfitting.

Note that random forests and boosted trees are not different in terms of model, the difference is how they are trained. For the training part itself, the functions f_k need to be learnt. These functions each contain the structure of the associated tree and the leaf scores. It is not easy to train all the trees at once. Instead, an additive strategy can be used by fixing what we have learned, and adding only one new tree at a time. Since only one new tree is added at each step, the step index of the process is equal to the current number of involved trees k.

Let $\hat{y}_i^{(k)}$ be the predicted value of y_i at step k ($i = 1, ..., n$), the procedure involves:

$\hat{y}_i^{(0)} = 0$
$\hat{y}_i^{(1)} = f_1(X_i) = \hat{y}_i^{(0)} + f_1(X_i)$
$\hat{y}_i^{(2)} = f_1(X_i) + f_2(X_i) = \hat{y}_i^{(1)} + f_2(X_i)$
...
$\hat{y}_i^{(K)} = \sum_{k=1}^{K} f_k(X_i) = \hat{y}_i^{(K-1)} + f_K(X_i)$

At each step, a natural improvement consists in adding the tree that optimizes the objective. Each successive $\hat{y}_i^{(k)}$ can be considered as an unweighted adjustment of the previous prediction.

If MSE is considered as loss function, this objective becomes:

$$obj^{(k)} = \sum_{i=1}^{n} (y_i - (\hat{y}_i^{(k-1)} + f_k(X_i)))^2 + \sum_{j=1}^{k} \Omega(f_j) \tag{1.38}$$

which can be rewritten as:

$$obj^{(k)} = \sum_{i=1}^{n} (2(\hat{y}_i^{(k-1)} - y_i) f_k(X_i) + f_k(X_i)^2) + \Omega(f_k) + constant \tag{1.39}$$

The later equation then becomes the optimization goal for the new tree. Note that Xgboost can support custom loss functions and can generally optimize every loss function, including logistic regression and weighted logistic regression.

For the regularization step, the complexity of the tree $\Omega(f)$ needs to be defined. In order to do so, let us first refine the definition of the tree $f_k(X)$ as:

$$f_k(X) = w_{k,q(X)}, w \in R^T \tag{1.40}$$

where w is the vector of obtained scores on leaves, q is a function assigning each initial data point to the corresponding leaf and T is the number of leaves in the tree k.

In Xgboost, the complexity is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2} \Lambda \sum_{j=1}^{T} w_j^2 \tag{1.41}$$

There is more than one way to define the complexity, but this specific one works well in practice ([Chen T. and Guestrin, 2016]). The regularization is one part most tree packages treat less carefully, or simply ignore. This was

because the traditional treatment of tree learning only emphasized improving impurity, while the complexity control was left to heuristics.

Finally, note that this regularization can be indirectly connected to the use of ridge/lasso penalty terms (Section 1.4.4.1) and to the sPLS and L-sOPLS algorithms' purpose (Sections 1.4.4.2 and 1.4.5) and can lead to sparse results by only taking into account final relevant features (interpretable as relevant biomarkers).

Indeed, a benefit of using boosting is that after the boosted trees are constructed, it is relatively straightforward to retrieve importance scores for each variable. Generally, importance provides a score that indicates how useful or valuable each feature was in the construction of the boosted decision trees within the model. The more an attribute is used to make key decisions with decision trees, the higher its relative importance.
This importance is calculated explicitly for each variable in the dataset, allowing variables to be ranked and compared to each other. Practically, this importance is calculated for a single decision tree by the amount that each attribute split point improves the performance measure, weighted by the number of observations the node is responsible for. The chosen performance measure may then be the purity (Gini index) for instance ([Friedman et al., 2001]). The variable importances are then averaged across all of the decision trees within the model.

An additional feature selection step, leading to the final use of a subset of relevant variables, then finally allows "sparse" solutions (for an illustration, see `http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html`).

Regularized Xgboost was tested using the *xgboost* R package (`https://cran.r-project.org/web/packages/xgboost/index.html`). Some results are available in Section 1.6.

1.5 Overview of the used datasets

In this section, all the data sets used in examples and simulations during this thesis are listed. Subsequent details and all the technical acquisitions and preprocessing parameters are fully available in Paper I, II or III respectively.

In Paper I, two data sets were used to demonstrate the interest of the 2D COSY spectral additional information compared to 1D ^{1}H-NMR in terms of

Metabolomic Informative Content. These two data sets are summarized in Table 1.1. Both of them included COSY and 1D ^{1}H-NMR measurements.

In Paper II, two data sets were used to illustrate the biomarker discovery issue in metabolomics studies and were used to train the (O)PLS, sPLS and the innovative L-sOPLS algorithms. The first one implied 1D ^{1}H-NMR spectral measurements, the second one was a genomic data set which consisted in Reverse Transcription-quantitative Polymerase Chain Reaction (RT-qPCR) results from miRNAs expression analysis in groups of patients suffering or not from endometriosis (as response vector y). These two data sets are summarized in Table 1.2.

Finally, in Paper III, two other data sets were used to illustrate the GPL and Vectorization 2D workflows and to perform L-sOPLS on 2D spectral data for the first time. These two data sets are summarized in Table 1.3. Both of them implied different COSY and 1D ^{1}H-NMR measurements.

Excepted the citrate/hippurate data set in Paper II and the genomic data set also used in Paper II, all these data sets were supplied by the Center for Interdisciplinary Research on Medicines (CIRM), Université de Liège (ULg). Many thanks to Pascal de Tullio and Justine Leenders. The citrate/hippurate set was supplied by Eli Lilly Belgium. The endometriosis genomic data set was supplied by the GIGA-Cancer laboratory, also from ULg. Thanks to Carine Munaut.

Table 1.1: Data sets used in Paper I.

Medium	**Group signal to retrieve (y)**	**Noisy factors**	**# of measures**	**Characteristics**
Mixture cell culture	4 mixtures	sampling (3), time replicates (3) bacterial degradation, frosting/defrosting	36	Unknown degree of separability a priori
Human serum	4 blood donors	days (8), replicates, permutations, different delays	32	Uncontrolled

1.6 Results and discussions

In this section, all the results, chronologically obtained in Papers I to III, are discussed. Some additional results, linked with the use of Random Forests

Table 1.2: Data sets used in Paper II.

Medium	**Group signal to retrieve (y)**	**Noisy factors**	**# of measures**	**Characteristics**
Urine of rats	2 groups linked with concentrations of spiked products (citrate, hippurate)	sampling (6), days (3), time replicates	36	Analysis of balanced and unbalanced cases, controlled
Human serum	2 groups of people (suffering or not from endometriosis)		120 individual measures	Few variables, very tough discrimination

Table 1.3: Data sets used in Paper III.

Medium	**Group signal to retrieve (y)**	**Noisy factors**	**# of measures**	**Characteristics**
Urine	3 urine donors	days (4), urine dilution (2)	24	Uncontrolled
Human serum	4 groups linked with concentrations of spiked products (threonine, glutamate)	days (3), replicates (3)	36	Controlled

and the Xgboost algorithm for feature selection (see Section 1.4.6) or involving faster COSY experiments (see Chapter 5), are also presented.

First, **PAPER I** is aimed at introducing the Global Peak List (GPL, Section 1.2.2.1) approach for handling a set of 2D COSY spectra included in a whole experimental design, at introducing the Metabolomic Informative Content (MIC, Section 1.3.2) concept for evaluating and comparing the quality of the available data sources and finally at demonstrating that the COSY second dimension brings an informational added value compared with conventional 1D ^{1}H-NMR spectra.

For that purpose, two data sets (see Table 1.1) were considered: a mixture cell culture data set containing multi-sources noise (sampling, replicates, bacterial degradation of the samples, ...); and a more complex human serum data set involving four blood donors. In each case, individual 2D COSY spectra have been collected and submitted to the GPL workflow according to different scenarii: using spectral intensities or binary positions of peaks (presence or absence of peaks), different adapted distance measures and different values for the number of decimal(s) during the 2D soft bucketing step (see Section 1.2.2.1). Also in each case, a group factor was defined in a y response vector

and had to be retrieved via the unsupervised clustering algorithms gathered in the MIC process.

The main 2D results are shown in Table 1.4.

Table 1.4: Numerical clustering performances according to different versions of COSY spectra. The column "Dec." denotes the number of decimal(s) for the data in the GPL matrix; T = number of observations in the corresponding GPL; DI = Dunn Index; DBI = Davies-Bouldin Index; RI = Rand Index; ARI = Adjusted Rand Index (paper I).

FIRST DESIGN : Mixture cell culture								
Signal	Dec.	Algorithm	Distance	T	DI	DBI	RI	ARI
Positions	1	Ward	Jaccard	909	0.712	1.466	**0.914**	**0.811**
Positions	1	Ward	Ochiai	909	0.712	1.466	**0.914**	**0.811**
Positions	2	Ward	Jaccard	2348	**0.857**	1.688	0.757	0.420
Positions	2	Ward	Ochiai	2348	**0.827**	1.688	0.757	0.420
Positions	3	Ward	Jaccard	3250	**0.893**	1.722	0.601	0.189
Positions	3	Ward	Ochiai	3250	**0.892**	1.722	0.601	0.189
Intensities	1	Ward	Euclidean	909	0.298	**0.541**	**0.927**	**0.853**
Intensities	1	K-means	Euclidean	909	0.298	**0.541**	**0.927**	**0.853**
Intensities	2	Ward	Euclidean	2348	0.239	1.249	0.669	0.609
Intensities	2	K-means	Euclidean	2348	0.311	1.356	0.657	0.501
Intensities	3	Ward	Euclidean	3250	0.439	1.540	0.588	0.425
Intensities	3	K-means	Euclidean	3250	0.434	1.614	0.597	0.439
SECOND DESIGN : Human serum								
Signal	Dec.	Algorithm	Distance	T	DI	DBI	RI	ARI
Positions	1	Ward	Jaccard	1106	0.796	1.569	**0.937**	**0.825**
Positions	1	Ward	Ochiai	1106	0.722	1.569	**0.937**	**0.825**
Positions	2	Ward	Jaccard	4172	**0.945**	1.800	0.772	0.401
Positions	2	Ward	Ochiai	4172	**0.899**	1.800	0.772	0.401
Positions	3	Ward	Jaccard	6686	**0.981**	1.792	0.650	0.171
Positions	3	Ward	Ochiai	6686	**0.963**	1.792	0.650	0.171
Intensities	1	Ward	Euclidean	1106	0.419	**0.643**	**0.932**	**0.804**
Intensities	1	K-means	Euclidean	1106	0.419	**0.643**	**0.932**	**0.804**
Intensities	2	Ward	Euclidean	4172	0.689	1.592	0.789	0.422
Intensities	2	K-means	Euclidean	4172	0.706	1.742	0.735	0.288
Intensities	3	Ward	Euclidean	6686	0.778	1.668	0.578	0.055
Intensities	3	K-means	Euclidean	6686	0.765	1.886	0.647	0.135

Globally, the results are very promising with many RI (and ARI) greater than 0.7, particularly with bucketed data when one or two decimals in the GPL matrices are used. Thus, unlabeled individuals are well grouped together, with no prior knowledge, and this according to the above-mentioned presence of numerous potential noise factors.

Considering first the position-based partitions, particular groups are already well isolated in the majority of cases. With the pre-processed intensities-based

partitions, the four mixtures are generally well recovered by the algorithms in spite of the sampling procedure and time repetitions. For the first design, the best clustering result is obtained with the one-decimal GPL matrix (RI=0.927 and ARI=0.853), thus underlying the importance of the "bucketing" step. The conclusion is the same for the second design concerning the blood donors (RI=0.932 and ARI=0.804).

More generally, these results show that Dunn and Davies-Bouldin indexes are subject to significative variations between experiments and are not very satisfactory for some combinations (several DI less than 0.5 and DBI greater than 1.5 in Table 1.4). This means that obtained clusters are not always compact and well separated. However, to judge if the clusters conform to the reality, i.e. if members in a cluster correspond to members of initial groups, Rand and Adjusted Rand indexes prevail because they are built on the degree of agreement and disagreement between groups (here between the reality and each of the clustering partitions).

Consequently, and based on the two experimental designs, the clustering results show that COSY spectra appear to be statistically robust and contain informative signal which helps to succeed in distinguishing groups of different spectra. In other words, **COSY spectra are enough informative so that the signal connected to the initial groups can be distinguished from the noise**. Working on positions gives satisfactory results but not better ones than working on intensities. Moreover, these results show that the 2D "bucketing" step is important and probably necessary to solve misalignment problems (as it is the case in 1D). This additional information can be profitable for further robust statistical analysis, as biomarker discovery.

Besides these convincing performances of COSY spectra, it remains to demonstrate that the additional information contained in 2D COSY spectra via the existence of correlation cross peaks is relevant and crucial by improving the quality of the clustering results. Taking into consideration the corresponding pre-processed 1D ^{1}H-NMR data sources, 1D results are shown in Table 1.5 for the two experimental designs. Note that the MIC clustering process was performed on raw 1D spectral intensities (normalized), as well as on intensities rounded to one and two decimal(s) for comparison.

If one compares these results with the best Rand and Adjusted Rand indexes in Table 1.4, it appears clearly that these indexes are mainly higher when the clustering algorithms are performed on 2D COSY spectra. Resulting partitions are better when using COSY data sources as they are more in agreement with the real initial groups.

Table 1.5: Numerical clustering performances for 1D data. The column "Dec." denotes the number of decimal(s) for the data in the GPL matrix; T = number of observations in the corresponding GPL; DI = Dunn Index; DBI = Davies-Bouldin Index; RI = Rand Index; ARI = Adjusted Rand Index. (Paper I)

FIRST DESIGN : Mixture cell culture								
Signal	Dec.	Algorithm	Distance	T	DI	DBI	RI	ARI
Intensities	1	Ward	Euclidean	201	**0.927**	**0.120**	**0.908**	**0.807**
Intensities	1	K-means	Euclidean	201	0.560	**0.463**	**0.835**	0.667
Intensities	2	Ward	Euclidean	2001	0.712	**0.426**	0.717	0.516
Intensities	2	K-means	Euclidean	2001	0.601	0.739	0.732	0.476
Raw intensities		Ward	Euclidean	199967	0.362	1.048	0.606	0.544
Raw intensities		K-means	Euclidean	199967	0.284	1.182	0.410	0.196
SECOND DESIGN : Human serum								
Signal	Dec.	Algorithm	Distance	T	DI	DBI	RI	ARI
Intensities	1	Ward	Euclidean	207	**0.981**	**0.688**	0.704	0.233
Intensities	1	K-means	Euclidean	207	0.334	**0.663**	0.702	0.230
Intensities	2	Ward	Euclidean	2054	**0.901**	0.986	0.704	0.233
Intensities	2	K-means	Euclidean	2054	0.635	1.119	**0.740**	**0.290**
Raw intensities		Ward	Euclidean	205034	0.670	1.074	0.588	0.017
Raw intensities		K-means	Euclidean	205034	0.447	1.321	0.675	0.145

Thus, this demonstrates, for the two experiments, that **the COSY additional spectral dimension does provide relevant information to improve the clustering results, the amount of captured signal and, consequently, the Metabolomic Informative Content (MIC) performances**.

In **PAPER II**, the biomarker discovery issue in metabolomics is generally approached and the novel L-sOPLS algorithm (see Section 1.4.5) is introduced in order to combine the orthogonality (Section 1.4.3) and the sparsity (Section 1.4.4) attractive properties.

For that purpose, two experimental designs (see Table 1.2) were considered: a controlled design based on a pool of urine of rat samples spiked with two products (and considering balanced and unbalanced cases); and a genomic human serum data set characterized by a small number of descriptors and a very tough discrimination between the two groups of interest (people suffering or not from endometriosis). For the first design, only 1D ^{1}H-NMR pre-processed data were used.

Starting from the ^{1}H-NMR urine of rat design, balanced and unbalanced cases correspond to doses of spiked products. In particular, two products were considered for spiking and determining groups of samples: hippurate and citrate.

Fig.1.26 shows the balanced and unbalanced designs for hippurate (see Paper II for more details).

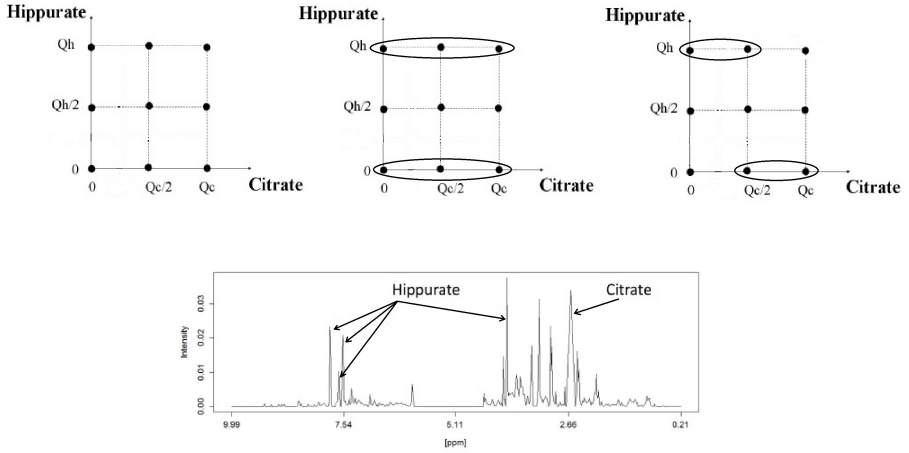


Figure 1.26: (a) The whole urine experimental design; (b) The hippurate balanced case; (c) The hippurate unbalanced case; (d) A typical urine spectrum with spiked citrate and hippurate.

Practically, two main sub-data sets were considered to obtain results, with balanced and unbalanced hippurate doses. 600 descriptors, all potential spectral biomarkers in X (without metadata or medical extra information), are available to explain the membership in a group (the 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 explain this discrimination.

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 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.

Results are presented in Table 1.6.

Table 1.6: (O)PLS and sparse results for the ^{1}H-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) (Paper II).

THE HIPPURATE BALANCED CASE					
		PLS coefficients $q = 5$	OPLS b_{OPLS} $q = 4 + 1$	sPLS selection $q_{pred} = 5$, $\lambda_1 = 0.6$	L-sOPLS selection $q_{ortho} = 4$, $q_{pred} = 3$, $\lambda_1 = 0.72$
Biomarkers (in ppm)	Zone	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 UNBALANCED CASE					
		PLS coefficients $q = 5$	OPLS b_{OPLS} $q = 4 + 1$	sPLS selection $q_{pred} = 5$, $\lambda_1 = 0.6$	L-sOPLS selection $q_{ortho} = 4$, $q_{pred} = 3$, $\lambda_1 = 0.73$
Biomarkers (in ppm)	Zone	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)**
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 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 1.6).

Thus, **L-sOPLS provides (very) sparse models with a better predictive power than (O)PLS**.

In Table 1.6, the first two columns show 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 1.6 for the 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 from the 600 initial input variables. All the hippurate and citrate peaks are 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.1.26(c)). Furthermore, the selected citrate descriptors are well associated with negative sparse coefficients and localized at the center of the citrate peak.

For the unbalanced hippurate case's sPLS and L-sOPLS models, Table 1.7 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**.

When results from sPLS and L-sOPLS are compared in Table 1.6 and Table 1.7, better predictive performances for L-sOPLS can be explained by the use of more targeted constraints, strictly focused on y. PLS and sPLS do not contain an orthogonalization step, so that the explanations (in term of variability) of X and y are performed together. With the additional orthogonalization, OPLS and L-sOPLS emphasize the explanation of y at the expense of the explanation of variabilities into X. When using penalty terms for sparse solutions, and by using the q_{ortho}-deflated matrix X_d as input (see Section 1.4.5), L-sOPLS further enhances this trend.

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 1.7) 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 in the model, the number of selected biomarkers also increases.

Table 1.7: 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 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 (Paper II).

THE HIPPURATE UNBALANCED CASE										
	q_{pred} for sPLS					q_{pred} for L-sOPLS (with $q_{ortho} = 4$)				
λ_1	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)**

Additional graphical results are available in Paper II in order to visualize the calculated loadings and coefficients of each model (PLS, OPLS, sPLS and L-sOPLS) and the obtained discrimination between the groups for the unbalanced hippurate case.

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.

Still in the context of the citrate/hippurate data set of Paper II, the jForest (Section 1.4.6.1) and Xgboost (Section 1.4.6.2) ensemble trees algorithms were also tested for descriptor selection. As described previously, these two methods imply the calculation of inner importance measures for each variables in the process and allow to highlight the more relevant ones. The Gini index was used for Xgboost regularization. Some results are presented in Table 1.8. Table 1.8 is built in a similar way than Table 1.6 (see details in Paper II).

One can see in this table that jForest and Xgboost are successful in highlighting the accurate ppm zones for both the hippurate balanced and unbalanced cases. However, in comparison with the optimal L-sOPLS solution found

Table 1.8: Variable selection provided by the jForest and Xgboost algorithms with importance measures and comparison with the optimal sPLS and L-sOPLS results (context of Paper II).

THE HIPPURATE BALANCED CASE					
Biomarkers (ppm)	Zone	Optimal sPLS	Optimal L-sOPLS	jForest	Xgboost
2.478		No	No	No	No
2.593	Citrate	No	No	No	No
2.609	Citrate	No	No	No	No
2.626	Citrate	No	No	No	No
2.642	Citrate	No	No	No	No
3.017		**Yes**	No	No	No
3.050		**Yes**	No	No	No
3.279		**Yes**	No	**Yes**	No
3.295		**Yes**	No	**Yes**	**Yes**
3.442		**Yes**	No	No	No
3.801		No	No	No	No
3.981	**Hippurate**	**Yes**	**Yes**	**Yes**	**Yes**
3.997	**Hippurate**	**Yes**	**Yes**	**Yes**	**Yes**
6.055		No	No	**Yes**	No
7.558	**Hippurate**	**Yes**	**Yes**	**Yes**	**Yes**
7.574	**Hippurate**	**Yes**	**Yes**	**Yes**	**Yes**
7.639	**Hippurate**	**Yes**	No	**Yes**	**Yes**
7.656	**Hippurate**	**Yes**	**Yes**	**Yes**	**Yes**
7.836	**Hippurate**	**Yes**	**Yes**	**Yes**	**Yes**
7.852	**Hippurate**	**Yes**	**Yes**	**Yes**	**Yes**

THE HIPPURATE BALANCED CASE					
Biomarkers (ppm)	Zone	Optimal sPLS	Optimal L-sOPLS	jForest	Xgboost
2.478		No	No	No	No
2.560	Citrate	**Yes**	No	No	No
2.576	Citrate	**Yes**	No	**Yes**	**Yes**
2.593	Citrate	**Yes**	**Yes**	**Yes**	**Yes**
2.609	Citrate	**Yes**	**Yes**	**Yes**	**Yes**
2.625	Citrate	**Yes**	**Yes**	**Yes**	**Yes**
2.642	Citrate	**Yes**	No	**Yes**	No
2.658	Citrate	**Yes**	No	No	No
3.017		**Yes**	No	No	No
3.279		**Yes**	No	**Yes**	No
3.295		No	No	No	No
3.442		**Yes**	No	No	No
3.801		No	No	No	No
3.981	**Hippurate**	**Yes**	**Yes**	**Yes**	**Yes**
3.997	**Hippurate**	**Yes**	No	**Yes**	**Yes**
6.055		No	No	No	No
7.558	**Hippurate**	**Yes**	**Yes**	**Yes**	**Yes**
7.574	**Hippurate**	**Yes**	**Yes**	**Yes**	**Yes**
7.639	**Hippurate**	**Yes**	No	**Yes**	**Yes**
7.656	**Hippurate**	**Yes**	No	**Yes**	**Yes**
7.836	**Hippurate**	**Yes**	**Yes**	**Yes**	**Yes**
7.852	**Hippurate**	**Yes**	**Yes**	**Yes**	**Yes**

previously, these two methods seem to be less sparse by selecting more final biomarkers.

In the RT-qPCR data set described in Table 1.2 and detailed in Paper II, $m = 19$ 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 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 sparse results obtained via sPLS and L-sOPLS are fully available in Paper II, along with the optimal parameters and the obtained minimal LOO-RMSEP. As for the previous data set, the global minimal LOO-RMSEP is obtained with a L-sOPLS model providing just seven final biomarkers. Moreover, 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, it is shown that **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**.

Finally, **PAPER III** can be viewed as a larger and concrete application of the GPL workflow, of the MIC concept and of the L-sOPLS algorithm (for the first time applied on 2D data) in order to evaluate and compare multiple 1D and 2D NMR data sources. In this paper, the Vectorization workflow (Section 1.2.2.2) was also introduced to handle and pre-process 2D COSY spectra.

For that purpose, two experimental designs (see Table 1.3) were considered: an uncontrolled design based on urine samples in which the signal corresponds to the three urine donors (noisy factors are urine dilution and different days of measurement); and a controlled design based on human serum samples in which two products (threonine and glutamate) were spiked to form sub-groups of samples. In this second design, threonine and glutamate were chosen because their peaks overlap with other molecules' peaks in 1D ^{1}H-NMR representations. Moreover, different days of measurements and replicates add noise in this design.

The goal was 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.

For both experimental designs in paper III, the MIC indexes were calculated on the basis of different degrees of pre-processing of the initial COSY spec-

tra. 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 (see Section 1.2.2.1). The Vectorization approach was also taken into account in the process, along with pre-processed ^{1}H-NMR spectra. The main results are described in Table 1.9. See also Appendix for dimensionalities purpose.

