

Uncovering informative content in metabolomics data: *From pre-processing of ^1H NMR spectra to biomarkers discovery in multifactorial designs.*

A thesis submitted in partial fulfilment of the requirements for the degree of Docteur en Sciences, UCLouvain, 2019

MANON MARTIN

Thesis committee:

Prof. Bernadette Govaerts, Supervisor (LIDAM, UCLouvain, Belgium)

Prof. Catherine Legrand, President (LIDAM, UCLouvain, Belgium)

Prof. Rainer von Sachs (LIDAM, UCLouvain, Belgium)

Prof. Michel Verleysen (ICTEAM, UCLouvain, Belgium)

Prof. Paul H. C. Eilers (Erasmus University Medical Center, The Netherlands)

Prof. Pascal de Tullio (CIRM, ULiège, Belgium)

Acknowledgements

Bien entendu, le fait qu'une thèse puisse se construire et que son auteure vive au mieux cette période est en grande partie lié à son environnement direct, y compris de travail.

Je tiens tout d'abord à remercier Bernadette, ma professeure, promotrice de mémoire et de thèse, pour ces nombreuses heures de réunion à parler statistiques (et voyages, randos, courses d'orientation, etc.). Je n'aurais pas pu espérer mieux comme encadrement, tu as toujours été présente pour me donner les clés et me permettre d'avancer en toute circonstance, tout en me laissant assez de latitude pour faire ce que je voulais et me permettre de développer les compétences qui m'attiraient le plus. Merci du fond du coeur !

I also wish to thank Paul H.C. Eilers, Pascal de Tullio, Michel Verleysen, Rainer Von Sachs and Catherine Legrand, some of them members of my thesis committee and all members of my thesis jury, for their interest in this work, involvement in the follow-up meetings and in my two defences, along with time spent proofreading and many helpful suggestions. Your help was very precious to improve this manuscript.

Ensuite, il va de soit que je remercie le SMCS, sans qui je n'aurais peut-être pas fait mes premiers pas dans la recherche. Mes collègues de bureau et de couloir: Séverine, Aurélie, Lieven, Céline, Alain, Catherine, Nathalie, Vincent, Marie-Pascale, Sabrina, Arnaud et Matthieu, merci pour votre bonne humeur et pour m'avoir fait confiance en me confiant des tâches de consultances et de formations. Les réunions du mardi, team buildings, repas annuels et tous les autres bons moments partagés ensemble me manqueront.

Merci également à tous mes collègues doctorants de l'institut sans qui l'ambiance n'aurait pas été la même: Alexandre, Anna, Aurélie, Baptiste, Benjamin, Cédric,

Florian, Gilles, Hélène, Maïlis, Michel, Mickaël, Nathalie, Nicolas, Rebecca, Sophie, Sylvie, Vincent, ainsi que tous les autres. Vous m'avez permis de rendre cette période travail (relativement solitaire pour un doctorant) d'autant plus agréable. Merci pour toutes ces pauses, ces temps de midi, ces parties de badminton, ces verres du vendredi et ces soirées partagées !

Merci plus généralement aux autres professeurs de l'institut que j'ai eu la chance de côtoyer pendant mes années de master et de thèse, pour votre bienveillance et le super encadrement dont j'ai bénéficié durant mes années de master.

Pour tous ces échanges enrichissants et ces mémorables souvenirs de conférences, je remercie nos collègues de Liège (Justine, Pascal, Matthieu et Ariana). Ces conférences n'auraient pas été les mêmes sans votre présence, c'est certain !

Je tiens également à remercier l'ensemble de l'équipe administrative de l'institut : Nancy, Maguy, Sophie, Nadja and Tatiana, pour leur disponibilité et écoute à toute épreuve. Il en va de même pour le service informatique, qui a toujours réagi en un temps record pour toutes mes demandes.

Merci également aux membres de l'équipe AHDIDA (en espérant que le hH soit bien placé :)) pour tous ces échanges constructifs et l'esprit de partage qui s'est dégagé de ces réunions.

Enfin, bien évidemment, je ne remercierai jamais assez mes proches pour leur support et encouragements, en toute occasion. Sans eux je ne serais peut-être pas arrivée au bout de cette aventure. Merci David, particulièrement durant cette dernière année qui fut bien remplie pour tous les deux !

Contents

1	Introduction	1
1.1	Foreword	1
1.2	The omics family and metabolomics in health sciences	4
1.2.1	Omics technologies	4
1.2.2	Metabolomics	4
1.3	A typical metabolomics study	8
1.3.1	Typical workflow	8
1.3.2	Where do statisticians help?	12
1.4	Multivariate data analysis for untargeted 1D NMR spectra	12
1.4.1	Data characteristics	12
1.4.2	The curse of dimensionality in classification	14
1.4.3	Explorative analysis	14
1.4.4	Supervised analysis with one categorical variable	18
1.4.5	Supervised analysis for more advanced designs	20
1.5	The question of spectral repeatability	23
1.6	Datasets used in this thesis	25
1.7	Content and contributions	26
1.8	Notes to the readers	31
1.8.1	On mathematical notation	31
1.8.2	On the bibliographic references	31
2	PepsNMR	33

2.1	Introduction	33
2.1.1	Literature review	33
2.1.2	Objectives	34
2.1.3	Content	35
2.2	Methodology	35
2.2.1	PepsNMR workflow presentation	35
2.2.2	Details on the pre-processing steps	38
2.2.3	Other PepsNMR features and functions	54
2.3	Quality criteria for PepsNMR evaluation	54
2.4	Results and discussion	56
2.4.1	Variance structure	56
2.4.2	Clustering and classification performance	57
2.5	Conclusions	61
2.6	Software availability	62
2.6.1	Bioconductor	62
2.6.2	Workflow4Metabolomics	63
2.6.3	Use PepsNMR in practice	63
2.7	Datasets	64
2.7.1	Human Serum	64
2.7.2	Human Urine	64
2.7.3	Spectral pre-processing and outlier removal	65
3	PLS-derived methods for FS	67
3.1	Introduction	67
3.1.1	Feature selection in metabolomics	67
3.1.2	Objectives	69
3.1.3	Content	70
3.2	The Semi-artificial urine database	70
3.3	Methods	71
3.3.1	Multiple t-tests	72
3.3.2	Partial Least Squares	73
3.3.3	Orthogonal Partial Least Squares	82
3.3.4	Sparse PLS	88
3.3.5	Light Sparse OPLS	92
3.3.6	Summary of the feature selection procedures	93
3.4	Meta-parameters tuning	93
3.5	Application of the methods to a sub-sample	95
3.6	Evaluation of the classifiers performances	102

CONTENTS

3.6.1	Introduction	102
3.6.2	MPs stability	104
3.6.3	Size of the selection sets for automatic threshold methods	105
3.6.4	ROC curves and classification accuracies	105
3.6.5	General assessment of features accuracy	109
3.6.6	Selected features inspection on the spectral scale	114
3.6.7	Feature selection stability	122
3.7	Conclusions	125
Appendices		129
3.A	Other results with $n = 60$	129
3.A.1	MPs stability	129
3.A.2	Size of the selection sets for automatic threshold methods	130
3.A.3	Classification accuracy	131
3.A.4	General assessment of features accuracy	131
3.A.5	Selected features inspection on the spectral scale	134
3.A.6	Feature stability	138
4	ASCA & other multivariate techniques for multifactorial designs	141
4.1	Introduction	141
4.1.1	Analysis of designed experiments with high dimensional responses	141
4.1.2	ASCA ⁺ , an extension to unbalanced experimental designs	144
4.1.3	Objectives	144
4.2	ASCA ⁺ and related methods	145
4.2.1	Overview	145
4.2.2	Fixed effects ANOVA in the General Linear Model notation	146
4.2.3	Effect matrix decomposition	147
4.2.4	Outputs	148
4.2.5	PARAFASCA	150
4.2.6	AComDim	155
4.2.7	AMOPLS	158
4.3	Data	162
4.4	Results	164
4.4.1	GLM decomposition	164
4.4.2	Components data decomposition and measure of component importance	164
4.4.3	Components interpretation	167

4.4.4	Significance tests of model effects	168
4.4.5	Scores	171
4.4.6	Loadings	177
4.5	Conclusions	180
Appendices		183
4.A	PARAFASCA algorithm	183
4.B	AComDim algorithm	185
4.C	Factor mixing in ACOMDIM and AMOPLS	186
5 LiMM-PCA		189
5.1	Introduction	189
5.1.1	Motivations and objectives	190
5.1.2	Content	190
5.2	Limitations of ASCA ⁺ when random factors & effects are included in the model	191
5.3	The Linear Mixed Model	193
5.3.1	Model terms definition and justification	193
5.3.2	General notations and assumptions	193
5.3.3	Estimation	194
5.3.4	Test of effects significance	194
5.4	LiMM-PCA: combining ASCA ⁺ and LMM for advanced designs and modelling	195
5.4.1	STEP 1: PCA orthogonalisation and dimension reduction of the response matrix	197
5.4.2	STEP 2: Parallel mixed modelling	197
5.4.3	STEP 3: Effect matrix decomposition	198
5.4.4	STEP 4a : Quantification of effects importance	199
5.4.5	STEP 4b : Testing effects significance	200
5.4.6	STEP 5 : Visual representation of the effect matrices	201
5.5	Case studies	202
5.5.1	Candy sensory dataset	203
5.5.2	Variance components analysis of metabolomics data	214
5.6	Conclusions and perspectives	224
Appendices		227
5.A	The model fitting process in LMM	227
5.B	Effective model dimension in the presence of random effects	229
5.B.1	Theory	229

CONTENTS

5.B.2 Results for the two case studies	230
5.C Histograms and quantiles of true and bootstrapped (R)LLR statistics	235
5.D ANOVA E(MS) and F-test table for the metabolomics dataset	240
6 General conclusions	241
Bibliography	246
General Appendices	266
A Abbreviations	267
B Mathematical notations	271
C Datasets	277

CHAPTER 1

Introduction

1.1 Foreword

Modern high-throughput technologies producing a tsunami of data are the today norm in life sciences rather than the exception. These biological data have multiple sources of unwanted variation: sampling, day, lab. variation, etc. While the number of samples remains dictated by budgetary limits or difficulties to hire targeted experimental samples, the number of measured variables has grown exponentially. So are the related computational and statistical challenges to process and extract meaningful information from these data tables. In this data- and technology-driven science, statisticians are asked to directly meet the needs and face the realities of the whole metabolomics data treatment chain.

Due to the high level of multidisciplinary competencies that are required in this field, metabolomics data processing and analysis is usually performed with an unquestioned use of mainstream software/methods (e.g. proprietary pre-processing software or the use of data analysis workflow in commercial software), mainly working as a complete black box.

This thesis addresses topical issues for the current trends for metabolomics data treatment. It suggests practical solutions with regards to ^1H NMR data pre-processing and biomarker search for simple or multifactorial designs and “short and wide” data tables. Figure 1.1 illustrates the manuscript outline on the scale of a metabolomics study. All the methodological developments presented in

this thesis were applied to a various set of experimental datasets, mostly from NMR metabolomics.

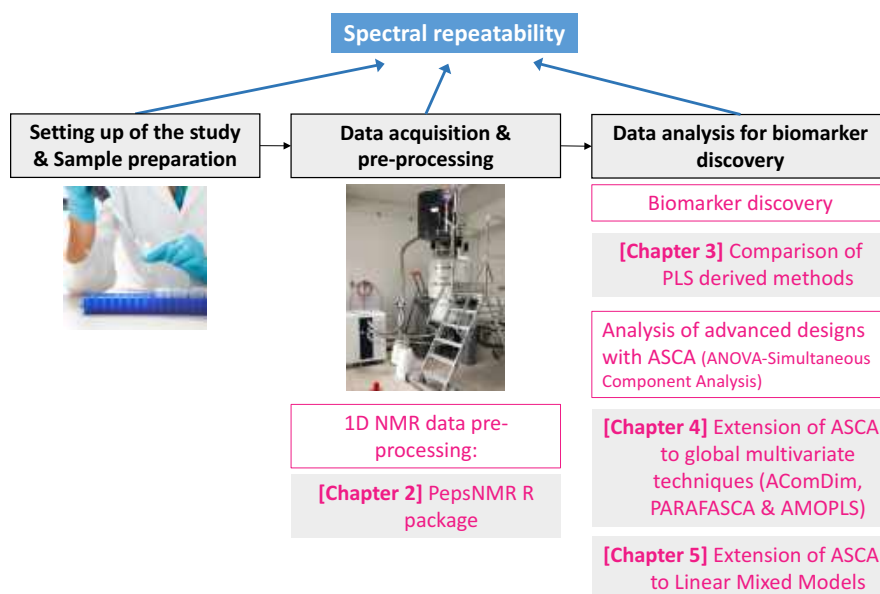


Figure 1.1 – Thesis content on the scale of a metabolomics study. Note that Chapter 1 is the present introductory chapter.

Specific to ^1H NMR, the first part of this work aimed at implementing and publishing as an **R package**, a **comprehensive** and **modular pre-processing workflow**. This package, **PepsNMR**, which is based on a pre-existing Matlab package, involves **common** and **innovative methodological steps** (e.g. solvent suppression, baseline removal or spectral realignment) and complementary tools to transform the FID signal into an interpretable spectrum and to further validate the **pre-processing quality**. This whole pre-processing strategy has been successfully compared to a classic scheme through quantitative quality criteria measuring the information recovery. The high modularity of this package makes it possible to combine/compare it with other well performing or new tools and methods.

The following parts of this thesis are focussed on various data analysis techniques targeted towards metabolomics applications, although a straightforward generalisation to other kinds of data that share the same characteristics (multicollinearity, low (*high*) number of samples (*variables*)) is possible.

The second chapter analyses deeply **classic** and **sparse (O)PLS methods** when it is intended to apply them in the scope of biomarkers search. They are first presented and explained in a unified fashion. Then, given the current lack in the literature, they are compared for **feature selection** in an **innovative evaluation process**. For this last purpose, a semi-artificial dataset is used where the exact **biomarkers location is known in advance** and different realistic conditions are investigated (signal to noise levels and sample sizes). Emphasis has been placed to (1) use numerous resampled sub-datasets to average the methods performances (2) train appropriately the supervised models to avoid their overfit to the data and (3) compare the feature selection techniques with several evaluation measurements, which, for a part of them, integrate the prior information from the true features location.

The last two chapters highlight and propose new developments of **ASCA**, a technique to analyse high-dimensional data issued from multifactorial designs, which is based on classic univariate ANOVA methods. ASCA is first **compared to analogous global methods (AComDim, AMOPLS and PARAFASCA)**, all being already applied in the context of metabolomics studies. Then, it is extended to the **General Linear Models (GLM)** and several enhancements of these methods are proposed. Here, an effort has been made to motivate all the methodological aspects of the applied techniques, for example the effect matrix augmentation, which is a cornerstone for each of them, although never well motivated in the literature. In the first instance, and again due to a lack in the current literature, these alternative methodologies are explained - sometimes reformulated in more simple terms - in a unified fashion, and compared to each other, based on an experimental dataset. This chapter also serves as an introduction to the final one which extends ASCA to deal with **random effects** by replacing GLM with **Linear Mixed Models**. The entire framework had to be adapted consequently and new methodological aspects were developed to enable the comparison of both fixed and random sources of variation: test of the effects significance, effect matrices augmentation in the presence of more than one source of random variation, etc.

The question of **spectral repeatability** and how to measure it has also been investigated in a transversal way through this work: (1) by quantifying the quality of spectral profiles that are pre-processed by different strategies and (2) by comparing the effects of fixed and random sources of variability on those spectra.

Finally, note that this manuscript can be viewed as a **practical handbook** directed towards both experimental researchers and statisticians, which is enriched by **online accessible software and data**, in the scope of making research reproducible.

1.2 The omics family and metabolomics in health sciences

1.2.1 Omics technologies

Global systems biology, including complex mammalian systems, is challenging in the health context (e.g. new infectious agents, antibiotic resistance or even the increased prevalence of cancer). Given the inter-relation between multiple gene processes and environmental and lifestyle factors, high-throughput analytical technologies and techniques have been necessary for the measurement and modelling of the multilevel biomolecular structure and metabolic processes, from the cell to the entire organism (Nicholson *et al.*, 2007).

Therefore, research in life and medical sciences is more and more driven by the study of such omics data. Analysing the human genome, the protein profile or the metabolic dynamism indeed provides a lot of information to help the understanding of disease causes and their behaviour.

Omics approaches are referred as the “collective characterisation and quantification of pools of biological molecules that translate into the structure, function, and dynamics of an organism” (Simó *et al.*, 2014). The development of these numerous and complementary technologies (*cf.* Figure 1.2) in systems biology such as metabolomics, along with genomics¹, transcriptomics² and proteomics³ (see e.g. Manzoni *et al.* (2016) for extended definitions) has a growing importance thanks to a revolution in the molecular biology techniques since the 1990s.

Nevertheless, with the rise of the omics sciences, a countless number of new biomarkers have been discovered but only a hundred are validated and current in routine clinical practice (Poste, 2011). Similarly, a recent review by Parker *et al.* (2018) showed that from the original diagnostic value of a molecular- or “omics”-based test, only a few of these tests had made progress toward clinical application within a decade in the citing literature.

1.2.2 Metabolomics

Metabolomics is a quite young and multidisciplinary scientific toolbox. It is defined as the global and unbiased study of small molecules (<1 kiloDaltons) in a biofluid, tissue, organ or organism (*cf.* Table 1.1, Beyoğlu & Idle, 2013). It provides a snapshot of the metabolic dynamics in an organism, a tissue or cells that reflect

¹Designed to sequence, assemble, and analyse the function and structure of genomes.

²Simultaneous determination of gene expression changes between subjects using micro-array technology.

³Study of protein changes in a cell or tissue

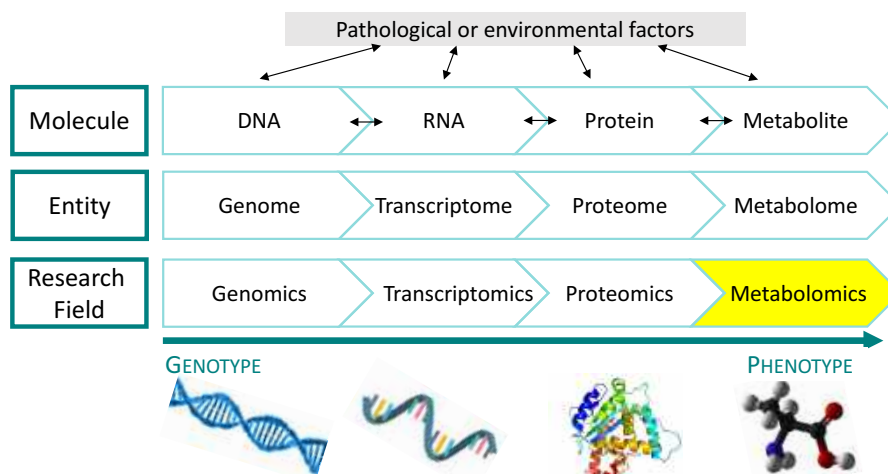


Figure 1.2 – The “omics” technologies. Changes in metabolite concentrations can occur due to genetic alterations, disease, or environmental influences. The figure shows that the metabolome is the most predictive of the phenotype. Double arrows represent alterations with interfering external factors.

the response of living systems to both pathophysiological stimuli and/or genetic modification (Emwas *et al.* , 2013). This field of investigation enables to measure the end products of the cellular metabolism of both endogenous and exogenous substrates.

The original purpose of metabolomics is biomarker discovery, *i.e.* identify metabolites that are correlated with various diseases and environmental exposures (Johnson *et al.* , 2016). It must be stressed that a single identified metabolite is usually not specific of a particular metabolic pathway. Hence, major challenges in this field consists in demonstrating the physiological roles and interactions of such molecules and their involvement in specific metabolic networks, as well as establish how the metabolites level impacts different phenotypic outcomes (Johnson *et al.* , 2016). Note that in this manuscript, the term *biomarker* relates to identified metabolites that could potentially, and after validation, be used as biomarkers of *e.g.* a disease.

Compared to genomics or proteomics, metabolomics give direct signatures of the biochemical reactions occurring, hence it is closely related to the individual phenotype (Patti *et al.* , 2012), which makes it of primary interest to understand induced biological mechanisms in the life sciences. It definitely has other advantages in the data collection such as non invasive and quick data sampling and provides rich and dynamic information. However these data need complex

treatment. Furthermore, the metabolic profile changes continuously and is the product of all the other entities and external factors.

Table 1.1 – Definitions of *metabo*- terms.

Metabolism	Metabolism is the set of life-sustaining chemical reactions and transformations within the cells of living organisms.
Metabolite	A metabolite is a small molecule (< 1 kDa) involved in metabolism either as a product of metabolism or necessary for metabolism. Metabolites represent the end products of gene expression and protein synthesis together with the cells response to environmental stresses (Emwas <i>et al.</i> , 2013).
Metabolomics	Metabolomics is part of the metabolome analysis framework. It is the scientific study of chemical reactions and transformations involving metabolites. More formally, it has been defined as the global and unbiased survey of the complement of small molecules (<1 kDa) in a biofluid, tissue, organ or organism (Beyoğlu & Idle, 2013).

Only semantic distinctions arise between words of metabonomics and metabolomics, as their purpose is mainly to describe the same process: the search for metabolites. This means that all or certain metabolites present in the sample are analysed and it gives an overall analysis of the metabolome of interest. Hence, metabolomics covers extensive and simultaneous profiling techniques of multi-levels metabolites together with their systematic and temporal changes induced by various elements such as diet, lifestyle, environment, gender, age, physical activity, genetic effects or pharmaceutical effects on the entire organism. It provides a means to investigate the whole integrated system responses (Nicholson *et al.* , 2007).

Analytical methodologies are separated into *targeted* and *untargeted* approaches (Patti *et al.* , 2012). Targeted metabolomics refers to a list of a priori known metabolites that need to be detected and quantified in biological samples, usually to study a particular biochemical pathway of interest. On the other hand, the untargeted approach has a broader scope than targeted analyses. It refers to the simultaneous measurement of as many metabolites as possible, without any bias. It does offer a global and simultaneous snapshot of the metabolites that are produced at various levels of our metabolism, and for which their presence is influenced by a variety of internal and external factors. Compared to the targeted strategy, it produces much more complex and high dimensional datasets to handle. Note that in this work, untargeted metabolomics will be investigated.

Compared to the other omics sciences, the study of the whole metabolome provides several advantages (Kell, 2007; Rousseau, 2011), including:

- Its downstream procedure: changes in the metabolome are amplified compared to modifications in the transcriptome and the proteome.
- Metabolic profiling is cheaper and provides a higher throughput in comparison with proteomics and transcriptomics, thus rendering it feasible to examine large numbers of samples under a wide range of conditions.
- Metabolites have a central place in the organisation of living systems and metabolomics is able to give a more exhaustive biological summary of perturbations in biofluids, which is also closest to the phenotype⁴.

Metabolomics is gaining widespread importance for drug development and clinical trials by extending the understanding of the pharmacology and toxicology of new drug candidates. The rationale to investigate for the development of metabolomic biomarkers is mainly due to the fact that biomarker studies predicting if a new drug is safe and effective can be carried out in a shorter lapse of time and at a lower expense than traditional types of clinical measurements (Beyoğlu & Idle, 2013).

The principal fields of application of metabolomics are disease diagnosis along with other clinical applications (e.g. nutritional studies or genetic differences), as well as drug development including preclinical development and patients preselection (Nicholson *et al.* , 2007; Beyoğlu & Idle, 2013; Emwas *et al.* , 2013):

1 Disease diagnosis

Searching for a specific disease biomarker that can be used for human disease diagnosis (e.g. cancer, diabetes, renal diseases, etc.) to detect and diagnose the severity of the disease, track, and monitor disease progression.

2 Phenotypic and physiological effects study

To understand the external and intrinsic causes interacting (e.g. male/female, age, diet, etc.) and responsible for modifications of a biochemical profile leading to the inter-individual variability.

3 Preclinical drug development

The use of metabolomics in the early stages of preclinical drug development can avoid potential subsequent costly failure of the candidate in the late-stage. In

⁴Defined as: "the set of observable characteristics of an individual resulting from the interaction of its genotype with the environment." in <https://www.lexico.com/en/definition/phenotype>.

this area, the metabolome study can be helpful for: the discovery of new enzyme targets, pharmacology⁵, toxicology, etc.

4 Clinical drug development

Metabolomic biomarkers have a recognised promising role to play as surrogate endpoints⁶ in clinical development for an effective evaluation of efficiency and toxicity, especially when allowing to take into account the variability and heterogeneity of patients and disease subtypes.

The metabolome investigation is challenging on many levels, especially from the chemical point of view, for analytical technologies as well as the application of bioinformatics (*cf.* Cambiaghi *et al.* (2016) for a complete review of bioinformatics tools in this field) and chemometrics to extract and maximise the compiled multivariate information in the data to help identify high numbers of metabolites. Its main related challenges are indeed the processing of such large datasets, which requires powerful signal processing and statistical tools (Alonso *et al.* , 2015).

1.3 A typical metabolomics study

1.3.1 Typical workflow

Two major goals in metabolomics are: (1) to identify metabolites that characterise the state of an organism and (2) to measure their dynamics under varying conditions (Cambiaghi *et al.* , 2016). A typical metabolomic study for biomarker identification in a two-class problem using the ¹H NMR analytical platform is introduced here. Various steps of data manipulation including acquisition, pre-processing and analysis are indeed essential to successfully complete a metabolomic study (Lindon & Nicholson, 2008; Rousseau, 2011; Smolinska *et al.* , 2012; Beyoğlu & Idle, 2013; Emwas *et al.* , 2013; Cambiaghi *et al.* , 2016) and can be seen in Figure 1.3.

First, one needs to set up the study, by the description of the goals and the study design to clearly define the biological reaction(s) of interest. If the outcome is qualitative, spectral biomarkers will be defined as spectral regions which discriminates between groups of individuals. Statistical design of experiment

⁵Metabolomics can be used to acquire personalised healthcare and determine the mode of action of many drugs by predicting the therapeutic efficacy and adverse drug response on an individual patient basis.

⁶A biomarker that is used as a substitute for clinical endpoints.

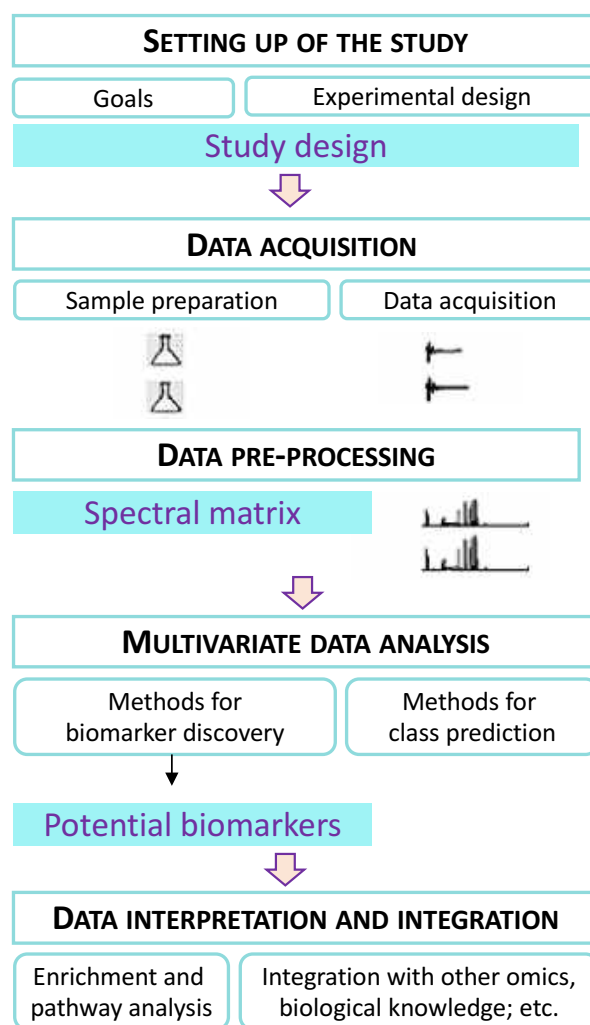


Figure 1.3 – Summary of the different steps in a classic metabolomic study.

enables to maximise the information content while keeping their number low thanks to the set up of a small number of experiments with appropriate factors varying systematically. It leads to a design matrix describing all the available information about experimental conditions linked to each spectrum to be produced.

Second, one needs to gather, store, prepare, acquire and pre-process the data by first collecting biofluids samples (e.g. urine or blood plasma/serum), cell or tissue extracts under the design conditions. The concentration of metabolites in the sample is assumed to be modified according to the presence or absence of the biological reaction(s) under study for each individual. Once the samples are ready, data can be acquired through different separation techniques, and each of them will chemically characterise different parts of the metabolome. Sample preparations and data acquisition are of critical importance for the outcome quality and vary with the analytical platform involved. These analytical technologies are used to obtain the composition of metabolites in the samples, called spectral data. The two most common platforms are Nuclear Magnetic Resonance (NMR) spectroscopy and Mass Spectrometry, which often requires an additional separation step through Gas Chromatography (GC) or Liquid Chromatography (LC) and Ultra-Pressure LC (UPLC).

Though Mass Spectrometry is more sensitive, NMR is a fast, and highly reproducible technique, widely used, and providing detailed information on molecular structure, for pure compounds and in complex mixtures. It can also give information on absolute or relative concentrations of metabolites. This technique exploits the magnetic properties of a definite nuclei: in a magnetic field, NMR active nuclei will absorb energy at a certain frequency, which is characteristic of the isotope and will then radiate this energy. This property can be used to determine from the irradiated energy the structure and concentration of the molecules present in the sample. Whenever the protons within molecules are used, this platform is called ^1H NMR technology, it is the one used in this thesis to acquire spectral data. Note that a typical 1D ^1H NMR spectrum contains 32.768 points and that one molecule can be represented by multiple peaks in the spectrum. See Emwas *et al.* (2019) for a complete review on NMR Spectroscopy for metabolomics data.

Although the researcher's interest lies in the analysis of the final spectra, it is essential for the raw data to be adequately pre-processed as its quality will affect the information recovery during data analysis. This is true for all analytical platforms.

In NMR, the resulting signal is a complex sinusoidal and decaying signal called the Free Induction Decay (FID). It has to be modified by Fourier Transform (FT) into a set of averaged regions called descriptors leading to a multivariate ^1H

NMR set of spectra. Other pre-processing procedures need to be applied to the spectra. They include at least for NMR spectra: baseline correction⁷, peak alignment⁸, binning⁹ and normalisation¹⁰.

Third, one needs to analyse the pre-processed data and search for spectral biomarkers or alterations, thus altered metabolites in link with the biological reaction under study. Although 2D NMR can also be very helpful to identify potential biomarkers in such studies (Guenec & Giraudeau, 2014), the present work will focus on 1D ¹H NMR.

A typical ¹H NMR metabolomic study will generate complex ¹H NMR spectra with a high number of points and where each variable – called *spectral descriptor* – represents one or an addition of several metabolites present in the sample. Since the concentrations of some metabolites are altered in a specific portion of spectra due to the biological reaction, the spectral profiles are varying accordingly. Note that spectral biomarkers¹¹ refer to spectral alterations where molecular biomarkers refer to its molecular structure interpretation.

Multivariate statistical methods (PCA, PLS, non-linear methods, etc.) can be addressed to analyse and interpret complex ¹H NMR spectra, and they are also helpful for biomarker discovery. Unsupervised methods are convenient to explore the overall structure of the data without prior knowledge on the sample origin while supervised methods use such information with the advantage that they provide information about which variables indicate differences between classes or other kind of variation in the response of interest.

Whenever spectral biomarkers are identified, spectral libraries can be used to recognise the corresponding specific compounds and provide an appropriate molecular interpretation.

Last, data need to be interpreted in relation to their bio-chemical context, through enrichment¹² and pathway¹³ analyses. Metabolomic data can also be integrated with other heterogeneous omics data or other biological knowledge.

⁷To remove baseline distortions.

⁸Due to the presence of peaks shift between spectra that can be due to e.g. instrumental factors, changes of temperature or pH, etc.

⁹Divides the spectra into segments in order to reduce the dimension of the problem since the original number of variables can be up to tens of thousands.

¹⁰The purpose is to have comparable samples in accounting for the inter-sample variation due e.g. to different overall concentrations of the samples.

¹¹Simplified to "biomarker" in this thesis since the purpose is to solely detect spectral biomarkers.

¹²Search for groups of functionally related metabolites that are over and/or under-expression in order to identify coordinated changes among them and identify a biological pathway to be investigated.

¹³Description and visualisation of the interactions among metabolites to recognise pathways influencing a biological process of interest.

1.3.2 Where do statisticians help?

Statisticians should indeed be included at all stages of the study, and mostly at the early stages, when the experimental design is being decided and will be decisive for the quality of the produced final results. They can be asked to design the study, but also to compute the sample sizes, or quantify the sources of variability before the study. They could also help in the development of mathematical and statistical pre-processing tools to recover meaningful information from the raw data and evaluate and compare the quality of the pre-processed data before data analysis. Undoubtedly, they should also be involved in the data analysis procedure for biomarkers finding, predictive purposes, extract meaningful information from a complex design, multi-omics integration, etc.

1.4 Multivariate data analysis for untargeted 1D NMR spectra

Classic data analysis in metabolomics (Worley & Powers, 2013; Ren *et al.*, 2015a) has tight connections with the field of chemometrics (van der Greef & Smilde, 2005), and particularly by the use of linear multivariate projection methods such as PCA, PLS, etc. These techniques have the advantage to provide powerful visual representations of the data or model results and to offer simple interpretations for the non-statistician researchers.

1.4.1 Data characteristics

The observed raw FID originally contains tens of thousands of points that includes a meaningful signal, but also uncontrolled variability and instrumental artefacts. Once pre-processed, spectral data tables are composed of n observations row-wise, or spectral profiles, representing metabolites' peaks, characterised by a series of m variables column-wise (the spectral descriptors) that are expressed on a quantitative scale of intensities, as depicted in Figure 1.4. The obtained data table (usually with $m = [250, 750]$) remains short and wide with $m \gg n$. Theoretical independency between those variables is violated due to their dimensionality, besides other biological and inherent spectral data construction causes.

Compared to classic multivariate datasets, the main characteristics of metabolomics data that will impede their analysis are:

- the low number of observations compared to the number of variables, due to expensive data collection and acquisition,

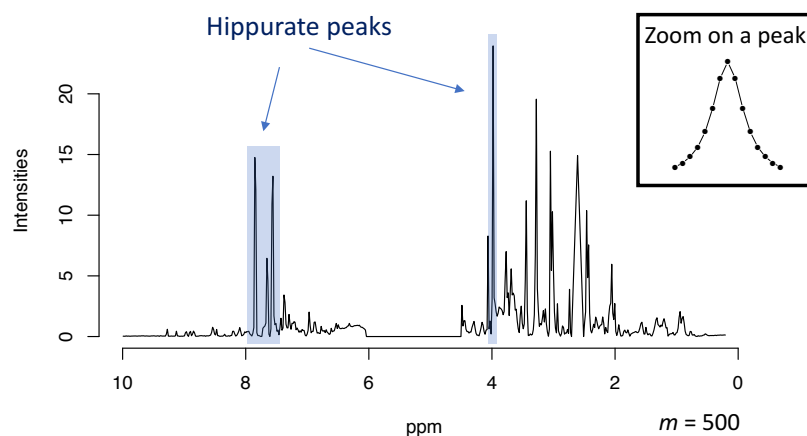


Figure 1.4 – An example of typical spectral metabolomic profile of a urine sample containing 500 variables or spectral descriptors. Here are highlighted all the peaks characterising the hippuric acid, or hippurate, a spiked molecule from one of the datasets used in this thesis. A zoom on a theoretical peak reveals that adjacent descriptors describe a same peak, hence are correlated to each other.

- the intrinsic biological variability,
- the presence of instrumental noise and artefacts (e.g. peak shifts among spectra),
- the need of an elaborated pre-processing strategy,
- the complex correlation structure: one metabolite is represented by different descriptors and peaks. Furthermore, there exist biological correlation between the metabolites.

Therefore these data are short and wide but also scattered and with a low true rank. Furthermore, due to the high number of measured variables and the complex correlation structure, variables are multi-collinear.

Yet, and besides their high information content, they have interesting properties, such as the low rate (or absence) of missing values, the fact that they usually originate from a controlled experimental design and are well structured data.

In general, a metabolomic dataset with a case/control problem also involves a class vector as an indicator of the sample state of interest (e.g. control vs treatment).

1.4.2 The curse of dimensionality in classification

One could think that always adding more variables to the dataset would increase its information content. However, it is actually the other way around at some point as it deteriorates the model performances, of a classifier for example. This counter-intuitive aspect of multivariate data modelling is referred as the *curse of dimensionality* (Spruyt, 2014).

As pointed out by Keogh & Mueen (2017), and typically in the Euclidean space, adding extra dimensions with a fixed amount of observations will indeed lead to an exponential increase in the space volume. These observations will be located far away from the center and distance measures are becoming more and more meaningless. This has multiple implications.

A direct result of the sparseness induced by this curse is the overfitting issue of the classifier that prevents its generalisation to unseen datasets. To preserve model performances, the number of observations should indeed grow proportionally to the increase in the number of variables, which is rarely the case.

Another consequence is that the notion of “nearest” points (*i.e.* proximity or similarity) disappears. All the points are located at an equidistant¹⁴ and increasing distance between each other, at the corners of the variables space. Dissimilarity metrics from lower dimensional rank therefore cannot be transposed in highly dimensional spaces.

Hopefully, in real terms, observations are not uniformly spread throughout the variables space and are rather concentrated in substructures of lower rank. In this case, dimension reduction techniques can be used in practice (Domingos, 2012).

Nevertheless, in this context, some multivariate techniques do generalise better than others, and are therefore recommended to use. Dimension reduction techniques, such as PCA, decrease the dimensionality of the data by combining the variables into a reduced set of new variables. In the case of supervised methods, cross-validation is routinely applied to prevent model overfit. Feature selection algorithms and sparse classifiers are another way of dealing with such wide and short data.

1.4.3 Explorative analysis

Unsupervised statistical methods are meant for a first glance at pre-processed data in an explorative fashion. To the contrary of *supervised* methods that uses additional information from the design such as the class membership, *unsupervised* means that spectral profiles are modelled without any other knowledge about the observations.

¹⁴The ratio of an object's nearest neighbour over its farthest neighbour ≈ 1 in such high dimensional space.

The standard workhorse of chemometrics for explorative purposes is PCA. The leading idea is to derive new variables from the data matrix, sort them according to an information criterion, and eventually obtain the best explanation about the sources of variance included in the data. One advantage of this approach is the fact that no specific assumption is needed except the one that the mean vector and the covariance matrix exist (Härdle & Simar, 2007). Furthermore, its implementation is simple, it provides a unique analytical solution that optimises an explicit criterion and the results are easy to handle and interpret as they represent the data in a smaller dimension (Wehrens, 2011).

Based on the spectral decomposition theorem, PCA projects a data matrix, say $\mathbf{X}$ of size $(n \times m)$, onto a set of new, unobserved, a latent variables, the Principal Components (PC) that are linear transformations of the original variables. Its objective is to summarise $\mathbf{X}$ by a few orthogonal PCs, that are as informative as possible. $\mathbf{X}$ is decomposed as the product of scores $\mathbf{T}$ of size $(n \times a)$ – or observations coordinates in the new space – and loadings $\mathbf{P}$ of size $(m \times a)$ – the weights of the original variables to construct the model – plus $\mathbf{E}$, the residual matrix: $\mathbf{X} = \mathbf{TP}' + \mathbf{E}$ where $a \leq \min(m, n)$. PCs are constructed component-wise such that the first one represents the maximum variation axis, the second PC represents the second maximum variation axis being orthogonal to the first one, etc. An illustrative example of the scree plot and scores and loadings plots resulting from a PCA analysis is given in Figure 1.5. A scree plot indicates the number of PCs needed to capture the major part of information. From the scores plots, indications on the variance structure (clusters, outliers, etc.) of the data can be found, and related, via the loadings, to the main variables responsible for this arrangement.

Practical considerations on the scaling of such data prior to their data analysis should be made. *Scaling* is defined as data transformations that are applied column-wise (by variables). Typically, the covariance (*i.e.* centring the variables) and correlation (*i.e.* centring and scaling up to unit variance the variables) matrices are the metrics used to carry out the decomposition of the dataset. A series of other scaling methods such as Pareto, VAST or range scalings can also be applied depending on the type of data to pre-process and the biological information one wants to investigate (see van den Berg *et al.*, 2006, for a multi-pre-processing comparison). The use of the covariance matrix has a major drawback, which is a high sensitivity to the arbitrary choice of the variables units¹⁵. Notwithstanding, the use of the correlation matrix when all the variables have the same unit, which is the case for ¹H NMR data, not only amounts to use an arbitrary unit of measurement (Jolliffe, 2005) but also can mask the importance of highly discriminant variables and increase the weight of minor

¹⁵The first PC will be dominated by the variables with the largest variance.

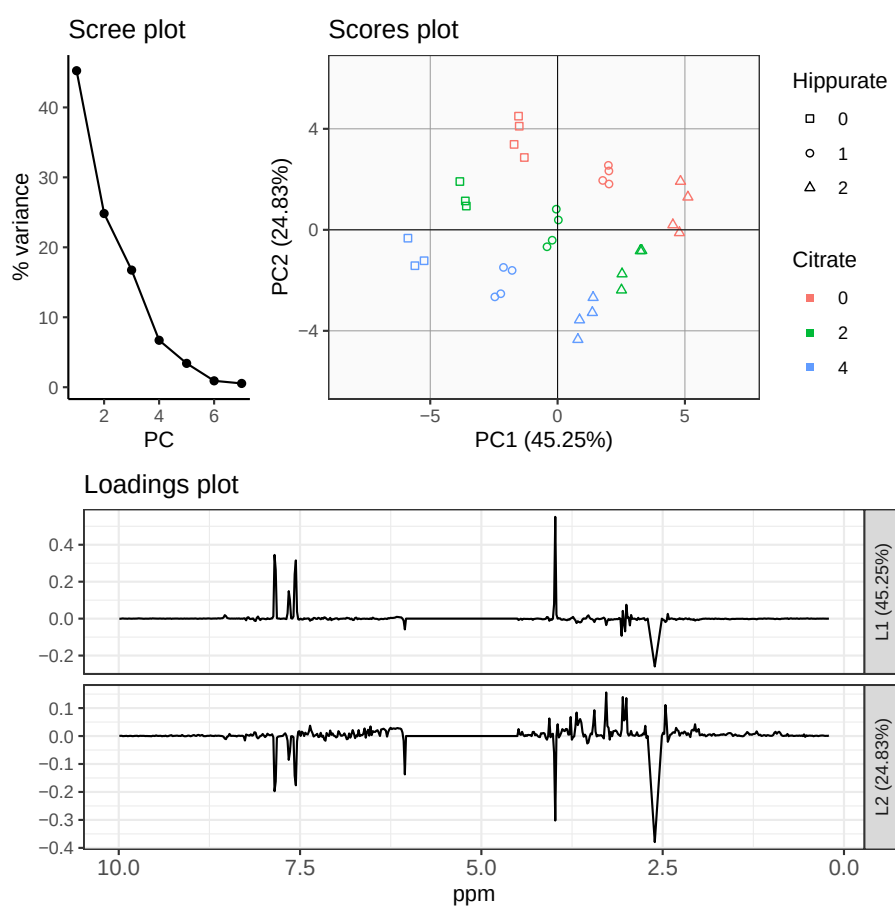


Figure 1.5 – Example of PCA scree, scores and loadings plots. Contrarily to traditional representation, the loadings are here represented as a line plot as a pseudo spectral profile.

variables associated with noise (Khalheim, 1985). Thus the advocated strategy for metabolomic NMR data analysis is to only center the variables so that multivariate projection methods rely on the covariance eigenstructure of the data.

A similar projection method that also decomposes the data into newly estimated components, is Independent Component Analysis (ICA) (Hyvärinen & Oja, 2000). This technique uses an alternative criterion, the maximum entropy, to build new components or sources that are as independent as possible by maximising the deviation from multivariate normality. ICA is better suited for non-gaussian variables, a frequent characteristic of biological sources of variance (Rousseau, 2011). Furthermore, from the point of view of metabolomics, biofluids (e.g. serum or urine) can be seen as a mixture of a certain amount of unrelated metabolites and ^1H NMR spectra, a weighted sum of individual spectrum of each metabolite that is present. The ICA method should therefore be able to retrieve multiple sources expressing the intensity of independent metabolites.

Another common exploratory approach is the clustering (Wehrens, 2011; Ren *et al.*, 2015a). The two most common algorithms that are in general applied are either hierarchical or k -means clustering. Clustering is directly aimed at recovering the “natural classes”, usually of observations, in the dataset at hand. This is achieved on the basis of (dis)similarity measures between the samples. Hierarchical clustering will organise the observations in a nested tree-like structure, or dendrogram. At first, every sample represents a different cluster ($\# \text{ clusters} = n$), and the two closest clusters are gathered together, based on a distance (Euclidean, Manhattan, etc.) matrix between samples and a *linkage*¹⁶ method (e.g. Ward). At this stage $n - 1$ clusters remain. The two next clusters to be aggregated are the next two that are the closest to each other, and so forth, until all the observations are in a same unique cluster. The choice of the final partition is dictated by the resulting dendrogram (distant clusters should not be aggregated together) or additional biological information on the observations. k -means clustering needs prior information on the number of clusters that have to be created out of the data. Then, an iterative procedure will lead to a single partition of the data into k non-overlapping clusters, and each observation belongs to the nearest identified centroid. Different quantitative quality indices can be derived from the recovered cluster partitioning, either to measure the homogeneity of the constituted groups or to compare the obtained partition to a known class membership.

¹⁶The way to measure the distance between two different groups.

1.4.4 Supervised analysis with one categorical variable

In supervised approaches, additional knowledge is introduced in the modelling process, generally a vector containing the class membership of the individual spectra.

Two typical applications guiding the supervised data analysis are:

1. Biomarkers discovery: find combinations of molecules (biomarkers) that differentiate the two groups and understand corresponding chemical pathways.
2. Diagnostic kit development: develop predictive models to assign a group membership to a subject on the basis of the discovered biomarkers.

PCA's principle is based on the largest axes of variation in the data. In the context of biomarker discovery though, the major part of the variation is usually due to uninteresting causes (Buydens, 2013a), highlighting the limits of such explorative tools. By orienting the model construction directly towards the class separation, it should lead to clearer results in that direction. Every supervised method needs Meta-Parameters (MP) to be adjusted to adequately represent the data, without an overfit of the model to keep good generalisation properties.

Partial Least Squares regression (PLS) is arguably the most widely used multivariate supervised method in metabolomics. It corresponds to the regression extension of unsupervised PCA. Its mission is to well represent $\mathbf{X}$ and $\mathbf{y}$, but also their relationship. To that extend, its objective function is oriented towards the maximisation of their covariance, *i.e.* both their variances and correlation. When used in a discrimination context, and known as PLS-DA, the class membership is coded into a $\mathbf{Y}$ -matrix of dummy variables that describes the class of each observation in the dataset. A series of derived algorithms do exist, such as Orthogonal-PLS (Trygg & Wold, 2002a), Sparse-PLS (Chun & Keleş, 2010), Kernel-PLS (Rosipal & Trejo, 2001), etc. However, these tools also have some of the following limitations: they are adapted to simple designs (with one binary or multinomial class label), do not have direct feature selection techniques nor statistical testing framework, and their linearity is not always adapted for biological data. Other, possibly non-linear, methods, mostly derived from the machine learning literature such as RF, SVM or NN could also be applied. Though they are mainly developed for predictive purposes and do not provide advanced visual representations of the results, some of them enable to quantify variable importance for biomarker discovery.

A major challenge when dealing with supervised classifiers is to avoid an overfit of the model to the data by selecting appropriate MPs and validating its performances based on an appropriate criterion (Wehrens, 2011). In practice, this validation is

achieved most of the time by k -fold Cross-Validation (CV): within a loop of size k , the dataset, called the *training set*, is split at random in sets of equally sized k subsamples. At each iteration, one different subsample is removed from the *training set*. The model is trained based on this reduced *training set* with observations from $(k - 1)$ folds and used to predict the subsample left aside. A validation criterion is calculated across the CV loops and the trained model with the best performances is selected out of the candidates with corresponding MPs. If the sample size is large enough, one can afford to leave aside a *test set* that will evaluate the final trained model performances based on a set of unseen data. Figure 1.6 illustrates the dangers of an overfit in the results interpretation of a PLS-DA model. Here, a matrix of size (40×400) was randomly generated from a $N(0, 1)$ distribution and a class vector was set arbitrarily to 1 for half of the observations and 0 to the other half. Although the scores plot depicts a perfect separation of the samples and the quality of the model based on the fitted model (R^2) is excellent, the data carries no meaningful information. The cross-validated R^2 (Q^2) genuinely reveals the overfit with negative values.

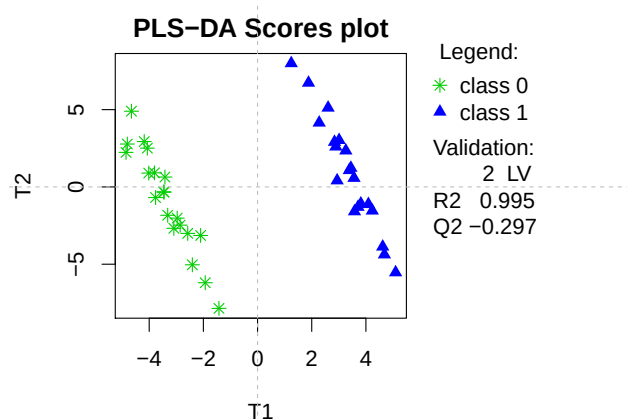


Figure 1.6 – A word of warning on the overfit of PLS-DA.