Table 1.9: MIC results for the two designs involved in Paper III, according to the choice of different pre-processing workflows (GPL and vectorization for 2D COSY or ^{1}H-NMR)

FIRST DESIGN - 3 urine donors									
Data	ACD/Lab	GPL	MIC indexes (Ward / K-means)				Inertia		Columns
	threshold	decimals	Dunn	Davies-Bouldin	Rand	Adj-Rand	Btw (%)	Wth (%)	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	GPL	MIC indexes (Ward / K-means)				Inertia		Columns
	threshold	decimals	Dunn	Davies-Bouldin	Rand	Adj-Rand	Btw (%)	Wth (%)	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

For the first experimental design, **the indexes tend to show that the best informative content and separability are reached by log-transformed 2D GPL approaches involving small matrices (with one or two decimals for the coordinates)**. Thus, the same conclusion is reached as in Paper I (see Table 1.4).

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.

For the second design, the results in Table 1.9 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).

Note that the threonine/glutamate experimental design proposes conditions of low variabilities between samples, which is doubtless close to real conditions.

When considering biomarker discovery results, PLS, sPLS and L-sOPLS were 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. Again, each model was 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 couple formed by the number of predictive component(s) q and the penalty term λ_1 is 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 Table 1.10 and Table 1.11.

Ideally, the goal was 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 1.9). The corresponding PLS, sPLS and L-sOPLS results are shown in Table 1.10 (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

Table 1.10: PLS, sPLS and L-sOPLS biomarker selection for ^{1}H-NMR data (threonine-glutamate design of Paper III).

		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

decisions concerning these biomarkers are also shown (Yes/No to indicate if the biomarker is selected, and the corresponding final sparse coefficient if yes).

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 initial MIC and PCA results (shown in Paper III) do not provide a smart separation between the groups of spectra. Furthermore, the use of sparse techniques, and in particular L-sOPLS, leads to a lower value of the RMSEP of the model, thus significantly improving the predictive power (from 0.1755 to 0.0759).

One can also see in Table 1.10 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 sources in order to ideally highlight relevant correlation peaks outside the diagonal. These cross peaks would then confirm and reinforce 1D results. The search for non-diagonal biomarkers based

Table 1.11: PLS, sPLS and L-sOPLS non-diagonal biomarker selection applied on the best 2D GPL solution in terms of MIC measures (threonine-glutamate design of Paper III).

		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

on the best 2D COSY solution (i.e. GPL with threshold = 0.05 and number of decimal = 1) is summarized in Table 1.11. Even if all the information was taken into account, biomarkers located on the diagonal are not displayed in Table 1.11, 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 then displayed.

One can see in Table 1.11 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 (remember that 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 tended to promote the use of some 2D GPL as a priority (see Table 1.9), PLS, sPLS and L-sOPLS had to be applied

on these data source first. So, relevant 2D correlation peaks were quickly found in a blind and non-supervised way as biomarkers, and the application of sparse algorithms on ^{1}H-NMR data served only as confirmation studies (Paper III).

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

All these results and considerations tend to promote a wider use of 2D COSY spectra in metabolomics studies because of better informative abilities and better performances for biomarker discovery compared to ^{1}H-NMR spectra. Among all these assets, a major problem still exists and can largely slow down such a generalized use of 2D spectra (COSY or others): the acquisition times, which are often prohibitive and non-acceptable when considering a lot of samples.

That is why faster COSY acquisition experiments are described in details and tested in **Chapter 5**: the single-scan COSY and the use of Non-Uniform Sampling (NUS). It is demonstrated in Chapter 5, which is a draft for a fourth paper to submit, that these two faster COSY acquisition techniques provide very similar MIC results than the initial conventional COSY (see Table 1.12).

Table 1.12: MIC results for conventional COSY NS4, COSY NS1 and COSY NUS50 spectra, according to the choice of different pre-processing workflows (GPL and Vectorization)

Experiment	Data	ACD/Lab threshold	GPL decimals	MIC indexes (Ward / K-means) Dunn	Davies-Bouldin	Rand	Adj-Rand	Inertia Btw (%)	Wth (%)	Columns in X
COSY NS4	GPL	0.02	1	0.8088 / 0.8088	1.5696 / 1.5696	**1.00 / 1.00**	**1.00 / 1.00**	41.01	58.99	828
COSY NS4	GPL	0.02	2	0.7917 / 0.7917	1.6943 / 1.6943	**1.00 / 1.00**	**1.00 / 1.00**	28.57	71.43	1954
COSY NS4	GPL	0.02	3	0.8131 / 0.8366	1.8565 / 1.7369	0.5942 / 0.5797	0.1744 / 0.2088	12.12	87.88	3994
COSY NS4	GPL	0.05	1	0.8792 / 0.8612	1.3635 / 1.3239	**1.00 / 1.00**	**1.00 / 1.00**	**42.62**	**57.38**	484
COSY NS4	GPL	0.05	2	0.8699 / 0.8699	1.6074 / 1.6074	**1.00 / 1.00**	**1.00 / 1.00**	31.76	68.24	1126
COSY NS4	GPL	0.05	3	0.8013 / 0.7642	1.8484 / 1.858	0.5942 / 0.6413	0.1743 / 0.2966	13.84	86.16	2332
COSY NS4	COSY vectorization			0.7555 / 0.758	0.6224 / 0.7166	0.9407 / **1.00**	0.8605 / **1.00**	**91.75**	**8.25**	65536
COSY NS1	GPL	0.02	1	0.8592 / 0.6763	1.3456 / 1.1932	**1.00 / 1.00**	**1.00 / 1.00**	**55.36**	**44.64**	337
COSY NS1	GPL	0.02	2	0.7543 / 0.7543	1.5941 / 1.5941	**1.00 / 1.00**	**1.00 / 1.00**	35.71	64.29	895
COSY NS1	GPL	0.02	3	0.7332 / 0.7359	1.8511 / 1.8802	0.5889 / 0.5889	0.153 / 0.153	15.68	84.32	1859
COSY NS1	GPL	0.05	1	0.623 / 0.623	1.3715 / 1.3715	**1.00 / 1.00**	**1.00 / 1.00**	**56.24**	**43.76**	201
COSY NS1	GPL	0.05	2	**0.9902 / 0.9902**	1.4361 / 1.4361	0.6996 / 0.6996	0.4038 / 0.4038	37.77	62.23	509
COSY NS1	GPL	0.05	3	0.8248 / 0.6065	1.639 / 1.8308	0.6245 / 0.8142	0.2834 / 0.5869	16.81	83.19	1065
COSY NS1	COSY vectorization			0.5733 / 0.6442	0.6226 / 0.7163	**1.00 / 1.00**	**1.00 / 1.00**	**91.38**	**8.62**	65536
COSY NUS50	GPL	0.02	1	0.8024 / 0.8024	1.3264 / 1.3264	**1.00 / 1.00**	**1.00 / 1.00**	**56.25**	**43.75**	348
COSY NUS50	GPL	0.02	2	0.7531 / 0.7531	1.5941 / 1.5941	**1.00 / 1.00**	**1.00 / 1.00**	36.15	63.85	996
COSY NUS50	GPL	0.02	3	0.7221 / 0.7221	1.9022 / 1.9022	0.8007 / 0.8007	0.5657 / 0.5657	13.25	86.75	2102
COSY NUS50	GPL	0.05	1	0.6019 / 0.5689	1.172 / 1.4532	**1.00 / 1.00**	**1.00 / 1.00**	50.07	49.93	203
COSY NUS50	GPL	0.05	2	0.6205 / 0.8603	1.5559 / 1.5977	0.8261 / 0.7102	0.6102 / 0.4103	31.89	68.11	574
COSY NUS50	GPL	0.05	3	0.6564 / 0.6176	1.5095 / 1.9596	0.4964 / 0.5942	0.093 / 0.0987	14.00	86.00	1167
COSY NUS50	COSY vectorization			0.6783 / 0.6783	0.8518 / 0.8518	**0.9447 / 0.9447**	0.8693 / 0.8693	**82.14**	**17.85**	65536

Thus, the demonstrated advantage of the COSY data source over the 1D ^{1}H-NMR one is still verified. Consequently, using techniques like the single scan or the Non-Uniform Sampling can really open the door to a more massive use of 2D COSY spectra in metabolomics studies and can facilitate the analyses of larger cohorts of samples.

1.7 Appendix of Chapter 1: comparisons of MIC quality measures when considering different dimensionalities for the descriptors

In Section 1.3.2.3, when MIC clustering quality measures were introduced, different matrices with different dimensionalities have to be assessed and compared between each other. An important question may arise: can these measures (inertia, Dunn, Rand, ...) be really compared from a dimension to another (with objects of same sample size n, but having different numbers of descriptors m) ? Indeed, higher dimensionality may induce a reduction in relevance for some indexes due to the curse of dimensionality.

Consider the mixture cell culture data set of Paper I (see Table 1.1). Four mixtures are present in the experimental design and 36 individual samples are available. Consider corresponding 1D ^{1}H-NMR pre-processed and bucketed spectra with 250 buckets, such that the global 1D spectral matrix is a (36×250) matrix. Let us call this matrix M. In order to explore this dimensionality issue, four experiments were performed. They all compare spectral matrices with increasing number of columns $m = 250$ to 16000. Four sets of matrices were built as follows:

- Experiment I : increase the number of columns of M by adding repeated blocks to M. Objects with $m = (250, 500, 1000, 2000, 4000, 8000, 16000)$ are considered such that the global matrix $M_G = \begin{bmatrix} M & M & M & ... \end{bmatrix}$. Corresponding MIC results are available in Table 1.13.

- Experiment II : increase the column dimensionality by multiplying M by several random error matrices $M.E_1^i$. E_1^i is chosen to simulate 5% noise according to random log-normal deviates. Here, M_G is then defined as $M_G = \begin{bmatrix} M.E_1^1 & M.E_1^2 & M.E_1^3 & M.E_1^4 & ... \end{bmatrix}$.

  Objects with $m = (250, 500, 1000, 2000, 4000, 8000, 16000)$ are considered and corresponding MIC results are available in Table 1.14.

- Experiment III : increase the column dimensionality by multiplying M by E_2^i in a similar way than in Experiment II, but with an amount of random noise of 25% of the original signal. M_G is then defined as

  $M_G = \left[M.E_2^1 \quad M.E_2^2 \quad M.E_2^3 \quad M.E_2^4 \quad ...\right]$.

  Objects with $m = (250, 500, 1000, 2000, 4000, 8000, 16000)$ are considered and corresponding MIC results are available in Table 1.15.

- Experiment IV : increase the column dimensionality by considering different bucketing schemes on the original 1D spectral data. Different bucketed objects with $m = (250, 500, 1000, 2000, 4000, 8000, 16000)$ are taken into account. Corresponding MIC results are available in Table 1.16.

In Table 1.13, one can see that replicating strictly redundant blocks, containing the same information, does not affect MIC criteria calculations. Thus, MIC criteria remain strictly identical when the number of columns (m) increases without proper increase in the dimensionality.

Table 1.13: MIC results for Experiment I

Columns	MIC indexes (Ward / K-means)				Inertia	
in X	Dunn	Davies-Bouldin	Rand	Adj-Rand	Btw (%)	Wth (%)
250	0.3523 / 0.2677	0.6233 / 0.5944	0.8714 / 0.8428	0.6564 / 0.6138	91.71	8.29
500	0.3523 / 0.2677	0.6233 / 0.5944	0.8714 / 0.8428	0.6564 / 0.6138	91.71	8.29
1000	0.3523 / 0.2677	0.6233 / 0.5944	0.8714 / 0.8428	0.6564 / 0.6138	91.71	8.29
2000	0.3523 / 0.2677	0.6233 / 0.5944	0.8714 / 0.8428	0.6564 / 0.6138	91.71	8.29
4000	0.3523 / 0.2677	0.6233 / 0.5944	0.8714 / 0.8428	0.6564 / 0.6138	91.71	8.29
8000	0.3523 / 0.2677	0.6233 / 0.5944	0.8714 / 0.8428	0.6564 / 0.6138	91.71	8.29
16000	0.3523 / 0.2677	0.6233 / 0.5944	0.8714 / 0.8428	0.6564 / 0.6138	91.71	8.29

Table 1.14 shows MIC results when introducing successive noisy deviations in the data blocks (by simulating a 5% random noise in each (36×250) new block). In this case, one can see that homogeneity indexes are very stable (Dunn, Davies-Bouldin and between inertia). Between inertia results decrease gradually and very slowly every time a block is added. Thus, these measures remain robust when a little amount of noise is considered. Moreover, accuracy indexes (Rand and Adjusted-Rand) are not impacted at all in this context.

Table 1.15 shows MIC results when introducing a bigger amount of noise in the data (by simulating a 25% random noise in each (36×250) new block). In this case, homogeneity measures are not any more stable. Between inertia results decrease gradually and rapidly every time a block is added.

Consequently, one can consider that inertia measures can not be really compared when the increase of m happens in the presence of an important noise level. On the other hand, accuracy indexes remain very stable, which means

Table 1.14: MIC results for Experiment II

Columns	MIC indexes (Ward / K-means)				Inertia	
in X	Dunn	Davies-Bouldin	Rand	Adj-Rand	Btw (%)	Wth (%)
250 (M)	0.3523 / 0.2677	0.6233 / 0.5944	0.8714 / 0.8428	0.6564 / 0.6138	91.71	8.29
250	0.3189 / 0.2972	0.7136 / 0.6621	0.8714 / 0.8428	0.6564 / 0.6138	90.85	9.15
500	0.3318 / 0.3120	0.6993 / 0.6627	0.8714 / 0.8428	0.6564 / 0.6138	90.52	9.49
1000	0.3526 / 0.3233	0.7262 / 0.6666	0.8714 / 0.8428	0.6564 / 0.6138	90.31	9.69
2000	0.3628 / 0.3380	0.7694 / 0.7078	0.8714 / 0.8428	0.6564 / 0.6138	90.01	9.99
4000	0.3762 / 0.3603	0.7867 / 0.7122	0.8714 / 0.8428	0.6564 / 0.6138	89.74	10.26
8000	0.3895 / 0.3707	0.8083 / 0.7326	0.8714 / 0.8428	0.6564 / 0.6138	89.46	10.54
16000	0.3986 / 0.3805	0.8242 / 0.7456	0.8714 / 0.8428	0.6564 / 0.6138	89.20	10.80

that the partition of the individuals into clusters remains robust and interpretable even with a higher level of noise.

Table 1.15: MIC results for Experiment III

Columns	MIC indexes (Ward / K-means)				Inertia	
in X	Dunn	Davies-Bouldin	Rand	Adj-Rand	Btw (%)	Wth (%)
250 (M)	0.3523 / 0.2677	0.6233 / 0.5944	0.8714 / 0.8428	0.6564 / 0.6138	91.71	8.29
250	0.3514 / 0.3514	1.1632 / 1.1632	0.8777 / 0.8777	0.6609 / 0.6609	89.20	10.80
500	0.4175 / 0.4124	1.3848 / 1.3719	0.8524 / 0.8397	0.5824 / 0.5594	73.86	26.14
1000	0.4433 / 0.3919	1.4088 / 1.3678	0.8428 / 0.8397	0.5619 / 0.5454	69.26	30.74
2000	0.4931 / 0.4157	1.3882 / 1.8376	0.8497 / 0.8524	0.5876 / 0.6019	65.18	34.82
4000	0.5064 / 0.5064	1.6829 / 1.6829	0.8497 / 0.8497	0.5791 / 0.5791	60.81	39.19
8000	0.6210 / 0.5051	1.2915 / 1.7483	0.8474 / 0.8565	0.5696 / 0.5991	57.01	42.99
16000	0.6168 / 0.5421	1.3604 / 1.7735	0.8587 / 0.8714	0.5996 / 0.6363	54.25	45.75

Finally, Table 1.16 is more devoted to compare different levels of bucketing applied on the 1D data matrix M. In this case, very different from the others, the same real data is bucketed and there is no artificial noise addition.

One can find that results observed in Tables 1.4, 1.5 (Paper I), 1.9 (Paper III) and 1.12 are again confirmed via this larger 1D study: the indexes tend to show that the best informative content and separability are reached by objects involving small matrices (i.e. a strong bucketing step, in bold in Table 1.16).

Table 1.16: MIC results for Experiment IV

Columns	MIC indexes (Ward / K-means)				Inertia	
in X	Dunn	Davies-Bouldin	Rand	Adj-Rand	Btw (%)	Wth (%)
250	0.3523 / 0.2677	0.6233 / 0.5944	0.8714 / 0.8428	0.6564 / 0.6138	91.71	8.29
500	0.3039 / 0.2456	0.6158 / 0.6853	0.8365 / 0.8175	0.5690 / 0.5357	89.18	10.82
1000	0.3227 / 0.3227	0.6215 / 0.6215	0.7809 / 0.7809	0.4291 / 0.4291	73.39	26.61
2000	0.4913 / 0.4788	0.5821 / 0.5796	0.7730 / 0.8064	0.4734 / 0.5369	65.37	34.63
4000	0.3574 / 0.4667	1.0624 / 1.0838	0.7809 / 0.7889	0.4291 / 0.4583	55.75	44.25
8000	0.5399 / 0.5399	1.2302 / 1.2302	0.7889 / 0.7889	0.4583 / 0.4583	52.12	47.88
16000	0.5390 / 0.5390	1.2887 / 1.2887	0.7889 / 0.7889	0.4583 / 0.4583	49.70	50.30

In conclusion, this appendix shows that the comparison of MIC clustering quality measures when different numbers of features/descriptors (m) are considered has to be done with caution, more especially for homogeneity indexes and inertia (accuracy indexes are more stable) and when a significant amount of noise exists in the data.

However, the pre-processing steps and tools discussed in this thesis are intended to decrease and control this amount of noise during metabolomics analyses.

CHAPTER 2

PAPER I : Statistical treatment of 2D NMR COSY spectra in metabolomics

Metabolomics (2015) 11:1756–1768
DOI 10.1007/s11306-015-0830-7

ORIGINAL ARTICLE

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

Baptiste Féraud[1,2] · Bernadette Govaerts[1] · Michel Verleysen[2,3] · Pascal de Tullio[4]

Received: 3 October 2014 / Accepted: 13 July 2015 / Published online: 23 July 2015

This Chapter 2 includes the full text of Paper I, which is entitled *"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"* (Féraud, B., Govaerts, B., Verleysen, M., and De Tullio, P., Metabolomics, 11(6), 1756-1768 (2015)).

It is aimed at introducing the Global Peak List (GPL, Section 1.2.2.1) approach for handling a set of 2D COSY spectra included in a whole experimen-

tal design, at introducing the Metabolomic Informative Content (MIC, Section 1.3.2) concept for evaluating and comparing the quality of the available data sources and finally at demonstrating that the COSY second dimension brings an informational added value compared with conventional 1D ^{1}H-NMR spectra.

Abstract

Compared with the widely used ^{1}H-NMR spectroscopy, two-dimensional NMR experiments provide more sophisticated spectra which should facilitate the identification of relevant spectral zones or biomarkers in metabolomics. This paper focuses on ^{1}H-^{1}H COSY (COrrelation SpectroscopY) spectral data. In spite of longer inherent acquisition times, it is commonly accepted by users (biologists, healthcare professionals) that the introduction of an additional dimension probably represents a huge qualitative step for investigations in terms of metabolites identification. Moreover, it seems natural that more information leads to more predictive power. But, until now, very few statistical studies clearly proved this assumption. Therefore a fundamental question is "Is this supplementary information relevant?". In order to extend the statistical properties developed for 1D spectroscopy to the challenges raised by 2D spectra, a rigorous study of the performances of COSY spectra is needed as a prerequisite. Having introduced new pre-processing concepts, such as the Global Peak List or an ad hoc 2D "bucketing", this paper presents an innovative methodology based on multivariate clustering algorithms to evaluate this question. Numerical clustering quality indexes and graphical results are proposed, based both on the spectral presence or absence of peaks (binary position vectors) and on peak intensities, and through different levels of spectral resolution. The second goal of this paper is to compare clustering performances obtained on COSY and on ^{1}H-NMR spectra, with the aim of understanding to what extent the COSY spectra carry more Metabolomic Informative Content (MIC) about the signal than 1D ones. The methodology is applied to two real experimental designs involving different groups of spectra (which define the signal): a 4-mixture cell culture media containing various supervised metabolites and a complex human serum based design. It is shown that COSY spectra appear to be statistically powerful and, in addition, provide better clustering results than corresponding ^{1}H-NMR when using unlabeled information. Consequently, additional information appears to be relevant for metabolomics applications.

Keywords: COSY spectra - Metabolomic Informative Content - Peak lists - Multivariate clustering algorithms - Real data

2.1 Introduction

Widely used for finding bio-active metabolites in living organisms and to study the impact of various biotic and abiotic effects, metabolomics studies [Nicholson et al., 2002] were initialy intended to search for metabolites interpretable as "biomarkers" for a given disease or toxicity. These discriminant metabolites are usually tracked in partial or complete spectral images, linked with biofluids (serum, urine) or tissues, using adequate univariate or multivariate statistical techniques. It is clear that the sampling preparation, spectral acquisition and data pre-processing have a crucial impact on the global quality of the data and on the chances to finally discover relevant biomarkers. Furthermore, an accurate statistical analysis will rarely compensate "dirty" initial data. Many acquisition and pre-processing tools are available and it is often difficult to objectively and quantitatively evaluate which ones will provide the more informative data in a given context. This motivated the development of the methodology described in this paper when data are organized in groups.

In this context, proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy generates spectral profiles describing the metabolite composition of collected samples. A comparison of several spectra of metabolites in various specific states permits a preliminary graphical and qualitative investigation of changes in metabolite composition inherent to the presence of a "stressor". However, the complexity of ^{1}H-NMR spectra and the number of spectra (of samples) usually available in metabolomics studies require a semi-automated data analysis. In addition, systematic differences between samples are often hidden behind biological noise and/or behind peak shifts.

To avoid these peak shifts and the usual overlappings of potentially independent signals, the use of two-dimensional homonuclear or heteronuclear experiments has gained attention these last years to identify metabolites. However, these tools are often left behind due to longer, and sometimes prohibitive, acquisition times. In this paper, COSY spectra are investigated [Aue et al., 1976] [Keeler, 2010] [Akitt and Mann, 2000]. COSY consists in a correlation-based method for determining which signals arise from neighboring protons (usually up to four bonds). Correlations appear when there is spin-spin coupling between protons (i.e. correlation between two or more nearby chemical processes). Several works exist in the literature in order to get around the time-consumming acquisition problem, specifically for multidimensional COSY data (for example Ultrafast-COSY [Giraudeau, 2010] [Queiroz et al., 2013] or PALSY-COSY [Vega et al., 2010]). Compared to ^{1}H-NMR experiments for instance, the introduction of an additional dimension should allow a better representation of metabolites, a better predictive power and a better biomarker identification. Several works are already based on 2D-COSY or on faster extentions of COSY,

such as in [Xi et al., 2007], [Yun et al., 2013] or [Le Guennec et al., 2014]. But, until now, very few statistical studies clearly tackled the potential superiority of 2D spectra (these studies mainly concern the group of Bruschweiler, see for instance [Bruschweiler and Bingol, 2011] and [Bruschweiler et al., 2014]). This lack leads to a central and fundamental question: is this supplementary information really relevant in the context of metabolomics analyses? If it is the case, the interest of using two-dimensional methods would be legitimated, even at the price of potentially superior acquisition times.

In this paper, the innovative notion of "Metabolomic Informative Content" (MIC) is central. The intention is to systematically evaluate the amount of captured information (i.e. the extent to which signals are captured) compared with the noisy part through several spectra. It goes away from the more classic version of repeatability in spectrometry which involves only one spectrum at a time. Generally, a measure can be considered as the sum of a signal (useful information) and noise. In many metabolomics studies, the signal is controlled and often linked with the existence of different mixtures, different samples or different groups of people (people affected by a disease vs. healthy people, people affected by a disease at different levels, ...). And the noisy part of the information is generally provoked by the characteristics of the design, the methods of acquisition, the temporal dimension (repetitions during time, deterioration of the samples, etc...) and the processing. The purpose would be to measure the predominance of the signal with regard to noise, but it is not obvious in a non-supervised context. A way is to suppose that only the useful information can be "repeatable" and that the noise is independent and identically distributed, with $signal \perp\!\!\!\perp noise$. With this hypothesis, evaluating the Metabolomic Informative Content gives an idea if $signal >> noise$ or not. In other words, for any set of spectra, it allows a direct comparison between different spectral tools and can also be seen as an evaluation of predictive ability in this paper.

According to this context and by using peak lists data sets, the first goal is to show that COSY spectra are successful, i.e. that COSY spectra allow to capture the main part of the information connected to the signal(s). The second goal is to demonstrate that COSY spectra are "more successful" than ^{1}H-NMR corresponding spectra and, by doing so, to demonstrate the utility and the importance of the additional second dimension. To reach these goals and to obtain quantitative responses, a multivariate clustering approach is selected and applied on unlabeled spectra in order to recover the signal. Two clustering algorithms (hierarchical Ward algorithm and K-Means algorithm) are used, as well as multiple combinations between different distance measures, between the

use of binary positions vectors (presence or absence of a peak) and of intensity vectors and between different spectral resolutions.