1.4.5 Supervised analysis for more advanced designs

Generalities on ASCA

Metabolomics data are intrinsically correlated and generally produced by another multivariate set of variables, collected in what is defined as the *design matrix*. Design factors usually involve the presence of a treatment, but other sources of biological or technical variability in the data are often measured, leading to more advanced designs such as longitudinal, cross-over, variance components etc., where multiple possible factors can intervene (treatment, time, gender, age, etc.). If one wanted to explain more advanced multifactorial designs than just a class membership, one could rely on classic statistical tools by considering spectral descriptors as the dependent variables or response, and the design matrix as the independent variables or predictors, and apply Multivariate Analysis Of VAriance (MANOVA) or Multivariate-General Linear Model (M-GLM, Johnson *et al.* , 2002) on these data. Figure 1.7 illustrates this shift in paradigm from classic chemometrics methods to statistical multivariate regression.

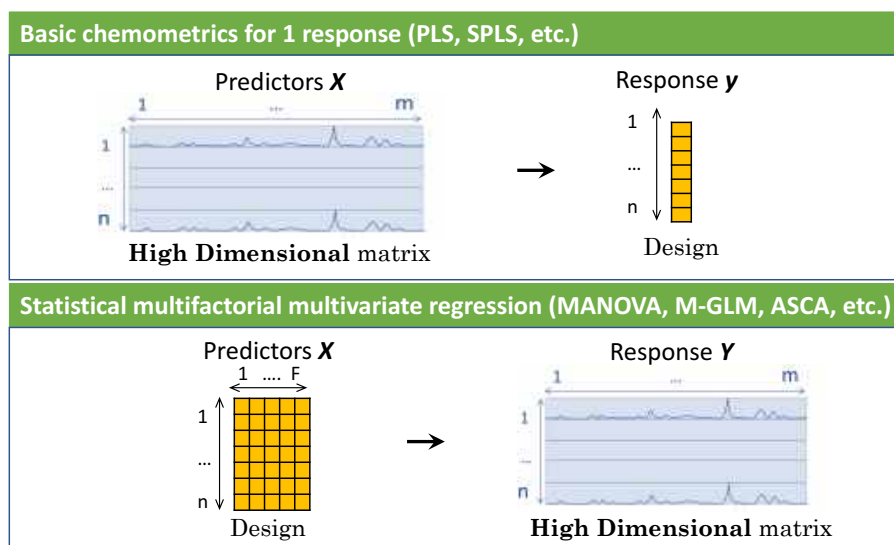


Figure 1.7 – What response and predictive variables means in statistical multifactorial multivariate regression versus in a usual chemometrics approaches with a binary class response.

Nevertheless, methods from the traditional regression toolbox still cannot be directly applied to such mega-variate and collinear data tables. The ASCA framework (Smilde *et al.* , 2005; d. B. Harrington *et al.* , 2005), which is based

on two well known methods from chemometrics – PCA – and statistics – ANOVA –, leads to promising results to solve this issue in the omics sciences. By combining and taking advantage of the dimension reduction projection methods and classic statistical modelling, it enables to decipher the main sources of variability in the generated response matrix and offers attractive graphical representations of the factors' effect. Note that, contrarily to PLS, ASCA is not meant for predictive but rather for explanatory purposes, *i.e.* explain the main source of variation in the data.

A conceptual graphical representation of ASCA for a 2 factor crossed ANOVA model is given in Figure 1.8: The idea is first to decompose, via an ANOVA applied to each column of the spectral matrix, the total variation in the spectral matrix into *effect matrices* and according to the different sources from the design (modelling step). Then, to study the interesting parts of the model by applying PCA on those matrices, either pure or *augmented* (see the following sub-section motivating this augmentation).

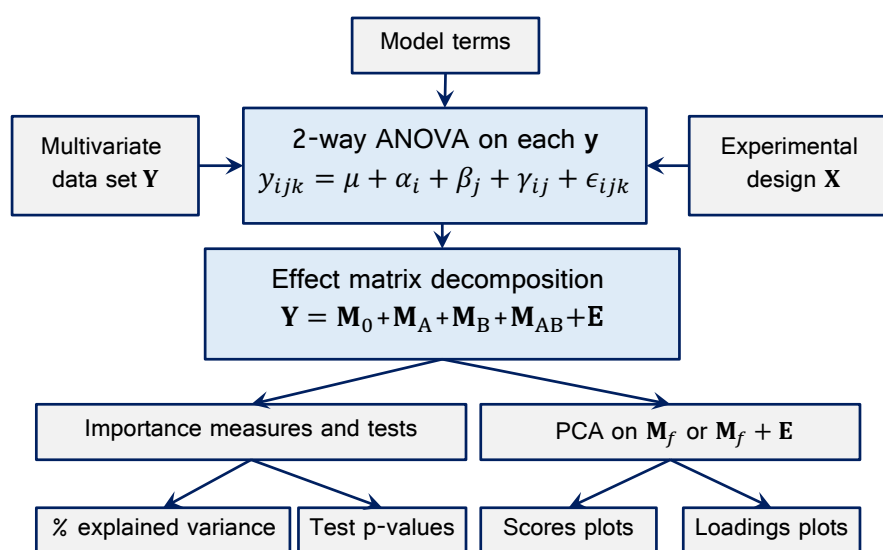


Figure 1.8 – Overview of ASCA for a 2 factor crossed ANOVA model.

Several variants have since been published, and in particular: ASCA⁺ for unbalanced designs where ANOVA is replaced by GLM (Thiel *et al.*, 2017), AComDim, the combination of ANOVA and Common Dimensions (Jouan-Rimbaud Bouveresse *et al.*, 2011), PARAFASCA, the PARallel Factor - Simultaneous Component Analysis (Jansen *et al.*, 2008) or ANOVA Multi-block

Orthogonal Partial Least Squares (AMOPLS, Boccad & Rudaz, 2016). Note that these methods will be further investigated in Chapters 4 and 5.

Incentive for the effect matrix augmentation

A core aspect of the ASCA framework is the effect matrix augmentation, *i.e.* the addition of the matrix of residuals to an effect matrix before performing PCA and deriving the scores from this augmented matrix. This augmentation is used for a visual representation of an effect of interest –which is particularly interesting in chemometrics to summarise the informative content of a multivariate dataset – but also to perform permutation tests on those scores (Thiel, 2015) to assess the effects significance.

One advantage is that, through this augmentation, non significant effects melt away in the noise and one can extract only the important effects from the model. Another is that we can assess, for one specific effect, how the different levels are clustered and differentiated from each other in the reduced PC space that includes not just the coefficients' mean values but also the natural variability around them, which improves the interpretation of the impact of this effect on the response matrix.

Although not justified on statistical grounds in the original papers describing ASCA (Smilde *et al.*, 2005) and APCA (d. B. Harrington *et al.*, 2005), this augmentation is intuitively justified on the grounds of the ANOVA F-tests. Zwanenburg *et al.* (2011) explains the principle to add the residuals matrix in ASCA in order to compare the factor levels with the natural variation. d. B. Harrington *et al.* (2005) also justifies this augmentation the following way: *"In ANOVA, the variance of each factor is compared to the residual error so that significance may be tested. In the ANOVA-PCA approach, the residual error is combined with each factor and subjected to PCA. Examination of the observation score plots from ANOVA-PCA provides a clear picture of significance as well as the relationship among the levels of each factor"*. In their article, they even suggest to visually assess the statistical significance by plotting a 95% confidence intervals around the means for levels of each factor, *i. e.* data ellipsoids designed to capture the observed variation. Liland *et al.* (2018) suggest in turn to use model ellipsoids in the scores plot in order to integrate the uncertainty in the modelling.

Marini *et al.* (2015) also inspected, in the context of ANOVA-TP, which residuals should be used for an effect of interest, either to visually inspect in its scores plot both the between- and the within-class variances, or to assess its significance via a permutation test, similarly to Anderson & ter Braak (2003). They argue that the choice of the residuals depends on the structure of the experimental design but also on the nature of the factors (fixed or random). They rely on the ANOVA

table of the model of interest with the Expected Mean Squares (E(MS)) and the corresponding F-tests. In this table, they inspect which term is contributing to the numerator and denominator of the E(MS) to decide which eligible random term should be used to test an effect of interest. They further suggest to use Hasse diagrams to easily identify those appropriate residual terms, rather than having to derive the ANOVA table for more advanced designs.

In this thesis, we tried to justify on the same grounds than in Marini *et al.* (2015) this effect matrix augmentation by making parallels with the ANOVA theory in the univariate case (see in particular Sections 4.2.4, 5.2 and 5.4.6). To a certain extend, we tried to generalise this univariate reasoning to a multivariate setting, especially in the case of a linear mixed model *i.e.* when multiple sources of random variation are present. In addition, we suggest to use a correction factor based on the effective model dimension (Eilers, 2018) to correct the effect matrix augmentation with the appropriate degrees of freedom.

This augmentation has advantages but also shortcomings, depending on the context in which it takes place. For example, in AComDim (see Section 4.2.6), it will be seen that the residuals matrix is added as many times as the number of terms included the model. A non negligible negative aspect is that the percentage of variance explained by the residuals is clearly overestimated and the scores and loadings linked to the effects may be slightly impacted. However, it also helps to the model stability and interpretation: this augmentation implies that the first common dimension corresponds by construction to the residuals and non significant model terms are hidden into the noise such that only the important ones are automatically extracted. A more general limitation of this augmentation is that the ANOVA theory underlying this augmentation is based on the hypothesis of a balanced design.

1.5 The question of spectral repeatability

Quantifying the spectral quality is far from trivial because these data are so complex that it is not possible to directly and visually assess spectral quality. Furthermore, one could be interested in the quantification of the different sources or variability from a dataset as well.

Although mentioned in the work of Rousseau (2011) and Féraud *et al.* (2015), the statistical perspective of spectral reproducibility/repeatability and quality control is not yet well studied in metabolomics. Other suggested quality assessment measures are mostly relying on the computation of descriptive statistics (means, standard coefficients, etc.) or on visual inspection, mostly of PCA scores plots, and pairwise comparisons (Teahan *et al.* , 2006; Dumas *et al.* , 2006; Parsons *et al.* , 2009).

Of course, a variety of methodological precautions are to be taken to ensure qualitative experimental samples. Practitioners are indeed aware that the spectral quality will determine the reliability of the structural information that is extracted in the subsequent data analysis steps (Schorn *et al.*, 2004). However, a multitude of different spectra can be generated on the basis of a single sample by verifying the: acquisition technique (1D vs 2D), NMR spectrometer, pre-processing strategy¹⁷ or sample preparation.

In order to evaluate quality of a set of spectra, questions of interest could be:

- How to evaluate and compare the quality of two spectral matrices coming from the same biological samples?
- Besides the treatment of interest, what are the other main sources of variability in the data and how to quantify them?
- Are the data of good quality for the results to be generalised?
- How to quantify the spectral repeatability/ reproducibility?

For this last point indeed, a case-study of interest is the assessment of spectral repeatability/reproducibility. Let define those terms according to Table 1.2.

Table 1.2 – Definition of spectral repeatability - reproducibility - replicability. What is measured when the data are generated from the same design and come from:

	Same samples	Different samples
Same experimental conditions	repeatability	reproducibility
Different experimental conditions	reproducibility	replicability

From this table and for the sake of clarity, we can therefore define more precisely the repeatability and reproducibility of the data. The *repeatability* is the variability caused by repeated measurements on the same sample, or within a series of identical subsamples, and under identical conditions. Whereas, the *reproducibility* is the variability caused by the modification of an analytical condition or different biological samples coming from a same population in the study.

In a broader scope of spectral quality assessment, Rousseau (2011) establishes in her work three questions of interest, namely: *the comparison of the variability of groups of spectra*, *the identification and quantification of the spectral sources of variability*, and *the comparison of groups of spectra with respect to their “signal” to*

¹⁷Since there is no absolute rule of which pre-processing step should be applied, much less the optimality of their methods' parameters.

noise ratio. To answer to those questions, different tools are proposed and applied: PCA with groups identification and inertia computation, as well as descriptive statistics and point-wise mixed modelling by deriving the variance components, the interclass correlation coefficient and a Signal to Noise Ratio (SNR) index.

Another study in Féraud *et al.* (2015) compares the quality of QC (Quality Control) samples with a known class label that were generated by 1D ^1H NMR and 2D COSY experiments. To achieve this comparison, they suggest the use of the Metabolomic Informative Content (MIC). The idea behind this criterion is to compare, on the basis of a single experimental sample, different conditions of data production. The underlying hypothesis is that the produced spectral profiles are a sum of repeatable signal and independent and identically distributed (i.i.d.) random variation or noise. Different unsupervised techniques (PCA and clustering indices) will blindly evaluate the information recovery of the signal from the noisy spectral data. Of course, besides the sole data generation variability, other (biological or technical) sources of variability can intervene and interfere with the source of variability of primary interest (e.g. case-control label).

Finally, the previously described ASCA framework offers in this context an alternative solution to measure the percentage of variation linked to each potential effect of the design and test its effect significance. The interest of this approach is the strong statistical grounds on which it is built that enables to quantify and test for the presence of these potential sources of variability.

In this thesis are explored these two last methodologies in an attempt to quantify spectral quality: the MIC criteria to evaluate different pre-processing strategies and the use of ASCA and linear mixed models to assess and quantify the sources or spectral variability e.g. in a repeatability/reproducibility study.

1.6 Datasets used in this thesis

The methodological developments of this thesis have been applied to several real life experimental datasets in order to grasp all the potential difficulties that could be faced during their treatment. These datasets are described in a nutshell within each chapter. A complete description is then available for each of them in the form of an index card, in the global Appendix C and they are all available at <https://github.com/ManonMartin/thesisMaterial>.

The databases and their related chapters are:

- The Human Serum database (Chapter 2)
- The Human Urine database (Chapter 2)
- The Semi-artificial Urine database (Chapter 3)

- The Citrate-Hippurate Urine database (Chapter 4)
- The Repeatability-Reproducibility database (Chapter 5)
- The Candies database (Chapter 5)

1.7 Content and contributions

With this background in mind, the objective of this thesis is to suggest methods and statistical tools to support the existing methodology of metabolomics data treatment. More specifically, this thesis intervenes at two major steps of the metabolomic study that are prone to new methodological developments to maximise the generated informative content from the data: **a successful pre-processing strategy** for the raw data, as well as applying **statistical methods that will reduce the final dataset dimensionality, taking into account its characteristics**. Also, it should be emphasised that the data analysis in this work has a broader scope to deal with any short and wide dataset, potentially generated by a multifactorial design, and where collinear variation is present among the variables. A strong commitment in this thesis has been brought to the **valorisation** of the statistical and computational tools that were developed, and their related datasets. All the computations were performed in R (version 3.6.0). As this scientific project is part of a wish to make these tools operational, **R packages** and **projects** are made available for each research axis. It is also linked to a will of engagement to bend towards **reproducible research** where computational aspects can be replicated and used complementarily through a series of different implemented solutions to generate a flux of analysis as efficient as possible for a given dataset at hand.

Metabolomic ^1H NMR data pre-processing

The first research axis of this thesis has the objective of delivering a broad pre-processing strategy dedicated to the production of interpretable NMR spectra from FID sources that can be further passed to data analysis. While the challenges related to data analysis are transversal, metabolomics, as the other omics, remains mostly a technology-driven science with specific data acquisition technologies. Understanding these acquisition aspects is necessary to improve the information recovery from the final spectra. The first contribution consists in the writing, implementation and publication of an R package called **PepsNMR or Packaged Extensive Pre-processing Strategy for NMR data**. Originally existing as a Matlab package called BUBBLE (Vanwinsberghe, 2005), this R package attempts to cover all aspects of NMR data pre-processing with classic

and non or semi-parametric methods. Through all its development trajectory, the package has been applied to various datasets and by various users with visual inspection of the results, to validate (*i.e.* test and improve its functionalities, find bugs and fix them, meet new use-cases and new needs) and maintain the package. From its original version in BUBBLE, improvements, new methods and steps have been added to PepsNMR, such as for bucketing, phase correction, normalisation, zero-filling etc. Spectral quality control has also been studied based on the MIC concept from Féraud *et al.* (2015), in order to compare with quantitative quality criteria (*vs.* visual inspection) different pre-processing strategies, and more specifically to compare between an “ordinary” pre-processing and PepsNMR pre-processing. Thanks to its modularity, the toolbox can be combined to other existing, sometimes better performing, tools from the literature, but also be integrated to existing processing workflows. To that regards, some steps of this package are implemented in Workflow4Metabolomics (<https://galaxy.workflow4metabolomics.org>; Guitton *et al.* , 2017), which is based on the Galaxy framework (<https://usegalaxy.org/>). The release version of PepsNMR is maintained on Bioconductor (Huber *et al.* , 2015). Another dedicated package stores the heavy (raw) datasets and is called PepsNMRData.

Chapter 2 reviews in details all the pre-processing steps available in PepsNMR, with a recommended application sequence order that is suggested based on previous experience or necessary states of the data being pre-processed. It further applies the whole strategy to two experimental datasets, the Human Serum and the Human Urine datasets. In the second part of this chapter, the fully pre-processed datasets are compared to datasets that were pre-processed in a classic fashion. To be able to compare both strategies, several indices derived from the MIC concept are compared to each other and demonstrate the power of PepsNMR.

End products:

Publication This chapter has been published as a journal article:

R packages The development version of PepsNMR is available on Github: <https://github.com/ManonMartin/PepsNMR>. The release version is published on Bioconductor: <https://bioconductor.org/packages/PepsNMRData/>. The package has mandatory *help* pages for functions and datasets, as well as a *vignette*, a long format case study. The *ad hoc* package, PepsNMRData, is also available on Github and Bioconductor (<https://bioconductor.org/packages/PepsNMR/>).

R codes from this chapter The R codes that were used for this chapter are available at <https://github.com/ManonMartin/thesisMaterial>.

W4M The nine first steps of PepsNMR are implemented in Workflow4Metabolomics (Guitton *et al.*, 2017) for NMR pre-processing.

Comparison of PLS-derived methods for biomarker discovery

Biomarker discovery in metabolomics is one of the main purposes of data analysis. In this field, classic linear multivariate projection methods such as PLS and OPLS from the chemometrics literature are usually applied to the spectral data to decipher which spectral descriptors, and consequently which metabolites, are varying across the classes of individuals. A first limitation occurs when using such methods as they seem less appropriate than sparse classifiers in the high dimensionality context (Chun & Keleş, 2010). Their sparse equivalent should therefore lead to a more appropriate feature selection solution. Furthermore, the effectiveness of their use and a measurement of their performances as feature selection methods is sorely lacking in the existing literature: when applied to experimental datasets, true biomarkers are inherently unknown, and the researcher has no clue that what he/she selected as potential interesting features are actual true biomarkers of a biological change. In contrast, this contribution evaluates and compares, on the basis of known discriminant features, different widespread PLS techniques used for feature selection, in an original resampling procedure and with various evaluation criteria.

Chapter 3 first reviews t-statistic, PLS, OPLS and their sparse counterparts (SPLS and LSOPLS) in a classification context and further applies them to the Semi-artificial Urine dataset where half of a set of control samples are artificially altered to mimic biological variation induced by an external stimuli. Half of these alterations are in turn located in a noise-less spectral area whereas the other half is located in a noisy spectral region. Hence, this dataset has the advantage that exact location of biomarkers is known in advance. The statistical principles of these methods and their applied algorithms are described in a unified form, as well as the related feature selection particularities. Keys to interpret their main results are also provided. Then, they are illustrated on the basis of a sub-sample from the Semi-artificial Urine dataset. The third part of this chapter evaluates the feature selection performances on the grounds of a double validation loop with resampling and CV. Different sample sizes and spectral regions where the true features are located enable to study feature recovery through various possible conditions.

From the models, several criteria are measured and averaged to compare the methods such as: feature recovery accuracy based on ROC curves, features and MPs stability, etc. The pros and the cons related to each methodology are

discussed. Results showed that none of these methods was always better than the others regarding the features recovery and stability, but PLS-vip with a selection threshold fixed at 1 had the overall best performance for true feature recovery.

End products:

Publication This chapter has been published as a Discussion Paper at ISBA, UCLouvain:

The implemented (LS)OPLS methods were also employed in Féraud *et al.* (2017).

R package The R package MBXUCL (<https://github.com/ManonMartin/MBXUCL>) also contains all the applied methods, in addition to other common multivariate methods (PCA, clustering).

R codes from this chapter The R codes that were used for this chapter are available at <https://github.com/ManonMartin/thesisMaterial>.

Analysis of advanced designs

Data tables from experimental designs with multiple response variables and a large number of responses with respect to the number of observations are becoming more common, particularly in omics and chemometrics fields. Traditional multivariate modelling approaches are not suitable for this type of data and more complex approaches like ASCA and ANOVA Principal Component Analysis (APCA) have been developed to analyse such data. They combine modelling and matrix decomposition based on an ANOVA model with a multivariate step conducted by Principal Component Analysis (PCA). With regard to the analysis of such advanced designs, contributions are proposed in this thesis by modifying the original ASCA⁺ methodology at its two main stages. First, by extending the dimension reduction step to other global dimension reduction techniques, all already applied to metabolomics data. This adaptation is motivated by the fact that other methods than PCA could provide supplementary information or other visualisation tools. Second, by modifying the modelling step and extending GLM to Linear Mixed Models (LMM) for more advanced case-studies applications. The incentive for this second modification is to allow more flexible modelling to analyse complex designs requiring the use of random effects, and identify and quantify reliably the fixed or random sources of variability.

ASCA⁺ compared to other global methods

Chapter 4 is a joint work with S. Guisset (Guisset *et al.* , 2019). In this chapter are compared ASCA and APCA to three other and less known techniques combining ANOVA and a multivariate step: PARallel Factor - Simultaneous Component Analysis (PARAFASCA), ANOVA Common Dimensions (AComDim), and ANOVA Multi-block Orthogonal Partial Least Squares (AMOPLS). The main advantages of these three methods are their ability to explore in a single and global procedure all (or a part) of the model effects, extract automatically the most important ones, interpret them visually and quantify their respective importance and significance. In this chapter, these approaches are presented in a common framework to make their comparison easier. Some enhancements are introduced and they are also extended to unbalanced experimental designs thanks to a GLM version of the matrix decomposition. Then, the three methods are applied to the Citrate-Hippurate Urine dataset with a three-factor unbalanced experimental design. Quantitative and visual results are compared for the three techniques and show that they are all suitable for the analysis of a complex model, but differ in the type of outputs and the ease of interpretation of the results.

End products:

Publication This chapter has been published as a journal article:

R package This work also contributed to the creation of an R package for ASCA⁺-related methodology and graphical outputs that is currently under development.

R codes from this chapter The R codes that were used for this chapter are available at <https://github.com/ManonMartin/thesisMaterial>.

LiMM-PCA

Although already extended to unbalanced datasets, ASCA⁺ is still constrained by the exclusive use of fixed effects in the model. It has indeed not yet been extended to more advanced designs involving random factors, being typically involved in longitudinal, hierarchical or repeatability-reproducibility studies. Chapter 5 extends ASCA⁺ by replacing the General Linear Model (GLM) by the Linear Mixed Model (LMM) and innovative adaptations to all the steps in the methodology are suggested accordingly in this chapter. Taking into account the error structure of the data indeed leads to more accurate data modelling and

more generalisable results. The suggested methodology, called LiMM-PCA, is applied to two experimental case studies (the Candies and the Repeatability-Reproducibility databases) that highlight the benefits of this approach as it leads to a refined data analysis with interesting inferential properties, while preserving the powerful visualisation outputs produced by ASCA.

End products:

Publication This chapter has been published as a Discussion Paper:

It has also been submitted for publication as a journal article in *Journal of Chemometrics*.

R codes from this chapter The R codes that were used for this chapter are available at <https://github.com/ManonMartin/thesisMaterial>.

Conclusions

Eventually, Chapter 6 is closing this work with general conclusions and potential leads for future research in this applied life science domain.

1.8 Notes to the readers

1.8.1 On mathematical notation

Within and between these parts of the thesis, an effort has been made to uniformise the notations across the numerous methods explained, while, at the same time, trying to remain faithful to common writing and notations in the scientific literature related to each of them. Thus, mathematical notations may differ across the chapters and an index table summarises the main specific notations per chapter in the global Appendix B. Another consequence is that estimators are for some chapters lacking their “hat” notation.

1.8.2 On the bibliographic references

Note that bibliographic references are given in once at the end of this thesis.

CHAPTER 2

PepsNMR for ^1H NMR metabolomic data pre-processing

2.1 Introduction

2.1.1 Literature review

The most common analytical techniques used to generate metabolomic data are Gas Chromatography-Mass Spectrometry (GC-MS), Liquid Chromatography-MS (LC-MS) and proton Nuclear Magnetic Resonance (^1H NMR) Spectroscopy. They work complementarily, each having its own purpose and limitations. The physical basis of NMR is to use the spin properties of specific nuclei (typically ^1H or ^{13}C) that can absorb distinct energy levels when placed in an external magnetic field. A 90° radio-frequency pulse then rotates the magnetisation in the magnetic field. Following this pulse excitation, the magnetisation returns to the equilibrium position during the relaxation phase, and the induced decaying signal, called the Free Induction Decay (FID), is recorded with high resolution (Ren *et al.* , 2015b). Despite its lower sensitivity at low magnetic fields, the use of ^1H -NMR to analyse biofluids do have silver linings: it requires low cost per sample and minimal preparation. It is generally untargeted, thus unbiased, non-destructive, non-invasive, reproducible and fast (Kosmides *et al.* , 2013; Smolinska *et al.* , 2012; Gebregiworgis & Powers, 2012).

Analytical and experimental variation that is unrelated to the biological perturbation under study deserve particular attention in metabolomic studies. This variation originates from sample collection, preparation and conservation, as

well as data acquisition, and intra-individual (e.g. nutrients) or inter-individual (e.g. genetic) variations (Maher *et al.*, 2007), especially in human studies (Lenz & Wilson, 2007). When incomplete or inappropriate pre-processing methods are applied, this variation can lead to inconclusive results or the discovery of false biomarkers during data analysis. Pre-processing, here defined as in the field of chemometrics as all data editing done before data analysis (Liland, 2011), is used not only to enhance the Signal-to-Noise Ratio (SNR) but also to correct for or minimise instrumental artefacts and appropriately transform the data into interpretable spectral profiles that better represent the biological system under study.

Good data pre-processing is important and an additional very important challenge is the availability of effective software for data preparation/pre-processing. Such software can be categorised as freeware, commercial or in-house. Freeware need their code and methodology to be “peer-reviewed” nevertheless the last two possibilities also have their drawbacks: commercial software can be expensive while designing in-house solutions need a lot a time to be built and are short-lived. Another key point is the will to reduce the costly time investment linked both to bioinformatics tools learning and coding for new practitioners and the necessity to switch between various software to obtain a complete data processing workflow (Worley & Powers, 2014).

Commercial software such as TopspinTM, Chenomx or MestReNova remain the most used solutions for NMR data pre-processing but they do not provide advanced editing steps, lack modularity and require iterative, manual and arbitrary adjustments. These interventions will thus not necessarily improve the data quality and can impact the performance of the multivariate data analysis conducted afterwards. On the other hand, free software that are mostly open-source exist for ¹H NMR data processing (see Alonso *et al.* (2015) for a review of the more important ones). However, their pre-processing tools remain basic and do not offer a global strategy dedicated to the chemometric approach¹ for further data analysis (Engel *et al.*, 2013). Besides, the multiplicity of pre-processing scenarios and methods (Buydens, 2013b), which may arbitrarily affect the final conclusions, would benefit from clear recommendations guiding the user’s choice of when and how methods should be applied.

2.1.2 Objectives

Given the importance of the pre-processing step and in contrast to existing alternatives, we introduce the R package PepsNMR (or *Packaged Extensive Pre-processing Strategy for NMR data*), for the (semi-automatic) pre-processing

¹Defined as the approach where the compounds of interest are identified after the statistical analysis of spectral intensities that led to the identification of significant spectral features.

of 1D ^1H NMR FID metabolomic datasets. The objective of this software is to provide an exhaustive and flexible workflow to deal with typical features of raw ^1H NMR data and cover the pre-processing (the cleaning of the raw instrumental data) and pre-treatment (the transformation of the cleaned data for data processing) steps as defined by Goodacre *et al.* (2007). After several steps, the proposed pre-processing strategy transforms FID data into a binned spectral table that will enable data analysis using classic chemometric tools (Ren *et al.*, 2015b) to identify global trends in spectral profiles.

Innovative methods are combined with well known algorithms in order to correct for instrumental artefacts and irrelevant biological variability, to increase the SNR, appropriately transform the data domain/scale and reduce its dimensionality. Besides these innovative aspects, one major benefit of this package lies in its modularity: it can be used in combination to other existing tools, and new functionalities/methods can be added in an effective and easy way.

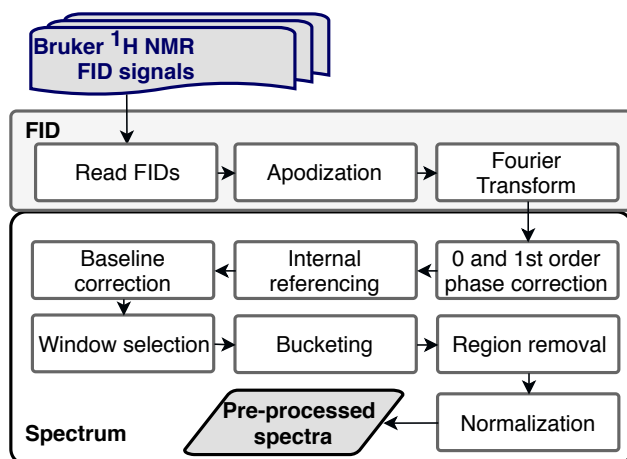
2.1.3 Content

Section 2.2 presents the pre-processing methodology: its motivation, the description of each step, and the package features. Section 2.3 briefly describes the quantitative criteria used to assess the quality of the pre-processed data. Then Section 4.4 illustrates and discusses the use of PepsNMR on the basis of two case studies. Thereafter, the conclusion summarises and brings forward key points of the methodology (Section 2.5) and finally, Sections 2.6 and 2.7 describe software implementation and the datasets used in Section 4.4.

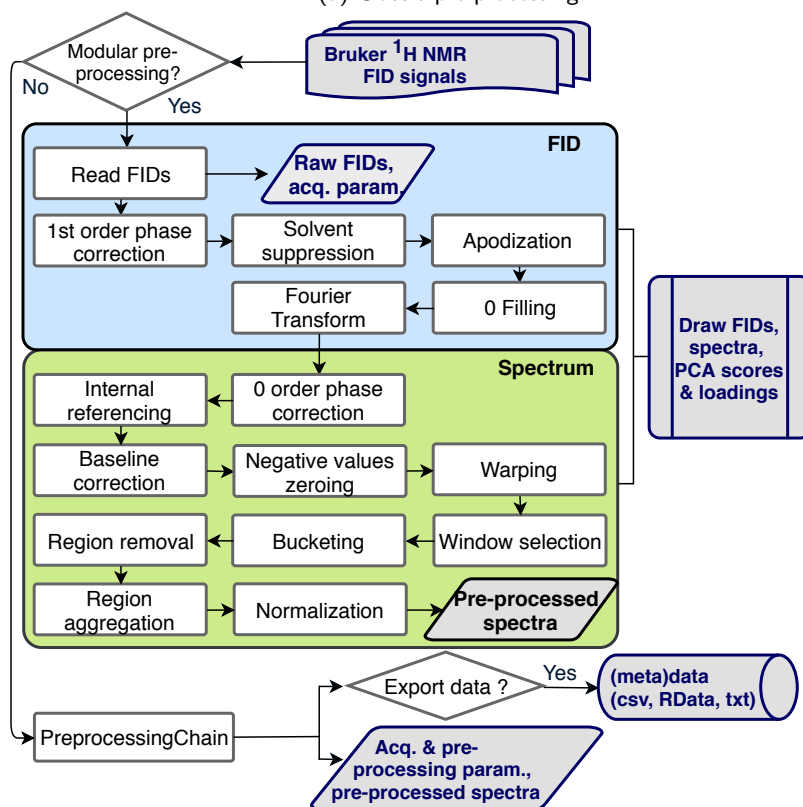
2.2 Methodology

2.2.1 PepsNMR workflow presentation

PepsNMR originates from an initially unpublished Matlab program called “Bubble” developed by Eli Lilly and one of the authors (Paul H.C. Eilers) (Rousseau, 2011; Vanwinsberghe, 2005). The package works as an open system with full flexibility: it provides a complete series of pre-processing functions to be run sequentially or independently to transform the FIDs into the final spectra. Raw data in Bruker format can be imported directly, different methods and settings are available in most pre-processing steps and visualisation functions are available to monitor the process. The pre-processing package is depicted in a flowchart in Figure 2.1b, where it can be compared to a classic pre-processing (Figure 2.1a). The complete sequence of R functions is detailed below in the suggested order.



(a) Classic pre-processing.



(b) PepsNMR pre-processing.

Figure 2.1 – Classic and PepsNMR ^1H NMR pre-processing flowcharts with available functions, inputs and outputs.

- GroupDelayCorrection()** This function corrects the frequency dependent — or first order — phase error due to the Group Delay induced by the Bruker's digital filter.
- SolventSuppression()** The variability of solvent residuals in the spectrum (especially for water) can mask informative signals from molecules of interest. This step removes the solvent signal (usually water) in the time domain using a Whittaker smoother.
- Apodization()** Apodization (mainly) aims to increase the SNR of the spectrum and/or its resolution. It multiplies the FID signal by a positive signal, often decaying. Several apodization functions are available in PepsNMR.
- ZeroFilling()** Zero Filling improves the visual representation of the spectra by adding a series of trailing 0 at the end of the signals.
- FourierTransform()** The Fourier transform translates the complex FID signal in the time domain into a complex spectrum expressed in frequency (Hz). This function further calibrates the frequency scale and converts it into chemical shifts (ppm).
- ZeroOrderPhaseCorrection()** In order to remove the frequency-independent phase shift, the spectrum is rotated to maximise the positiveness of its real part.
- InternalReferencing()** An internal compound (e.g. TMSP) is used for the chemical shift scale referencing.
- BaselineCorrection()** In order to remove the spectral baseline, this function uses asymmetric least squares with a roughness penalty.
- NegativeValuesZeroing()** Imperfect phase or baseline correction can lead to small residual negative values in the spectrum that cannot be interpreted. These negative values are thus set to zero.
- Warping()** The first global alignment solution provided in PepsNMR is a warping algorithm. It applies a warping function to spectra based on a reference spectrum. This warping function is a combination of polynomials and P-Splines that stretch/compress the chemical shift axis to increase the similarity between the transformed spectrum and the reference spectrum.
- WindowSelection()** This function selects the informative part of the spectrum to be kept for data analysis.
- Bucketing()** The second solution for minimising peak misalignment across spectra and for reducing the dimension of the data is soft bucketing. This function uses trapezoidal or rectangular interpolation to deliver equally-spaced bins with a specified length or number.
- RegionRemoval()** This function sets specific spectral regions of no interest in the study to zero.
- ZoneAggregation()** It is useful to apply targeted data reduction for molecules with an unstable chemical shift (e.g. citrate). This routine aggregates a pre-defined interval of descriptors into a single symmetrical peak.

Normalisation() This function normalises intensities in the spectrum to avoid unwanted inter-sample variation.

2.2.2 Details on the pre-processing steps

This Section describes in a nutshell the methodological features of the pre-processing steps. For each R function, underlying mathematical and statistical concepts along with their algorithmic implementation are explained and discussed.

Raw signal

Each FID (Figure 2.2) recorded with the ^1H NMR instrument in the time domain is a complex decaying signal $s = s_x + is_y$ of m data points (where s_x and s_y are respectively the real and imaginary parts of the signal). It represents the net sum of distinct signal components from different proton nuclei. The theoretical form of each single signal can be written as a function of its frequency offset ν , amplitude s_0 and transverse relaxation time T with an exponential decay over time t : $s(t) = s_0 \exp(i2\pi\nu t) \exp(-\frac{t}{T})$. The function `ReadFids()` imports raw Bruker files into two matrices: the complex FID signal matrix and the Bruker acquisition parameters needed for the pre-processing steps. Note that this exponentially decaying function in the time domain will give a Lorentzian line shape in the frequency domain.

Phase errors and Group Delay suppression

For a perfectly phased signal, its real part starts at a maximum value and its imaginary part starts at 0. The corresponding spectrum has a positive real part in absorptive mode and an imaginary part in dispersive mode with both positive and negative peaks. For technical reasons (Bruker digital filter, incorrect magnetisation, impossibility to excite nuclei in the sample and directly measure its response, etc.) the real and imaginary parts of the signal result in a mixture of both absorptive and dispersive mode line shapes (Keeler., 2002). These lines need to be phase corrected. The total phase error Φ of a signal is usually written as a sum of frequency-independent (φ_0) and frequency-dependent ($\varphi_1(\nu)$) phase errors, $\Phi = \varphi_0 + \varphi_1(\nu)$, that are corrected respectively with the zero and first order phase corrections. The initial non-phased signal is: $s = s_{\text{phased}} \exp(i\Phi)$.

This first pre-processing step aims to partly correct the first order phase error introduced by the Bruker digital filter.

The latter has multiple advantages (higher quality spectral baselines to SNR and effective dynamic range improvements, enhanced reduction of potential sources of folded signals, etc.). However, it leads to a group delay, *i.e.* a pre-acquisition delay before the start of acquisition. Hence, the real part of the FID does not have a

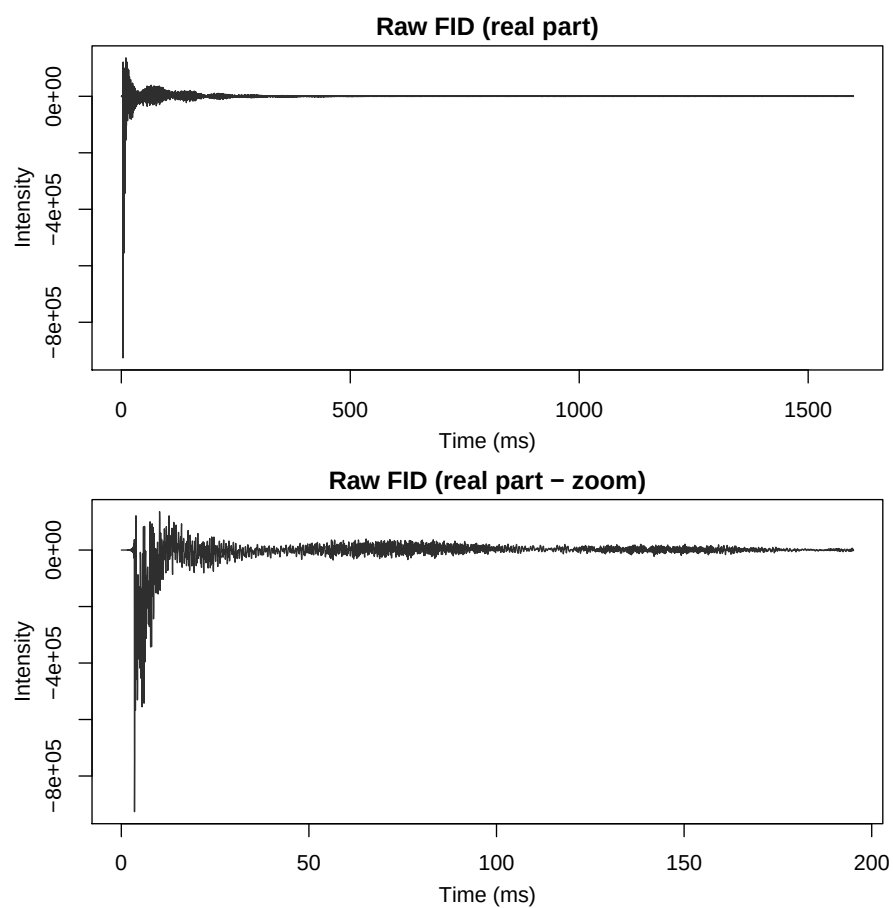


Figure 2.2 – The raw FID.

maximum value at $t = 0$. This delay, usually of tens of data points, is illustrated in Figure 2.3.

In PepsNMR, the delay is removed from the FID based on properties of the Discrete Fourier Transform (DFT) (Siebert, 1986): $\frac{1}{m} \sum_{\nu=0}^{m-1} f(\nu) \exp(-i2\pi\nu\tau/m) = \{s((t - \tau) \bmod m)\}$. In order to remove the delay of $\tau > 0$ samples in the time domain, which is called the Group Delay, one can apply a circular shift, which is a property of the DFT. This circular shift is identical to multiplying the corresponding spectrum f by the linear term $\exp(-i2\pi\nu\tau/m)$ *i.e.* applying a frequency-dependent linear phase shift in the frequency domain.

The signal is first Fourier transformed, then multiplied by the phase shift, and finally back-transformed with the inverse Fourier transform. This step is mandatory in the common case of a non-integer Group Delay. The function `GroupDelayCorrection()` removes the first order phase shift with τ recovered from the spectrometer acquisition parameters and returns the processed FIDs.

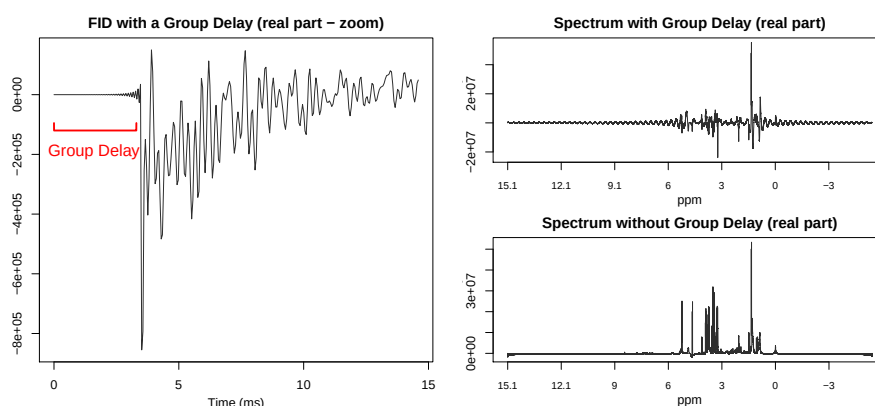


Figure 2.3 – The Group Delay recorded before the signal acquisition has to be removed from the FID.

Solvent signal suppression

Water is the most common solvent for biofluid samples. It has a rather broad resonance, and its chemical shift varies according to solution conditions. Usual solvent suppression techniques aim to reduce the magnitude of the solvent resonance before the NMR signal reaches the receiver. Three different approaches are (Claridge, 2016): pre-saturation of the water resonance, production of zero net excitation of the water resonance and destruction of the water resonance with pulsed field gradients. Nevertheless, residual peaks of water are still observed and

their intensity and instability can mask other interesting compounds. Moreover, and based on our experience, solvent suppression improves the efficiency of subsequent correction steps such as the baseline correction.

PepsNMR solves this common issue with non-parametric modelling and the subtraction of the residual water signal directly from the FID. Under the assumption that water is the main compound of the analysed samples, a discrete penalised least squares with second-order differences presented by Eilers (2003) and based on the Witthaker's graduation method (Whittaker, 1923), is used to separately model the real and imaginary parts of the water signal.

This method estimates the smoothed series w that minimises the Q criterion:

$$Q = \sum_{j=1}^m (s_j - w_j)^2 + \lambda \sum_{j=3}^m (\Delta^2 w_j)^2 \quad (2.1)$$

This expression is a balanced combination of two conflicting objectives, *i.e.* fidelity to the data and smoothness. The first term is the sum of squares of differences measuring the lack of fit of w to the real or imaginary signal s on the m data points. The second term penalises the series' roughness with λ the penalty parameter multiplied by the sum of squared second order differences (defined as $\Delta^2 w_j = w_j - 2w_{j-1} + w_{j-2}$), which quantifies the roughness. The larger λ is, the smoother the water signal. An analytical solution is provided by Eilers (2003) and the use of sparse matrices greatly reduces the computational time for large series such as ^1H NMR signals. The function `SolventSuppression(Fid_data, lambda.ss = 1e6, ...)` estimates the water signal for both the real and the imaginary parts of the FID and returns the processed FIDs and the real and imaginary water signals. `lambda.ss` accounts for the parameter λ and its default value (1e6) is a practical choice, based on empirical results. Nevertheless, Eilers (2003) and Frasso & Eilers (2015) describe several ways to tune λ in order to optimise the smoothing: either visually, by cross-validation or using V-curves with R code available for the latter.

Apodization

Apodization multiplies the FID by a signal that changes its shape and emphasises distinct spectral portions (Claridge, 2016) to enhance the sensitivity and/or the resolution of the spectrum. Due to its nature, NMR signal intensity decays over time whereas noise has random fluctuations with constant amplitude. Thus, the SNR is higher at the beginning of the FID and declines over time. An increase in sensitivity (higher SNR) can thus be achieved by multiplying the FID by a decaying signal. However, a faster decay of the FID (already dependent on the decay time T) implies broader spectral peaks line widths, lowering their height and thus their resolution since the area under the peaks remains unchanged. Therefore,

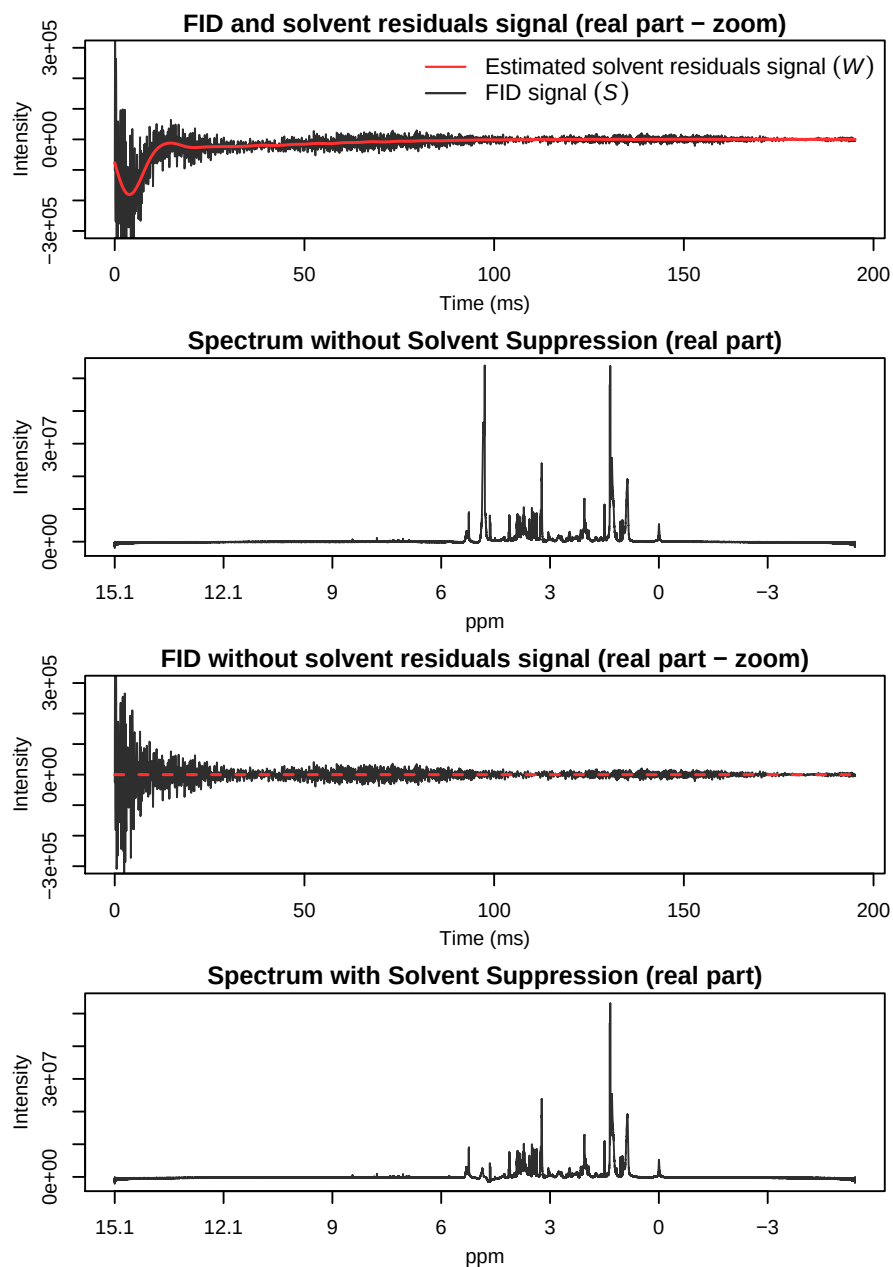


Figure 2.4 – The solvent residuals signal is first estimated with a smoothing function and then subtracted from the FID.

optimal trade-off parameters for the apodization signal are needed to combine both of these objectives. A classic apodization step applies an exponentially decaying signal principally to enhance the SNR but other apodization factors, such as a concave signal, can be used to enhance both sensitivity and resolution.

Several types of apodization functions are provided by PepsNMR with the option `type.apod` in `Apodization(Fid_data, type.apod = "exp", ...)`: decreasing exponential or gaussian functions, Hanning window, etc.. By default PepsNMR multiplies the FID by a decaying exponential function: $h(t) = \exp(-LB * t)$ with LB the Line Broadening parameter. If LB increases, the SNR increases at the expense of the spectral resolution (broader line widths). Usual values in proton NMR for LB found in the literature are 0.3 for the NOESY presat pulse sequence and -0.01 for the CPMG presat pulse sequence. LB should not exceed the value of 1 to avoid information loss.