The paper is organized as follows. Section 2.2 provides a detailed description of the two experimental designs used to illustrate the methodology and reach the goals. These experimental designs are both involving real data: the first one implies repeated measurements of 4-mixture cell culture media containing various supervised metabolites, and the second one, a human serum based design with time sampling repetitions and multiple measurement permutations. In each case, COSY spectra and ^{1}H-NMR spectra are collected together for further comparisons. Section 2.3 provides a description of the different pre-processing steps: construction of a "Global Peak List", construction of binary positions vectors and intensity vectors, symmetrisation, water peak deletion, outlier deletion, etc... In this section, it is also explained how the spectral resolution is controlled (and the size of the data sets) by performing a so-called "bucketing" based on changes in the number of decimals in the global matrix. Clustering algorithms and associated distance measures are also detailed in this section. Section 2.4 contains the results: first, an analysis of the COSY spectra performances based on both positions (presences or absences of peaks) and intensities. Numerical outputs and a dendrogram illustrate these results. The comparisons between 1D and 2D spectra are also discussed in Section 2.4. Finally, a general conclusion and further works are given in Section 2.5.

2.2 Materials: COSY spectra, experimental designs and acquisition parameters

In this section, the second dimension in COSY spectra will motivate the main goal of this paper. Details about the two real experimental designs used to illustrate the methodology are provided. And finally, a technical explanation of ^{1}H-NMR and COSY spectra acquisition is proposed in Section 2.2.2.

2.2.1 Experimental designs

The methodology presented in this paper is applied on two real data sets, obtained from two designed experiments.

2.2.1.1 First design: cell culture media

The first design is based on four different mixtures (four cell culture media containing various levels of different metabolites like fetal bovine serum, amino acids, vitamins, proteins, etc...): DMEM/F12, MEM, RPMI/1640 and DMEM.

500μl of cell culture media were supplemented with 200μl of deuterated phosphate buffer (0.1M, pH=7.4) containing 1% of sodium azide and 10μl of TMSP (10 mg/ml). The solutions were then transferred into 5mm NMR tubes before NMR measurement. Three samples per mixture were collected, and three repeated measures were performed on each sample. All samples were subject to freezing (-20°C) and defrosting steps before te analysis, with real risks of degradation and bacterial contamination because of the duration of the 2D analysis process.

In this design, signal corresponds to the four initial mixtures and noise arises from the sampling, time replicates, risks of degradation and other acquisition and condition parameters. 36 measures are finally available, corresponding to 36 COSY spectra and 36 corresponding peak lists. These peak lists are $(t_i \times 3)$ matrices: with t_i the number of detectable peaks (whose values are above a machine-designed threshold) in sample i ($i = 1, ..., 36$). The three columns in these matrices correspond to the coordinate, or chemical shift, on the first axis (ppm), the coordinate on the second axis (ppm) and the raw intensity of the peak. The 36 corresponding ^{1}H-NMR spectra were also collected. The total number of spectra is called n in the remainder of the paper.

2.2.1.2 Second design: human serum

The second experimental design is based on human serum. Peripheral blood was collected in serum separating tubes (Greiner). Sera were distributed into 0.5ml aliquots and stored at -80°C just after sampling. Four blood donors were engaged for the study. For each collected sample, 500 μl of serum, defrosted just before the analysis, were supplemented with 200μl of deuterated phosphate buffer (0.1M, pH=7.4) containing 1% of sodium azide and 30μl of TMSP (10 mg/ml). The solutions were then transferred into 5mm NMR tubes before NMR measurement. The design consists in eight days of measurements with replicates within each day and multiple permutations according to a latin hypercube sampling (LHS) method [Iman, 2008] (in order to avoid confusion between donors and times of analysis). For each day, the four donors samples were analyzed and provided four COSY spectra and four ^{1}H-NMR spectra. Spectral techniques (1D or 2D COSY) have not been applied at the same moment of the day, thus creating different delays before spectral measurement.

Finally, eight measures/spectra/peak lists are obtained per donor, which corresponds to 32 measures in all (32 1D spectra and 32 COSY spectra). Again, peak lists are from particular interest and correspond to $(t_i \times 3)$ matrices with t_i the number of detectable peaks in the sample, or spectra, i ($i = 1, ..., 32$).

2.2.2 Spectra acquisition

1D and 2D spectra were recorded at 298K on a Bruker Avance spectrometer operating at 500.13MHz for the proton signal acquisition. The instrument was equipped with 3 channels and with a 5mm triple resonance (HCN) cryoprobe with a Z-gradient and automatic tuning and matching system (ATMA). All samples were locked using deuterated water and shims were tuned using autotune (Topshims) and samples are measured manually without spinning. Due to the nature of the samples, a presaturation sequence was used in all the experiments in order to minimize the water signal. All data were referenced to internal sodium 3-trimethylsilyl-2,2,3,3-d4-propionate (TMSP) at 0.00 ppm chemical shift (all spectra are calibrated with regard to TMSP).

According to sample type, the ^{1}H-NMR spectra were acquired using either a 1D noesypresat sequence (cell culture mixture) or a CPMG relaxation-editing sequence with presaturation (human sera). Upon the presence of proteins in serum, the use of a sequence with a T2 filter (CPMG) greatly improves the baseline.

The noesypresat experiment (pulse sequence noesygppr1d supplied by Bruker) used a RD-90°-t1-90°-tm-90°-sequence with a relaxation delay of 4s, a mixing time of 100ms and a fixed t1 delay of 20μs (spectral window of 10000 Hz). The water suppression pulse was placed during the relaxation delay (RD). The number of transients was typically 32. The acquisition time was fixed to 3.2769001s and a quantity of four dummy scans was chosen. FIDs were collected in 64 K time data points. The data were processed with the Bruker Topspin 2.1 software with a standard parameter set. The phase and baseline corrections were performed manually over the entire spectral range and a line broadening of 0,3 Hz was applied.

The CPMG experiment (pulse sequence cpmgpr1d supplied by Bruker) used a RD-90°-(t-180°-t)$_n$-sequence with a relaxation delay (RD) of 2s, a spin echo delay (t) of 400μs and the number of loops (n) equal to 80. The water suppression pulse was placed during the relaxation delay (RD) and a spectral window of 10245 Hz. The number of transients was typically 32. The acquisition time was fixed to 3.982555s and a quantity of four dummy scans was chosen. FIDs were collected in 64 K time data points. The data were processed with the Bruker Topspin 2.1 software with a standard parameter set. The phase and baseline corrections were performed manually over the entire spectral range without line broadening.

Gradient enhanced magnitude COSY experiment (pulse sequence cosygpprqf supplied by Bruker) with a presaturation during relaxation delay was used for 2D measurements. Spectra were collected with 4096 points in t2 and 300 points in t1 over a sweep width of 10 ppm, with six scans per t1 value. The

acquisition times were fixed to 0.2557028s in F2 and 0.0187212 in F1. The resulting COSY spectra were processed in Topspin 2.1 using standard methods, with sine-squared apodization in both dimensions and zero filling in t1 to yield a transformed 2D dataset of 2048 by 2048 points.

Individual peak lists were then extracted using ACD/Labs 12.00 (ACD/NMR processor, freeware). For each 2D spectra, a $(t_i \times 3)$ matrix provides the (X, Y) coordinates of the t_i peaks observed in the spectrum and their related intensities in a third column.

2.3 Methods: global peak list, pre-processing steps and clustering analysis

In this section, all data manipulations and pre-processing steps, both for standard metabolomics studies and for the particular purpose of this paper, are detailed. Then, a discussion on how to control the data resolution and, consequently, the size of the databases is proposed. Finally, clustering algorithms and related distance measures are also described in detail.

2.3.1 Global Peak List matrix

When one wants to perform simultaneous data analysis of a set of COSY spectra expressed in peak lists, a first requirement is to gather them together in a global object. The solution proposed here consists in building a so-called "Global Peak List" (GPL) matrix from the n $(t_i \times 3)$ individual peak list matrices available for each of the n 2D spectra. This $(T \times N)$ GPL matrix includes, as rows, the T pairs of coordinates that appear in at least one of the individual spectra. The $N = 2 + 2n$ columns include first the two peak coordinates columns, and then, for each individual spectrum, two columns corresponding to the observed peak intensities and to a deduced binary number (1 or 0) indicating if the peak appears or not in the spectrum (i.e. corresponds then to a strictly positive or to a null intensity). For the first design, the GPL is a (3250×74) matrix $(74 = 2 + 2 * 36)$. For the second one, the GPL is a (6686×66) matrix $(66 = 2 + 2 * 32)$. The GPL matrix can also be viewed as a combination of three matrices: a $(T \times 2)$ matrix of coordinates, a $(T \times n)$ matrix of intensities I and a $(T \times n)$ matrix of positions P (presence or absence).

In this paper, both spectral intensities and peak positions are considered of particular interest. Working on the signals' positions or, in other words, on the simple existence of signals is motivated by a biological justification: a signal, or a particular metabolite, can be observed or not for a particular donor or in

a particular media. If a signal is present in detectable quantity, this presence or absence is supposed to be stable, whereas intensities are variable from one measure to another, according to potential uncontrolled factors. Working on positions, or absence/presence, can then be seen as a qualitative approach; working on intensities can be seen as a more quantitative approach as it is directly linked with concentrations.

2.3.2 Pre-processing steps

The following pre-processing steps have been implemented on the Global Peak List matrices:

- **Symmetrisation.** By construction, all COSY spectra have to be symmetrical around the diagonal (see Section 2.2). Consequently, all other points or peaks are artifacts and have to be removed. It mainly concerns negative intensities and typical "crosses" which arise when choosing a wrong baseline and/or when some signals are abnormally too intense [Marion, 2016].

- **Water zone deletion.** Water is not of interest for metabolomics investigations. In COSY spectra, water peaks are concentrated in a square area between 4.5 and 5.5 ppm. Intensities and positions inside this area were simply removed by assigning them a zero value. A light loss of information can occur concerning metabolites present in this zone. More advanced methods also exist, see for example [Mao and Ye, 1997].

- **Normalization of the intensities.** Vectors of recoded intensities are simply obtained by applying a constant sum (CS = 1) normalization [Craig et al., 2006] after water zone deletion. This method is also known as 1-Norm [Rasmussen et al., 2011].

- **Outlier deletion.** Some spectra can contain unexpected extreme signal values due, for example, to exceptional external factors. These outliers should be identified before data analysis and removed because they could be too influent. On both position and recoded intensities vectors, Principal Component Analysis (PCA) were applied for graphical identification of outliers, and Mahalanobis distances were calculated between raw spectra and between group centered spectra. In the context of the first experimental design, three spectral outliers were found and ignored for upcoming analysis (because simultaneously suspected via positions and intensities). In the same way, three spectra were suspected to be outliers for the second design (see Fig.2.1 for an illustration).

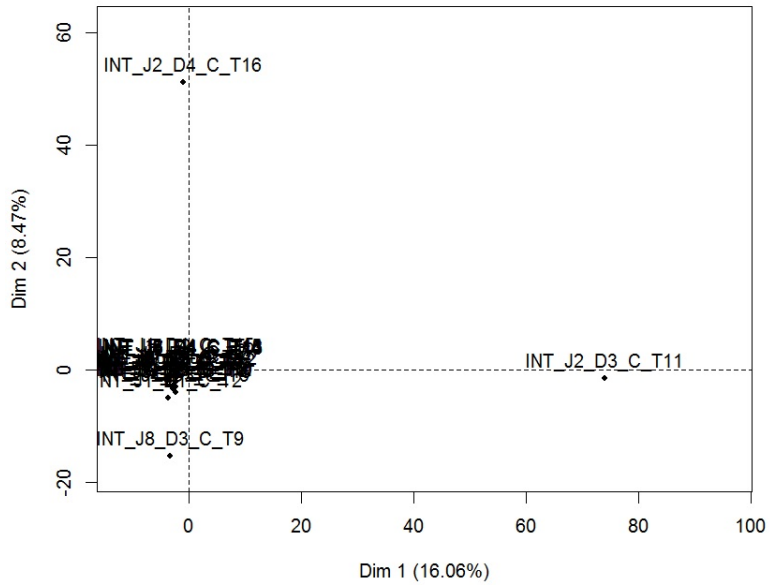


Figure 2.1: Graphical identification of outliers using PCA's score plot. Here, three spectral outliers are detected based on intensities vectors: Day 2 (J2) at Time 16 (T16) for donor 4 (D4), Day 2 (J2) at Time 11 (T11) and Day 8 (J8) at Time 9 (T9) for donor 3 (D3).

2.3.3 Control of data resolution

In ^{1}H-NMR spectroscopy, bucketing tools are common and widely used to control the spectral resolution and/or to overcome the misalignments problem. Classical and more advanced bucketing methods have already shown their usefulness for ^{1}H-NMR spectra [Rousseau, 2011] [Sousa et al., 2013]. In this paper, a bucketing step adapted to 2D-COSY is proposed in order to control the size of the database and, consequently, the resolution level of the two-dimensional spectra. Practically, a variation of the number of decimals of the coordinates is 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. In this paper, three, two and one decimal cases are tested for analysis. Using this method, the width of the COSY peaks is adjusted and the resolution and size of the databases are also adjusted simultaneously. Furthermore, intermediate resolutions can of course be computed in the same way.

Table 2.1: Dimensions of the GPL matrices according to different resolutions

	One decimal	Two decimals	Three decimals
First design	(909×74)	(2348×74)	(3250×74)
Second design	(1106×66)	(4172×66)	(6686×66)

For positions, the aggregation process leads to another dummy variable and results from a simple function: a peak is considered in a bucket if at least one peak is present at the lower resolution level. For example, 1 and 1 lead to 1, 1 and 0 lead to 1 and 0 and 0 lead to 0 (absence everywhere). Finally, for the two experimental designs, the GPL matrices have the dimensions described in Table 2.1 after "bucketing" and before outlier deletion.

2.3.4 Clustering algorithms and distance measures

This section introduces a methodology able to quantify and compare with adequate indexes the Metabolomic Informative Content (in a sense of advantageous signal capture compared to noise) of different sets of spectra. This methodology is applied in Section 2.4 to the two designs (each organized in four groups: mixtures or blood donors) in order to compare the MIC of 1D and 2D COSY spectra.

For 2D experiments, several combinations between intensities and position vectors, and between the use of one, two or three decimal(s) GPL matrices is tested. Using all this information, an intuitive way to evaluate the MIC of COSY spectra consists in non-supervised multivariate clustering (blind, with no a priori labeled information). The key idea is quite simple: if one manages to well separate and recover the four initial mixtures starting from the 36 unlabeled spectral measures of the first design, and/or if one manages to well separate and recover the four blood donors starting from the 32 unlabeled spectral measures of the second design, the goal is reached. It would mean that the signal, specific to each group of spectra, is sufficiently captured, despite the noise due to the time repetitions, the sampling methods, the freezing and defrosting periods of the samples, the risks of bacterial contamination, the potential environmental changes, etc... In other words, the objective is to verify that the within-cluster (intra) variance is minimized and that the between-cluster (inter) variance is maximized during data acquisition. Clustering is a natural and accurate tool to check this problem.

Two well-known clustering algorithms are considered: the hierarchical Ward algorithm and the K-Means algorithm (for intensities only in this second case).

- **The Ward's Algorithm** [Ward, 1963] [Murtagh and Legendre, 2011] is a commonly used procedure for forming hierarchical groups of mutually exclusive subsets. It is particularly useful for large-scale studies (ideally with more than 100 objects) when a precise optimal solution for a specified number of groups is not practical. Given n objects, this procedure reduces them to $n-1$ mutually exclusive sets by considering the union of all possible $n(n-1)/2$ pairs and selecting the union having the minimal dissimilarity measure or aggregation index. Initially, each object is assigned to its own cluster and then the algorithm proceeds iteratively, at each stage joining the two most similar clusters, continuing until there is just a single cluster. At each stage distances between clusters are recomputed by the Lance-Williams dissimilarity update formula [Murtagh and Legendre, 2011] [Murtagh and Contreras, 2012]. In this work, the *hclust* R (http://www.R-project.org) function is used, which performs hierarchical clustering and proposes a set of dissimilarity measures.

- **K-means clustering** [McQueen, 1967] is a vector quantization method, originally from signal processing, that is popular for cluster analysis in machine learning. K-means clustering aims at partitioning the n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. The most common algorithms use an iterative refinement and alternates between assignment steps and update steps [McKay, 2003]. They can only be applied on intensities here because means can not be properly defined for binary variables. The *kmeans* R function is used, allowing the joint use of the Hartigan-Wong, Lloyd and McQueen algorithms [Hartigan and Wong, 1979] [Lloyd, 1957, 1982].

Of course, both algorithms need the choice of an appropriate similarity measure, or distance, to quantify the neighborhood relation between two objects.

Clustering on binary positions vectors is first considered here. Specific similarity measures such as Ochiai, Jaccard, Dice or Russel-Rao are needed to capture the binary specificity [Plasse et al., 2007]. To avoid redundancy, only two of them are used in this paper: the Jaccard and Ochiai ones. Given p binary (0=absent; 1=present) attributes, like the positions in this paper, these similarity measures between any two objects X and Y of a library are built from a general contingency table, counting the number of common attributes. Let us call a the number of common attributes when $X = Y = 1$, b when

$X = 1$ and $Y = 0$, c when $X = 0$ and $Y = 1$ and finally d when $X = Y = 0$. On this basis, the Jaccard and Ochiai similarity measures have both a range of 0 to 1 and are defined as follows:

$$Jaccard(X,Y) = \frac{a}{a+b+c} = \frac{|X \cap Y|}{|X \cup Y|} \tag{2.1}$$

$$Ochiai(X,Y) = \sqrt{\frac{a}{a+b}}\sqrt{\frac{a}{a+c}} = \frac{|X \cap Y|}{\sqrt{|X|.|Y|}}. \tag{2.2}$$

Obviously, two spectra with many common peaks are supposed to provide high measures. The Jaccard measure can be interpreted as the size of the intersection divided by the size of the union of the sample sets; and the Ochiai measure as a geometric mean of the probabilities that if one object has the attribute, the other object has it too. Because product increases weaker than sum when only one of the terms grows, Ochiai will be really high only if both of the two proportions (probabilities) are high. It implies that the objects must share a great part of their attributes to be considered similar by Ochiai. Notice that the Ochiai coefficient can be considered as being identical to the cosine similarity index [Holliday et al., 2002].

Clustering algorithms on recoded intensities are also implemented and use the classic euclidean distance as similarity measure. In Cartesian coordinates, if $X = (x_1, x_2, ..., x_p)$ and $Y = (y_1, y_2, ..., y_p)$ are two objects in $\mathbb{R}^p$, then the euclidean distance from X to Y, or from Y to X, is given by:

$$d(X,Y) = d(Y,X) = \sqrt{\sum_{i=1}^{p}(y_i - x_i)^2} = ||Y - X||_2 . \tag{2.3}$$

In this paper, the number of clusters is set to $K = 4$ in order to correspond to the initial number of mixtures (first design) and to the initial number of blood donors (second design). Multiple combinations between Ward/K-Means algorithms, between intensities/positions vectors, between one/two/three decimals for the resolution of the GPL matrices (and between Ochiai/Jaccard measures for the binary part of the clustering) were experimented; results are discussed in Section 2.4.

2.3.5 Numerical indexes to evaluate the quality of the clustering results

To allow an objective and numerical evaluation of the quality of the clustering results, four indexes have been used: Dunn, Davies-Bouldin, Rand and

Adjusted Rand indexes. The two first evaluate the homogeneity of clusters regardless of the initial groups content; the two last measure the correctness of the non-supervised clustering process according to these initial groups.

The **Dunn index (DI)** [Dunn, 1973] corresponds to the ratio between the smallest distance between observations not in the same cluster and the largest intra-cluster distance. Let $\mathbf{C} = (C_1, ..., C_K)$ be a particular clustering partition of n objects into K disjoint clusters. Let Δ_m be the maximum distance between observations in the cluster C_m. The Dunn index is computed as:

$$DI_{\mathbf{C}} = \min_{C_k, C_l \in \mathbf{C}, C_k \neq C_l} \left\{ \frac{\min_{i \in C_k, j \in C_l} dist(i,j)}{\max_{C_m \in \mathbf{C}} \Delta_m} \right\}. \tag{2.4}$$

In this work, the evaluation of inter and intra-cluster distances is based on euclidean, Ochiai or Jaccard metrics according to the nature of the data.

The **Davies-Bouldin index (DBI)** [Davies and Lampen, 1993] is an internal evaluation scheme, where the validation of how well the clustering has been done is evaluated using quantities and features inherent to the dataset. Let C_i be a cluster. Let x_j be a p dimensional feature vector of the objects assigned to cluster C_i, A_i the medoid associated with C_i and $|C_i|$ the cardinality of C_i. Medoids are preferred compared with centroids in order to be able to work on binary positions. With these preliminary definitions, let Γ_i be the euclidean measure of variation within this cluster:

$$\Gamma_i = \sqrt{\frac{1}{|C_i|} \sum_{j=1}^{|C_i|} ||x_j - A_i||_2} \, .$$

The euclidean example is detailed here but note that many other distance metrics can be used. Again, in the case of binary variables, the euclidean distance is replaced by the Ochiai and Jaccard ones. Let also $\gamma(C_i, C_j)$ be a measure of separation between cluster C_i and cluster C_j, and $a_{i,l}$ the l^{th} element of A_i, $(l = 1, ..., L)$. One can write:

$$\gamma(C_i, C_j) = d(A_i, A_j) = ||A_i - A_j||_2 = \sqrt{\sum_{l=1}^{L} |a_{i,l} - a_{j,l}|^2} \, .$$

On this basis, the Davies-Bouldin index is defined as follows:

$$DBI_{\mathbf{C}} \equiv \frac{1}{K} \sum_{i=1}^{K} \max_{j:i \neq j} \left\{ \frac{\Gamma_i + \Gamma_j}{\gamma(C_i, C_j)} \right\}. \tag{2.5}$$

Finally, the **Rand index (RI)** and the **Adjusted Rand index (ARI)** [Hubert and Arabie, 1985] are measures of the similarity between two data clusterings, used to evaluate the quality of the classification. From a mathematical point of view, the Rand index is related to the accuracy, but is applicable even when class labels are not directly used. Given again a set of n objects $X = \{x_1, \ldots, x_n\}$ and two partitions of X to compare, $C^1 = \{C_1^1, \ldots, C_{K_1}^1\}$, a partition into K_1 subsets, and $C^2 = \{C_1^2, \ldots, C_{K_2}^2\}$, a partition into K_2 subsets, let us define the following notations:

- m_a as the number of pairs of elements in X that are in the same set in C^1 and in the same set in C^2,
- m_b as the number of pairs of elements in X that are in the same set in C^1 and in different sets in C^2,
- m_c as the number of pairs of elements in X that are in different sets in C^1 and in the same set in C^2,
- m_d as the number of pairs of elements in X that are in different sets in C^1 and in different sets in C^2.

The Rand index is defined as:

$$RI = \frac{m_a + m_d}{m_a + m_b + m_c + m_d} \tag{2.6}$$

Intuitively, $m_a + m_d$ can be considered as the number of agreements between C^1 and C^2, and $m_b + m_c$ as the number of disagreements. The Rand index has a value between 0 and 1, with 0 indicating that the two data clusters do not agree on any pair of points and 1 indicating that the data clusters are exactly the same.

The Adjusted Rand index (ARI) is the corrected-for-chance version of the Rand index. Though the Rand Index may only yield a value between 0 and 1, the Adjusted Rand Index can yield negative values if the index is less than the expected index. Following the same notations m_a, m_b, m_c and m_d, it is defined as [Santos and Embrechts, 2009]:

$$ARI = \frac{\binom{n}{2}(m_a + m_d) - [(m_a + m_b)(m_a + m_c) + (m_b + m_d)(m_c + m_d)]}{\binom{n}{2}^2 - [(m_a + m_b)(m_a + m_c) + (m_b + m_d)(m_c + m_d)]} \tag{2.7}$$

Note that DI, RI and ARI have to be as high as possible to represent good clustering partitions. DBI has to be as small as possible.

2.4 Results and discussion

In this section, clustering methods and related quality indexes are used to assess the performance of different versions of COSY spectra. The first upcoming subsection proposes numerical outputs based on the indexes defined in Section 2.3.5 for both designs and for combinations of experiments described in Section 2.3.3 and Section 2.3.4. Then, in Section 2.4.2, comparisons between clustering results based on 2D COSY spectra and on corresponding ^{1}H-NMR spectra are shown. This section will allow to demonstrate on the two designs that, indeed, COSY spectra include additive relevant information for spectral classification as compared to ^{1}H-NMR ones (providing consequently a better MIC). By doing this, a statistically-proved response to the initial key question (does supplementary information include relevant information?) is provided.

2.4.1 Results for COSY spectra

The numerical results are available in Table 2.2 for the first experimental design and for the more complex second one. Note that only a part of all the considered combinations are shown to avoid too much redundancy and to improve clarity.

Globally, the results are very promising with many RI (and ARI) greater than 0.7, particularly with bucketed data when one or two decimals in the GPL matrices are used. Unlabeled individuals are well grouped together, with no prior knowledge, and this according to the presence of numerous potential noise factors (sampling, time replicates, degradation, changes in temperature, etc...).

Considering first the position-based partitions, particular groups are already well isolated in the majority of cases (for instance, the third cell culture mixture in the first design was known to be significantly different from the others). An example is given in Fig.2.2(A).

With the recoded intensities-based partitions, the four mixtures are generally well recovered by the algorithms in spite of the sampling procedure and time repetitions. For the first design, the best clustering result is obtained with the one-decimal GPL matrix, thus underlying the importance of the "bucketing" step: Fig.2.2(B) shows that there is only one error during the blind clustering process and RI and ARI indexes are maximized (RI=0.927 and ARI=0.853 in Table 2.2). The conclusion is the same for the second design concerning the blood donors (RI=0.932 and ARI=0.804 in Table 2.2).

More generally, Dunn and Davies-Bouldin indexes are subject to significative variations between experiments and are not very satisfactory for some combinations (several DI less than 0.5 and DBI greater than 1.5 in Table 2.2).

Table 2.2: Numerical clustering performances according to different versions of COSY spectra. The column "Dec." denotes the number of decimal(s) for the data in the GPL matrix; T = number of observations in the corresponding GPL; DI = Dunn Index; DBI = Davies-Bouldin Index; RI = Rand Index; ARI = Adjusted Rand Index.

FIRST DESIGN								
Signal	Dec.	Algorithm	Distance	T	DI	DBI	RI	ARI
Positions	1	Ward	Jaccard	909	0.712	1.466	**0.914**	**0.811**
Positions	1	Ward	Ochiai	909	0.712	1.466	**0.914**	**0.811**
Positions	2	Ward	Jaccard	2348	**0.857**	1.688	0.757	0.420
Positions	2	Ward	Ochiai	2348	**0.827**	1.688	0.757	0.420
Positions	3	Ward	Jaccard	3250	**0.893**	1.722	0.601	0.189
Positions	3	Ward	Ochiai	3250	**0.892**	1.722	0.601	0.189
Intensities	1	Ward	Euclidean	909	0.298	**0.541**	**0.927**	**0.853**
Intensities	1	K-means	Euclidean	909	0.298	**0.541**	**0.927**	**0.853**
Intensities	2	Ward	Euclidean	2348	0.239	1.249	0.669	0.609
Intensities	2	K-means	Euclidean	2348	0.311	1.356	0.657	0.501
Intensities	3	Ward	Euclidean	3250	0.439	1.540	0.588	0.425
Intensities	3	K-means	Euclidean	3250	0.434	1.614	0.597	0.439
SECOND DESIGN								
Signal	Dec.	Algorithm	Distance	T	DI	DBI	RI	ARI
Positions	1	Ward	Jaccard	1106	0.796	1.569	**0.937**	**0.825**
Positions	1	Ward	Ochiai	1106	0.722	1.569	**0.937**	**0.825**
Positions	2	Ward	Jaccard	4172	**0.945**	1.800	0.772	0.401
Positions	2	Ward	Ochiai	4172	**0.899**	1.800	0.772	0.401
Positions	3	Ward	Jaccard	6686	**0.981**	1.792	0.650	0.171
Positions	3	Ward	Ochiai	6686	**0.963**	1.792	0.650	0.171
Intensities	1	Ward	Euclidean	1106	0.419	**0.643**	**0.932**	**0.804**
Intensities	1	K-means	Euclidean	1106	0.419	**0.643**	**0.932**	**0.804**
Intensities	2	Ward	Euclidean	4172	0.689	1.592	0.789	0.422
Intensities	2	K-means	Euclidean	4172	0.706	1.742	0.735	0.288
Intensities	3	Ward	Euclidean	6686	0.778	1.668	0.578	0.055
Intensities	3	K-means	Euclidean	6686	0.765	1.886	0.647	0.135

This means that obtained clusters are not always compact and well separated. However, to judge if the clusters conform to the reality, i.e. if members in a cluster correspond to members of initial groups, Rand and Adjusted Rand indexes prevail because they are built on the degree of agreement and disagreement between groups (here between the reality and each of the clustering partitions).

In conclusion, and based on the two experimental designs, the clustering results show that COSY spectra appear to be statistically robust and contain informative signal which helps to succeed in distinguishing groups of different spectra. In other words, COSY spectra are enough informative so that the signal connected to the initial groups is distinguished from the noise. Working on

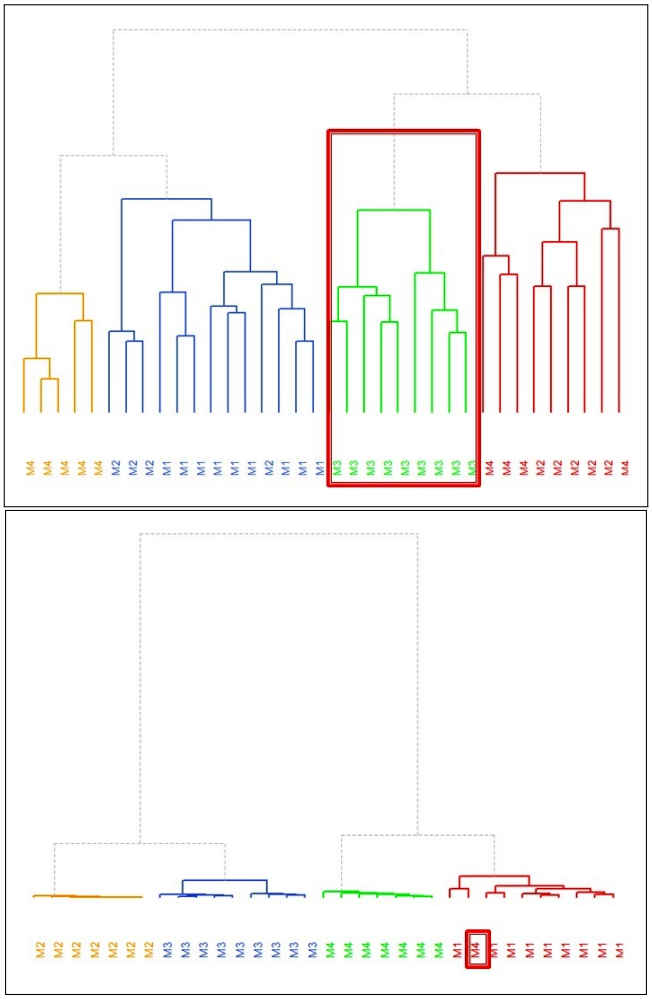


Figure 2.2: (A) Illustrative dendrogram of the clustering on positions (first design, Ward algorithm with Ochiai distance, dec = 2, before outlier deletion). M1 to M4 denote the four different initial mixtures. Different colours are attributed to different clusters; (B) Illustrative dendrogram of the clustering on intensities (first design, Ward algorithm, dec = 1, after outlier deletion).

positions gives satisfactory results but not better ones than working on intensities. Moreover, these results show that the 2D "bucketing" step is important and probably necessary to solve misalignment problems (as it is the case in 1D). This additional information can be profitable for further robust statistical analysis, as biomarker discovery.

2.4.2 Comparisons with corresponding ^{1}H-NMR spectra

Besides the convincing performances of COSY spectra, this section intends to demonstrate that the additional information contained in 2D spectra (compared to 1D) is relevant and crucial by improving the quality of the clustering results. It is always difficult to directly compare objects of different dimensions. To be able to compare ^{1}H-NMR and COSY spectra, some pre-processing steps are necessary to upgrade the one-dimensional spectra before performing clustering.

2.4.2.1 Upgrade of ^{1}H-NMR spectra

Some very powerful and complete tools are available to pre-process, transform and interpret ^{1}H-NMR data, for example in [Xia et al., 2008] [Vanwinsberghe, 2005]. Here, the goal is not to focus on sophisticated pre-processing methods for ^{1}H-NMR data, but just to upgrade them in order to obtain comparable pre-processing levels for both 1D and COSY data. To do so, one-dimensional spectra were subject to the following steps after phase and baseline correction:

- elimination of negative intensities,
- deletion of the water spectral zone between 4.5 and 5.5 ppm,
- application of the constant sum transformation,
- "bucketing" by using the same number of decimals (for the ppm coordinates) than for the COSY experiments,
- outlier deletion.

In a general way, it is quite intuitive that 2D spectra require less pre-processing to obtain an acceptable vizualisation of the signal and of potential biomarkers. Finally, note that the reasoning based on the positions makes no sense when working with 1D spectra. Consequently, clustering processes were only performed on 1D normalized intensities (raw intensities, "bucketed" with two decimals and with only one decimal).

2.4.2.2 Comparisons

The clustering results obtained with ^{1}H-NMR spectra are available in Table 2.3 for the two designs.

Table 2.3: Numerical clustering performances for 1D data. The column "Dec." denotes the number of decimal(s) for the data in the GPL matrix; T = number of observations in the corresponding GPL; DI = Dunn Index; DBI = Davies-Bouldin Index; RI = Rand Index; ARI = Adjusted Rand Index.

FIRST DESIGN								
Signal	Dec.	Algorithm	Distance	T	DI	DBI	RI	ARI
Intensities	1	Ward	Euclidean	201	**0.927**	**0.120**	**0.908**	**0.807**
Intensities	1	K-means	Euclidean	201	0.560	**0.463**	**0.835**	0.667
Intensities	2	Ward	Euclidean	2001	0.712	**0.426**	0.717	0.516
Intensities	2	K-means	Euclidean	2001	0.601	0.739	0.732	0.476
Raw intensities		Ward	Euclidean	199967	0.362	1.048	0.606	0.544
Raw intensities		K-means	Euclidean	199967	0.284	1.182	0.410	0.196
SECOND DESIGN								
Signal	Dec.	Algorithm	Distance	T	DI	DBI	RI	ARI
Intensities	1	Ward	Euclidean	207	**0.981**	**0.688**	0.704	0.233
Intensities	1	K-means	Euclidean	207	0.334	**0.663**	0.702	0.230
Intensities	2	Ward	Euclidean	2054	**0.901**	0.986	0.704	0.233
Intensities	2	K-means	Euclidean	2054	0.635	1.119	**0.740**	**0.290**
Raw intensities		Ward	Euclidean	205034	0.670	1.074	0.588	0.017
Raw intensities		K-means	Euclidean	205034	0.447	1.321	0.675	0.145

If one compares these results with Rand and Adjusted Rand indexes in Table 2.2, it appears clearly that these indexes are mainly higher when the clustering algorithms are performed on COSY spectra. Resulting partitions of the spectra are better for COSY: they are more in agreement with the real initial groups. Thus, this proves, for the two experiments, that the additional spectral dimension does provide relevant information to improve the clustering results and, consequently, to improve the amount of captured signal and the spectral informative content (MIC).

2.5 Conclusion and further works

In this article, an advanced clustering approach is proposed to evaluate and quantify the "Metabolomic Informative Content" (as defined in the introduction) of NMR spectral data in metabolomics. It is applied on both ^{1}H-NMR and COSY spectra, using both qualitative input (binary positions, linked with the absence or presence of a peak or metabolite) or quantitative inputs (intensities). More precisely, the choice of this clustering approach is to be related to the presence of initial groups in the signal. In the two real experimental

designs detailed in Section 2.3.1, four different cell culture mixtures and blood samples from four different donors were respectively involved. The elements of these groups were "mixed" via several noisy factors: sampling, time replicates, delays, deterioration, risk of bacterial contamination, etc... The goal of the clustering processes was to blindly recover these initial groups using all the individual unlabeled spectra. And the final quality of these processes informs us about the quantity of captured signal and can be finally viewed as a measure of MIC.

This work shows that COSY spectra allow to well discriminate groups linked with different signals, proving that they are successful in spite of the noisy factors. It also shows that the clustering processes perform better with COSY spectra than with ^{1}H-NMR ones, thus confirming a higher MIC for the two-dimensional spectra.

More precisely, this work first shows some pre-processing steps for the data analyst in order to handle 2D COSY data, including the construction of the Global Peak List (GPL) matrix and an ad hoc 2D bucketing step which allows to control the data resolution. It then demonstrates that COSY spectra provide very satisfying clustering results: in other words, in spite of the multiple noise factors, the informative part of the signal can be well discovered and, finally, the final clusters are mainly in concordance with the initial groups. The clustering methodology also highlights that 2D bucketing, by aggregating data according to reduced numbers of decimals for the ppm coordinates, helps to improve the amount of captured information when using COSY data.

Furthermore, this paper also provides a comparison between COSY and ^{1}H-NMR spectra, based on the quality of respective clustering results. It is demonstrated that COSY spectra provide more metabolomic informative content (MIC) about the groups than corresponding upgraded ^{1}H-NMR spectra, thus proving the importance and the relevance of the additional dimension of COSY to go further in NMR-based metabolomics studies. Note that this methodology can be generalized for comparing any spectrometric tools when groups are present and of interest in the experimental design.

These promising results have to be confirmed on faster versions of COSY experiments, in order to deal with more competitive acquisition times. The methodology can also be applied to heteronuclear two-dimensional spectroscopy tools, like HSQC spectra. Ultimately, once the MIC of a given 2D tool or technology is verified, the research of discriminating zones or biomarkers, which represents the main objective of most metabolomics studies, can benefit from the advantageous use of 2D-NMR spectra instead of the more traditional ^{1}H-NMR ones (for instance, in PLS-DA, OPLS-DA or more advanced machine learning techniques).

Softwares The raw data were firstly processed with the Bruker Topspin 2.1 software. Peak lists were extracted using ACD/Labs 12.00 (ACD/NMR processor). For manipulating the 1D and COSY data (pre-processings steps), generating the GPL matrices and performing the clustering processes, the R software environment was exclusively used (http://www.R-project.org).

Acknowledgements This work was supported by the FNRS from which P. de Tullio is senior research associate. Support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy) is also gratefully acknowledged.

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

Compliance with ethical requirements This study analyzes collected data which involved human participants who had provided informed consent.

Local bibliography

[Akitt and Mann, 2000] Akitt J. W., Mann B. E., NMR and Chemistry (Manual), Cheltenham UK, Stanley Thornes. p. 287 (2000).

[Aue et al., 1976] Aue W.P., Bartholdi E., Ernst R.R., Two-dimensional spectroscopy. Application to nuclear magnetic resonance, J. Chem. Phys. 64, pp. 2229-2246 (1976).

[Bruschweiler and Bingol, 2011] Bruschweiler R., Bingol K., Deconvolution of Chemical Mixtures with High Complexity by NMR Consensus Trace Clustering, Anal. Chem., 83 (19), pp. 7412-7417 (2011).

[Bruschweiler et al., 2014] Bruschweiler R., Bingol K., Bruschweiler-Li L., Li D.-W., Customized Metabolomics Database for the Analysis of NMR ^{1}H-^{1}H TOCSY and ^{13}C-^{1}H HSQC-TOCSY Spectra of Complex Mixtures, Anal. Chem., 86 (11), pp. 5494-5501 (2014).

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

[Davies and Bouldin, 1979] Davies D.L., Bouldin D.W., A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1 (2), pp. 224-227 (1979).

[Dunn, 1973] Dunn J.C., A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, Journal of Cybernetics, 3 (3), pp. 32-57 (1973).

[Giraudeau et al., 2009] Giraudeau P., Remaud G., Akoka S., Evaluation of Ultrafast 2D NMR for Quantitative Analysis, Anal. Chem., 81 (1), pp. 479-484 (2009).

[Hartigan and Wong, 1979] Hartigan J. A., Wong M. A., A K-means clustering algorithm, Applied Statistics 28, pp. 100-108 (1979).

[Holliday et al., 2002] Holliday J.D., Hu C.Y., Willett P., Grouping of coefficients for the calculation of Inter-Molecular Similarity and Dissimilarity using 2D fragment bit-strings, Combinatorial Chemistry and High Throughput Screening 5, Number 2, pp. 155-166 (2002).

[Hubert and Arabie, 1985] Hubert L., Arabie P., Comparing partitions, Journal of Classification 2 (1), pp.193-218 (1985).

[Iman, 2008] Iman R. L., Latin hypercube sampling, John Wiley and Sons, Ltd (2008).

[Jacobsen, 2007] Jacobsen N. E., NMR spectroscopy explained: simplified theory, applications and examples for organic chemistry and structural biology, John Wiley and Sons, Ltd (2007).

[Keeler, 2010] Keeler J., Understanding NMR Spectroscopy (2nd ed.), John Wiley and Sons, pp. 190-191 (2010).

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

[Lloyd, 1957, 1982] Lloyd S. P., Least squares quantization in PCM, Technical Note, Bell Laboratories, IEEE Transactions on Information Theory 28, pp. 128-137 (1957, 1982).

[McKay, 2003] MacKay D., An Example Inference Task: Clustering, Information Theory, Inference and Learning Algorithms, Cambridge University Press, pp. 284-292 (2003).

[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).

[Mao and Ye, 1997] Mao X., Ye C., Phase-shift presaturation for water peak suppression in biomolecular NMR experiments, Science in China. Series C, Life sciences, 40 (4), pp. 345-350 (1997).

[Marion and Bax, 1988] Marion D., Bax A., Baseline distortion in real-fourier-transform NMR spectra, Journal of Magnetic Resonance (1969), 79(2), pp. 352-356 (1988).

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

[Murtagh and Contreras, 2012] Murtagh F., Contreras P., Algorithms for hierarchical clustering: an overview, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1), pp. 86-97 (2012).

[Nicholson et al., 2002] Nicholson J., Connelly J., Lindon J.C., Holmes E., Metabonomics: a generic platform for the study of drug toxicity and gene function, Nature Reviews Drug Discovery, 1, pp. 153-161 (2002).

[Plasse et al., 2007] Plasse M., Niang N., Saporta G., Villeminot A., Leblond L., Combined use of association rules mining and clustering methods to find relevant links between binary rare attributes in a large data set, Computational Statistics and Data Analysis, 52(1), 596-613 (2007).

[Queiroz et al., 2013] Queiroz Junior L. H. K.; Ferreira A. G., Giraudeau P., Optimization and practical implementation of ultrafast 2D NMR experiments. Quím. Nova [online], vol.36, n.4, pp. 577-581 (2013).

[Rasmussen et al., 2011] Rasmussen L.G., Savorani F., Larsen T.M., Dragsted L.O., Astrup A., Engelsen S.B., Standardization of factors that influence human urine metabolomics, Metabolomics, 7(1), pp. 71-83 (2011).

[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).

[Santos and Embrechts, 2009] Santos J. M., Embrechts M., On the use of the adjusted rand index as a metric for evaluating supervised classification, Artificial Neural Networks, ICANN 2009, Springer Berlin Heidelberg, pp. 175-184 (2009).

[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).

[Vanwinsberghe, 2005] Vanwinsberghe J., Bubble: development of a matlab tool for automated ^{1}H-NMR data processing in metabonomics, Master's thesis, Université de Strasbourg (2005).

[Vega et al., 2010] Vega-Vazquez M., Cobas J.C., Martin-Pastor M., Fast multidimensional localized parallel NMR spectroscopy for the analysis of samples, Magn Reson Chem, 48(10): pp. 749-752 (2010).

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

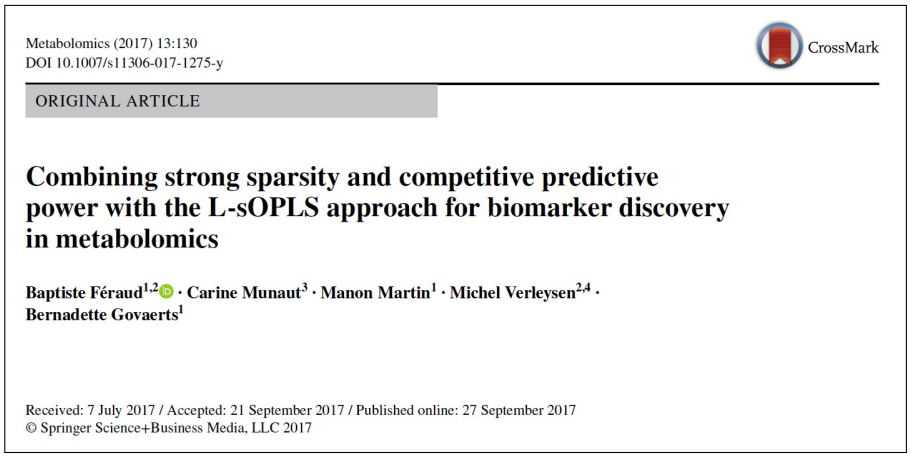
[Xi et al., 2007] Xi Y., deRopp J.S., Viant M., Woodruff D., Yu P., Automated screening for metabolites in complex mixtures using 2D COSY NMR spectroscopy, Metabolomics, Vol. 2, No. 4, pp. 221-233 (2007).

[Xia and Wishart, 2010] Xia J., Wishart D., MetPA: a web-based metabolomics tool for pathway analysis and visualization, Bioinformatics, 26(18), pp.2342-2344 (2010).

[Yun et al., 2013] Yun K., Sunghyouk P., Jongheon S., Dong-Chan O., Application of ^{13}C-labeling and ^{13}C-^{13}C COSY NMR experiments in the structure determination of a microbial natural product, Archive of Pharmacal Research, DOI 10.1007/s12272-013-0254-8 (2013).

CHAPTER 3

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



Metabolomics (2017) 13:130
DOI 10.1007/s11306-017-1275-y

CrossMark

ORIGINAL ARTICLE

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

Baptiste Féraud[1,2] · Carine Munaut[3] · Manon Martin[1] · Michel Verleysen[2,4] · Bernadette Govaerts[1]

Received: 7 July 2017 / Accepted: 21 September 2017 / Published online: 27 September 2017

This Chapter 3 includes the full text of Paper II, which is entitled *"Combining strong sparsity and competitive predictive power with the L-sOPLS approach for biomarker discovery in metabolomics"* (Féraud, B., Munaut, C., Martin, M. et al., Metabolomics, 13: 130. doi.org/10. 1007/s11306-017-1275-y (2017)).

In this original paper, the biomarker discovery issue in metabolomics is generally approached and the novel L-sOPLS algorithm (see Section 1.4.5) is in-

troduced in order to combine the orthogonality (Section 1.4.3) and the sparsity (Section 1.4.4) concepts.

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 ^{1}H-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 - ^{1}H-NMR data - RT-qPCR data

3.1 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.

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 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 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 3.2 provides a detailed description of the two real data sets used to perform the algorithms: a ^{1}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 Section 3.3. Results on both data sets are then shown and discussed in Section 3.4. Finally, a general conclusion and description of further works are given in Section 3.5.

3.2 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.

3.2.1 Spectral ^{1}H-NMR data set based on urine

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

3.2.1.1 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.3.1(a) and a typical urine spectrum with spiked citrate and hippurate is provided in Fig.3.1(d). 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.3.1(b) and Fig.3.1(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. Intuitivelly, 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 unsparse 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,

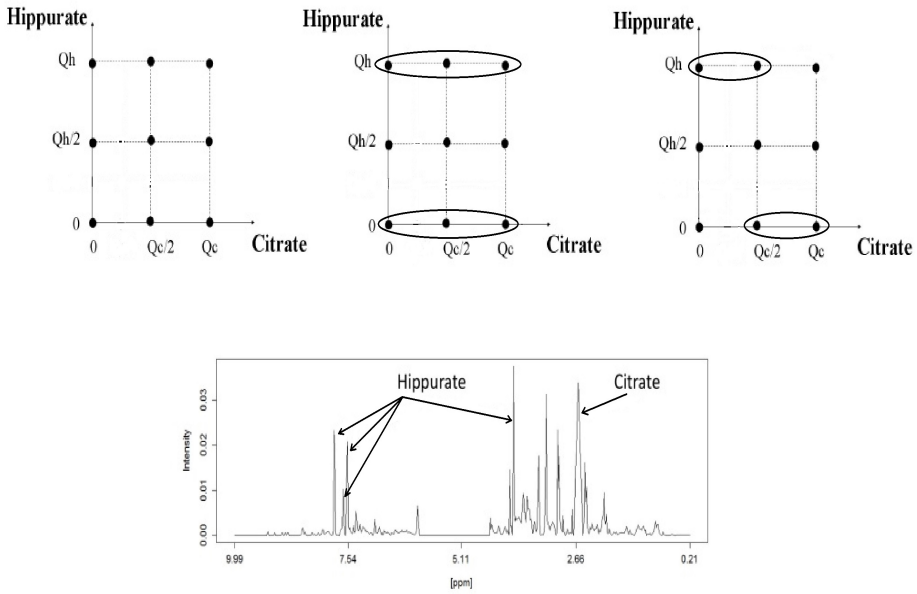


Figure 3.1: (a) The whole urine experimental design; (b) The hippurate balanced case; (c) The hippurate unbalanced case; (d) A typical urine spectrum with spiked citrate and hippurate.

the individual spectra, are identified by the way of the citrate and hippurate levels of concentration.

3.2.1.2 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.

3.2.2 RT-qPCR genomic data set

In this subsection, a description of a second data set is provided. It consists in RT-qPCR (Reverse Transcription-quantitative Polymerase Chain Reaction) 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.

3.2.2.1 Description and motivations

Endometriosis is principally 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.

3.2.2.2 Acquisition

Total RNA, including miRNA, was extracted from 200μl of serum using the miRNeasy Serum/Plasma Kit (Qiagen) according to the manufacturer's 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 manufacturer's 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 minutes, followed by 45 cycles at 95°C for 15 seconds, 55°C for 30 seconds, and 70°C for 30 seconds. All reactions were run in duplicate.