Zero filling

When applying zero filling with the function `ZeroFilling()`, a vector of zeros is appended at the end of each FID. This number is rounded to the nearest 2^x value to ease the upcoming Fourier Transform of the FIDs. By default, the number of points is increased to twice its value during this process. Zero filling does not improve the spectral resolution but will lead to better visually defined lines in the spectra.

Fourier transform and chemical shift scale

At this stage, the recorded signal in the time domain S is ready to be converted into a spectrum in the frequency domain f . The Fourier Transform (FT) provides a general solution to transforming the complex signal sequence into a complex spectral sequence. It extracts each signal and translates it into peaks in a spectrum with specific heights, positions and widths that depend on the amplitude of the signal, the frequency of the corresponding compound and the relaxation time T (Keeler., 2002).

PepsNMR achieves this transformation with the function `FourierTransform()`, which applies the Discrete Fourier Transform (DFT) for uniformly spaced samples with the Fast Fourier Transform (FFT) algorithm.

The DFT can be written as:

$$f_k = \sum_{j=0}^{m-1} s_j \exp(-2i\pi k j / m) \quad (2.2)$$

for $k = 0, \dots, m - 1$.

The scale of the $\mathbf{f}$ vector has yet to be defined. PepsNMR first expresses it in Hertz on the basis of the spectrometer acquisition parameters with the following formula: $\nu_j = \nu_S + \Delta_\nu \times (j - (m - 1)/2)$ where ν_S is the transmitter frequency (called SFO1 in the Bruker acquisition file) that is the absolute value of the frequency in MHz used to excite the observed nucleus. It is located at the center of the spectral window and Δ_ν represents the difference in frequency between 2 data points and is calculated from the spectral window width (or sweep width). Note that ν_S is a sum of ν_B , the basic transmitter frequency (set automatically depending on the spectrometer and the observed ^1H nucleus) and ν_O , the transmitter frequency offset, tuned by the user to shift the center of the spectral window from ν_B to ν_S .

The scale in Hertz is then converted into a chemical shift vector δ expressed in parts per million (ppm), a dimensionless scale that is unrelated to the magnetic field strength of the spectrometer (Keeler., 2002). A chemical shift represents the frequency ν_j relative to ν_B , where the 0 ppm is located at the ν_B frequency. The conversion equation is: $\delta_j = 10^6 \times \left(\frac{\nu_j - \nu_B}{\nu_B} \right)$.

A spectral matrix is then created where each row corresponds to a spectral profile and each column represents a spectral intensity at a specific chemical shift. By convention, the chemical shift along the horizontal axis is set as decreasing towards the right.

Zero order phase correction

For the technical reasons explained above, the spectrum can still exhibit a zero order phase shift error of a certain angle φ_0 that can be expressed as $\mathbf{f} = \mathbf{f}_{\text{phased}} \exp(i\varphi_0)$, where $\mathbf{f}_{\text{phased}}$ is the perfectly phased spectrum. This phase shift is constant for all signal vectors and thus independent of the spectral frequencies.

Based on the principle that the real part of the perfect spectrum $\mathbf{f}_{\text{phased}}$ has a pure absorptive mode, an optimal angle φ_0 can automatically be found in PepsNMR by maximising an adequate criterion of the positiveness of the real part of the spectrum.

PepsNMR proposes two positiveness criteria measured on a range r of descriptors: it can be quantified by the maximisation of the highest positive peak or by the *rms* criterion that represents the ratio between the sum of squares of positive intensities and the sum of squares of all intensities in the spectrum:

$$\varphi_0 = \arg \max_{\phi \in [0, 2\pi]} \frac{\sum_{j \in \mathbf{R}} (\mathbf{f}_j^\phi)^2 \times \text{ind}_j}{\sum_{j \in \mathbf{R}} (\mathbf{f}_j^\phi)^2} \quad (2.3)$$

where $f^\phi = f \exp(-i\phi)$ is the real part of the rotated spectrum with a correction angle $-\phi$ and $\text{ind}_j = 1$ if $f_j^\phi > 0$ and 0 otherwise. The user can decide on which range r to maximise the positiveness criterion since particular areas that can be difficult to phase (especially the spectral water region) can be excluded. The function `ZeroOrderPhaseCorrection(RawSpect_data, type.zopc = "rms", ...)` rotates each spectrum in the spectral matrix and returns the processed spectra with their rotation angles. In addition, a vector of angles can be directly provided by the user with the argument `type.zopc = "manual"`.

Note that residuals of first order phase shift can still be present. They are currently not removed by PepsNMR but are negligible after the Group Delay removal.

Referencing with an internal standard

In order to be more reliable, the scale can be referenced to a known standard, *i.e.* an internal reference compound whose chemical shifts are minimally influenced by external factors (*e.g.* 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 Trimethylsilyl-propionate (TMSP) or 4,4-Dimethyl-4-silapentane-1-sulfonic acid (DSS). The chemical shift of these standard peaks is referenced to 0 ppm by convention but other reference compounds may have a non-null ppm value.

In PepsNMR, the function `InternalReferencing(RawSpect_data, shiftHandling = "cut", method = "max", range = "near0", ...)` proposes two ways to locate the reference compound peak in each spectrum within a range of intensities (`range`) with the `method` argument: it selects either the maximum intensity or the first peak in the search range higher than a predefined threshold.

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, a not entirely linear electronic detection process or calibration errors from the 180° pulse (Euceda *et al.*, 2015). It is critical to remove these artefacts 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 have been suggested that estimate the baseline to be removed from the spectra. They mainly involve iterative penalised least squares methods, robust baseline estimation, iterative polynomial fitting, local means or wavelets transforms (Liland *et al.*, 2010; Marion, 2016).

The Asymmetric Least Squares Smoothing (AsLS) algorithm (Eilers & Boelens, 2005) in PepsNMR allows flexible baseline estimation with fast and reproducible

computations. It is a nonparametric penalised method that combines asymmetric weighting of deviations from the trend with a penalty on any non-smooth behaviour of the estimated series.

This generalisation of the Whittaker smoother provides a robust baseline estimate with no artefacts at the end of the spectrum, where a polynomial fit (Liland, 2011).

The baseline estimator z of a signal f , is found by minimising the objective function (2.4), similar to Equation (2.1), where the first term represents the lack of fit to the data, measured by asymmetric least squares, and the second term is the roughness penalty. The asymmetric weighting will favour positive corrected intensities as positive deviations are less weighted than negative ones.

$$Q = \sum_{j=1}^m \psi_j (f_j - z_j)^2 + \lambda \sum_{j=3}^m (\Delta^2 z_j)^2 \quad (2.4)$$

In Equation (2.4), ψ_j is a specific weight for the frequency j : if $f_j > z_j$, $\psi_j = p$, otherwise $\psi_j = (1 - p)$ where $p \in [0, 1]$ is close to 0.

The function `BaselineCorrection(RawSpect_data, lambda.bc=1e7, p.bc=0.05, ...)` starts by initialising z . Then, from Equation (2.4)), we compute a (new) smoother z and from $f_j - z_j$ follow new weights ψ_j . This estimation procedure is iterated until these estimated weights don't change anymore (Eilers, 2004; Eilers & Boelens, 2005). It then subtracts the estimated baseline function from the spectrum: $f^* = f - z$.

An illustration of this process is given in Figure 2.5. The larger λ is, the smoother the baseline will be and the smaller p is, the more the baseline will tend to be both below and close to the spectral line (containing random noise and signal). Based on experience, Eilers & Boelens (2005) recommend the following ranges of parameters values: $0.001 \leq p \leq 0.1$ and $10^2 \leq \lambda \leq 10^9$. Often, a visual inspection is sufficient for tuning these parameters. Nonetheless, these authors provide ad hoc computation procedures in their article to optimise their values.

Since this baseline correction algorithm can create artificial positive intensities in the water region (Rousseau, 2011), a previous solvent suppression step is recommended and a subsequent step is dedicated to the removal of uninformative regions (e.g. water) to cope with this behaviour.

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.

In PepsNMR, this operation is done with the function `NegativeValuesZeroing()`.

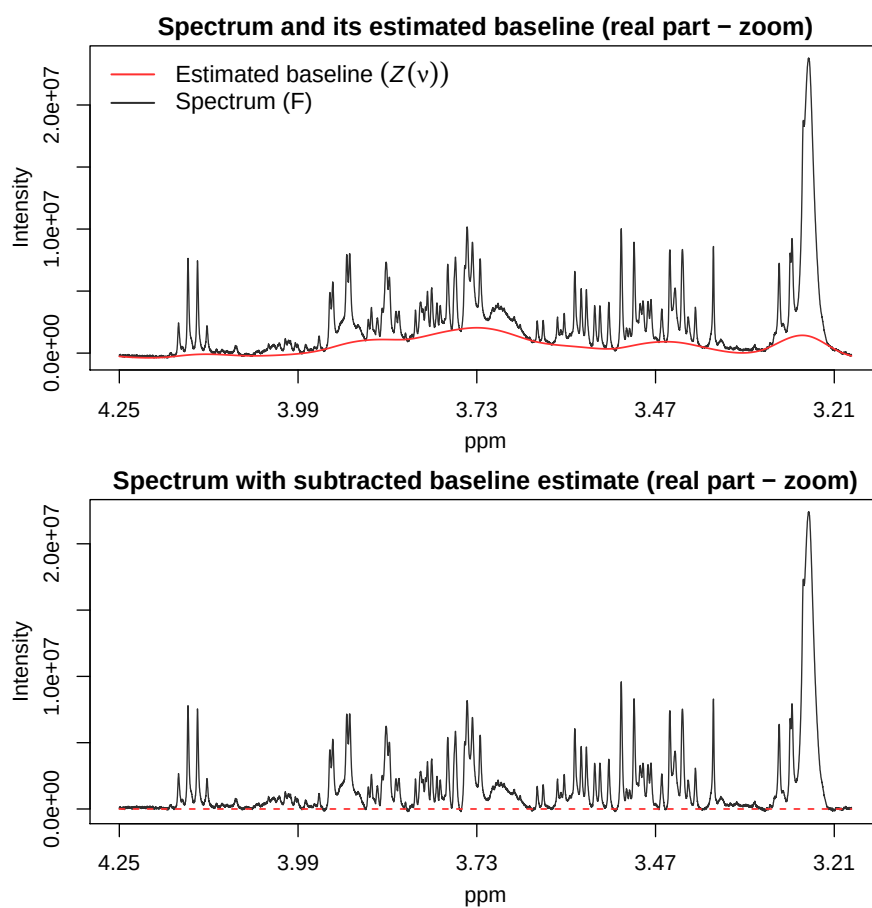


Figure 2.5 – The baseline is first estimated and then removed from the spectrum.

Warping

In metabolomic applications, due to the nature of biological samples and variation 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 bucketing², and both methods are included in PepsNMR. 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) and Vu & Laukens (2013) 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), Dynamic Time Warping (DTW), Correlation Optimised Warping (COW), interval correlation shifting (icoshift), Cluster-based Peak Alignment (CluPA), etc.

The warping method implemented in PepsNMR is inspired by different elements of time warping as developed by Eilers (2004) and van Nederkassel *et al.* (2006). It applies a warping function $w(\nu)$ to a normalised spectrum f , i.e. a distortion function of the ppm axis that combines a polynomial term and a penalised B-splines (P-splines) term. The warping function $w(\nu)$ is defined as:

$$w(\nu) = \sum_{k=0}^K \beta_k \nu^k + \sum_{l=1}^L \alpha_l B_l(\nu) \quad (2.5)$$

The first term is a polynomial of order K with β the corresponding polynomial coefficients vector. The second term is a linear combination of B-splines, with L the number of B-splines and α_l the coefficients for the l^{th} B-spline $B_l(\nu)$ to weight the splines for the fitted curve. The B-spline curves are constructed from polynomial pieces constructed from adjacent and predefined knots and smoothly joined together. They represent base functions for non-parametric regression.

The normalised and now distorted spectral profile $f(w(\nu))$ is interpolated from the discrete warping function.

The warping principle consists of building $w(\nu)$ such that the distance between a warped spectrum $f(w(\nu))$ and the reference spectrum g is minimised. The parameters are found by the iterative minimisation of the following penalised least squares objective function:

$$Q = \sum_{j=1}^m (g_j - f_j(w(\nu)))^2 + \lambda \sum_{l=3}^L (\Delta^2 \alpha_l)^2 + \kappa \sum_{l=1}^L \alpha_l^2 \quad (2.6)$$

²Data reduction technique where the values are integrated over discrete spectral areas (Craig *et al.*, 2006).

where the unknowns are β_k and α_l . Two penalty terms are applied to the B-splines coefficients (Eilers & Marx, 1996). The first one is based on second order differences $\Delta^2\alpha_l$ and tunes the B-splines smoothness with parameter λ whereas the second one is a Ridge penalty applied to regularise the coefficients (*i.e.* control their variance) with parameter κ . The function `Warping()` warps the spectra with the following default parameter values empirically chosen based on the authors' experiments: $K = 3$, $L = 40$, $\lambda = 0.01$ and $\kappa = 0.0001$. An illustration of warped spectra is given in Figure 2.6.

Each spectrum must be first normalised and then denormalised after warping. Normalisation is defined as a row operation on the spectral matrix lines for making the samples more directly comparable with each other. The choice of a reference spectrum among all profiles is not straightforward. With no prior knowledge and no pooled samples, a robust reference selection process, available in PepsNMR, selects the spectrum that minimises the sum of squared distances with all other spectra before or after warping. Another possibility in PepsNMR is to refer to a manually defined sample, such as an artificial sample with known metabolic concentrations. Note that when the samples come from different classes, using one reference sample per class can be misleading and is likely to induce artificial between-class variability (Bloemberg *et al.*, 2013).

Spectral window selection

In order to focus the data analysis on particularly informative spectral areas, the function `WindowSelection()` trims the spectral window between right and left chemical shift bounds (by default: 0.2 ppm and 10 ppm), 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 m original spectral intensities into m^b 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 points, usually from $m > 15000$ to $m^b < 1000$.

In this step, there is a trade-off between keeping spectral information and removing peaks shifts as well as decreasing the total number of variables. Among possible binning methods, PepsNMR's bucketing proposes two integration options, either

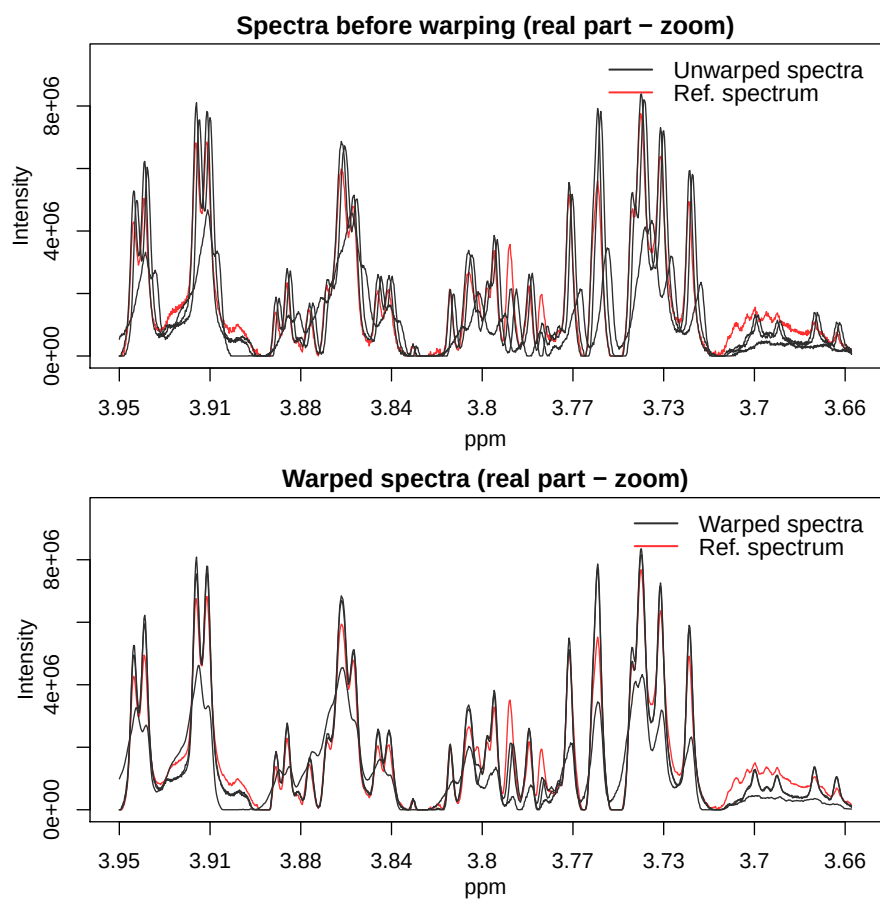


Figure 2.6 – Spectra are horizontally warped to remove peak shifts and to more accurately compare intensity differences.

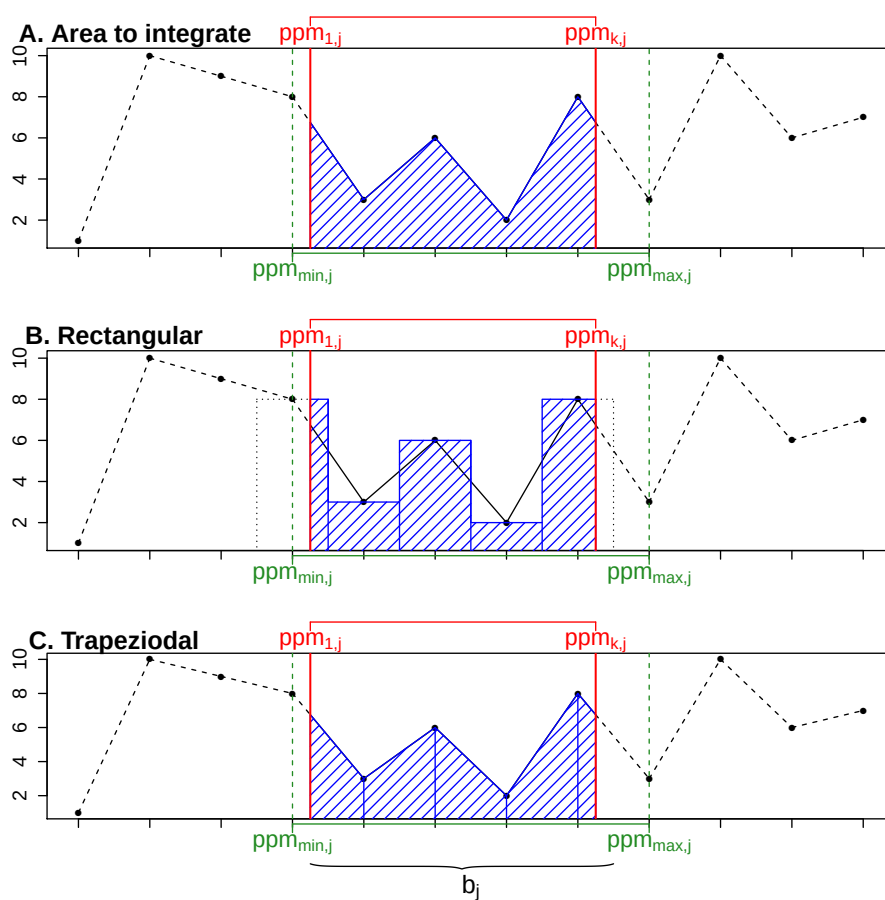


Figure 2.7 – Illustration of the target area to be integrated and the two bucketing methods (B. and C.) in PepsNMR.

trapezoidal or rectangular, with equally sized buckets and generalised to cut the original x -axis at any chosen location.

The interpolation is described as follow: (1) the number of buckets m^b is first decided and an equidistant width for each interval is computed as: m/m^b ; (2) for each bucket j of length k , minimum, maximum and middle ppm values are identified, *resp.* $ppm_{1,j}$, $ppm_{k,j}$, $z_j = \frac{ppm_{1,j} + ppm_{k,j}}{2}$; (3) within each bucket, intensities are integrated and the result is a vector of m^b data points; (4) each spectral point or bucket is now identified with its z_j value. The integration is either trapezoidal or rectangular and generalised to cut the original x -axis at any chosen location.

The integration in step (3) is based on a generalised version of the trapezium rule (Bourne, 2015) that enables to cut the original x -axis at any chosen location. Figure 2.7 illustrates its principle: the bucket j starts from $ppm_{1,j}$ to $ppm_{k,j}$ and the trapezium rule is applied between $ppm_{min,j}$ and $ppm_{max,j}$, *i.e.* original data points just before and just after the bucket limits. Note that depending on m^b and the original m , bucket intervals do not necessarily cut in the middle of two initial data points. Then the integrated area between $ppm_{min,j}$ and $ppm_{1,j}$ and between $ppm_{k,j}$ and $ppm_{max,j}$ are subtracted from the total integrated area. Given the trapezium rule, the value of the bucket is:

$$y_j = \int_{ppm_{1,j}}^{ppm_{k,j}} f(ppm) d ppm = \frac{1}{2}k \left[(f(ppm_1) + f(ppm_k)) + 2 \left(\sum_{i=2}^{i=k-1} f(ppm_i) \right) \right] - A$$

where A is the sum of areas to subtract between $ppm_{min,j}$ and $ppm_{1,j}$ and between $ppm_{k,j}$ and $ppm_{max,j}$.

In the package, the function `Bucketing(Spectrum_data, mb = 500, ...)` reduces the spectra to m^b buckets. In the advised procedure, the number of buckets varies from 500 for urine spectra to 750 buckets for serum spectra.

Region removal

Included by default in the selected spectral window, 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). By default, the function `RegionRemoval()` removes urea, maleic acid and water (4.5 – 6.1 ppm) peaks for urine spectra, whereas the water resonance area (4.5 – 5.1 ppm) is only removed by default for serum spectra.

Region aggregation

Strong variations in peak position are particularly present for ionised metabolites such as citrate, taurine, succinate or lactate (Bloemberg *et al.*, 2013) and cannot be totally avoided through sample preparation procedures (*e.g.* pH buffer), constrained warping methods or soft bucketing.

To avoid this unwanted source of variation, targeted data reduction can be applied with the function `ZoneAggregation()` in PepsNMR. Peaks are appropriately reshaped with the aggregation technique suggested in Rousseau (2011): the total intensity in an $[a, b]$ ppm interval $\sum_{j=a}^{j=b} f_j$ is distributed across this area so that the metabolite has a unique symmetrical triangular peak and null intensities at the border. A future version of the package will include the possibility of aggregating intensities into a Lorentzian curve.

Normalisation

Normalisation is defined as a row operation (Craig *et al.*, 2006) $f_i \times c_i$ where each spectrum i is multiplied by a constant term c_i . 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. This variation is especially present in urine samples that do not possess homeostasis properties such as serum samples do, but other technical issues can also create this unwanted variation. Various normalisation methods coexist in the literature (see for example Craig *et al.* (2006), Smolinska *et al.* (2012) or Marion (2016)).

The Constant Sum Normalisation (CSN) technique is advised for sera, cell media or biopsy samples where large differences between groups of patients are not expected. With this method, $c_i = \frac{m^b}{\sum_{j=1}^{m^b} f_{ij}}$. The mean intensity for each spectrum becomes m^b . The Probabilistic Quotient Normalisation (PQN) (Dieterle *et al.*, 2006) — specially designed for urine spectra and for distinguishing the effects of overall and local metabolite concentration changes on the spectrum — is another popular method included in PepsNMR. Other normalisation techniques available are: by the median, by the first quartile or by the intensity of an endogenous metabolite, usually creatinine for urine samples whose concentration is then assumed to be constant across samples under certain conditions (Tang *et al.*, 2015). Guidance on the choice of normalisation approach based on the type of sample analysed can be found in Wu & Li (2016).

2.2.3 Other PepsNMR features and functions

For the sake of reproducibility, PepsNMR gives the possibility to modify, view and save spectra together with their acquisition parameters as well as pre-processing parameters during or after the whole pre-processing workflow. These different features are explained below.

The `PreprocessingChain()` function A wrapper for data import and pre-processing steps. When working with this function, the pre-processed spectral intensities and acquisition parameters matrices can be exported (.RData; .txt or .csv) as well as a text file with all the pre-processing input parameters to ensure traceability.

Functions parameters For most pre-processing steps, different input parameters can be modified to adjust the methods to every dataset. Default values are in general assigned to most of those parameters in the package but can also be tuned by the user according to visual support, this article's guidelines and the related *help* support in R. They can also be optimised with an appropriate experimental design and/or objective quality criteria as explained below in Section 2.3.

Functions order Some pre-processing blocks can be switched or bypassed with respect to the advised order, while some corrections should precede other steps (e.g. solvent suppression before baseline correction or warping before bucketing) and several are compulsory (e.g. Fourier transform or phase corrections). As for their parameters, quality criteria can guide the choice of an appropriate function order.