3.3 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.

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

Partial Least Squares (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 $W = (w_1, ..., w_q)$ from successive optimization

problems. For the iterative computation of each component $k = 1, ..., q$, the PLS goal can be written as follow:

$$\max_{w} \left\{ corr^2(y, Xw_k) var(Xw_k) \right\} \equiv \max_{w} \left\{ w_k' X_k' y y' X_k w_k \right\} \tag{3.1}$$

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

Alternatively, this criterion can be written:

$$\min_{w} \left\{ -w_k' M w_k \right\} \tag{3.2}$$

$$\text{subject to } w_k' w_k = 1, \text{ where } M = X' y y' X.$$

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, CPPLS (Canonical Powered Partial Least Squares

[Indahl et al., 2009]), SIMPLS (Straightforward Implementation of a Modification of PLS [De Jong, 1993]), NIPALS (NonLinear Iterative Partial Least Squares [Wold H, 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: $\boldsymbol{r_1 = X'y}$
- Iterations from $k = 1$ to q :
  1. Weights normalization: $\boldsymbol{w_k} = \frac{\boldsymbol{r_k}}{\sqrt{\boldsymbol{r_k' r_k}}}$
  2. Scores calculation: $\boldsymbol{t_k = X w_k}$
  3. Scores normalization: $\boldsymbol{t_k} = \frac{\boldsymbol{t_k}}{\sqrt{\boldsymbol{t_k' t_k}}}$
  4. Loadings calculation: $\boldsymbol{p_k = X' t_k}$
  5. Deflation step: $\boldsymbol{r_{k+1} = r_k - p_k (p_k' p_k)^{-1} p_k' r_k}$.
- Choice of q, the optimal number of components of the PLS model using an adequate validation criterion. The RMSEP (Root Mean Square Error of Prediction, [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 independent validation/test set is very often too expensive and not possible is most metabolomics studies.
- Matrices $\boldsymbol{T = XW}$ and $\boldsymbol{P = X'T}$ are obtained, based on the optimal q.
- Final PLS coefficients are calculated as $\boldsymbol{b_{PLS} = W(T'T)^{-1}T'y}$ 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, ^{1}H-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.

3.3.2 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] [Weljie et al., 2006]) 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 H, 1975] predictions are equal to SIMPLS predictions, described in Section 3.3.1, 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 $t_{ortho_i} p'_{ortho_i}$ [Wold, Trygg et al., 2001]. 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. $\boldsymbol{\tilde{w}_k = \frac{X'_k y}{y' y}}$

3. X-weights normalization to $\| \tilde{w}_k \| = 1$: $\boldsymbol{\tilde{w}_k = \frac{\tilde{w}_k}{\sqrt{\tilde{w}_k' \tilde{w}_k}}}$
4. X-scores computation: $\boldsymbol{t_k = X_k \tilde{w}_k}$
5. X-loadings computation: $\boldsymbol{p_k = \frac{X_k' t_k}{t_k' t_k}}$
6. Calculations of the orthogonal components' weights: $\boldsymbol{\tilde{w}_{ok} = p_k - \frac{\tilde{w}_k' p_k}{\tilde{w}_k' \tilde{w}_k} \tilde{w}_k}$, then normalized to $\boldsymbol{\tilde{w}_{ok} = \frac{\tilde{w}_{ok}}{\sqrt{\tilde{w}_{ok}' \tilde{w}_{ok}}}}$
7. Computation of orthogonal scores: $\boldsymbol{t_{ok} = X_k \tilde{w}_{ok}}$
8. And computation of orthogonal loadings: $\boldsymbol{p_{ok} = \frac{X_k' t_{ok}}{t_{ok}' t_{ok}}}$
9. Deflation step on X: $\boldsymbol{X_{k+1} = X_k - t_{ok} p_{ok}'}$
10. A whole deflated matrix can be extracted at the end of the iterations such that:

$$X_d = X - T_o P_o' = X - X_o \tag{3.3}$$

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

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

1. $\boldsymbol{\tilde{w}_d = \frac{X_d' y}{y' y}}$, then $\boldsymbol{\tilde{w}_d = \frac{\tilde{w}_d}{\sqrt{\tilde{w}_d' \tilde{w}_d}}}$
2. $\boldsymbol{t_d = X_d \tilde{w}_d}$
3. $\boldsymbol{p_d = \frac{X_d' t_d}{t_d' t_d}}$

- OPLS coefficients are obtained by:

$$b_{OPLS} = \tilde{w}_d (p_d' \tilde{w}_d)^{-1} \frac{y' t_d}{t_d' t_d} \tag{3.4}$$

- 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 stricly interpretable in the same way as b_{PLS}. So, b_{OPLS} can be transformed to match with X as follow:

$$b_{OPLScorr} = (I_m - (\tilde{W}_o (P_o' \tilde{W}_o)^{-1} P_o')) b_{OPLS} \tag{3.5}$$

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

$$\hat{y}_{OPLS} = X_d . b_{OPLS} = X . b_{OPLScorr} \tag{3.6}$$

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:

$$\hat{y}_{OPLS(q-1)} = \hat{y}_{PLS(q)} \tag{3.7}$$

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. 3.7), 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.

3.3.3 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 $m >> n$ 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 VIP (Variable Importance in the Projection criterion [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. 3.2)), the sPLS algorithm of Chun and Keles, used in this paper, provides an efficient implementation based on the LARS objective function:

$$\min_{w,c} \{(-\beta w_k' M w_k) + (1-\beta)(c-w_k)' M (c-w_k) + \lambda_1 \|c\|_1 + \lambda_2 \|c\|_2\} \quad (3.8)$$

$$\text{subject to } w_k' w_k = 1 \text{ for } k = 1, ..., q, \text{ where } M = X' y y' X.$$

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. 3.8) 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:

$$\min_{w} \{(-\beta w_k' M w_k) + (1-\beta)(c-w_k)' M (c-w_k)\} \quad (3.9)$$

$$\text{subject to } w_k' w_k = 1 \text{ for } k = 1, ..., q.$$

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:

$$\min_{c} \{(R'c - R'w_k)'(R'c - R'w_k) + \lambda_1 \|c\|_1 + \lambda_2 \|c\|_2\} \quad (3.10)$$

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 Equation (Eq. 3.10) 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. 3.8) 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 (Eq. 3.8) with $M = X_k' y y' X_k$,
  2. Update $\Pi = \Pi_k$ as $\{i : \hat{w_{ki}} \neq 0\} \cup \{i : b_{PLS_{ki}} \neq 0\}$,
  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.

3.3.4 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 Gerven 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 Section 3.3.2, to follow the process until the construction and the recovery of the data matrix X_d (deflated by the y-orthogonal components, see (Eq. 3.3)) 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 Section 3.3.2) 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 $\boldsymbol{X_d = X - X_o}$.

3. Application of a SIMPLS-sPLS model (see Section 3.3.3) using X_d as input. This implies the inner identifications of $\boldsymbol{M_d = X_d' y y' X_d}$ and $\boldsymbol{R_d = X_d' y}$ instead of M and R in (Eq. 3.9) and (Eq. 3.10). 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 lenght of b^* is obviously smaller or equal than the lenght 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 Section 3.3.3).

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 3.1).

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.

3.4 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.

3.4.1 Results for the ^{1}H-NMR urine spectral data set

Starting from the ^{1}H-NMR urine of rat design, two main sub-data sets are considered, with balanced and unbalanced hippurate doses (see Section 3.2.1.1). 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 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 though 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 3.1). Thus, L-sOPLS provides (very) sparse models with a better predictive power than (O)PLS.

In Table 3.1, 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 3.1 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 3.2) from the 600 initial input variables. All the hippurate and citrate peaks are 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.3.1(c)). Furthermore, the selected citrate descriptors are well associated with negative sparse coefficients and localized at the center of the citrate peak.

Figure 3.2 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 3.3 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

Table 3.1: (O)PLS and sparse results for the ^{1}H-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					
		PLS coefficients $q = 5$	OPLS b_{OPLS} $q = 4 + 1$	sPLS selection $q_{pred} = 5$, $\lambda_1 = 0.6$	L-sOPLS selection $q_{ortho} = 4$, $q_{pred} = 3$, $\lambda_1 = 0.72$
Biomarkers (in ppm)	Zone	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 UNBALANCED CASE					
		PLS coefficients $q = 5$	OPLS b_{OPLS} $q = 4 + 1$	sPLS selection $q_{pred} = 5$, $\lambda_1 = 0.6$	L-sOPLS selection $q_{ortho} = 4$, $q_{pred} = 3$, $\lambda_1 = 0.73$
Biomarkers (in ppm)	Zone	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)**
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)**

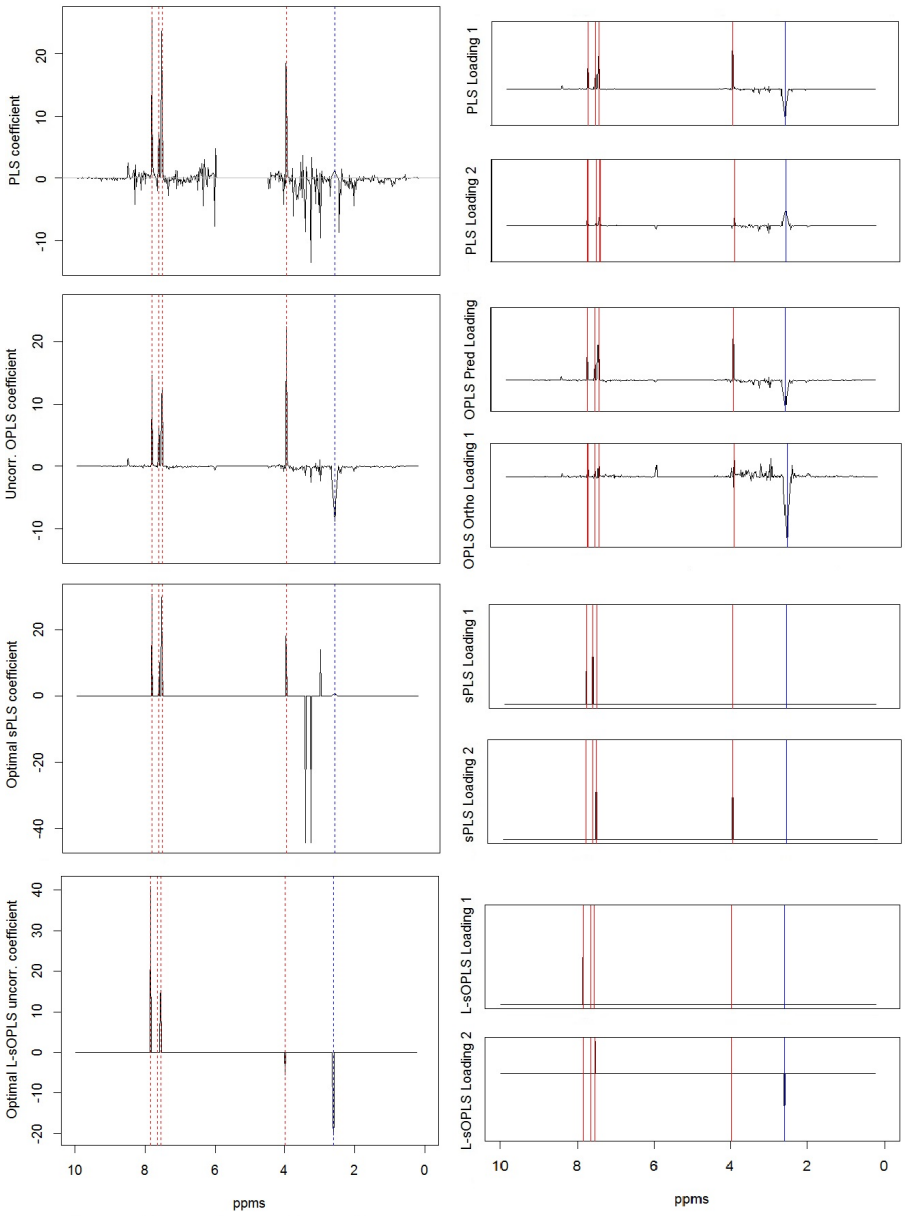


Figure 3.2: (O)PLS, sPLS and L-sOPLS coefficients (left) and two first loadings (right) for the unbalanced hippurate case.

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.3.3). Finally, the day effect does not seem to be very relevant for the group classification (the numbers are not always ordered in Fig.3.3).

For the unbalanced hippurate case's sPLS and L-sOPLS, Table 3.2 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 3.2) 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.

3.4.2 Results for genomic RT-qPCR data

In the RT-qPCR data set described in Section 3.2.2, $m = 19$ 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 binary vector to predict is the presence/absence of the endometriosis pathology.

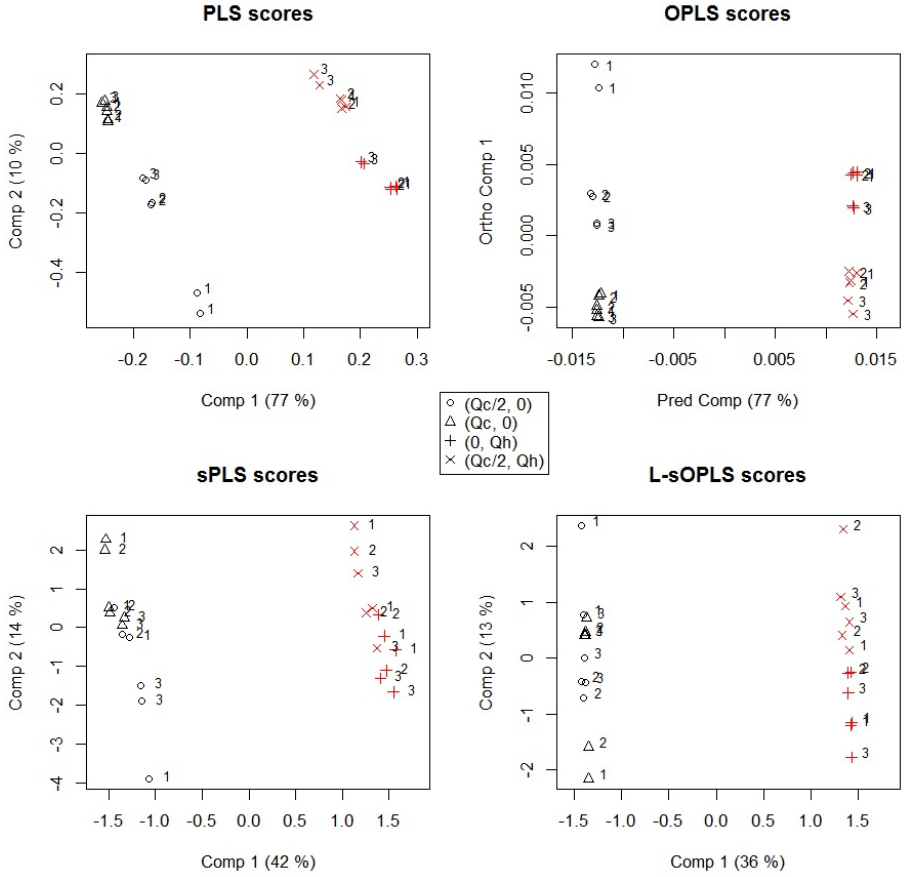


Figure 3.3: (O)PLS, sPLS and L-sOPLS scores for the unbalanced hippurate case. Colours discriminate Group 1 and Group 2, symbols refer to the citrate and hippurate concentrations and numbers refer to the day of measurement.

Table 3.2: 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 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.

THE HIPPURATE UNBALANCED CASE										
	q_{pred} for sPLS					q_{pred} for L-sOPLS (with $q_{ortho} = 4$)				
λ_1	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)**

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 3.3, 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.

Table 3.3: (O)PLS and sparse results for the RT-qPCR data. The six higher coefficients in absolute value are highlighted in bold for PLS and OPLS

	PLS coefficients $q = 2$	OPLS b_{OPLS} $q = 1 + 1$	sPLS selection $q_{pred} = 2$, $\lambda_1 = 0.68$	L-sOPLS selection $q_{ortho} = 1$, $q_{pred} = 1$, $\lambda_1 = 0.67$
miRNAs	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

Two clear primary biomarkers are identified by the four algorithms: log_miR_x2 and log_miR_x3 (Table 3.3). These miRNAs are among the higher coefficients found by PLS/OPLS and are also selected by the sparse solutions.

3.5 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 ^{1}H-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, consequently, lighter cross-validated optimal models to practitionners 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}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 [Feraud et al., 2017]) 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.

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.

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

Compliance with ethical requirements 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 B4122012 15082-1267) and all patients gave their informed consent.

Local bibliography

[Abdi, 2010] Abdi H., Partial least squares regression and projection on latent structure regression (pls regression), Wiley interdisciplinary reviews: Computational Statistics, 2(1), 97-106 (2010).

[Afanador et al., 2013] Afanador N.L., Tran T.N., Buydens L., 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 (2013).

[Barker and Rayens, 2003] Barker M., Rayens W., Partial Least Squares for discrimination, Journal of chemometrics, 17(3), 166-173 (2003).

[Bartel, 2009] Bartel D. P., MicroRNAs: target recognition and regulatory functions, cell, 136(2), 215-233 (2009).

[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).

[Chapman and Saad, 1997] Chapman A., Saad Y., Deflated and augmented Krylov subspace techniques, Numerical linear algebra with applications, 4(1), 43-66 (1997).

[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).

[De Jong, 1993] De Jong S., SIMPLS: an alternative approach to partial least squares regression, Chemometrics and intelligent laboratory systems, 18(3), 251-263 (1993).

[Efron et al., 2004] Efron B., Hastie T., Johnstone I., Tibshirani R., Least Angle Regression, Annals of Statistics, 32(2), 407–499 (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).

[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).

[Gabrielsson et al., 2006] Gabrielsson J., Jonsson H., Airiaub C., Schmidt B., OPLS methodology for analysis of pre-processing effects on spectroscopic data, Chemometrics and Intelligent Laboratory Systems, 84(1-2), 153-158 (2006).

[Geladi and Kowalski, 1986] Partial Least Squares regression: a tutorial, Analytica chimica acta, 185, 1-17 (1986).

[Giudice and Kao, 2004] Giudice L. C., Kao L. C., Endometriosis Lancet. 2004; 364: 1789–99. CrossRef, PubMed, Web of Science®Times Cited, 423 (2004).

[Hastie et al., 2015] Hastie T., Tibshirani R., Wainwright M., Statistical learning with sparsity: the lasso and generalizations. CRC Press (2015).

[Hoskuldsson, 1988] Hoskuldsson A., PLS regression methods, Journal of chemometrics, 2(3), 211-228 (1988).

[Indahl et al., 2009] Indahl U. G., Liland K. H., Næs T., Canonical partial least squares: a unified PLS approach to classification and regression problems, Journal of Chemometrics, 23(9), 495-504 (2009).

[Jung et al., 2010] Jung Y., Lee J., Kwon J., Lee K.S., Ryu D.H., Hwang G.S., Discrimination of the geographical origin of beef by ^{1}H-NMR-based metabolomics, Journal of agricultural and food chemistry, 58(19), 10458-10466 (2010).

[Lai, 2002] Lai E. C., Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation, Nature genetics, 30, 363 (2002).

[Lê Cao et al., 2008] Lê Cao K.A., Rossouw D., Robert-Granié C., Besse P., A sparse PLS for variable selection when integrating omics data. Statistical applications in genetics and molecular biology, 7(1) (2008).

[Lu et al., 2014] Lu B., Castillo I., Chiang L., Edgar T.F., Industrial PLS model variable selection using moving window variable importance in projection. Chemometrics and Intelligent Laboratory Systems, 135, 90-109 (2014).

[Mevik and Cederkvist, 2004] Mevik B. H., Cederkvist H. R., 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 (2004).

[Munoz-Romero et al., 2015] Munoz-Romero S., Arenas-García J., Gómez-Verdejo V., Sparse and kernel OPLS feature extraction based on eigenvalue problem solving, Pattern Recognition, 48(5), 1797-1811 (2015).

[Nisenblat et al., 2016] Nisenblat V., Bossuyt P. M., Shaikh R., Farquhar C., Jordan V., Scheffers C. S., ... and Hull M. L., Blood biomarkers for the non-invasive diagnosis of endometriosis, The Cochrane Library (2016).

[Rousseau, 2011] Rousseau R., Statistical contribution to the analysis of metabonomic data in ^{1}H-NMR spectroscopy (Doctoral dissertation, Université Catholique de Louvain, Belgium) (2011), permalink: `http://hdl.handle.net/2078.1/75532`.

[Stenlund et al., 2008] Stenlund H., Gorzsas A., Persson P., Sundberg B., Trygg J., 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 (2008).

[Tapp and Kemsley, 2009] Tapp H. S., Kemsley E. K., Notes on the practical utility of OPLS, TrAC Trends in Analytical Chemistry, 28(11), 1322-1327 (2009).

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

[Van Gerven and Heskes, 2010] van Gerven M. A. J., Heskes T., Sparse orthonormalized partial least squares. In Benelux Conference on Artificial Intelligence (2010).

[Wehrens, 2011] Wehrens R., Chemometrics with R: multivariate data analysis in the natural sciences and life sciences, Springer Science and Business Media, 155-165 (2011).

[Weljie et al., 2011] Weljie A.M., Bondareva A., Zang P., Jirik F.R., ^{1}H-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 (2011).

[Wiklund et al., 2008] Wiklund S., Johansson E., Sjostrom L., Mellerowicz E., Edlund U., Shockcor J.P., Gottfries J., Moritz T., Trygg J., Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models, Analytical Chemistry, 80(1), 115-122 (2008).

[Wold H, 1975] Wold H., Path models with latent variables: The NIPALS approach (pp. 307-357). Acad. Press., 1975.

[Wold, Trygg et al., 2002] Wold S., Trygg J., Berglund A., Antti H., Some recent developments in PLS modeling, Chemometrics and Intelligent Laboratory Systems, 58(2), 131-150 (2002).

[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).

[Zou and Hastie, 2005] Zou H., Hastie T., Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320 (2005).

CHAPTER 4

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

This Chapter 4 includes the full text of Paper III, which is entitled *"Two data pre-processing workflows to facilitate the discovery of biomarkers by 2D NMR metabolomics"* (Féraud, B., Leenders, J., Martineau, E. et al.) and was submitted to Metabolomics in 2018 (June).

It can be viewed as a larger and concrete application of the GPL workflow, of the MIC concept and of the L-sOPLS algorithm (for the first time applied on 2D data) in order to evaluate and compare multiple 1D and 2D NMR data sources. In this paper, the Vectorization workflow (Section 1.2.2.2) was also introduced to handle and pre-process 2D COSY spectra.

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).

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

4.1 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 popularity 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., 2017] 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 4.2 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., 2017]) and the biomarker discovery statistical models (PLS, sPLS, L-sOPLS) are detailed in the methodological Section 4.3. Results for

both data sets are then shown and discussed in Section 4.4. Finally, a general conclusion is given in Section 4.5.

4.2 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.

4.2.1 First experimental design: urine donors

4.2.1.1 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).

4.2.1.2 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 4.2.3.

4.2.2 Second serum based experimental design: spiked products

4.2.2.1 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.

4.2.2.2 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 4.2.3.

4.2.3 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 noesypresat pulse sequence with presaturation for urine samples and a CPMG relaxation-editing sequence with presaturation for serum samples. The 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.

4.3 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., 2017] and Section 4.3.1) and another which implies the vectorization of the initial spectral matrices (Section 4.3.2). 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., 2017] are presented to answer this question (Section 4.3.3).

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 4.3.4).

4.3.1 The Global Peak List (GPL) approach

This approach is fully detailed in [Feraud et al., 2017] and is structured as illustrated in Fig.4.1 (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`).

4.3.2 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

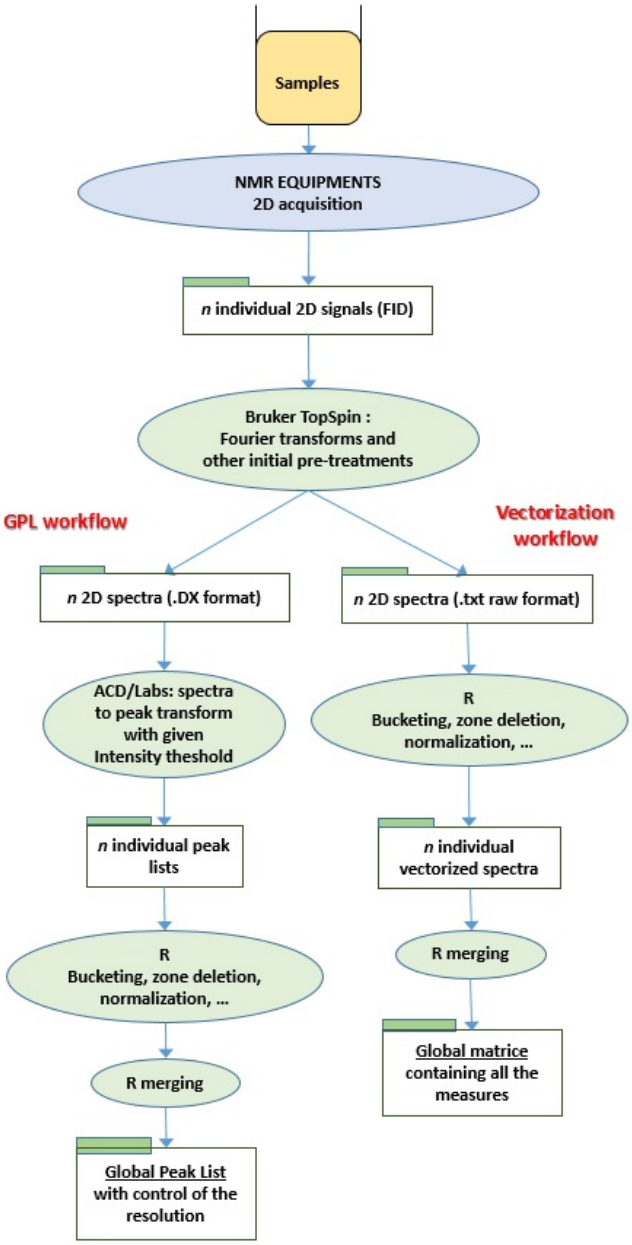


Figure 4.1: The GPL and vectorization approaches workflows for 2D experiments: steps and resulting data files.