Visualisation tools After each step, data can quickly be visualised to check the effects of the applied pre-processing method(s) and detect unexpected or abnormal results that could mislead data analysis, but also to spot outliers and global trends (Liland, 2011). PepsNMR provides different ggplot2-themed plotting possibilities with the `Draw()` function to represent either the signal/spectral intensities or PCA scores and loadings. They can be directly displayed in the R interface or saved as png/pdf to be recovered later.

2.3 Quality criteria for PepsNMR evaluation

In the same perspective but beyond the scope of this article is the importance of relying on objective quality criteria to assess and quantify pre-processing efficiency leading to spectra with more informative content. The development and wide use of such quality parameters, as named by Engel *et al.* (2013), are still lacking and should be investigated more intensively across the scientific

literature. Indeed, sub-optimal pre-processing strategies can deteriorate relevant spectral information and affect study conclusions. However, currently there is no global consensus on which selection criteria other than visual inspection, should be used to test pre-processing efficiency. In this paper, the evaluation of PepsNMR performance is done as suggested by Engel *et al.* (2013): the decision relies on a robust chemometric approach consisting of different quantitative quality parameters that aggregate into a super-parameter. This approach can either be used to compare the pre-processing efficiency of competing strategies or to appropriately calibrate the PepsNMR function parameters and order.

This approach requires a validation dataset to be collected for every dataset under study in order to properly compare and calibrate the pre-processing strategies based on quantitative quality criteria. Indeed these criteria measure the spectral repeatability from a designed validation dataset with technical and/or biological replicates (e.g. subsamples from the same biological unit analysed at different times or involving different dilutions). The spectral repeatability is defined as the ratio between the repeatable, structured group "signal" from the replicates and i.i.d. noise under the hypothesis that the sampled spectra are composed of both structured signal and random noise. Note that this validation dataset should be similar but independent of the dataset of interest in the study.

In the next Section, quality criteria are measured for distinct pre-processing strategies that are applied on to identical validation dataset with known replicates class to rank their efficiency. These criteria are derived from unsupervised as well as supervised chemometric tools. They are gathered under the Metabolomic Informative Content (MIC) concept (Féraud *et al.*, 2015) and include Total Variance Decomposition (TVD), PCA, clustering and PLS-DA related criteria. Applying a Sums of Squares decomposition on the spectral matrix breaks down the total variance into two complementary parts: the variance between the groups and the variance within the groups of observations. For clustering, the (adjusted) Rand indexes measure the true class recovery efficiency and should be maximised, while the Dunn and Davies-Bouldin indexes measure the clustering homogeneity, and they have to be respectively maximised and minimised (formula details can be found in Féraud *et al.* (2015)). PLS-DA builds a latent structure with Latent Variables (LV) to extract the variance components from the data but instead of maximising the spectral data matrix variance as PCA does, it maximises the variance of both the spectral data matrix (the regressors) and the class labels (the response) and their correlation. Its cross-validation criterion Q^2 assesses model efficiency in terms of class prediction.

The incentive to use unsupervised methods (PCA and hierarchical clustering with Ward method) is to contrast the amount of class information that we are able to recover in the data after implementing different pre-processing pipelines, provided that the between-group variation is the main source of variability in the data. This information recovery is a blind process: no prior information is provided about the class labels. The supervised classification algorithm (PLS-DA) completes the performance analysis since it constitutes standard multivariate data analysis that a researcher would use to discriminate between different classes of spectra.

2.4 Results and discussion

In this Section, the advanced pre-processing methods in PepsNMR are compared to a manual data pre-processing with TopspinTM AMIX on the grounds of two typical biofluid media: human serum (HSerum) and human urine (HUrine) datasets. To that end, the data tables are fully pre-processed either with PepsNMR or with Topspin/AMIX. The HSerum dataset comes from 4 different blood donors whereas the HUrine dataset is measured from 3 different urine donors. The both of them are designed in such a way that the biological information in between-donor variability should be the main factor affecting spectral shape. Noisy parameters that were intentionally added to the data to reproduce standard analytical conditions are: day of measurement, sample replicates and sample dilution. More details on these data, their acquisition and pre-processing are available in Section 2.7.

2.4.1 Variance structure

If the pre-processing steps manage to accurately transform the data and correct for technical sources of variation, the main source of variability in the two case studies would correspond to the biological differences between the donors.

As explained in Section 2.3, TVD and PCA are first applied to characterise and quantify the amount of biological information recovered. Variance decomposition results are given in Table 2.1. For the two datasets, the variance between the groups of samples is larger with PepsNMR than with manual pre-processing. This difference in values for HUrine is small (+13.18%), since the manual pre-processing is already effective, but a real improvement can be observed for the HSerum data (+67.42%): a larger amount of TVD is captured by the variability among the groups rather than between them. Since a larger between-group variance implies better group separation, PepsNMR performs better than the manual alternative.

Table 2.1 shows that the sum of Principal Components (PC) 1 and 2 is similar for the two pre-processing approaches for the HSerum dataset but the group structure is only captured by PC2 for the manual pre-processing, as can be seen

in the scores plot in Figures 2.8 and 2.9. The main source of variation (79.53 %) has another origin: significant peak shifts are observed in the PC1 loadings plot (Figure 2.10). They are hiding the information of interest. Scores plot of PC1 and PC2 for the HUrine dataset shares the same structure, regardless of the pre-processing mode. Spectral profiles are closely tied together and groups are clearly separated. PC1 discriminates between donor 3 and the two others, whereas PC2 discriminates patient 1 from donor 2. Moreover, the PepsNMR workflow for this dataset recovers more informative variability in the two first components (+12.52 %) than the manual pre-processing.

Dataset	Pre-processing	Inertia		PCA variance	
		Between (%)	Within (%)	PC1 (%)	PC2 (%)
HSerum	manual	21.44	78.56	79.53	10.96
HSerum	PepsNMR	88.86	11.14	73.01	16.75
HUrine	manual	82.14	17.86	67.07	15.75
HUrine	PepsNMR	95.32	4.68	82.77	12.57

Table 2.1 – Repeatability is partly assessed by variance decomposition methods: total variance decomposition (% of between and within variance) and PCA (cumulated % of variance for the first two PCs).

2.4.2 Clustering and classification performance

In this Section, the unsupervised and supervised classification algorithms (hierarchical clustering and PLS-DA) explained in Section 2.3 are applied to both datasets. Note that PCA is applied to the data matrix prior to clustering in order to reduce the dataset dimensionality. The number of clusters corresponds to the natural structure of the data: the number of donors. The number of LVs for PLS-DA is chosen based on the cross-validated Root-Mean-Square Error (RMSE): 4 for HSerum and 2 for HUrine. Accurate classification and homogeneity of groups are examined to evaluate the unsupervised tool while the predictive quality criterion Q^2 is used to rate the PLS-DA performance. In the case of a null model, $Q^2 \leq 0$, and with a perfectly discriminating model, $Q^2 = 1$. Results from these two approaches can be visualised in Table 2.2.

Clustering-related indexes quantify either the similarity between two partitions (the true class and the one obtained with blind clustering) or the cluster homogeneity. For the HSerum, higher Rand and Adjusted Rand indexes for PepsNMR clearly indicate a better recovery of the true classes of samples, as well as better group homogeneity according to the Dunn index. Less importantly, even though the Davies-Bouldin index shows a better group homogeneity for the manual pre-processing, the grouping does not represent correctly the classes. As for the variance structure analysis, the two pre-processing techniques for the

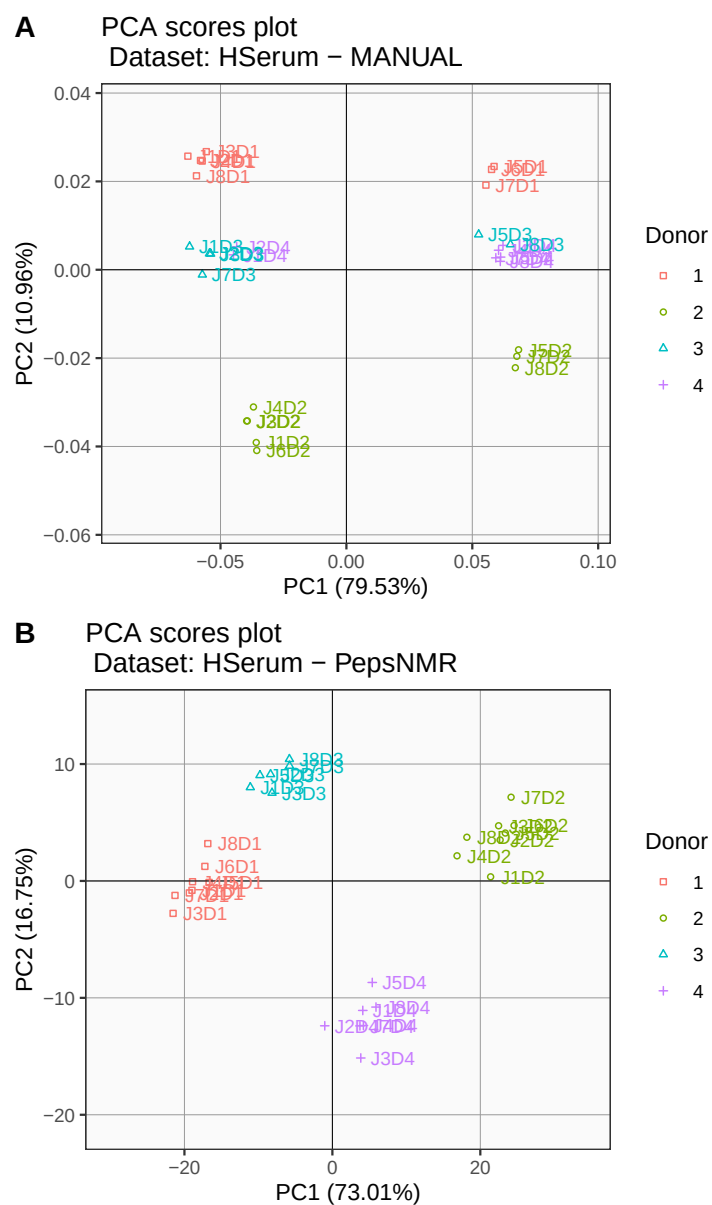


Figure 2.8 – PCA scores plot of PC1 and PC2 for the HSerum (A-B) dataset and the two pre-processing strategies (manual and PepsNMR). The colours represent the donors' identity.

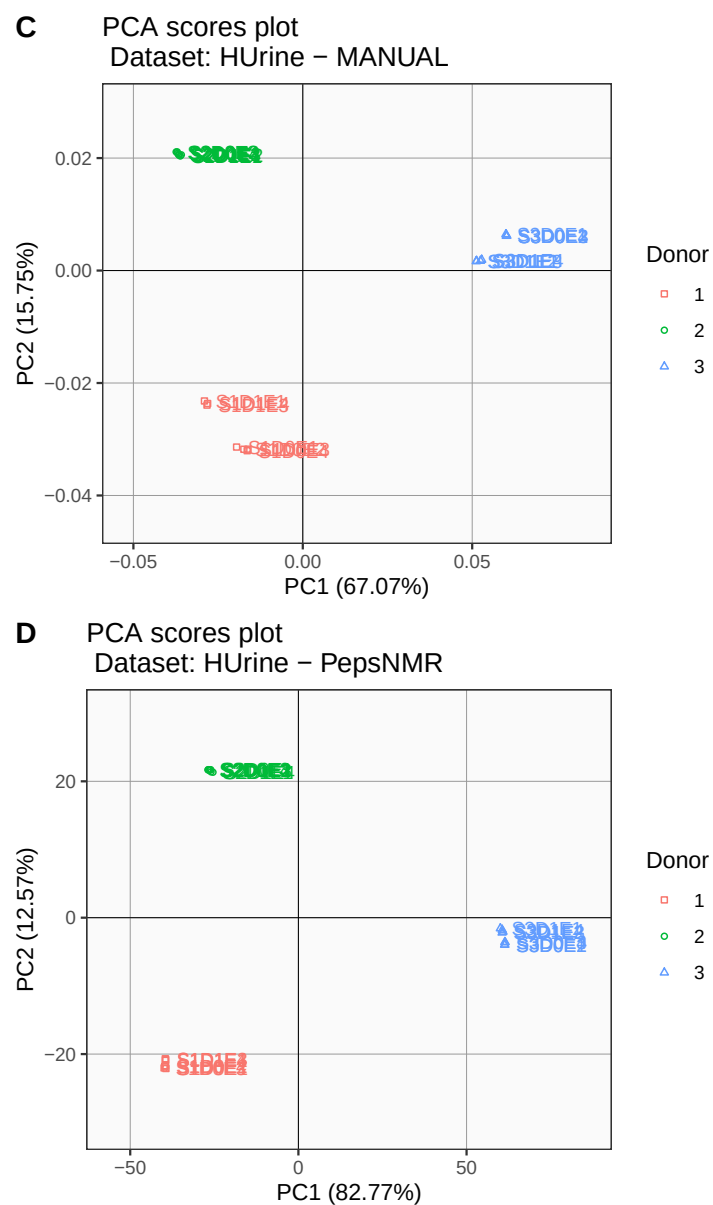


Figure 2.9 – PCA scores plot of PC1 and PC2 for the HUrine (C-D) dataset and the two pre-processing strategies (manual and PepsNMR). The colours represent the donors' identity.

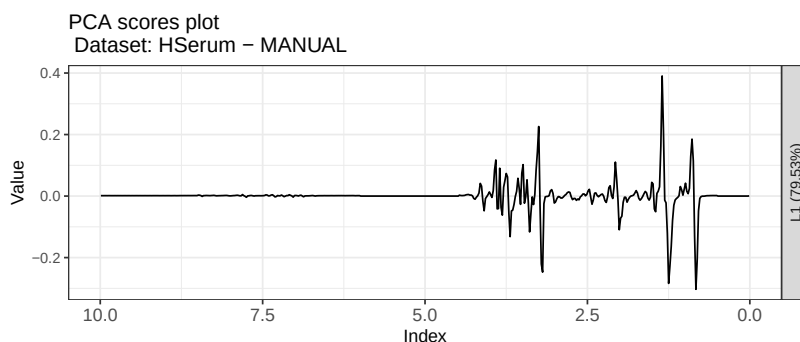


Figure 2.10 – First PCA loadings plot of the manually pre-processed HSerum dataset.

HUrine dataset lead to analogous clustering results illustrated here by a perfect categorisation of samples. Nonetheless, clusters with the PepsNMR pre-processed datasets do have better homogeneity criteria.

Findings with the PLS-DA are consistent with the previous results: the classification algorithm has a very bad cross-validation performance for the manually pre-processed human serum dataset since the Q^2 is negative. With the PepsNMR strategy, the performance criterion is substantially improved. For the urine dataset, Q^2 is almost 1, reflecting the excellent class prediction for both pre-processing strategies.

Dataset Pre-processing	Ward hierarchical clustering				PLS-DA Q^2
	Rand	Adj. Rand	Dunn	D-B	
HSerum manual	0.64	0.14	0.62	0.56	-0.53
HSerum PepsNMR	0.88	0.72	0.87	0.68	0.79
HUrine manual	1	1	0.83	0.69	0.98
HUrine PepsNMR	1	1	1.38	0.32	1

Table 2.2 – Clustering/classification results for repeatability assessment: indexes derived from unsupervised hierarchical clustering ((Adjusted) Rand, Dunn and Davies-Bouldin (D-B) indexes) and a supervised PLS-DA validation criterion (the cross-validated coefficient of determination Q^2).

2.5 Conclusions

Spectroscopic data pre-processing is a keystone procedure for enhanced information recovery in metabolomics, and PepsNMR is the only existing **R package** able to comprehensively address a very-detailed series of pre-processing steps designed for 1D ^1H NMR data. The two designed experiments comparing PepsNMR to a classic pre-processing strategy confirm its **practical interest**: it has led to an increase in spectral repeatability and an enhanced predictive power in a classification context, suggesting an overall **better information recovery**. In addition, the package has major **built-in assets**: it is open-source, it has routine modularity, automation and objectivity, as well as structured documentation, and yields reproducible results and reporting.

Furthermore, this package should help to continuously improve the processing of metabolomic data characterised by a high degree of complexity, and in the future, complementary/alternative methods will be added to this first version.

Although it is unlikely that all the methods included in PepsNMR will always perform better than other existing ones in the literature, its **flexibility** makes it possible to combine parts of it to other available, sometimes more optimal tools for a greater efficiency. This is especially true for the domains of baseline correction and realignment algorithms, which are complex issues where a high number of new and sophisticated methods are presented in the literature. For example, other iterative penalised least squares methods developed to improve the weighting scheme, such as Adaptive iteratively reweighted Penalised Least Squares (Zhang *et al.* , 2010, airPLS) or Asymmetrically Reweighted Penalised Least Squares (Baek *et al.* , 2015, arPLS) deserve some attention as they have been reported in the literature as performing better than AsLS to estimate the baseline, especially in noisy areas (Baek *et al.* , 2015). Weighted P-splines regression with weights calculated from a mixture model (de Rooi & Eilers, 2012) would also deserve some attention as it avoids to choose the asymmetry parameter while preserving flexibility in the baseline estimation.

With regards to warping, segment-wise realignment techniques should also be investigated. In particular, two quite recent realignment techniques for NMR are the Interval correlation shifting (Savorani *et al.* , 2010, Icoshift), which is substantially faster than other algorithms and showed good realignment performances (Emwas *et al.* , 2018); and CluPA (Vu *et al.* , 2011), which uses hierarchical clustering to realign spectra segments in a recursive fashion and is already integrated to W4M. However, it should be noted that none of these current realignment techniques can yet handle the problem of peak order changes.

The R package is available in Bioconductor and some functionalities are already available in Workflow4Metabolomics (Guitton *et al.* , 2017). Last but not least,

the selection of optimal parameters as well as the set up of an optimal procedure for method selection is beyond the scope of this article. Methodological needs in this field (Engel *et al.* , 2013) open up vast research possibilities.

2.6 Software availability

PepsNMR has been first made available on GitHub and it is now available on the <https://bioconductor.org> repository (Huber *et al.* , 2015) and some functionalities are implemented on the <https://workflow4metabolomics.org> Galaxy platform.

Note that the additional R routines used to validate the pre-processing workflow can also be accessed in the MBXUCL R package located on GitHub (<https://github.com/ManonMartin/MBXUCL>). Datasets that were used in this article are available at various processing stages in an *ad hoc* package PepsNMRData, also available on Bioconductor at <https://bioconductor.org/packages/PepsNMRData>.

2.6.1 Bioconductor

Bioconductor project started in 2001. It is an open source software for bioinformatics with the aim of promoting tools for the analysis and comprehension of high-throughput omics data that leads to reproducible research. It is essentially based on the R statistical programming language and is an open development software where researchers are invited to contribute, share, collaborate and connect the tools that are necessary to achieve their research. These tools are categorised by *BiocViews*, which has four main categories:

- *Software*: dedicated to data analysis
- *Annotation Data*: associate omics data with biological metadata
- *Experiment Data*: illustrate software functionality based on a dataset
- *Workflow*: describe a workflow of several Bioconductor packages dedicated to a specific field of application

The software has both development and release versions: the release is updated twice a year, whereas the development version includes new features and new packages prior to their release. Note that during release, all package versions are updated accordingly.

All the source R code from the packages are maintained under git version control.

2.6.2 Workflow4Metabolomics

Workflow4Metabolomics (W4M) (Guitton *et al.*, 2017) is a web-interface dedicated to metabolomic data processing, analysis and annotation, where PepsNMR works complementarily with other developed tools (*cf.* Figure 2.11). The nine first PepsNMR functions (from data importation until zeroing of negative intensities) are available on this platform in order to read and pre-process raw FIDs before data analysis.

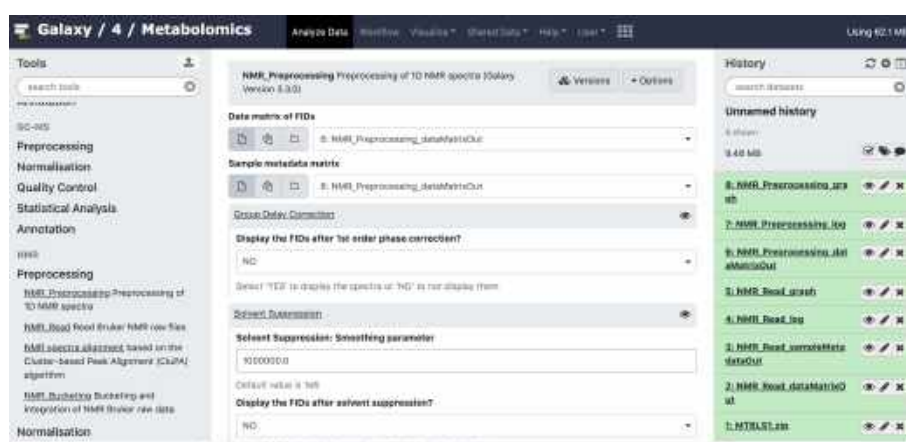


Figure 2.11 – PepsNMR integrated into Workflow4Metabolomics for NMR raw data read and pre-processing.

2.6.3 Use PepsNMR in practice

The release version of PepsNMR is provided via Bioconductor: <http://bioconductor.org/packages/PepsNMR>. The user can also access the devel version, either from GitHub (<https://github.com/ManonMartin/PepsNMR>) or from Bioconductor:

<https://www.bioconductor.org/packages/devel/bioc/html/PepsNMR.html>

The code to install the (release version of the) package is the following:

```
# install release version
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")

BiocManager::install("PepsNMR")
```

PepsNMR has common R packages characteristics: *help* pages indicate how to use the pre-processing functions and the datasets present in PepsNMRData by applying the function `help(x)` around a user function or dataset `x`. In addition, other sources of documentation can be found: a vignette with an extensive practical pre-processing application is attached to the package and is accessed with:

```
# vignette  
browseVignettes("PepsNMR")
```

A short ReadMe.md file on the GitHub homepage of the package also provide basic installation and operating instructions.

2.7 Datasets

The package is applied to two real datasets from well-known human biofluid: serum (HSerum) and urine (HUrine), to demonstrate its performance. These datasets were both designed to collect multiple measures from the same experimental unit and capture other technical sources of variation, allowing the comparison of inter- and intra-unit variability. These datasets are available upon request

2.7.1 Human Serum

In the HSerum dataset, a blood sample was collected from 4 different donors. For each sample, 8 sub-samples were measured across 8 days with one sub-sample from each donor per day and permutations according to a latin hypercube sampling method. The total number of available FID signals is then $4 \times 8 = 32$. Data were acquired with a 500 MHz Bruker Avance spectrometer equipped with a TCI cryoprobe and using a CPMG relaxation-editing sequence with pre-saturation (number of scans: 32). Spectra are labelled as: JxDx where Jx is the day of measurement and Dx is the donor label. More details about this dataset can be found in Féraud *et al.* (2015).

2.7.2 Human Urine

In the second dataset, morning urine from 3 different donors was collected and directly aliquoted in eppendorfs with the following procedure: for each donor, 4 aliquotes containing 400 μ l of urine each + 300 μ l of D2O buffer + 10 μ l of TMSP and 4 aliquotes containing 320 μ l of urine + 380 μ l of D2O buffer and 10 μ l of TMSP. Hence, each donor sample was divided into 2 sub-samples: one was kept pure and the other 25 % diluted. The 4 aliquotes of each dilution were analysed on 4 different days. For each day, the order of measurement was held

constant across the 6 sub-samples. 1D spectra were recorded on a Avance III NMR spectrometer operating at 700.28 MHz for the proton signal acquisition, and with the NOESYPR1D (1D Nuclear Overhauser effect spectroscopy with water pre-saturation) sequence (number of scans: 64). . The total number of collected FID signals is then: $3 \times 2 \times 4 = 24$. Spectra are labelled as: SxDx where x is the donor label and D is the dilution (0: no dilution; 1: 25 % diluted).

2.7.3 Spectral pre-processing and outlier removal

Both datasets were pre-processed with PepsNMR (version 1.3.2) and a combination of TopspinTM 3.1 and AMIX 3.9.14 for manual pre-processing. With PepsNMR, the full advised workflow was applied with default parameters. The region removal and zone aggregation steps were adapted given the type of medium. For region removal, the water (4.5 – 5.1 ppm) area was removed for the serum while the water, urea and maleic acid (4.5 – 6.1 ppm) areas were removed for the urine medium. The zone aggregation step was only applied to the urine dataset for the citrate peak (2.5 – 2.7 ppm). The manual pre-processing included the following sequence: exponential apodization, Fourier transform, first and zero order phase corrections, internal calibration, polynomial baseline correction, bucketing and spectral region removal.

Outliers were detected with the ROBPCA (Hubert *et al.* , 2005) approach, a robust PCA used to detect spectra with a high score distance and/or with a high orthogonal distance to the PCA space. Two outlying observations were removed from the HSerum dataset (J6D4 and J6D3) and one from the HUrine dataset (S1D1E2). Note that a missing observation in the manual pre-processing dataset (J4D3) was completely removed for data analysis.