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 PepsNMR's ([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.4.1 (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 correction, etc.). Recommended data file formats are also mentioned. For instance, the .dx format is more adapted to the ACD/Labs environment.

4.3.3 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 fully available, quality criteria to compare distinct preprocessing 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., 2017] 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., 2017]). 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.

4.3.4 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.

4.4 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 4.4.1 provides MIC indexes and Section 4.4.2 shows biomarker discovery results. In Section 4.4.2, 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]).

4.4.1 Metabolomic Informative Content results

For both experimental designs, the MIC indexes mentioned in Section 4.3.3 and described in [Feraud et al., 2017] 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 4.3.2) was also taken into account in the process, along with ^{1}H-NMR spectra. The main results are described in Table 4.1.

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.

Table 4.1: 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)

FIRST DESIGN - 3 urine donors									
Data	ACD/Lab	GPL	MIC indexes (Ward / K-means)				Inertia		Columns
	threshold	decimals	Dunn	Davies-Bouldin	Rand	Adj-Rand	Btw (%)	Wth (%)	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	GPL	MIC indexes (Ward / K-means)				Inertia		Columns
	threshold	decimals	Dunn	Davies-Bouldin	Rand	Adj-Rand	Btw (%)	Wth (%)	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

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.4.2, 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 4.1 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.4.2, 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 4.4.2 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 4.2.2.2 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).

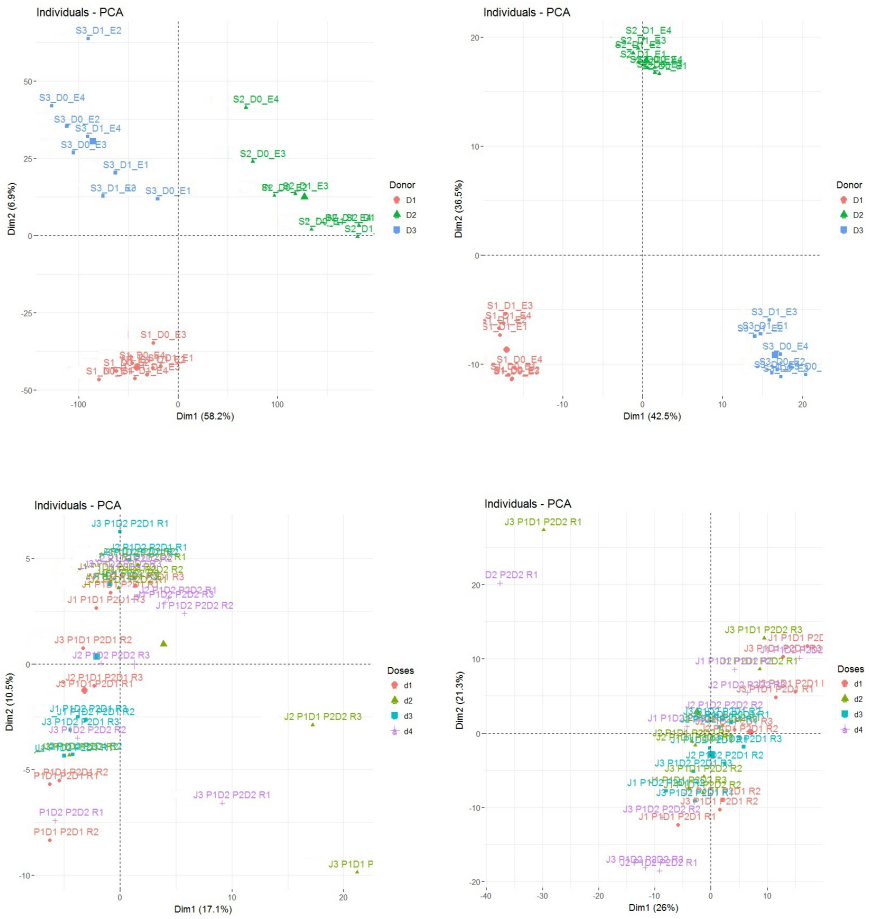


Figure 4.2: First two PCA factors for the best GPL in terms of MIC results (left) and for the corresponding pre-processed ^{1}H-NMR data (right). The upper individuals PCA graphs are related to the first urine design; the lower ones to the second threonine/glutamate design. Note again that each individual 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.

4.4.2 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.4.3.

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 4.1).

The PLS, sPLS and L-sOPLS results are shown in Table 4.2 (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).

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 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

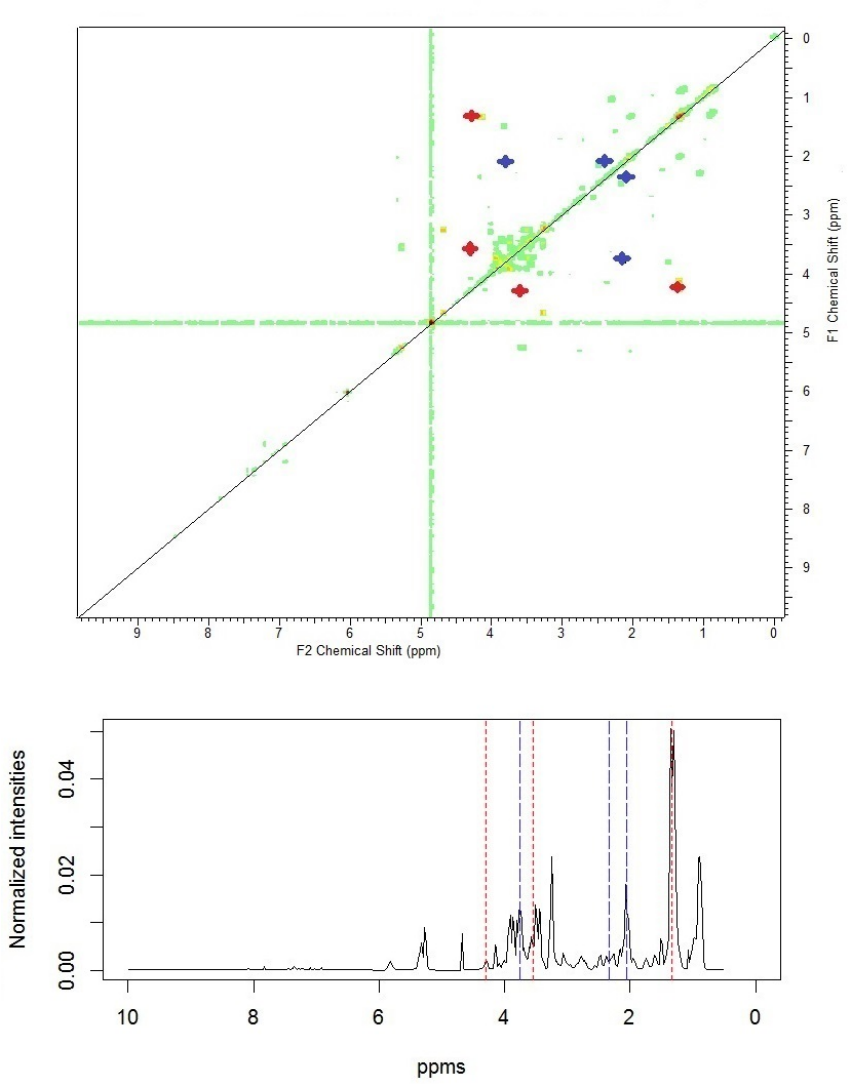


Figure 4.3: Threonine (in red) and glutamate (in blue) 2D correlation peaks in a raw COSY spectrum and corresponding peaks in a 1D NMR spectrum (red dashes for threonine and blue long dashes for glutamate).

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

Table 4.2: PLS, sPLS and L-sOPLS biomarker selection for ^{1}H-NMR data (threonine-glutamate design).

		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

= 0.05 and number of decimal = 1) is summarized in Table 4.3. 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.

One can see in Table 4.3 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).

Table 4.3: PLS, sPLS and L-sOPLS non-diagonal biomarker selection applied on the best GPL solution in terms of MIC measures (threonine-glutamate design).

		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

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 4.1), 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.

4.5 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 pre-processed 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, based on 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 4.3), even if the initial MIC and PCA results were not very optimistic (Table 4.1 and Fig.4.2).

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.

Local bibliography

[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), 407–499 (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 preprocessing to dimension reduction and statistical analysis. TrAC Trends in Analytical Chemistry, 30(6), 827–841 (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).

CHAPTER 5

Contributions of faster 2D COSY acquisition experiments to metabolomics: the Non-Uniform Sampling (NUS) and the single scan COSY (NS1)

This chapter concerns the potential use of faster experiments inspired by the initial 2D NMR COSY acquisition technique (see Section 1.1.3) and to determine their contributions in the context of metabolomics studies.

This chapter has to be considered as a draft and a basis for a further fourth submitted paper. It is resulting from an inter-laboratories collaboration involving Patrick Giraudeau and Estelle Martineau from the Université de Nantes (France), and Pascal de Tullio and Justine Leenders from ULg.

5.1 Motivations

It has been proved in Paper I and III that the use of 2D COSY spectral data can allow to reach a better level of Metabolomic Informative Content (MIC, see Section 1.3.2) compared to the traditional use of 1D ^{1}H-NMR spectra, thus taking advantage of the additional information contained in the 2D cross non-diagonal peaks. These gains in terms of MIC are then confirmed by subsequent better performances when searching for relevant final biomarkers to interpret (see again Paper I and III conclusions).

In terms of pre-processing, the main initial intuition of this thesis is by the way verified: the use the 2D COSY data sources allows to reach general metabolomics results which are at least as good as the results obtained with 1D data, and this with a less advanced and less complex level of pre-processing (Sections 1.2.1 to 1.2.2.3).

Among all these assets, a major problem still exists and can largely slow down a generalized use of 2D spectra (COSY or others) in metabolomics: the acquisition times. Actually, the acquisition of a single COSY spectrum requires from 20 minutes to an hour ([Delikatny et al., 1991]) depending on the spectrometer and on the acquisition parameters. These times and costs result in potentially several days of measurements for completing a whole experimental design.

Fortunately, some advances have been made recently in order to avoid this drawback. In this work, thanks to our collaboration with Patrick Giraudeau and Estelle Martineau from the Université de Nantes, two techniques are experimented: the Non-Uniform Sampling (NUS, [Barna and Laue, 1987] [Hoch et al., 2014]) performed on COSY data, and the use of a single scan COSY (noted NS1, [Frydman et al., 2002]).

Practically, NUS and NS1 data sources were obtained and tested in the context of the Paper III inter-laboratory study involving three urine donors (see the description of the data set in Section 1.5 and full details in Paper III). The experiment 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. 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 this regard, urine dilution and days factors can be considered as additional sources of noise and will not be commented separately in details.

The data were then pre-processed through the GPL (Section 1.2.2.1) and Vectorization (Section 1.2.2.2) 2D workflows, and MIC measures (Section 1.3.2) were calculated on them.

If the faster NUS and NS1 acquisition techniques allow to reach similar results as COSY in Paper III, the acquisition times issue will be solved concerning the use of 2D NMR data sources, thus opening the door to their potential more massive use in metabolomics studies. Thereby, in case of conclusive results and through shorter experimental durations, analyses of larger cohorts of samples may be considered and facilitated.

5.2 Introduction to the Non-Uniform Sampling (NUS) method

The duration of a two-dimensional NMR experiment is both directly proportional to the number of time increments in the indirect dimension t_1 (see Section 1.1.3), allowing this dimension to be correctly sampled, and the number of scans NS which is required to obtain a sufficient signal-to-noise ratio to achieve a needed accuracy. Therefore, these two parameters must be high enough to get the desired resolution and accuracy, leading to a significant increase of the total duration of 2D NMR experiments. Moreover, the experiments become more sensitive to spectrometer instabilities, which are responsible for the appearance of noise in the 2D spectra.

Thus, experiments involving a reduced number of time increments t_1 and/or a reduced number of scans NS are the only options to decrease the experimental duration. In order not to lose too much resolution and sensitivity, a first solution has been proposed in [Barna and Laue, 1987] to overcome this problem by introducing the concept of Non-Uniform Sampling (NUS). The principle of NUS consists in reducing the duration of 2D NMR experiments by only acquiring a fraction of the time increments t_1 in the indirect dimension.

In pratice, the Bruker Topspin[1] software randomly creates a sampling scheme (or a so-called "NUS List") which contains the different values of t_1 actually in use during the FID acquisition according to the desired NUS percentage (or "NUS Amount").

Fig.5.1(a) presents an example of a sampling scheme containing 24 time increments t_1 without using NUS, whereas Fig.5.1(b) shows an example of the same sampling scheme acquired with a random 50% NUS. A more complex and global process is shown in Fig.5.2.

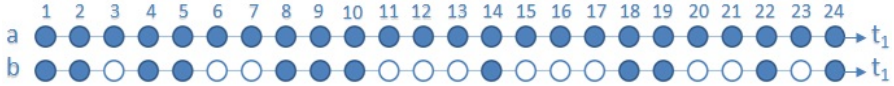


Figure 5.1: Examples of sampling schemes used for the acquisition of conventional 2D NMR experiments with t_1=24 without NUS (a) and with 50% NUS (b). The blue dots correspond to the t_1 values for which a FID will be acquired during the experiment, while the white dots represent the t_1 values for which no FID will be recorded (figure provided by Estelle Martineau).

[1] https://www.bruker.com/products/mr/nmr/nmr-software/nmr-software/topspin/overview.html

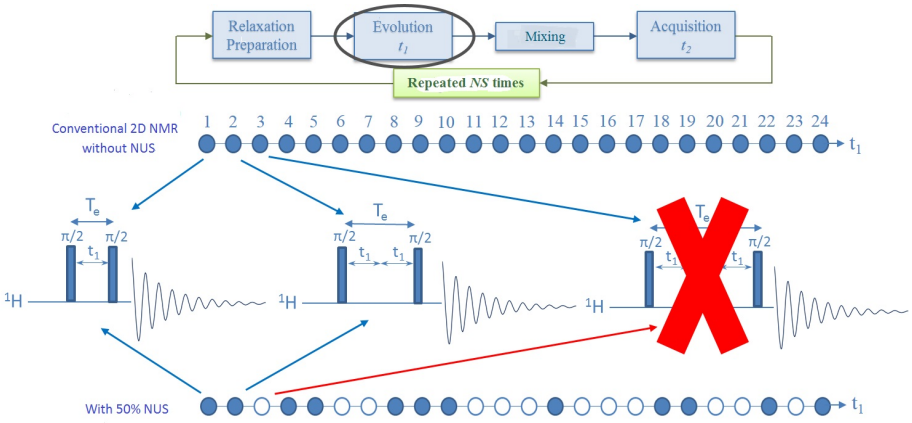


Figure 5.2: A more complex view of examples of sampling schemes used for the acquisition of conventional 2D NMR experiments with t_1=24 without NUS and with 50% NUS (figure provided by Estelle Martineau).

After data acquisition, the two-dimensional FID thus obtained can not be directly processed as for a conventional 2D NMR experiment. Indeed, intervals between two successive values of t_1 are not constant anymore due to the random non-acquisition of some FID (see Fig.5.1(b)).

As a result, the value "zero" is internally added to each of the missing FIDs. Then, a final 2D spectrum can be digitally reconstructed through the use of a non-linear data processing method. This spectral reconstruction is largely performed via the "Compressed Sensing" (CS, [Kazimierczuk and Orekhov, 2011]) algorithm which has been developed specifically for 2D NMR experiments.

In this context, the use of NUS in 2D NMR experiments can be motivated in two situations:

- Either the aim is to reduce the duration of experiments because of time constraints (for instance, too long spectrometers occupation times or analyses involving a very large cohort of samples). In this case, for a constant number of scans (NS, compared to conventional experiments), the percentage of NUS must be chosen in order to decrease the experimental duration without unduly deteriorating the spectral resolution;

- Or the aim is to enhance sensitivity and, therefore, to keep the duration of a conventional experiment. In this case, for a constant number of time increments (t_1, compared to conventional experiments), the percentage

of NUS must be selected in order to limit the loss of spectral resolution while increasing the number of scans (NS) to enhance sensitivity.

The purpose of this project, following the context of Paper III in metabolomics, aims at demonstrating the applicability of NUS for the analysis of some human urine samples according to the first goal mentioned above (i.e. reducing the experimental duration compared to conventional one without unduly deteriorating the spectral resolution).

5.3 Data acquisition parameters for the COSY NS1 and COSY NUS50 data sources

For full details about conventional COSY acquisitions and a description of the protocol for urine collection, see Paper III (sections 4.3.1 and 4.3.3).

For the particular acquisitions of interest here, all the samples were analyzed in Nantes on a Bruker Avance-III HD spectrometer operating at a ^{1}H frequency of 7000.13 MHz, equipped with an inverse ^{1}H/^{13}C/^{15}N/^{2}H cryogenically cooled probe and a Z gradient coil. Analyses were performed at 298 K.
COSY experiments were performed on each sample. The pulse sequence includes a water suppression scheme applied during the recovery delay $D1$ and the use of a gradient-based coherence selection.

First, samples were analyzed under the whole conventional experimental conditions developed in Paper III (called here "COSY NS4", because it involves four scans). Then, in order to significantly reduce the duration of the experiments, the acquisitions were performed with only a single scan ($NS = 1$, called here "COSY NS1") instead of four while the other parameters were kept identically between COSY NS1 and COSY NS4. Here, the sensitivity decreased but the signal-to-noise ratio remained sufficient to allow subsequent studies.

Finally, Non-Uniform Sampling was applied. In particular, some acquisition parameters were modified or added compared to COSY NS1:

- The percentage of NUS was set to 50% (the resulting data set being called "COSY NUS50"). This value was previously tested on experimental data in order to obtain a sufficient signal-to-noise ratio with only one scan ($NS = 1$) without losing too much resolution. Considering a fixed number of points t_1 (for example equal to 256), experiments without NUS, with 50% of NUS and with 25% of NUS were thereby tested. Spectra and

preliminary results based on signal-to-noise ratios are available in Fig.5.3. The noisy trails in the first dimension are slightly more important in the 25% NUS spectrum (right of Fig.5.3), thus implying lower signal-to-noise ratios for some signals compared to the other experiments. With 50% of NUS, the S/N ratio decreases for signals 3 and 4 with respect to the experiment without NUS and evolves little for the two other signals. This loss in sensibility remains acceptable considering that the experimental time was divided by two.

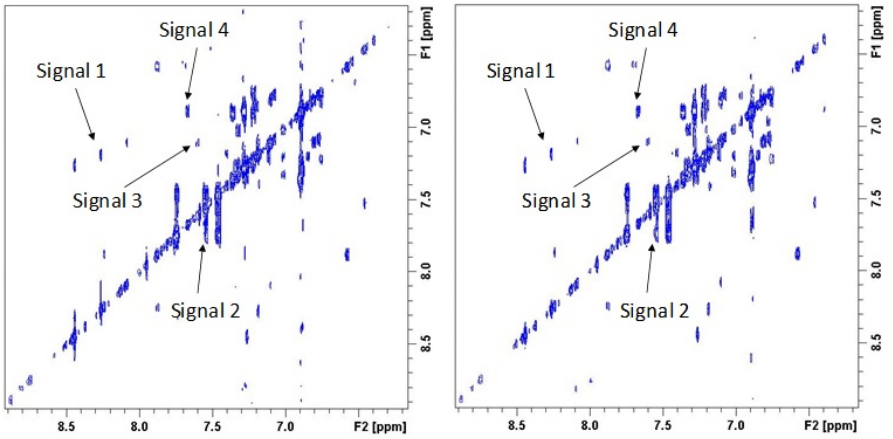
- The number of time increments t_1 in the indirect dimension went from 300 to 256. Note that only 128 increments are finally retained with the 50% NUS. These points were selected by generating a Poisson-gap scheme ([Hyberts et al., 2010], [Maciejewski et al., 2011]) which allows to limit the creation of artefacts and to avoid too large intervals between successive t_1 values.

- The recovery delay was set to 3 s. This value was previously evaluated in order to have a good compromise between sensitivity and experimental duration.

Let us remember, from Paper III, that the experimental design involves three different urine donors, two urine dilution levels and four days of acquisition. Eight measures are finally available for each donor. A total of 24 COSY NS4, 24 COSY NS1 and 24 COSY NUS50 individual spectra, all acquired in Nantes, were then collected and taken into consideration for metabolomics purpose. These spectra were pre-processed according to the data pipeline fully described in Paper III and in Sections 1.2, 1.3 and 1.4 of this thesis.

5.4 Methods: application of the 2D workflows (GPL and Vectorization) and of the MIC on the faster COSY data sources

In order to build global objects which can contain all the information from the 24 individual spectra, the Global Peak List (GPL) and the Vectorization 2D workflows are performed on the pools of COSY NS4, COSY NS1 and COSY NUS50 spectra. The GPL workflow is fully detailed in Section 1.2.2.1 and in Paper III. The Vectorization one is fully described in Section 1.2.2.2 and also in Paper III.



	without NUS (texp = 16 min)		50% NUS (texp = 8 min 30 s)		25% NUS (texp = 4 min 43 s)	
	S/N (col)	S/N (row)	S/N (col)	S/N (row)	S/N (col)	S/N (row)
Signal 1	59	65	45	77	57	70
Signal 2	35	62	31	47	20	34
Signal 3	30	25	21	17	6	5
Signal 4	125	90	72	79	57	61

Figure 5.3: Preliminary analyses of NUS amounts: experimental COSY spectra acquired with 50% NUS (left) and 25% NUS (right) and corresponding measured signal-to-noise ratios (from Estelle Martineau).

Once these global objects are built, their quality in terms of informative content (with respect to their ability to allow to retrieve the main signal, which correspond to the urine donors and to the y response vector) has to be evaluated via the Metabolomic Informative Content (MIC) indexes. The MIC concept is fully detailed in Section 1.3.2 and in Paper I. According to these MIC quality indexes, all the candidate data sources can then be ranked.

Finally, the best data source(s) can be advantageously selected to perform biomarker discovery, as described in Section 1.4 and in Papers II and III.

5.5 Results and discussion

Again, the first objective is to blindly retrieve the main signal of the experimental design which corresponds to the group membership (i.e. the urine donors) of the samples. PCA and MIC results are successively shown in this section.

5.5.1 Descriptive PCA results

For descriptive statistics, Fig.5.4 shows PCA results for four cases as input matrices: log-transformed GPL (with one decimal for the bucketing step and a ACD/Labs significance threshold of 0.02, see Section 1.2.2.1 for details) and Vectorization for COSY NS1 data, and GPL (using the same parameters) and Vectorization for COSY NUS50 data. One can see that these PCA results generally indicate a good a priori separability between the urine donors.

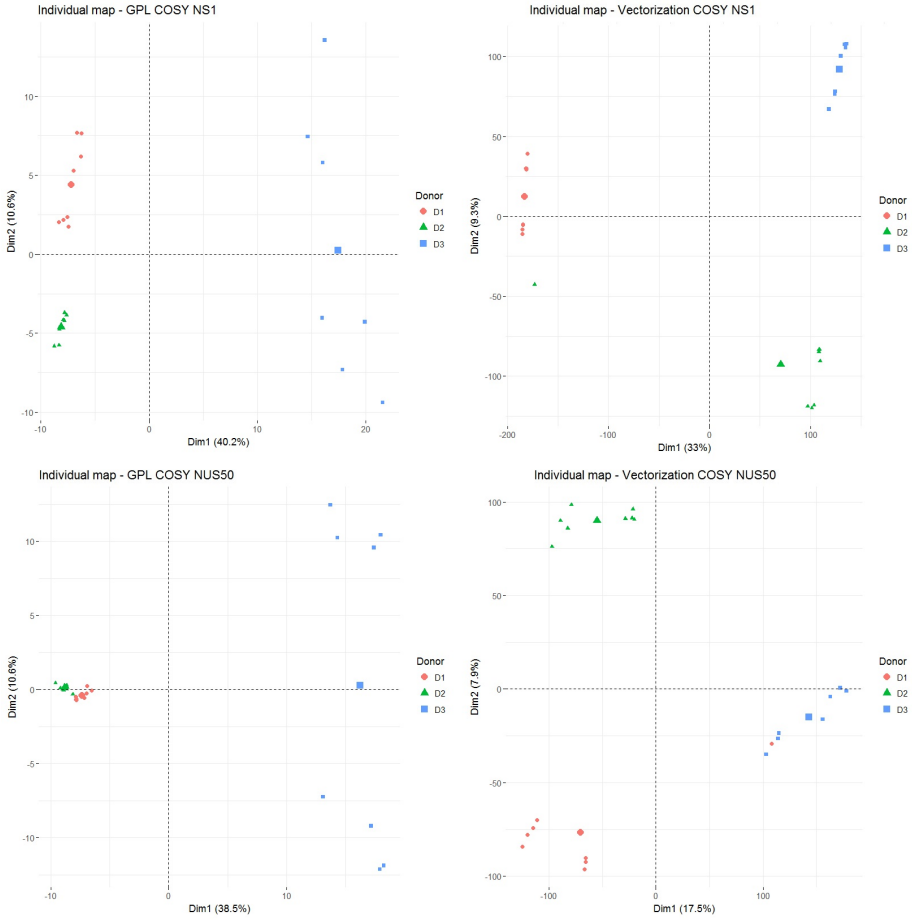


Figure 5.4: First two PCA factors for GPL and Vectorization of COSY NS1 data (top) and COSY NUS50 data (bottom).

It is also interesting to visualize the effects of the ACD/Labs significance threshold and of the number of retained decimals during the bucketing step

on PCA results based on the GPL matrices (Fig.5.5). When the number of decimal(s) increases and/or the significance threshold increases, the first and second urine donors tend to become more and more confused.

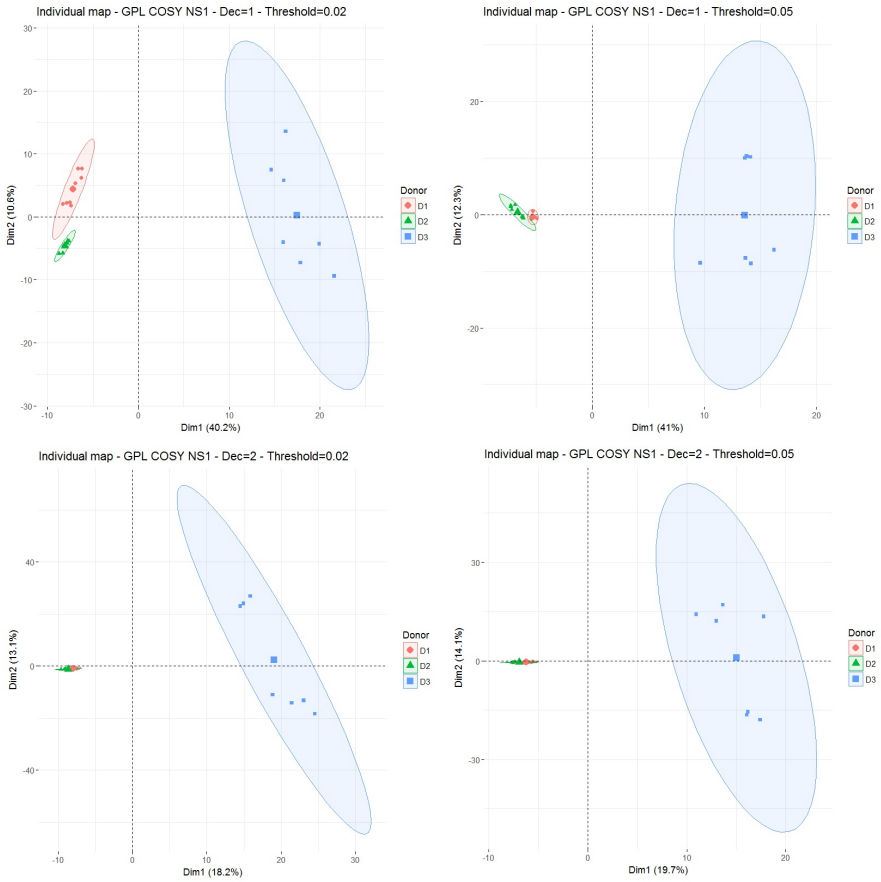


Figure 5.5: The GPL parameters (the number of decimals for bucketing and the ACD/Labs significance threshold) effects on PCA results (illustrated on the COSY NS1 spectral data).

Consequently, GPL using large buckets (with one decimal) and a conservative significance threshold should be favored to keep a sufficient amount of information according to the signal. This will be confirmed via the use of the MIC indexes in Section 5.5.2.

5.5.2 MIC results

In this section, the MIC results obtained with conventional COSY spectra presented in Paper III are supplemented by new MIC results on the COSY NS1 and COSY NUS50 spectral data sources detailed in Sections 5.2 and 5.3. All these results are shown in Table 5.1. Note that all GPL matrices are log-transformed before MIC calculations.

Table 5.1: MIC results for conventional COSY NS4, COSY NS1 and COSY NUS50 spectra, according to the choice of different pre-processing workflows (GPL and Vectorization)

Experiment	Data	ACD/Lab threshold	GPL decimals	MIC indexes (Ward / K-means) Dunn	Davies-Bouldin	Rand	Adj-Rand	Inertia Btw (%)	Wth (%)	Columns in X
COSY NS4	GPL	0.02	1	0.8088 / 0.8088	1.5696 / 1.5696	**1.00 / 1.00**	**1.00 / 1.00**	41.01	58.99	828
COSY NS4	GPL	0.02	2	0.7917 / 0.7917	1.6943 / 1.6943	**1.00 / 1.00**	**1.00 / 1.00**	28.57	71.43	1954
COSY NS4	GPL	0.02	3	0.8131 / 0.8366	1.8565 / 1.7369	0.5942 / 0.5797	0.1744 / 0.2088	12.12	87.88	3994
COSY NS4	GPL	0.05	1	0.8792 / 0.8612	1.3635 / 1.3239	**1.00 / 1.00**	**1.00 / 1.00**	**42.62**	**57.38**	484
COSY NS4	GPL	0.05	2	0.8699 / 0.8699	1.6074 / 1.6074	**1.00 / 1.00**	**1.00 / 1.00**	31.76	68.24	1126
COSY NS4	GPL	0.05	3	0.8013 / 0.7642	1.8484 / 1.858	0.5942 / 0.6413	0.1743 / 0.2966	13.84	86.16	2332
COSY NS4	COSY vectorization			0.7555 / 0.758	0.6224 / 0.7166	0.9407 / **1.00**	0.8605 / **1.00**	**91.75**	**8.25**	65536
COSY NS1	GPL	0.02	1	0.8592 / 0.6763	1.3456 / 1.1932	**1.00 / 1.00**	**1.00 / 1.00**	**55.36**	**44.64**	337
COSY NS1	GPL	0.02	2	0.7543 / 0.7543	1.5941 / 1.5941	**1.00 / 1.00**	**1.00 / 1.00**	35.71	64.29	895
COSY NS1	GPL	0.02	3	0.7332 / 0.7359	1.8511 / 1.8802	0.5889 / 0.5889	0.153 / 0.153	15.68	84.32	1859
COSY NS1	GPL	0.05	1	0.623 / 0.623	1.3715 / 1.3715	**1.00 / 1.00**	**1.00 / 1.00**	**56.24**	**43.76**	201
COSY NS1	GPL	0.05	2	**0.9902 / 0.9902**	1.4361 / 1.4361	0.6996 / 0.6996	0.4038 / 0.4038	37.77	62.23	509
COSY NS1	GPL	0.05	3	0.8248 / 0.6065	1.639 / 1.8308	0.6245 / 0.8142	0.2834 / 0.5869	16.81	83.19	1065
COSY NS1	COSY vectorization			0.5733 / 0.6442	0.6226 / 0.7163	**1.00 / 1.00**	**1.00 / 1.00**	**91.38**	**8.62**	65536
COSY NUS50	GPL	0.02	1	0.8024 / 0.8024	1.3264 / 1.3264	**1.00 / 1.00**	**1.00 / 1.00**	**56.25**	**43.75**	348
COSY NUS50	GPL	0.02	2	0.7531 / 0.7531	1.5941 / 1.5941	**1.00 / 1.00**	**1.00 / 1.00**	36.15	63.85	996
COSY NUS50	GPL	0.02	3	0.7221 / 0.7221	1.9022 / 1.9022	0.8007 / 0.8007	0.5657 / 0.5657	13.25	86.75	2102
COSY NUS50	GPL	0.05	1	0.6019 / 0.5689	1.172 / 1.4532	**1.00 / 1.00**	**1.00 / 1.00**	50.07	49.93	203
COSY NUS50	GPL	0.05	2	0.6205 / 0.8603	1.5559 / 1.5977	0.8261 / 0.7102	0.6102 / 0.4103	31.89	68.11	574
COSY NUS50	GPL	0.05	3	0.6564 / 0.6176	1.5095 / 1.9596	0.4964 / 0.5942	0.093 / 0.0987	14.00	86.00	1167
COSY NUS50	COSY vectorization			0.6783 / 0.6783	0.8518 / 0.8518	**0.9447 / 0.9447**	0.8693 / 0.8693	**82.14**	**17.85**	65536

One can see in Table 5.1 that the entire results are very close between the two faster COSY experiments and the conventional one. Indeed, GPL matrices with larger buckets (with 1 or 2 decimal(s) for the coordinates during the soft bucketing step) allow to reach perfect clustering results in almost all the cases, with Rand and adjusted-Rand indexes equal to 1. Matrices obtained via the 2D Vectorization workflow provide quite similar clustering performances and are also characterized by particularly good inertia measures (in bold in Table 5.1).

Thus, COSY NS1 and COSY NUS50 pools of spectra are very comparable with conventional COSY spectra in terms of Metabolomic Informative Content (MIC). No significant loss of information about the signal is observed when using faster acquisition methods for 2D COSY spectra.

Note that COSY NS1 and COSY NUS50 also allow to consider GPL matrices of smaller sizes (last columns of Table 5.1).

5.6 Conclusions

It has been proved in Paper I and III and during this thesis that 2D COSY spectra allow to capture a better level of Metabolomic Informative Content compared to the traditional use of 1D ^{1}H-NMR spectra. A more massive use of COSY in metabolomics (or -omics in general) studies is nevertheless slowed down by important and often non-acceptable acquisition times to handle a whole experimental multi-factors design.

However, these COSY acquisition times can be significantly reduced by imposing a smaller number of scans (NS) during the COSY experiment and/or by considering Non-Uniform Sampling (NUS) techniques which allow to consider a smaller number of time increments t_1.

In this draft, COSY NS1 spectra, using a single scan ($NS = 1$), and COSY NUS50 spectra, involving a desired NUS percentage of 50% with a single scan, are tested and compared with conventional COSY NS4 spectra in order to retrieve different urine donors (the signal) in a multi-factor design (described in Paper III).

It is demonstrated in Table 5.1 that the two faster COSY acquisition techniques provide very similar MIC results than the initial COSY. Thus, the proved advantage of the COSY data source over the 1D ^{1}H-NMR one is still verified. Consequently, using techniques like the single scan or the Non-Uniform Sampling can really open the door to a potential more massive use of 2D COSY

spectra in metabolomics studies and can facilitate the analyses of larger cohorts of samples.

To go further, other values can be tested for the NUS percentage amount in order to be more accurate. NUS would also be tested along with other values for the number of scans ($NS > 1$). Other values for the significance threshold in the GPL workflow, maybe more adapted to the faster acquisition methods, could also be tested in order to avoid the loss of potentially important cross peaks.

Local bibliography

[Barna and Laue, 1987] Barna, J. C., and Laue, E. D., Conventional and exponential sampling for 2D NMR experiments with application to a 2D NMR spectrum of a protein, Journal of Magnetic Resonance (1969), 75(2), 384-389 (1987).

[Delikatny et al., 1991] Delikatny, E. J., Hull, W. E., and Mountford, C. E., The effect of altering time domains and window functions in two-dimensional proton COSY spectra of biological specimens. Journal of Magnetic Resonance (1969), 94(3), 563-573 (1991).

[Frydman et al., 2002] Frydman, L., Scherf, T., and Lupulescu, A., The acquisition of multidimensional NMR spectra within a single scan. Proceedings of the National Academy of Sciences, 99(25), 15858-15862 (2002).

[Hoch et al., 2014] Hoch, J. C., Maciejewski, M. W., Mobli, M., Schuyler, A. D., and Stern, A. S.,Nonuniform sampling and maximum entropy reconstruction in multidimensional NMR. Accounts of chemical research, 47(2), 708-717 (2014).

[Hyberts et al., 2010] Hyberts, S. G., Takeuchi, K., and Wagner, G., Poisson-gap sampling and forward maximum entropy reconstruction for enhancing the resolution and sensitivity of protein NMR data. Journal of the American Chemical Society, 132(7), 2145-2147 (2010).

[Kazimierczuk and Orekhov, 2011] Kazimierczuk, K., and Orekhov, V. Y., Accelerated NMR spectroscopy by using compressed sensing. Angewandte Chemie, 123(24), 5670-5673 (2011).

[Maciejewski et al., 2011] Maciejewski, M. W., Mobli, M., Schuyler, A. D., Stern, A. S., and Hoch, J. C., Data sampling in multidimensional NMR:

fundamentals and strategies. In Novel Sampling Approaches in Higher Dimensional NMR (pp. 49-77). Springer, Berlin, Heidelberg (2011).

CONCLUSIONS

The starting point and the motivation of this thesis were based on a fact and on a very crucial intuition.

The fact is that a large majority of NMR-based metabolomics studies actually depend on 1D ^{1}H-NMR spectral data. These data are well-known and their pre-processing, which can rapidly and heavily become complex (see Section 1.2.1), is more and more sharp and well-mastered with best practices at different levels.

As well as ^{1}H-NMR data, 2D NMR experiments exist for many years. Theoretically, these experiments (and in particular COSY which is considered in this thesis, see Section 1.1.3) can supply more information as they get around the 1D overlapping and multiplicity of peaks problems via the occurence of 2D cross peaks (Section 1.1.2). Indeed, COSY spectra can avoid the superimposing of 1D peaks for a same molecule or for different correlated molecules.

Currently, the COSY data source is not widely used in metabolomics: there is no full automatic handling, no best practices for pre-processing, no adapted statistical modelling or tools and prohibitive acquisition times remain an important limitation.

In this context, our starting intuitive point assume that 2D spectra need a less advanced pre-processing level to express at least as much information as corresponding 1D ones. In other words, this places trust in the fundamental relevance and quality of the additional informative content inherent in the 2D spectral data sources.

Practically, and as a result, some particular pre-processing steps (such as baseline correction or warping), so crucial and complex when applied on ^{1}H-

NMR spectra, could be more easily and quickly by-passed when using 2D data sources.

In this thesis, three main contributions are proposed in order to provide a full process for using 2D COSY spectra in metabolomics studies. Note that every contribution can be generalized and extended to any 2D data source (see Section 1.1.3) having a similar pattern.

First, two pre-processing worflows are presented to handle individual 2D spectra and to create a global object which contains all the relevant information of a whole experimental design: the **GPL (Global Peak List, Section 1.2.2.1) and the Vectorization (Section 1.2.2.2) workflows**. Indeed, a considered experimental design can involve one or several factor(s) of interest, different days of measurement, time replicates, among others. According to these factors, a potentially huge number of individual samples can be treated, thus resulting in a potentially huge number of individual spectra. For statistical analyses and multivariate modelling, all these spectra have to be merged into such a global object (i.e. a global data matrix).
The GPL workflow provides a global matrix for which the dimensionality is partially controlled and allows dimension reduction by considering detectable peaks only (according to a devoted threshold). For its part, the Vectorization workflow provides a global matrix with a fully controlled dimensionality and takes into account all the initial information (because, like in 1D, there is no peak integration before bucketing).

Once GPL or Vectorization matrices are built, potentially using different data sources (1D NMR, 2D NMR, ...) and/or different tuning pre-processing parameters, the quality of all the available data sets has to be assessed and ranked.
For that purpose, the process proposed in this thesis is based on data sets organized into groups of subjects, on which signal/noise studies can be performed. Indeed, a proper group factor (linked with the presence of different urine donors, with different blood donors or with different doses of spiked products according to the encountered case studies (see Section 1.5)) has to be taken into particular consideration in the experimental design in order to discriminate groups of samples. Practically, this group factor is always considered as the main signal of the design, all the other factors being considered as noise factors.

Considering this particular signal, the novel **Metabolomic Informative Content** (MIC, see Section 1.3.2) concept is proposed. It gathers a pool of objective indexes to evaluate in what extent a specific data source allows to

retrieve and to discriminate the signal, in spite of the existence of numerous potential sources of noise (instrumental issues, experimental conditions, replicates, freezing/defrosting of samples, bacterial degradation of samples, etc...). Naturally, the presence of groups in the data leads to the use of unsupervised multivariate clustering algorithms and the MIC indexes finally evaluate both the homogeneity of the obtained clusters (inertia, Dunn and Davies-Bouldin indexes) and their accuracy compared with the true groups (Rand and adjusted Rand indexes).

Calculating MIC indexes for different data sources (1D, 2D, from GPL or Vectorization, ...) finally allows to rank these latter: some data sources contain a better Metabolomic Informative Content than others. Indeed, MIC indexes (or ASCA/APCA/ICA related methodologies, see Sections 1.3.3 and 1.3.4) ideally lead to the emphasis of a "best" data source, which allows to capture the signal (i.e. to discrimine the groups) in the best possible way.

This data source can then be chosen to advantageously perform further statistical studies to discover relevant biomarkers and to make further predictions.

For the biomarker discovery issue itself, PLS-based techniques have been explored in Sections 1.4.2 to 1.4.4, implying the notion of orthogonality (OPLS) or of sparsity (sPLS). The general use of penalty terms to obtain sparse solutions is developed in Section 1.4.4.1. If one wants to keep the OPLS interpretability advantages which consist of a withdrawal of the y-orthogonal effects relative to the data matrix X, and if one also wants to propose sparse results and lighter final models, a natural enhancement is to explore the possible combinations of OPLS and sPLS.

To this end, the third innovative proposal of this thesis consists in the development of the so-called **"Light-Sparse-OPLS" (L-sOPLS) model**. L-sOPLS consists in following the OPLS process until the construction and the recovery of a deflated X_d data matrix (matrix which is deflated by the y-orthogonal variations) and in applying in a second stage the sPLS algorithm using X_d as input. The best L-sOPLS model is obtained by optimizing, via successive cross-validated minimizations of the RMSEP criterion, the number of orthogonal component(s) q_{ortho} in a first step and the couple (q_{pred}, λ_1) formed by the number of predicted component(s) and the penalty term in a second step.

In this context, and using all these novel contributions, it has been shown in this thesis that:

- the GPL and Vectorization workflows are fully operational and allow to efficiently combine several individual 2D spectra (one per sample) into accurate global matrices. COSY spectra are considered in this thesis but

these two workflows can be applied on any 2D data sources having a similar pattern;

- by integrating only detectable peaks and proposing an inner soft bucketing step, the GPL workflow allows to consider final matrices of reduced dimensions. It is demonstrated on several data sets that MIC indexes provide better informative content and separability when using log-transformed 2D GPL scenarii involving small matrices (with one or two decimals for the coordinates);

- the novel MIC solution offers tools to compare the quality of spectral data matrices according to a particular group factor (i.e. the signal) to retrieve. For this purpose, no systematic solution existed before for 2D data diagnostic in a metabolomics context.

- 2D COSY spectra appear to be statistically robust and contain informative signal which helps to succeed in distinguishing groups of different spectra. So, COSY spectra are enough informative so that the signal connected to the initial groups can be well distinguished from the noise;

- working on binary positions (absence/presence of 2D peaks in a spectra) gives satisfactory results but not better ones than working on normalized intensities. However, according to the context and to the research objectives, it can be a very acceptable alternative;

- compared to the widely used 1D ^{1}H-NMR spectra, 2D COSY spectra, via their second dimension and correlation cross peaks, does provide additional relevant information to improve the clustering results, the amount of captured signal and, consequently, the MIC performances. Our starting intuition is then consolidated: 2D spectra actually need a less advanced pre-processing level to express at least as much information as corresponding 1D ones;

- when finally searching for relevant biomarkers, our L-sOPLS algorithm allows to provide sparse, orthogonal, cross-validated and more interpretable models. Sparsity, even very severe, goes hand in hand with better predictive ability when using L-sOPLS instead of non-sparse PLS or OPLS, or instead of sPLS;

- the application of L-sOPLS on 2D COSY data sources (pre-processed via GPL or Vectorization) allows to highlight the relevant molecules' fingerprints without any equivocation via non-diagonal biomarkers (corresponding to cross peaks), even if preliminary separability MIC and/or

PCA results are not optimistic. These cross peaks biomarkers would then totally confirm and reinforce 1D results;

- the existence of prohibitive acquisition times for COSY spectra, compared to 1D ones, could not be any more an issue by using faster acquisition techniques as the Non-Uniform Sampling (NUS) or the single scan (NS1). Indeed, these two techniques were submitted to MIC analyses and performed in a similar way than the conventional COSY one (see Chapter 5).

Among all these conclusions, which largely promote a more massive use of 2D spectra in metabolomics, some enhancements and perspectives can be mentioned.

For instance, the notion of sparsity can be approached and considered earlier in the process, by potentially trying sparse hierarchical clustering tools ([Witten and Tibshirani, 2010]) or sparse-PCA ([Witten et al., 2009], [Jolliffe et al., 2003]).

On the other hand, cross-validation choices can be more deeply assessed and controlled when obtaining optimal parameters for the PLS, OPLS, sPLS and L-sOPLS algorithms. In this thesis, a LOO cross-validation scheme was preferred for minimizing the RMSEP criterion.

A first work would be to check if similar results are obtained if a K-fold cross-validation scheme is implemented instead or if external test sets become available.

Also, it is known that feature selection via cross-validation in very high dimensionalities can imply potential new variabilities in the selected items. In this context, stability selection would be a very interesting and rich perspective of enhancement ([Meinshausen and Buhlmann, 2010], [Ye et al., 2012], [Shah and Samworth, 2013]).

More generally, one can considers to apply the whole workflow (GPL-MIC-biomarker discovery) on multi-table or multi-omics analyses, which directly and simultaneously combine different data sources ([Meng et al., 2016], [Hasin et al., 2017], [Akter et al., 2017]). For instance, a multi-table object would mix 1D and 2D data sources, or NMR and Mass data sources, etc... And a multi-omics object would mix metabolomics databases with metadata, genomics, proteomics (among others) ones.

Then, a further interesting work would consist in more deeply investigate the predictive performances of the L-sOPLS algorithm (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.

Finally, another further possible work would be to keep addressing the correlation or colinearity in the data that very often characterizes metabolomics spectral databases. The use of 2D COSY spectra instead of ^{1}H-NMR already improves this issue but several signals could still overlap, move together and/or be associated to a same molecule when considering very complex media.

So, an idea would be to take into account groups of features instead of individual features as inputs of the sparse algorithms. Methods like Group Lasso ([Yuan and Lin, 2007], [Friedman et al., 2010]), Sparse-Group Lasso ([Vincent and Hansen, 2014], [Rao et al., 2016]) or Overlay-Group Lasso ([Friedman et al., 2010]) could be tested in this context.

More directly applicable in the line of this thesis, Group-PLS and Sparse-Group PLS also exist in recent literature ([Liquet et al., 2015], [de Micheaux et al., 2017]) and could be advantageous.

GENERAL BIBLIOGRAPHY

IMPORTANT NOTE : this global bibliography contains all the references quoted in the introduction, Chapter I and conclusions. Other references are available through local bibliographies at the end of Chapters II to V.

[Akter et al., 2017] Akter, S., Wilshire, G., Davis, J. W., Bromfield, J., Crowder, S., Joshi, T., ... and Nagel, S. C., A multi-omics informatics approach for identifying molecular mechanisms and biomarkers in clinical patients with endometriosis. In Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on (pp. 2221-2223). IEEE (2017).

[Aue et al., 1976] Aue, W. P., Bartholdi, E., Ernst, R. R., Two-dimensional spectroscopy. Application to nuclear magnetic resonance, Journal of Chemical Physics. 64: 2229–46 (1976).

[Baraldi et al., 2009] Baraldi, E., Carraro, S., Giordano, G., Reniero, F., Perilongo, G., and Zacchello, F., Metabolomics: moving towards personalized medicine. Italian journal of pediatrics, 35(1), 30 (2009).

[Barupal et al., 2018] Barupal, D. K., Fan, S., and Fiehn, O., Integrating bioinformatics approaches for a comprehensive interpretation of metabolomics datasets. Current opinion in biotechnology, 54, 1-9 (2018).

[Bax and Freeman, 1981] Bax, A. D., and Freeman, R., Investigation of complex networks of spin-spin coupling by two-dimensional NMR, Journal of Magnetic Resonance (1969), 44(3), 542-561 (1981).

[Bax and Summers, 1986] Bax, A., and Summers, M. F., Proton and carbon-13 assignments from sensitivity-enhanced detection of heteronuclear multiple-bond connectivity by 2D multiple quantum NMR, Journal of the American Chemical Society, 108(8), 2093-2094 (1986).

[Beckonert et al., 2007] Beckonert, O., Keun, H. C., Ebbels, T. M., Bundy, J., Holmes, E., Lindon, J. C., and Nicholson, J. K., Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts, Nature protocols, 2(11), 2692 (2007).

[Benjamini and Hochberg, 1995] Benjamini, Y., and Hochberg, Y., Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, Journal of the Royal Statistical Society. Series B (Methodological) 57 (1), 289-300. doi:10.2307/2346101 (1995).