CHAPTER 3

Feature Selection in metabolomics with PLS-derived methods

3.1 Introduction

3.1.1 Feature selection in metabolomics

Compared to the budgetary limited number of experimental samples, modern high-throughput technologies produce an enormous amount of variables. In such wide and short data tables, computational and statistical challenges are substantial to extract all the meaningful information from them.

There is a myriad of variable selection techniques (see Guyon & Elisseeff (2003) for a comprehensive review of the different kinds of feature selection) that have been developed accordingly, some are domain-specific while others are more general. They all lead to a subset selection of relevant variables called *features*. One main purpose of Feature Selection (FS) is to identify potential biomarkers to answer the biological questions of interest. Furthermore, after their selection, these potential biomarkers should latter be double-checked by an independent and targeted confirmatory study. Note that in this chapter, *features* is defined as the set of variables selected by the discriminant techniques whereas *biomarkers* involve the whole metabolites' descriptors, e.g. 16 true features corresponds to 4 different true biomarkers.

Variable selection techniques can be distinguished into three main categories: filters, wrappers or embedded (Guyon & Elisseeff, 2003). They are either ranking the features or selecting a subset of them. *Filters* are independent of the classifier. They will assign a ranking to the variables that is based on a predefined univariate or multivariate importance measure. For their part, *wrappers* are classifier-specific. They evaluate several subsets of interest based on prediction performances. Since the number of potential subset partitions is too important to examine each of them, an iterative search procedure is implemented to only evaluate the most interesting ones (e.g. with backward/forward selection). The latter category, *embedded*, lies in between filters and wrappers: the selection procedure is integrated to the training procedure of the classifier as e.g. in LASSO regression.

In the area of feature selection for biomarker discovery, and depending on the domain of study, the researcher must first ask himself: "what do I want as features?" and "What is a good feature?". The algorithms are usually designed to select a minimal set of uncorrelated and relevant features that will increase the predictive performances of the model to detect e.g. a patient state. This approach is the *minimal-optimal problem* (Shi *et al.*, 2018). However another possibility exists. It selects all the relevant features e.g. to explain complex biological mechanisms with various varying parameters. This second approach is called the *all-relevant problem* (Shi *et al.*, 2018).

Beside their high dimensionality and inherent noise, metabolomics data are particularly challenging to be used for feature selection: they contain a high number of correlated features since metabolites are spanned across multiple descriptors and they can be correlated to each other. Both cases of feature selection (minimal-optimal and all-relevant) could therefore be interesting for biomarker discovery with these data. Moreover, it is well known that algorithms performances vary depending on the goals of the study and the dataset at hand (Determan Jr., 2015).

In this research domain, Partial Least Squares (PLS, Abdi, 2010) and Orthogonal PLS (OPLS, Trygg & Wold, 2002b) regression are the current preferred method to build a prediction model and perform feature selection.

Metabolomics is derived from analytical chemistry, hence traditional statistical tools such as PLS that were applied to this field - *i.e.* chemometrics - have been directly transposed to this young research field. In addition, the prediction efficiency is usually employed to assess model performances in metabolomics. Since multivariate calibration is the workhorse of chemometrics, it is easy to understand why evaluating such predictive performances are as much present in the literature, even in the context of feature selection. To explain the widespread use of PLS in metabolomics, Gromski *et al.* (2015) also argue that it comes from its availability in well-known statistical software packages, with an easy

implementation and directly interpretable statistics (e.g. the prediction accuracy), although it can lead to misleading results (*cf.* e.g. Figure 1.6) when used by non-experts.

A review of existing filter, wrapper and embedded selection approaches that have been combined with PLS is given in Mehmood *et al.* (2012). Among the most popular selection techniques are filters based on the regression coefficients or the Variable Importance in Projection (VIP). Other common approaches are embedded techniques where PLS is constrained by a form of regularisation.

3.1.2 Objectives

Recent articles suggested various feature selection approaches in metabolomics (see for example Rinaudo *et al.* (2016), Grissa *et al.* (2016) or Shi *et al.* (2018)). Nevertheless, they are all based on experimental datasets for which the exact identification of features cannot be exactly verified.

Similarly to Rousseau (2011) and Christin *et al.* (2013), the intent of this chapter is to compare in an original fashion the methods performance, based on a list of known discriminant features and under different realistic conditions (different sample sizes and different degrees of signal to noise ratio).

To that purpose, we used spectra derived from a large set of “control” samples from rat urine, and we artificially altered some of them. Using such baseline spectra renders possible a genuine extension to experimental data. In this work, resampling techniques are performed to evaluate the true performances from a statistical point of view and undercover the methods properties. It further enables to study, within and between methods, the stability of features and meta-parameters of the classifiers as well as the prediction accuracies. This evaluation scheme can directly be extended to other feature selection techniques.

We decided to compare prevailing methods used for feature selection in metabolomics, namely PLS and OPLS along with their most natural extensions to embedded techniques: Sparse-PLS (SPLS, Chun & Keleş, 2010) and Light Sparse OPLS (LSOPLS, Féraud *et al.* , 2017).

Knowing that classifier performance is heavily dependent on the dataset at hand, this chapter does not intent to give general recommendations on which supervised approaches should be routinely used. It rather gives insights on how the compared methods are behaving, on keys to select an appropriate method, and on how to evaluate them in the scope of feature selection.

3.1.3 Content

This chapter is organised as follows: Section 3.2 describes first the Semi-artificial database, then Section 3.3 reviews the methods that are used to find potential features of interest in the spectral profiles. The next part, Section 3.4 describes how the Meta Parameters (MP) are tuned to train the classifiers. Section 3.5 presents classic data analysis results by applying the previously described methods to one subsample. Afterward, Section 3.6 presents a generic evaluation of those methods, based on resampled sub-datasets and multiple criteria. Finally, Section 5.6 ends this chapter by concluding remarks.

3.2 The Semi-artificial urine database

The Semi-Artificial Urine database is fully described in the global Appendix C. In this dataset, artificial alterations were added to control (untreated) rat urine spectra in order to assess the quality of techniques to identify potential biomarkers when the biomarkers' location is known in advance. This raw dataset is issued from the COMET database (Lindon *et al.*, 2003). The final dataset is composed of a spectral data matrix with 960 samples and 500 spectral descriptors where half of the spectra were altered. The other half of the data are original spectra. It also includes a binary vector characterising the presence of these alterations and it corresponds to the class to be predicted. Details about pre-processing and alterations of the spectra are given in Appendix C.

The semi-artificial dataset involves 46 artificially modified descriptors defined as true biomarkers separated into 10 different areas of the spectrum (*cf.* Figure 3.1), some of these are correlated with each other, leaving a total of 6 independent groups of biomarkers. Furthermore, each spectrum is only altered by 2 out of the 6 randomly chosen independent biomarkers, to mimic various responses from one organism to another. Alterations on the left side of the spectra (ranging from 7.26 to 6.48 ppm) are located in a "noise-free" area and have a low intensity. In contrast, alterations present on the right side of the spectral window (6.48 to 3.37 ppm) are located in a noisy area but have a higher intensity.

Due to the added alteration, built-in correlation is then present between and within them: correlation between descriptors of the same peak but also correlated pairs of alterations since the addition of alterations was designed to always add 2 different biomarkers in each altered spectrum. This is intended to mimic correlation between different peaks of a same molecule or between peaks of correlated metabolites.

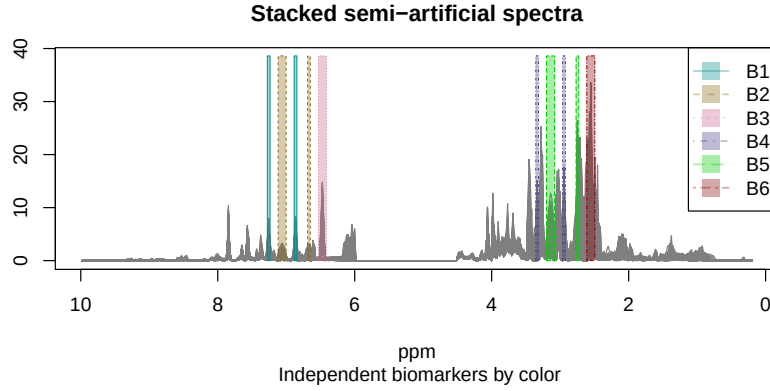


Figure 3.1 – Stacked representation of the spectral profiles in the Semi-Artificial Urine database. The 6 alterations' locations are outlined with the colour boxes.

3.3 Methods

Consider a data matrix $\mathbf{X}$ of size $(n \times m)$ with n observations and m variables, and a univariate response vector $\mathbf{y}$ of size $(n \times 1)$ representing a binary class indicator $\{0, 1\}$ in the context of a discrimination problem, e.g. coding for the presence/absence of alterations in the semi-artificial urine spectra. Note that to simplify the mathematical notations, we will now consider $\mathbf{X}$ and $\mathbf{y}$ as being their mean-centred-by-column equivalents.

Every method is described below with its statistical properties and algorithms, as well as how it performs feature selection in this context. Note that they are all presented in the case of a single $\mathbf{y}$ response vector but could easily be generalised to the multivariate case of a matrix response $\mathbf{Y}$.

From a single classifier, there can be multiple feature selection procedures. For each of them, we will define a *Feature Score* vector $\mathbf{B}$ of size $(1 \times m)$ as the value on which the decision to select or not the variables will be based. Then a decision rule will be defined to determine the set of retained variables as the active variables – the *features* – that will be contained in $\mathcal{A}$ of size $(1 \times m^*)$ with m^* the number of retained variables.

The distinction between filters and embedded methods will disappear altogether, in favour of a clearer homogenised definition of what is stored in $\mathbf{B}$ and the decision rules to determine the variables in $\mathcal{A}$. They will be explained in each methods section and resumed in Table 3.7.

In Section 3.6, they are later compared to each other through a resampling method with a double loop scheme.

3.3.1 Multiple t-tests

Description

Univariate t-tests are applied as benchmark method (as in Rousseau *et al.* (2008) or Christin *et al.* (2013)) to be compared against the multivariate classifiers. Though this is not a classification technique *per se*, the test statistic and its p-value are good indicators of the differences between two classes of observations for a given variable. This test is still widely applied in bioinformatics applications, typically to measure differential expression in Microarray Data (Jeanmougin *et al.* , 2010; Kuyuk & Ercan, 2017), single-cell RNA-Seq (Poirion *et al.* , 2016), etc.

The test statistic is applied to each spectral descriptor. It measures the difference in means of two different samples and, based on that value, one can formally test if the difference is significant.

Because several tests are performed simultaneously, the overall Type I error is inflated. Hence, the derived p-values need a correction factor for multiple tests. The Benjamini-Yekutieli (B-Y) approach (Benjamini & Yekutieli, 2001) controls the False Discovery Rate (FDR) under arbitrary dependency assumptions, which is appropriate in our case since spectral descriptors are correlated. It provides a less conservative control of Type I errors, on the contrary to procedures used to control the family-wise error rate such as the Bonferroni correction.

Statistical methodology

For each descriptor $i = 1, \dots, m$, the two-sample Welch's t-statistic for unequal variances is:

$$\delta_i = \frac{\bar{x}_{0i} - \bar{x}_{1i}}{\sqrt{s_{0i}^2/n_0 + s_{1i}^2/n_1}} ; \quad (3.1)$$

where n_0 and n_1 are the sample sizes of each class, $\bar{x}_{0j}$ and $\bar{x}_{1j}$ are the sample means, and s_{0j}^2 and s_{1j}^2 are the sample variances.

Under H_0 , the Welch's t-statistic follows a Student's t-distribution τ_ν with ν degrees of freedom approximated by the Welch-Satterthwaite equation¹. The corresponding p-values are calculated for each variable as: $\rho_i = 2P(\tau_\nu > |\delta_i|)$ with $i = 1, \dots, m$.

¹This equation estimates the effective pooled degrees of freedom of a linear combination of independent sample variances:
$$\frac{(\frac{s_{0i}^2}{n_0} + \frac{s_{1i}^2}{n_1})^2}{(\frac{s_{0i}^4}{n_0^2(n_0-1)} + \frac{s_{1i}^4}{n_1^2(n_1-1)})}$$

The B-Y FDR controlling procedure is the following: ρ is the vector of p -values. The B-Y rule to select significant test statistics consists in finding the index m^* such that:

$$m^* = \max \left\{ i : \rho_{(i)} \leq \frac{iq}{m \sum_{i=1}^m 1/i} \text{ for } i = 1, \dots, m \right\}; \quad (3.2)$$

where q is the FDR to be controlled (usually set to 10 or 5 %) and (i) corresponds to the index i of ordered ρ in descending order. It rejects the null hypotheses for ordered $i = 1, \dots, m^*$. Alternatively, the corrected p -value ρ_{fdr} can be directly compared to the FDR level q .

Feature selection

If based on the test statistic (t-stat), the Feature score is $B^{t-stat} = |\delta|$ and an arbitrary threshold m^* determines which descriptors are included in the selection set: $\mathcal{A}^{t-stat} = \{i : |\delta|_{(i)} \geq |\delta|_{(m^*)}\}$. On the other hand, if the decision rule relies on the p -value (t-fdr), the Feature score remains the same ($B^{t-fdr} = |\delta|$) but only variables with $\rho_{fdr} \leq 0.05$ are kept: $\mathcal{A}^{t-fdr} = \{i : \rho_{(i),fdr} \leq 0.05\}$.

3.3.2 Partial Least Squares

Description

The original Partial Least Squares (PLS) regression as well as the Nonlinear estimation by Iterative Partial Least Squares (NIPALS) algorithm were first introduced by Herman Wold in the 1960s-1970s (Geladi, 1988). They were later combined in Sjöström *et al.* (1983) to solve ill-conditioned linear regression problems with a high number of multi-collinear variables. PLS can be defined as a combinaison of dimension reduction and a prediction tool that is, as for Principal Component Analysis (PCA), aimed at extracting a small number of orthogonal Latent Variables (LV) – also called Predictive Components (PrC) here – that are linear combinations of the original variables to best summarise the available information. Unlike PCA that only maximises the X variance explained, here both explained variances of X and y are maximised, as well as the correlation between these two matrices.

Besides the original NIPALS algorithm, several others have been suggested in the literature (see Varmuza & Filzmoser (2016) for more details about the most common ones and Andersson (2009) for a comparison between various algorithms) such as Kernel PLS (Lindgren *et al.*, 1993), Canonical Powered PLS (Indahl *et al.*, 2009) or Statistically Inspired Modification of PLS (SIMPLS) (de Jong, 1993)

algorithms, all with varying scores and loadings properties and different ways to find the solution to the maximisation problem.

Here, a distinction arises between having a single or multivariate response (Boulesteix & Strimmer, 2006). In the case of a single response vector, the method is called PLS1. In contrast, to predict one or multiple responses, the two main variants are: PLS2 (obtained usually with NIPALS or Kernel-PLS algorithms) and SIMPLS. Both multivariate versions are equivalent, though PLS2 has different and less intuitive constraints on the X -weights (ter Braak & de Jong, 1998). SIMPLS can also be seen as the direct multivariate extension of PLS1, as it shares the same criterion and the same constraints (Boulesteix & Strimmer, 2006).

PLS is therefore a widely applied method for modelling relations between dependent (y) and independent (X) variables (see Mehmood & Ahmed (2016) for examples of applications). The technique is also popular and extensively used in chemometrics and metabolomics (Wold *et al.*, 2001; Antti *et al.*, 2004) where it has proven its usefulness and efficiency to underline metabolic changes in biofluids *e.g.* due to toxicity or disease process.

This technique was developed so as to better handle the overdetermined regression problems usually addressed with Principal Component Regression (PCR). In this situation, one needs to predict a set of dependent variables from a very large set of correlated predictors (Abdi, 2010), resulting in a much higher degree of multicollinearity, which is almost systematically the case in the field of metabolomics. Indeed, PLS is an improvement of the classic Multiple Linear Regression and PCR as the method is more robust in its parameters estimation (Geladi & Kowalski, 1986) and is explicitly built with the incentive to predict y . However, Wentzell & Vega Montoto (2003) demonstrated in their literature review and simulations based on complex chemical mixtures that, besides resulting in a reduced number of LV, PLS appeared to have no predictive advantage over PCR. Therefore, it seems that the popularity of PLS comes from a greater interpretability, in particular of the loadings, rather than an improved predictive ability.

Originally, PLS was not intended to solve classification problems. Nevertheless, Barker & Rayens (2003) showed that the method is well suited to fit that particular application. It gives good separation results emphasising their benefits compared to unsupervised methods such as PCA². In this case, the response variables are coded with binary values, typically $\{0, 1\}$. When the class response has $F > 2$ levels, y is converted into a series of F binary vectors $\tilde{y}_1, \dots, \tilde{y}_F$ gathered in a matrix of size $(n \times F)$, such that for $i = 1, \dots, n$ and $f = 1, \dots, F$:

²Since the dimension reduction of PLS is guided by the variability existing among groups rather than by the total variability of the data for PCA.

$$\tilde{y}_{if} = \begin{cases} 1 & \text{if } y_i = f \\ 0 & \text{otherwise.} \end{cases}$$

Statistical methodology

Principles

PLS is fundamentally a dimension reduction technique based on a basic latent structure of the data (Singular Value Decomposition), that is built for each block of data $\mathbf{X}$ and $\mathbf{y}$ and between them.

Its purpose is to find the final linear relationship between $\mathbf{X}$ and $\mathbf{y}$ such that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^{pls} + \mathbf{e}$ with the regression coefficients vector $\boldsymbol{\beta}^{pls}$ of size $(m \times 1)$ and an error vector $\mathbf{e}$ of size $(n \times 1)$. To that purpose, PLS will build a new orthogonal variables (the LV estimates), with corresponding X -scores, $\mathbf{t}_j$ with $j = 1, \dots, a$, which are linear combinations of the original $\mathbf{X}$ variables, as well as good predictors of $\mathbf{y}$. The model components are extracted sequentially: every iteration deflates the variability from the data that is associated with the extracted component. The remaining information is used to estimate the next component, etc.

$\mathbf{X}$ can be decomposed into scores, loadings and weights with the following relations:

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E}_X ; \quad (3.3)$$

where $\mathbf{T}$ is the $(n \times a)$ X -scores matrix, $\mathbf{P}$ is the $(m \times a)$ X -weights matrix. Their product is considered a good “summary” of $\mathbf{X}$ if one obtains small X -residuals $\mathbf{E}_{X(n \times m)}$.

In the scope of estimating orthogonal X -scores, $\mathbf{T}$ are defined as linear combinations of the original variables $\mathbf{X}$ with $\mathbf{W}_{(m \times a)}$ the X -weights matrix (also denoted as the set of *direction vectors*):

$$\mathbf{T} = \mathbf{X}\mathbf{W}. \quad (3.4)$$

Calculating those weights represents the key computation of the PLS methodology.

During the estimation process, $\mathbf{P}$ is estimated by OLS regression of $\mathbf{T}$ on $\mathbf{X}$: $\mathbf{P}' = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{X}$.

Similarly, $\mathbf{y}$ is defined as the addition of summarised information (through scores and loadings) and an error terms:

$$\mathbf{y} = \mathbf{U}\mathbf{q} + \mathbf{e}_y ; \quad (3.5)$$

where $\mathbf{U}$ is the $(n \times a)$ matrix of y -scores, $\mathbf{q}$ is the $(a \times 1)$ vector of y -loadings and $\mathbf{e}_y$ is the $(n \times 1)$ vector of y -residuals.

For the same technical reasons as in Eq. (3.4), the y -weights vector $\mathbf{c}$ of size $(1 \times a)$ needs to be introduced: $\mathbf{U} = \mathbf{y}\mathbf{c}$.

These weights are estimated by OLS regression of $\mathbf{T}$ on $\mathbf{y}$: $\mathbf{c} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{y}$.

The two data tables are linked by the following equation:

$$\mathbf{y} = \mathbf{T}\mathbf{c} + \mathbf{e} = \mathbf{X}\mathbf{W}\mathbf{c} + \mathbf{e} = \mathbf{X}\boldsymbol{\beta}^{pls} + \mathbf{e}. \quad (3.6)$$

And therefore, regression coefficients are estimated as:

$$\mathbf{b}^{pls} = \mathbf{W}\mathbf{c} = \mathbf{W}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{y}.$$

PLS can be seen as a constrained maximisation problem to derive direction vectors $\mathbf{w}_j$. In the case of one response vector, the objective function is:

$$\begin{aligned} \mathbf{w}_j &= \operatorname{argmax}_{\mathbf{w}} \mathbf{w}'\mathbf{X}'\mathbf{y}\mathbf{y}'\mathbf{X}\mathbf{w} \text{ for } j = 1, \dots, a ; \\ &= \operatorname{argmax}_{\mathbf{w}} \operatorname{cov}^2(\mathbf{X}\mathbf{w}, \mathbf{y}) \text{ for } j = 1, \dots, a ; \\ &= \operatorname{argmax}_{\mathbf{w}} \operatorname{cor}^2(\mathbf{X}\mathbf{w}, \mathbf{y}) \operatorname{var}(\mathbf{X}\mathbf{w}) \operatorname{var}(\mathbf{y}) \text{ for } j = 1, \dots, a ; \end{aligned} \quad (3.7)$$

Hence, it is clear that the direction vectors contained in $\mathbf{W}$ seek to capture most of the X -variance as well as a high correlation between $\mathbf{X}$ and $\mathbf{y}$.

Note that the space spanned by the direction vectors $\mathbf{w}_j$ is an orthogonal basis for the Krylov³ subspace: $K_a(\mathbf{X}'\mathbf{X}, \mathbf{X}'\mathbf{y})$. It implies that PLS estimator can be seen as the solution of the following optimisation problem (Rosipal & Krämer, 2006): $\mathbf{b} = \operatorname{argmin}_{\mathbf{b}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$, subject to $\mathbf{b} \in K_a(\mathbf{X}'\mathbf{X}, \mathbf{X}'\mathbf{y})$. PLS can therefore be viewed as a regularised least squares fit or shrinkage estimator in a similar way than Ridge (Krämer, 2007). The directions of the estimator that are responsible for a high variance are indeed shrunk. This introduces a bias, which is hopefully not too large and compensated by a lowered variance, hence a lowered Mean Square Error of the estimators. Therefore, compared to classic Ordinary Least Squares (OLS) estimators for $\boldsymbol{\beta}$, PLS leads to lowered estimated

³Krylov methods are iterative methods especially suited for wide and short data tables, used to solve a large system of linear equations by avoiding matrix-matrix computations (Ipsen & Meyer, 1998).

coefficients.

Details of the algorithms

Recall that in PLS, the properties of the scores, loading, etc. are subject to the choice of the applied algorithm. NIPALS and SIMPLS are reviewed here in details as they will be used to build our classifiers.

Here are a few remarks concerning these two algorithms:

- The scores vectors are usually orthogonal to each other but not necessarily normalised. Hence, $\mathbf{T}'\mathbf{T}$ is diagonal. In SIMPLS, they are orthonormal *i.e.* $\mathbf{t}_j^{simpls} \mathbf{t}_j^{simpls} = 1 \forall j = 1, \dots, a$ and $\mathbf{T}'^{simpls} \mathbf{T}^{simpls} = \mathbf{I}_a$.
- The loadings vectors $\mathbf{p}$ are usually not orthogonal to each other (if the scores are) and the norm of each $\mathbf{p}$ will depends on the algorithm.
- The SIMPLS loading matrix $\mathbf{P}^{simpls}$ can be obtained from the NIPALS one as follows: $\mathbf{P}^{simpls} = \mathbf{P}^{nipals} (\mathbf{T}'^{nipals} \mathbf{T}^{nipals})$.
- If the SIMPLS algorithm is used, the weight matrix $\mathbf{W}$ that can be applied to the original mean-centred $\mathbf{X}$ matrix can be directly obtained, and we omit the ^{simpls} notation in this case. If NIPALS is used, $\mathbf{W}^{nipals}$ is originally related to the deflated $\mathbf{X}$ matrix. To recover $\mathbf{W}$ from the original NIPALS weights matrix $\mathbf{W}^{nipals}$, the following formula is applied: $\mathbf{W} = \mathbf{W}^{nipals} (\mathbf{P}'^{nipals} \mathbf{W}^{nipals})^{-1}$.
- For univariate $\mathbf{y}$, NIPALS and SIMPLS are equivalent to each other (de Jong, 1993; ter Braak & de Jong, 1998).

NIPALS algorithm

NIPALS algorithm has been the first algorithm to solve the PLS problem. Scores and loadings are obtained iteratively component-wise with the deflation of the $\mathbf{X}$ matrix at the end of each iteration. The latter is a crucial step in order to obtain orthogonal scores vectors (Varmuza & Filzmoser, 2016).