[Billault et al., 2004] Billault, I., Mantle, P. G., and Robins, R. J., Deuterium NMR Used To Indicate a Common Mechanism for the Biosynthesis of Ricinoleic Acid by Ricinus c ommunis and Claviceps p urpurea. Journal of the American Chemical Society, 126(10), 3250-3256 (2004).

[Bloemberg et al., 2013] Bloemberg, T.G., Gerretzen,J., Lunshof, A., Wehrens, R., Buydens, L.M., Warping methods for spectroscopic and chromatographic signal alignment: a tutorial, Anal. Chim. Acta 781, 14-32 (2013).

[Bouveyron et al., 2016] Bouveyron, C., Latouche, P., and Mattei, P. A., Bayesian variable selection for globally sparse probabilistic PCA. arXiv preprint arXiv:1605.05918 (2016).

[Breiman, 2001] Breiman L., Random forests, Machine learning, 45(1), 5-32 (2001).

[Breunig et al., 2000] Breunig, M. M., Kriegel, H. P., Ng, R. T., and Sander, J., LOF: identifying density-based local outliers. In ACM sigmod record (Vol. 29, No. 2, pp. 93-104). ACM (2000).

[Brindle et al., 2003] Brindle, J. T., Nicholson, J. K., Schofield, P. M., Grainger, D. J., and Holmes, E., Application of chemometrics to 1 H NMR spectroscopic data to investigate a relationship between human serum metabolic profiles and hypertension. Analyst, 128(1), 32-36 (2003).

[Brown et al., 1977] Brown, F. F., Campbell, I. D., and Kuchel, P. W., Human erythrocyte metabolism studies by 1H spin echo NMR. FEBS letters, 82(1), 12-16 (1977).

[Buhlmann and Yu, 2006] Buhlmann, P., and Yu, B., Sparse boosting. Journal of Machine Learning Research, 7(Jun), 1001-1024 (2006).

[Chapman and Saad, 1997] Chapman, A., and Saad, Y., Deflated and augmented Krylov subspace techniques. Numerical Linear Algebra with Applications, 4(1), 43–66 (1997).

[Chen M. and Vigneau, 2016] Chen, M., and Vigneau, E., Supervised clustering of variables. Advances in Data Analysis and Classification, 10(1), 85-101 (2016).

[Chen T. and Guestrin, 2016] Chen T., Guestrin C., Xgboost: A scalable tree boosting system. arXiv preprint arXiv:1603.02754 (2016).

[Chun and Keles, 2007] Chun, H., and Keles, S., Sparse partial least squares regression with an application to genome scale transcription factor analysis. Madison: Department of Statistics, University of Wisconsin (2007).

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

[Claridge, 2016] Claridge, T. D., High-resolution NMR techniques in organic chemistry (Vol. 27). Elsevier (2016).

[Courant et al., 2014] Courant, F., Antignac, J. P., Dervilly-Pinel, G., and Le Bizec, B., Basics of mass spectrometry based metabolomics. Proteomics, 14(21-22), 2369-2388 (2014).

[Dalvit and Bovermann, 1995] Dalvit, C., and Bovermann, G., Pulsed field gradient one-dimensional NMR selective ROE and TOCSY experiments, Magnetic Resonance in Chemistry, 33(2), 156-159 (1995).

[Davies and Lampen, 1993] Davies, A. N., and Lampen, P., Jcamp-Dx for NMR. Applied spectroscopy, 47(8), 1093-1099 (1993).

[De Jong, 1993] De Jong S., SIMPLS: an alternative approach to partial least squares regression, Chemometrics and intelligent laboratory systems, 18(3), 251-263 (1993).

[de Micheaux et al., 2017] de Micheaux, P. L., Liquet, B., and Sutton, M., A unified parallel algorithm for regularized group PLS scalable to big data. arXiv preprint arXiv:1702.07066 (2017).

[Dettmer et al., 2007] Dettmer, K., Aronov, P. A., and Hammock, B. D., Mass spectrometry-based metabolomics, Mass spectrometry reviews, 26(1), 51-78 (2007).

[Dixon, 1984] Dixon, W. T., Simple proton spectroscopic imaging. Radiology, 153(1), 189-194 (1984).

[Duarte et al., 2014] Duarte, I. F., Diaz, S. O., and Gil, A. M., NMR metabolomics of human blood and urine in disease research. Journal of Pharmaceutical and Biomedical Analysis, 93, 17-26 (2014).

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

[Eilers, 2004] P.H.C. Eilers, Parametric time warping, Anal. Chem. 76 (2), 404-411 (2004).

[Eilers and Boelens, 2005] Eilers, P.H.C., Boelens, H.F., Baseline Correction with Asymmetric Least Squares Smoothing, Medical Centre Report (unpublished results), Leiden University (2005).

[Euceda et al., 2015] Euceda, L.R., Giskeodegaard, G.F., Bathen, T.F., Preprocessing of NMR metabolomics data, Scand. J. Clin. Lab. Investig. 75 (3), 193-203 (2015).

[Feng et al., 2016] Feng, Q., Liu, Z., Zhong, S., Li, R., Xia, H., Jie, Z., ... and Guo, Z., Integrated metabolomics and metagenomics analysis of plasma and urine identified microbial metabolites associated with coronary heart disease. Scientific reports, 6, 22525 (2016).

[Feraud et al., 2017] Feraud, B., Rousseau, R., de Tullio, P., Verleysen, M., and Govaerts, B., Independent Component Analysis and Statistical Modelling for the Identification of Metabolomics Biomarkers in 1H-NMR Spectroscopy. J Biom Biostat 8: 367. doi: 10.4172/2155-6180.1000367 (2017).

[Fiehn, 2002] Fiehn, O., Metabolomics - the link between genotypes and phenotypes. In Functional genomics (pp. 155-171). Springer, Dordrecht (2002).

[Frederich et al., 2016] Frederich, M., Pirotte, B., Fillet, M., and De Tullio, P., Metabolomics as a challenging approach for medicinal chemistry and personalized medicine. Journal of medicinal chemistry, 59(19), 8649-8666 (2016).

[Freund and Schapire, 1999] Freund, Y., and Schapire, R., A short introduction to boosting, Journal of japanese society for artificial intelligence, 14(5), pp. 771-780 (1999).

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

[Friedman et al., 2001] Friedman, J., Hastie, T., and Tibshirani, R., The elements of statistical learning (Vol. 1, No. 10). New York, NY, Springer series in statistics (2001).

[Giraudeau, 2010] Giraudeau, P., Résonance magnétique nucléaire bidimensionnelle quantitative. Editions universitaires européenne, p. 20 (2010).

[Giraudeau and Frydman, 2014] Giraudeau, P., and Frydman, L., Ultrafast 2D NMR: an emerging tool in analytical spectroscopy, Annual Review of Analytical Chemistry, 7, 129-161 (2014).

[Gu et al., 2009] Gu, I. Y., Billeter, M., Sharafy, M. R., Sorkhabi, V. A., Fredriksson, J., and Staykova, D. K., Parameter estimation of multidimensional NMR signals based on high-resolution subband analysis of 2D NMR projections, ICASSP 2009. IEEE International Conference, 497-500 (2009).

[Hasin et al., 2017] Hasin, Y., Seldin, M., and Lusis, A., Multi-omics approaches to disease. Genome biology, 18(1), 83 (2017).

[Hastie et al., 2015] Hastie, T., Tibshirani, R., and Wainwright, M., Statistical learning with sparsity: The lasso and generalizations. Boca Raton: CRC Press (2015).

[Hastie, Rosset et al., 2009] Hastie, T., Rosset, S., Zhu, J., and Zou, H., Multiclass adaboost. Statistics and its Interface, 2(3), 349-360 (2009).

[Henneges at al., 2009] Henneges, C., Bullinger, D., Fux, R., Friese, N., Seeger, H., Neubauer, H., ... and Kammerer, B., Prediction of breast cancer by profiling of urinary RNA metabolites using Support Vector Machine-based feature selection. BMC cancer, 9(1), 104 (2009).

[Hoerl and Kennard, 1970] Hoerl, A. E., and Kennard, R. W., Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67 (1970).

[Hoerl et al., 1975] Hoerl, A. E., Kennard, R. W., and Baldwin, K. F., Ridge regression: some simulations. Communications in Statistics-Theory and Methods, 4(2), 105-123 (1975).

[Hotelling, 1933] Hotelling, H., Analysis of a complex of statistical variables into principal components, Journal of Educational Psychology, 24(6), pp. 417-441 (1933).

[Hunter, 2009] Hunter, P., Reading the metabolic fine print: The application of metabolomics to diagnostics, drug research and nutrition might be integral to improved health and personalized medicine. EMBO reports, 10(1), 20-23 (2009).

[Hyvarinen, 1999] Hyvarinen, A., Fast ICA for noisy data using Gaussian moments. In Circuits and Systems, 1999. ISCAS'99. Proceedings of the 1999 IEEE International Symposium on (Vol. 5, pp. 57-61). IEEE (1999).

[Hyvarinen and al., 2004] Hyvarinen, A., Karhunen, J., and Oja, E., Independent component analysis (Vol. 46). John Wiley and Sons (2004).

[Jolliffe et al., 2003] Jolliffe, I. T., Trendafilov, N. T., and Uddin, M., A modified principal component technique based on the LASSO. Journal of computational and Graphical Statistics, 12(3), 531-547 (2003).

[Kearns and Valiant, 1989] Kearns, M., and Valiant, L., Crytographic limitations on learning Boolean formulae and finite automata, Symposium on Theory of computing. ACM. 21: 433-444 (1989).

[Keeler, 2010] Keeler, J., Understanding NMR Spectroscopy (2nd ed.), Wiley, pp. 184–187, 2010.

[Keogh and Mueen, 2011] Keogh, E., and Mueen, A., Curse of dimensionality, In Encyclopedia of machine learning (pp. 257-258), Springer, Boston, MA (2011).

[Kock et al., 1996] Kock, M., Reif, B., Fenical, W., and Griesinger, C., Differentiation of HMBC two-and three-bond correlations: A method to simplify the structure determination of natural products. Tetrahedron letters, 37(3), 363-366 (1996).

[Kramer, 2007] Kramer, N., An overview on the shrinkage properties of partial least squares regression. Computational Statistics, 22(2), 249-273 (2007).

[Krastanov, 2010] Krastanov, A., Metabolomic - The state of art, Biotechnology and Biotechnological Equipment, 24(1), 1537-1543 (2010).

[Kriegel et al., 2009] Kriegel, H. P., Kroger, P., Schubert, E., and Zimek, A., LoOP: local outlier probabilities. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 1649-1652). ACM (2009).

[Le Cao et al., 2008] Le Cao, K. A., Rossouw, D., Robert-Granie, C., and Besse, P., A sparse PLS for variable selection when integrating omics data. Statistical Applications in Genetics and Molecular Biology, 7(1), 35 (2008).

[Leenders et al., 2015] Leenders, J., Frederich, M., and De Tullio, P., Nuclear magnetic resonance: a key metabolomics platform in the drug discovery process. Drug Discovery Today: Technologies, 13, 39-46 (2015).

[Liland, 2011] Liland, K.H., Multivariate methods in metabolomics - from preprocessing to dimension reduction and statistical analysis, Trac. Trends Anal. Chem. 30 (6), 827-841 (2011).

[Liland et al., 2010] Liland, K.H., Almoy, T., and Mevik, B. H., Optimal choice of baseline correction for multivariate calibration of spectra, Appl. Spectrosc. 64 (9), 1007-1016 (2010).

[Lin et al., 2012] Lin, X., Yang, F., Zhou, L., Yin, P., Kong, H., Xing, W., ... and Xu, G., A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information. Journal of chromatography B, 910, 149-155 (2012).

[Lindon et al., 2003] Lindon, J. C., Nicholson, J. K., Holmes, E., Antti, H., Bollard, M. E., Keun, H., ... and Stevens, G. J., Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project. Toxicology and applied pharmacology, 187(3), 137-146 (2003).

[Liquet et al., 2015] Liquet, B., de Micheaux, P. L., Hejblum, B. P., and Thiebaut, R., Group and sparse group partial least square approaches applied in genomics context. Bioinformatics, 32(1), 35-42 (2015).

[Loria et al., 1999] Loria, J. P., Rance, M., and Palmer, A. G., A TROSY CPMG sequence for characterizing chemical exchange in large proteins. Journal of biomolecular NMR, 15(2), 151-155 (1999).

[Malz and Jancke, 2005] Malz, F., and Jancke, H., Validation of quantitative NMR. Journal of pharmaceutical and biomedical analysis, 38(5), 813-823 (2005).

[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).

[Marion, 2016] Marion, R., Pre-processing of NMR Spectra: Review and Evaluation of Baseline Correction, Normalization, Scaling and Transformation Methods, Master thesis (unpublished results), Ecole de Statistique, Biostatistique et Sciences Actuarielles, UCL (2016).

[Martin and Zekter, 1988] Martin, G. E, Zekter, A. S., Two-Dimensional NMR Methods for Establishing Molecular Connectivity. New York: VCH Publishers, Inc. p. 59 (1988).

[Martin et al., 2017] Martin, M., Legat, B., Leenders, J., Vanwinsberghe, J., Rousseau, R., Boulanger, B., ... and Govaerts, B., PepsNMR for the 1H-NMR metabolomic data pre-processing (No. UCL-Université Catholique de Louvain) (2017).

[Mehmood and al., 2011] Mehmood, T., Martens, H., Saebo, S., Warringer, J., Snipen, L., A Partial Least Squares algorithm for parsimonious variable selection. Algorithms for Molecular Biology, 6-27 (2011).

[Meiboom and Gill, 1958] Meiboom, S., and Gill, D., Modified spin-echo method for measuring nuclear relaxation times. Review of scientific instruments, 29(8), 688-691 (1958).

[Meinshausen and Buhlmann, 2010] Meinshausen, N., and Buhlmann, P., Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(4), 417-473 (2010).

[Meng et al., 2016] Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M., and Culhane, A. C., Dimension reduction techniques for the integrative analysis of multi-omics data. Briefings in bioinformatics, 17(4), 628-641 (2016).

[Mevik and Cederkvist, 2004] Mevik, B. H., and Cederkvist, H. R., 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 (2004).

[Munoz-Romero et al., 2015] Munoz-Romero, S., Arenas-Garca, J., and Gomez-Verdejo, V., Sparse and kernel OPLS feature extraction based on eigenvalue problem solving. Pattern Recognition, 48(5), 1797–1811 (2015).

[Norris and Lecavalier, 2010] Norris, M., and Lecavalier, L., Evaluating the use of exploratory factor analysis in developmental disability psychological research, Journal of autism and developmental disorders, 40(1), 8-20 (2010).

[Papadimitriou et al., 2003] Papadimitriou, S., Kitagawa, H., Gibbons, P. B., and Faloutsos, C., LOCI: Fast outlier detection using the local correlation integral. In Data Engineering, 2003. Proceedings. 19th International Conference on (pp. 315-326). IEEE (2003).

[Patti et al., 2012] Patti, G. J., Yanes, O., and Siuzdak, G., Innovation: Metabolomics: the apogee of the omics trilogy. Nature reviews Molecular cell biology, 13(4), 263 (2012).

[Paul and Dupont, 2015] Paul J., Dupont P., Inferring statistically significant features from Random Forests, Neurocomputing, 150, 471-480 (2015).

[Pauli, 2001] Pauli, G. F., qNMR—a versatile concept for the validation of natural product reference compounds. Phytochemical Analysis, 12(1), 28-42 (2001).

[Pearson, 1901] Pearson, K., On lines and planes of closest fit to systems of points in space, Philosophical Magazine, 2, pp. 559-572 (1901).

[Rao et al., 2016] Rao, N. S., Nowak, R. D., Cox, C. R., and Rogers, T. T., Classification With the Sparse Group Lasso. IEEE Trans. Signal Processing, 64(2), 448-463 (2016).

[Raykov et al., 2016] Raykov, Y. P., Boukouvalas, A., Baig, F., and Little, M. A., What to do when K-Means clustering fails: a simple yet principled alternative algorithm, PloS one, 11(9) (2016).

[Rinaudo et al., 2016] Rinaudo P., Boudah S., Junot C., Thevenot E. A., Biosigner: a new method for the discovery of significant molecular signatures from omics data, Frontiers in Molecular Biosciences, 3, 26, ISO 690 (2016).

[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).

[San Cristobal et al., 2013] San Cristobal, M., Sanchez, M. P., Mercat, M. J., Rohart, F., Liaubet, L., Tribout, T., ... and Laurent, B., Le métabolome, un moyen pour trouver de nouveaux biomarqueurs?. Viandes et Produits Carnés, 1-5 (2013).

[Schapire, 2003] Schapire, R. E., The boosting approach to machine learning: An overview, In Nonlinear estimation and classification, pp. 149-171, Springer, New York, NY (2003).

[Schroeder et al., 2007] Schroeder, F. C., Gibson, D. M., Churchill, A. C., Sojikul, P., Wursthorn, E. J., Krasnoff, S. B., and Clardy, J., Differential analysis of 2D NMR spectra: New natural products from a pilot-scale fungal extract library, Angewandte Chemie, 119(6), 919-922 (2007).

[Sevin et al., 2015] Sevin, D. C., Kuehne, A., Zamboni, N., and Sauer, U., Biological insights through nontargeted metabolomics. Current opinion in biotechnology, 34, 1-8 (2015).

[Shah and Samworth, 2013] Shah, R. D., and Samworth, R. J., Variable selection with error control: another look at stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75(1), 55-80 (2013).

[Silverstein et al., 2014] Silverstein, R. M., Webster, F. X., Kiemle, D. J., and Bryce, D. L., Spectrometric identification of organic compounds. John wiley and sons (2014).

[Smolinska et al., 2012] Smolinska, A., Blanchet, L., Buydens, L.M., Wijmenga, S.S., NMR and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review, Anal. Chim. Acta 750, 82-97 (2012).

[Song Q. and Liang, 2015] Song, Q., and Liang, F., A split-and-merge Bayesian variable selection approach for ultrahigh dimensional regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 77(5), 947-972 (2015).

[Song R. et al., 2016] Song R., Chen S., Deng B., Li L., eXtreme Gradient Boosting for Identifying Individual Users Across Different Digital Devices. International Conference on Web-Age Information Management (43-54). Springer International Publishing (2016).

[Spicer et al., 2017] Spicer, R., Salek, R. M., Moreno, P., Canueto, D., and Steinbeck, C., Navigating freely-available software tools for metabolomics analysis. Metabolomics, 13(9), 106 (2017).

[Steuer et al., 2007] Steuer, R., Morgenthal, K., Weckwerth, W., and Selbig, J., A gentle guide to the analysis of metabolomic data, Metabolomics, pp. 105-126, Humana Press (2007).

[Tapp and Kemsley, 2009] Tapp, H. S., and Kemsley, E. K., Notes on the practical utility of OPLS. TrAC Trends in Analytical Chemistry, 28(11), 1322–1327 (2009).

[Taylor et al., 2002] Taylor, J., King, R. D., Altmann, T., and Fiehn, O., Application of metabolomics to plant genotype discrimination using statistics and machine learning, Bioinformatics, 18(suppl. 2), S241-S248 (2002).

[Tenailleau et al., 2004] Tenailleau, E. J., Lancelin, P., Robins, R. J., and Akoka, S., Authentication of the origin of vanillin using quantitative natural abundance 13C NMR. Journal of agricultural and food chemistry, 52(26), 7782-7787 (2004).

[Tenenhaus, 1998] Tenenhaus, M., La régression PLS: Théorie et pratique, Editions Technip, Paris (1998).

[Thiel et al., 2017] Thiel, M., Feraud, B., and Govaerts, B., ASCA+ and APCA+: Extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs. Journal of Chemometrics, 31(6), e2895 (2017).

[Tibshirani, 1996] Tibshirani, R., Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Methodological), 267-288 (1996).

[Tibshirani et al., 2005] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K., Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(1), 91-108 (2005).

[Tofts and Wray, 1988] Tofts, P. S., and Wray, S., A critical assessment of methods of measuring metabolite concentrations by NMR spectroscopy. NMR in Biomedicine, 1(1), 1-10 (1988).

[Tredwell et al., 2016] Tredwell, G.D., Bundy, J.G., De Iorio, M., Ebbels, T.M.D., Modelling the acid/base 1H NMR chemical shift limits of metabolites in human urine. Metabolomics.12(10):152 (2016).

[Trygg and Wold, 2002] Trygg, J., and Wold, S., Orthogonal projections to latent structures (O-PLS). Journal of Chemometrics, 16(3), 119–128 (2002).

[van Gerven and Heskes, 2010] van Gerven, M. A. J., and Heskes, T., Sparse orthonormalized partial least squares. In Benelux conference on artificial intelligence (2010).

[Vincent and Hansen, 2014] Vincent, M., and Hansen, N. R., Sparse group lasso and high dimensional multinomial classification. Computational Statistics and Data Analysis, 71, 771-78 (2014).

[Vigneau and Qannari, 2003] Vigneau, E., and Qannari, E. M., Clustering of variables around latent components. Communications in Statistics-Simulation and Computation, 32(4), 1131-1150 (2003).

[Vigneau et al., 2015] Vigneau, E., Chen, M., and Qannari, E. M., ClustVarLV: An R Package for the Clustering of Variables Around Latent Variables. R Journal, 7(2), (2015).

[Vis et al., 2007] Vis, D.J., Westerhuis, J.A., Smilde, A.K., van der Greef, J., Statistical validation of megavariate effects in ASCA. BMC Bioinf. 8-322 (2007).

[Vu and Laukens, 2013] Vu, T.N., Laukens, K., Getting your peaks in line: a review of alignment methods for NMR spectral data, Metabolites 3 (2) (2013).

[Wang et al., 2015] Wang, Y., Yang, G., and Guo, L., A novel sparse boosting method for crater detection in the high resolution planetary image. Advances in Space Research, 56(5), 982-991 (2015).

[Weckwerth, 2003] Weckwerth, W., Metabolomics in systems biology. Annual review of plant biology, 54(1), 669-689 (2003).

[Wehrens, 2011] Wehrens R., Chemometrics with R: multivariate data analysis in the natural sciences and life sciences, Springer Science and Business Media, 155-165 (2011).

[Weljie et al., 2006] Weljie, A.M., Newton, J., Mercier, P., Carlson, E., Slupsky, C.M., Targeted pofiling: Quantitative analysis of 1H NMR metabolomics data. Analytical Chemistry. 78(13):4430–4442 (2006).

[Wiklund et al., 2008] Wiklund, S., Johansson, E., Sjostrom, L., Mellerowicz, E.J., Edlund, U., Shockcor, J.P., Gottfries, J., Moritz, T., Trygg, J., Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models, Analytical Chemistry, 80(1):115-22 (2008).

[Witten and Tibshirani, 2010] Witten, D. M. and Tibshirani, R., A framework for feature selection in clustering. Journal of the American Statistical Association, 105(490), 713-726 (2010).

[Witten et al., 2009] Witten, D. M., Tibshirani, R., and Hastie, T., A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics, 10(3), 515-534 (2009).

[Wold H., 1975] Wold H., Path models with latent variables: The NIPALS approach, 307-357, Acad. Press. (1975).

[Wold, 1995] Wold, S., PLS for multivariate linear modeling, Chemometric methods in molecular design, Vol 2. Wiley-VCH (1995).

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

[Wright et al., 1995] Wright, B., Greatbanks, D., Taberner, J., and Wilson, I. D., Comparison of three methods for water suppression in biofluid NMR: advantages of NOESYPRESAT. Pharmacy and Pharmacology Communications, 1(4), 197-199 (1995).

[Wu and Li, 2016] Wu, Y., Li, L., Sample normalization methods in quantitative metabolomics, J. Chromatogr. A 1430, 80-95 (2016).

[Xi et al., 2006] Xi, Y., de Ropp, J. S., Viant, M. R., Woodruff, D. L., and Yu, P.,. Automated screening for metabolites in complex mixtures using 2D COSY NMR spectroscopy. Metabolomics, 2(4), 221-233 (2006).

[Xi and Rocke, 2008] Xi, Y., and Rocke, D. M., Baseline correction for NMR spectroscopic metabolomics data analysis. BMC bioinformatics, 9(1), 324 (2008).

[Xia et al., 2008] Xia, J., Bjorndahl, T. C., Tang, P., and Wishart, D. S., MetaboMiner: semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. BMC bioinformatics, 9(1), 507 (2008).

[Xia and Wishart, 2016] Xia, J. and Wishart, D. S., Using MetaboAnalyst 3.0 for comprehensive metabolomics data analysis. Current protocols in bioinformatics, 14-10 (2016).

[Ye et al., 2012] Ye, J., Farnum, M., Yang, E., Verbeeck, R., Lobanov, V., Raghavan, N., ... and Narayan, V. A., Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. BMC neurology, 12(1), 46 (2012).

[Yuan and Lin, 2007] Yuan, M., and Lin, Y., On the non-negative garrotte estimator, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(2), 143-161 (2007).

[Zhou, 2013] Zhou, D. X., On grouping effect of elastic net. Statistics and Probability Letters, 83(9), 2108-2112 (2013).

[Zou and Hastie, 2005] Zou, H., Hastie, T., Regularization and Variable Selection via the Elastic Net, Journal of the Royal Statistical Society. Series B (statistical Methodology). Wiley. 67 (2), 301-20 (2005).