Its corresponding constraints to be applied to the optimisation function in Equ. (3.7) are:

Subject to $\mathbf{w}'_j(\mathbf{I}_m - \mathbf{W}\mathbf{W}^+)\mathbf{w}_j = 1$ and $\mathbf{t}'_j\mathbf{t}_k = \mathbf{w}'_j\mathbf{X}'\mathbf{X}\mathbf{w}_k = 0$ for $j = 1, \dots, a$ and $k = 1, \dots, j-1$;
 where $\mathbf{W}^+$ is the unique Moore-Penrose inverse of $\mathbf{W}$.

(3.8)

NIPALS algorithm for univariate $\mathbf{y}$ is detailed below:

Table 3.1 – NIPALS algorithm for a single response.

A. Extract PLS components		
(1)	Initialisation: $\mathbf{X}_1 = \mathbf{X}$	Start from the initial mean-centred data matrix $\mathbf{X}$ to be deflated. Recall that $\mathbf{y}$ is also mean-centred.
(2)	Iterate for $j = 1, \dots, a$:	
(2.1)	$\mathbf{w}_j = \mathbf{X}'_j\mathbf{y}(\mathbf{y}'\mathbf{y})^{-1}$	Calculate the weights vector $\mathbf{w}_j$ by regressing $\mathbf{X}_j$ on $\mathbf{y}$
(2.2)	$\mathbf{w}_j = \mathbf{w}_j(\mathbf{w}'_j\mathbf{w}_j)^{-1/2}$	Normalize $\mathbf{w}_j$
(2.3)	$\mathbf{t}_j = \mathbf{X}_j\mathbf{w}_j$	Calculate scores vector $\mathbf{t}_j$ as a linear combination of original variables
(2.4)	$\mathbf{p}_j = \mathbf{X}'_j\mathbf{t}_j(\mathbf{t}'_j\mathbf{t}_j)^{-1}$	Calculate corresponding loadings vector by regressing $\mathbf{X}$ on $\mathbf{t}_j$
(2.5)	$\mathbf{X}_{j+1} = \mathbf{X}_j - \mathbf{t}_j\mathbf{p}'_j$	Deflate matrix $\mathbf{X}_j$
B. Build global matrices		
	$\mathbf{T} = (\mathbf{t}_1, \dots, \mathbf{t}_a)$	The $(n \times a)$ matrix of scores
	$\mathbf{\Sigma} = (\mathbf{p}_1, \dots, \mathbf{p}_a)$	The $(m \times a)$ matrix of loadings
	$\mathbf{W}^{nipals} = (\mathbf{w}_1, \dots, \mathbf{w}_a)$	The $(m \times a)$ "deflated" form of matrix of weights
	$\mathbf{W}$	The $(m \times a)$ "undeflated" form of the weights matrix
	$\mathbf{W}^{nipals}(\sum' \mathbf{W}^{nipals})^{-1} =$	
C. Calculate the regression coefficients		
	$\mathbf{b}^{pls} = \mathbf{W}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{y}$	

Note that the NIPALS scores matrix $\mathbf{T}$ can be directly calculated from non deflated matrix $\mathbf{X}$ by using the "undeflated" form of the weight matrix: $\mathbf{T} = \mathbf{X}\mathbf{W}$.

SIMPLS algorithm

The strength of SIMPLS is that it is not just a series of iteration steps, as it was first developed to solve an optimisation problem. In that regard, SIMPLS has a direct statistical meaning. It has further advantages of computational speed and direct computation of T in terms of linear combinations of the X matrix, and we will now focus on this algorithm.

Its corresponding constraints on the optimisation function are:

$$\text{Subject to } \mathbf{w}_j' \mathbf{w}_j = 1 \text{ and } \mathbf{t}_j' \mathbf{t}_k = \mathbf{w}_j' \mathbf{X}' \mathbf{X} \mathbf{w}_k = 0 \text{ for } j = 1, \dots, a \text{ and } k = 1, \dots, j-1. \quad (3.9)$$

After the estimation of $\mathbf{w}_j$ and once the latent variables $\mathbf{t}_j$ are estimated, the loadings $\mathbf{p}_j$ can be calculated via OLS regression: $\mathbf{t}_j = \mathbf{X} \mathbf{p}_j + \mathbf{e}$.

The entire SIMPLS algorithm for univariate $\mathbf{y}$ is detailed here:

Note that an advantage of SIMPLS is that the weights $\mathbf{W}$ are directly related to $\mathbf{X}$ (and not calculated from the deflated matrices as in NIPALS) and the deflation Step (2.6) sets by construction the orthogonality constraint on the scores vector $\mathbf{t}_j$ to all previous ones extracted.

PLS results interpretation

The geometrical representation of PLS is depicted in Figure 3.2: the $\mathbf{X}$ matrix is projected onto an a -dimensional hyper-plane so as the coordinates of the projection, which turns out to be the X -scores, $\mathbf{t}_j$, appear to be predictors of $\mathbf{y}$.

The scores ($\mathbf{t}_j$) indicate how are projected the observations onto the latent variables and their (dis)similarities, while the loadings ($\mathbf{p}_j$) indicate the contribution of a particular $\mathbf{X}$ variable to the above-mentioned latent variable.

One can also inspect the percentage of information from the response $\mathbf{y}$ that is explained by each component of the model SSY_j . In the case of PLS1, $SSY_j = \mathbf{c}_j^2 \mathbf{t}_j' \mathbf{t}_j$ for $j = 1, \dots, a$.

Weight matrices $\mathbf{W}$ and $\mathbf{c}$ provide information about how the variables are combined in order to form the relation between $\mathbf{X}$ and $\mathbf{y}$. Furthermore, the X -weights $\mathbf{w}_j$ serve to express both “positive” and “compensation” correlations respectively between $\mathbf{X}$ and $\mathbf{y}$ and needed to predict $\mathbf{y}$ from $\mathbf{X}$.

Regression coefficients $\mathbf{b}$, which are independent of the algorithmic details of the method, give a global view summarising the influence of all PLS components over the response variable.

Table 3.2 – PLS1/SIMPLS algorithm for a single response.

A. Extract PLS components		
(1)	Initialisation: $s_1 = X'y$	Start from the cross-product of mean-centred X and y matrices to be deflated <i>i.e.</i> their initial covariance vector
(2)	Iterate for $j = 1, \dots, a$:	
(2.1)	$w_j = s_j(s_j's_j)^{-1/2}$	Calculate the weights vector w_j from the normalised s_j vector
(2.2)	$t_j = Xw_j$	Calculate scores vector t_j as a linear combination of the original variables
(2.3)	$t_j = t_j(t_j't_j)^{-1/2}$	Normalize t_j
(2.4)	$w_j = w_j(t_j't_j)^{-1/2}$	Adapt weights w_j to normalised scores
(2.5)	$p_j = X't_j(t_j't_j)^{-1} X't_j$	Calculate corresponding loadings vector by regressing X on t_j
(2.6)	$s_{j+1} = s_j - \sum_{i=1}^j (\sum_{i=1}^j s_i' s_i)^{-1} \sum_{i=1}^j s_i' s_j$	Deflate covariance vector s_{j+1} as the residuals of the regression of s_j on previous loadings $\sum_{i=1}^j p_i$ where $\sum_{i=1}^j = (p_1, \dots, p_j)$
B. Build global matrices		
	$T = (t_1, \dots, t_a)$	The $(n \times a)$ matrix of scores
	$P = (p_1, \dots, p_a)$	The $(m \times a)$ matrix of loadings
	$W = (w_1, \dots, w_a)$	The $(m \times a)$ matrix of weights
C. Calculate the regression coefficients		
	$b^{pls} = WT'y$	

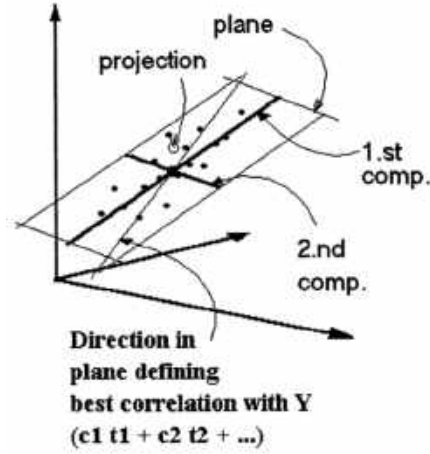


Figure 3.2 – Geometrical representation of a PLS model (Figure from Wold *et al.* (2001)).

Graphical representations of the weights w_j and c_j , loadings p_j , scores t_j and the regression coefficients b and combinations of them are therefore expected to allow a comprehensive interpretation of the model.

Finally, the residuals that are not explained by the model can be used for model diagnostics and outliers detection.

Feature selection

The first and more common Feature selection with PLS is based on the regression coefficients b^{pls} that can be employed to rank spectral descriptors: $B^{PLS-coef} = |b^{pls}|$. In this case, an arbitrary threshold is set to select the m^* higher absolute coefficients and $\mathcal{A}^{pls-coef} = \{i : |b^{pls}|_{(i)} \geq |b^{pls}|_{(m^*)}\}$. The disadvantage of this measure is the dependence to an arbitrary scale and variance of the $\mathbf{X}$ columns.

Another solution would be to rely on the Variable Importance in Projection (VIP) (Farrés *et al.*, 2015) value v . This is a well-known criterion applied in PLS-derived techniques for variable selection. It summarises the influence of an individual variable to build the model. They are measured on the basis of a weighted sum of the Sum of Squares for explained variation in the response (SSY). The VIP value for the i^{th} variable is given as:

$$v_i = \sqrt{\frac{m \cdot \sum_{j=1}^a [SSY_j \cdot (w_{ij} / \|\mathbf{w}_j\|)^2]}{SSY_{tot}}}; \quad (3.10)$$

where w_{ij} represents the X -weight of variable i and LV j , SSY_j is the Sum of Squares of the y -variance explained by the j^{th} LV and SSY_{tot} is the Total Sum of Squares of the dependent variable y : $SSY_{tot} = \sum_{j=1}^a SSY_j$.

Since $(w_{ij} / \| \mathbf{w}_j \|)^2$ represents the importance of the i^{th} variable in the construction of the j^{th} LV, the VIP value is a global measure of the contribution of this variable in the whole PLS model.

Given that the average of the squared VIP scores is equal to 1⁴, the rule of thumbs is to select variables having a VIP value greater than 1.

These VIP values will be used as well as features ranking. The decision rule to include a variable i in $\mathcal{A}$ is either based on a predefined number of features m^* to retain (pls-vip), or on the rule of thumbs threshold of 1 (pls-vip1) with $m^* = \max\{i : |v^{pls}| \geq 1\}$ for $i = 1, \dots, m$. More formally, $\mathcal{A}^{pls-vip} = \{i : |v^{pls}|_{(i)} \geq |v^{pls}|_{(m^*)}\}$ and $\mathcal{A}^{pls-vip1} = \{i : |v^{pls}| \geq 1\}$.

3.3.3 Orthogonal Partial Least Squares

Description

Multivariate tools and projection methods such as PLS are widely used in metabolomics, though they remain strongly affected by the occurrence of systematic orthogonal variations - also called structured noise - in $\mathbf{X}$ that is not associated with the response variable y . Indeed, the structured noise in the data matrix coming from spectrometric measures can affect the precision of a new sample prediction and deteriorate the model robustness at some point (Trygg & Wold, 2002b).

To overcome this issue, Wold *et al.* (1998) first developed the idea to build a filter that only removes from the spectral matrix $\mathbf{X}$ the portion that is orthogonal to the dependent variable and thus obtain more concise and interpretable models. In this first investigation, they developed a filter called Orthogonal Signal Correction (OSC). Several different algorithms have since been proposed and compared, in particular in Svensson *et al.* (2002). The latter concluded that OSC filters do not improve predictive abilities but do lead to more parsimonious models, with an easier interpretation of the filtered data. An additional and interesting advantage is the possibility to analyse the structured noise that is extracted from the data.

One well known OSC variant, called Orthogonal Projections to Latent Structures or Orthogonal PLS (OPLS) has emerged in chemometrics. It has been first introduced by Trygg & Wold (2002b) and can be seen as a preprocessing

⁴Proof: $\frac{\sum_{i=1}^m (v_i)^2}{\sum_{i=1}^m (w_{ij} / \| \mathbf{w}_j \|^2)} = \frac{m}{m} \cdot \sum_{j=1}^a \sum_{i=1}^m \frac{[SSY_j \cdot (w_{ij} / \| \mathbf{w}_j \|^2)]}{SSY_{tot}} = 1 \cdot \frac{\sum_{j=1}^a SSY_j}{SSY_{tot}} = 1$.
 $\sum_{j=1}^a \sum_{i=1}^m (w_{ij} / \| \mathbf{w}_j \|^2)^2 = 1 \cdot 1 \cdot \sum_{j=1}^a 1 = 1$; since $\sum_{i=1}^m (w_{ij} / \| \mathbf{w}_j \|^2)^2 = 1$.

technique or as part of an integrated filter in a classic PLS model. OPLS extends the original NIPALS-PLS regression algorithm, which is already known for its usefulness in multiple application areas thanks to the fact that it can cope with variables collinearity, noise in the data and can deal with a moderate amount of missing values. Indeed, a filtering step is added to remove the y -orthogonal variation from the X -scores. Thus, the variance of interest is extracted into a single predictive component from the variance that is unrelated (or orthogonal, defined by multiple Orthogonal Components (OC)) to the response variable. This approach results in a rotation of the standard PLS model so that the variance of interest is focused into the predictive component and the variance unrelated to the tested hypothesis is filtered into orthogonal components.

A comparison of geometrical representations of PLS and O-PLS models in a 2D-plane with two components can be seen on Figure 3.3 where the OPLS model is a rotation of the PLS components.

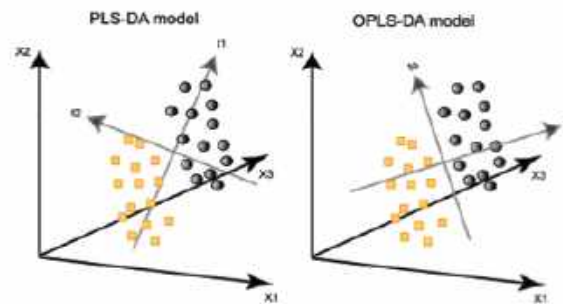


Figure 3.3 – Comparison between geometrical representations of PLS and OPLS models (Figure from Wiklund *et al.* (2008)). t_p represents the predictive component while t_o is a y -orthogonal component.

Similarly to PLS, OPLS can also be applied to discriminant problems (Bylesjö *et al.*, 2006). In this case, the between-class variation resides in the predictive component whereas the within-class variation is contained in the y -orthogonal components.

Statistical methodology

As for PLS, OPLS creates a new orthogonal variables that are the estimates of the latent variables and correspond to the X -scores, t_j , which are linear combinations of the original X variables for $j = 1, \dots, a$. Among the X -scores, only the last one is predictive, t_p , and hence a good predictor of y . The other X -scores t_{jo}

do not contain information from $\mathbf{y}$ and are designed as y -orthogonal X -scores for $j = 2, \dots, a - 1$.

The final predictive model is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^{opls} + \mathbf{e} = \mathbf{X}_p\boldsymbol{\beta}_p^{opls} + \mathbf{e}; \quad (3.11)$$

where $\mathbf{X}_p$ is the filtered $\mathbf{X}$ data matrix and $\boldsymbol{\beta}_p^{opls}$ are the predictive regression coefficients linked to $\mathbf{X}_p$.

As for PLS, the OPLS model can also be seen as a set of relations that define either separately $\mathbf{X}$ and $\mathbf{y}$ or their relationship to each other.

The information contained in the dependent variable $\mathbf{y}$ is used to break down the matrix $\mathbf{X}$ into three different blocks: one that contains the common variation, another that only involves variation only for $\mathbf{X}$, i.e. the structured noise, and the last part corresponding to the residual variance (Bylesjö *et al.*, 2006).

$$\mathbf{X} = \mathbf{t}_p\mathbf{p}_p' + \mathbf{T}_{ortho}\mathbf{P}_{ortho}' + \mathbf{E}_X; \quad (3.12)$$

with the predictive and orthogonal X -scores *resp.* $\mathbf{t}_p$ of size $(n \times 1)$ and $\mathbf{T}_{ortho}$ of size $(n \times (a - 1))$. Note that $\mathbf{T}_{ortho} \cdot \mathbf{y} = 0$, i.e. the orthogonal X -scores are orthogonal to $\mathbf{y}$. The predictive scores $\mathbf{t}_p$ are estimated on the basis of the predictive X -weights: $\mathbf{t}_p = \mathbf{X}\mathbf{w}_p$.

Similarly, $\mathbf{y}$ is defined as the addition of summarised information and an error terms:

$$\mathbf{y} = \mathbf{u}q + \mathbf{e}_y; \quad (3.13)$$

Where $\mathbf{u}$ is the $(n \times 1)$ vector of y -scores, q is here a constant term (due to the single $\mathbf{y}$ and single predictive component) and corresponds to the y -loading and $\mathbf{e}_y$ of size $(n \times 1)$ is the vector of y -residual.

Since $\mathbf{t}_p$ is a good predictor of $\mathbf{y}$, we have the following inter-block relationship (as in Eq. (3.6) for PLS):

$$\mathbf{y} = \mathbf{t}_p\mathbf{c} + \mathbf{e} = \mathbf{X}_p\boldsymbol{\beta}_p^{opls} + \mathbf{e} = \mathbf{X}\boldsymbol{\beta}^{opls} + \mathbf{e} \quad (3.14)$$

$\boldsymbol{\beta}_p^{opls}$ can now be estimated as: $\mathbf{b}_p^{opls} = \mathbf{w}_p\mathbf{c} = \mathbf{w}_p\mathbf{y}'\mathbf{t}_p(\mathbf{t}_p'\mathbf{t}_p)^{-1}$. To predict a new unfiltered centred observation, the regression coefficients need to be back-transformed as: $\mathbf{b}^{opls} = (\mathbf{I}_m - \mathbf{W}_o\mathbf{P}_o')\mathbf{b}_p^{opls}$, where $\mathbf{P}_o'$ and $\mathbf{W}_o$ will be described below.

OPLS algorithm

The OPLS algorithm is made of 2 different operating blocks : (1) the orthogonalisation of $\mathbf{X}$ by extracting step by step one or several orthogonal

components and (2) the derivation of predictive components based on the corrected predictive matrix $\mathbf{X}_p$ as in a regular PLS model. In the case of a univariate $\mathbf{y}$, only one predictive PLS component is extracted in the final model.

Its algorithm is based on slight modifications of the original NIPALS-PLS algorithm and is fully described in Tables 3.3 and 3.4.

Note that orthogonal scores $\mathbf{T}_o$ can be directly calculated from the non deflated matrix $\mathbf{X}$ by using the “undeflated” form of the orthogonal weight matrix: $\mathbf{T}_o = \mathbf{X}\mathbf{W}_o$ and $\mathbf{P}_o = \mathbf{X}_o'\mathbf{T}_o(\mathbf{T}_o'\mathbf{T}_o)^{-1}$.

OPLS results Interpretation

Loadings and scores can be interpreted similarly to PLS except that they are now separated into y -predictive and y -orthogonal categories of variations of the data matrix $\mathbf{X}$. Indeed, the OPLS-DA can effectively set apart the actual direction that discriminates $\mathbf{t}_p$ from the y -orthogonal directions $\mathbf{T}_{ortho}$, which renders the corresponding vector of loadings $\mathbf{p}_p$ straightforward to interpret.

Scores and loadings linked to the predictive variation indicate relationships between the spectra on the basis of this predictive component in the model (scores) as well as the contribution of a spectral descriptor in the construction of this component (loadings).

The vectors of orthogonal loadings aim at identifying the remaining structured variation that is associated with high intra-class variability. The orthogonal variation contained in $\mathbf{X}_{ortho}$ can further be decomposed by PCA (Trygg & Wold, 2002b) since the scores and loadings are no more related to the dependent variable $\mathbf{y}$. The PCA on orthogonal variation gives hints about the identity and potential causes of the encountered structured noise in the data.

As for PLS, the weight vectors $\mathbf{w}_p$ and $\mathbf{c}$ indicate how the variables are combined together to build up the link between $\mathbf{X}$ and $\mathbf{y}$. Note that the weight vector $\mathbf{w}_p$ of the OPLS-pretreated PLS model is equivalent to the first weight of the initial PLS model since $\mathbf{w}_p$ is the projection of the $\mathbf{X}$ matrix onto $\mathbf{y}$ (if single) and orthogonal columns in $\mathbf{X}$ do not influence this projection (Trygg & Wold, 2002b).

The regression coefficients vector $\mathbf{b}^{opls}$ of size $(m \times 1)$ is again referred as a measure of the influence of a particular variable over the response variable $\mathbf{y}$.

Eventually, the scatter plot of both covariance (Eq. 3.15) and correlation (Eq. 3.16) between the variables and the unique predictive component $\mathbf{t}_p$ which is called the *S-plot*, helps visualising the influence of a specific variable in the model construction.

For $i = 1, \dots, m$:

$$cov(\mathbf{t}_p, \mathbf{X}_i) = \frac{\mathbf{t}_p' \mathbf{X}_i}{n-1}; \quad (3.15)$$

Table 3.3 – Orthogonalisation step of the OPLS algorithm for a single response.

A.	Orthogonalisation step	Compute the part of $\mathbf{X}$ orthogonal to $\mathbf{y}$ where both matrices are mean-centred.
(1)	Initialisation: $\mathbf{X}_1 = \mathbf{X}$	Start from the initial mean-centred data matrix $\mathbf{X}$ to be deflated
(2)	Iterate for $j = 1, \dots, (a - 1)$	First 4 steps are identical than the NIPALS-PLS algorithm
(2.1)	$\mathbf{w}_j = \mathbf{X}_j' \mathbf{y} (\mathbf{y}' \mathbf{y})^{-1}$	Calculate weight vector $\mathbf{w}_j$ by regressing $\mathbf{X}_j$ on $\mathbf{y}$
(2.2)	$\mathbf{w}_j = \mathbf{w}_j (\mathbf{w}_j' \mathbf{w}_j)^{-1/2}$	Normalize $\mathbf{w}_j$
(2.3)	$\mathbf{t}_j = \mathbf{X}_j \mathbf{w}_j$	Calculate score vector $\mathbf{t}_j$ as a linear combination of original variables
(2.4)	$\mathbf{p}_j = \mathbf{X}_j' \mathbf{t}_j (\mathbf{t}_j' \mathbf{t}_j)^{-1}$	Calculate corresponding loading vector by regressing $\mathbf{X}_j$ on $\mathbf{t}_j$
(2.5)	$\mathbf{w}_{oj} = \mathbf{w}_j - (\mathbf{w}_j' \mathbf{p}_j) \mathbf{p}_j$	Calculate orthogonal weight vector as the residual of the regression of $\mathbf{w}_j$ on $\mathbf{p}_j$
(2.6)	$\mathbf{w}_{oj} = \mathbf{w}_{oj} (\mathbf{w}_{oj}' \mathbf{w}_{oj})^{-1/2}$	Normalize $\mathbf{w}_{oj}$
(2.7)	$\mathbf{t}_{oj} = \mathbf{X}_j \mathbf{w}_{oj}$	Calculate orthogonal score vector $\mathbf{t}_{oj}$
(2.8)	$\mathbf{p}_{oj} = \mathbf{X}_j' \mathbf{t}_{oj} (\mathbf{t}_{oj}' \mathbf{t}_{oj})^{-1}$	Calculate corresponding orthogonal loading vector by regressing $\mathbf{X}_j$ on $\mathbf{t}_{oj}$
(2.9)	$\mathbf{X}_{j+1} = \mathbf{X}_j - \mathbf{t}_{oj} \mathbf{p}_{oj}'$	Deflate matrix $\mathbf{X}_j$ by removing the orthogonal component j And go to step(1) to derive the next orthogonal component
B.	Build global matrices	Compute global matrices and $\mathbf{X}_p$ filtered which will be used to predict $\mathbf{y}$
	$\mathbf{T}_o = (\mathbf{t}_{o1}, \dots, \mathbf{t}_{o(a-1)})$	The $(n \times (a - 1))$ matrix of orthogonal scores
	$\mathbf{P}_o = (\mathbf{p}_{o1}, \dots, \mathbf{p}_{o(a-1)})$	The $(m \times (a - 1))$ matrix of orthogonal loadings
	$\mathbf{W}_o^{nipals} = (\mathbf{w}_{o1}, \dots, \mathbf{w}_{o(a-1)})$	The “deflated” form of the $(m \times (a - 1))$ matrix of orthogonal weights
	$\mathbf{W}_o = \mathbf{W}_o^{nipals} (\sum_{j=1}^a \mathbf{W}_j^{nipals})^{-1}$	The “undeflated” form of the orthogonal weight matrix
	$\mathbf{X}_o = \mathbf{T}_o \sum_{j=1}^a \mathbf{p}_j'$	The $(n \times m)$ orthogonalised $\mathbf{X}$ matrix
	$\mathbf{X}_p = \mathbf{X} - \sum_{j=1}^a \mathbf{X}_o$	The “filtered” $\mathbf{X}$ matrix without the part orthogonal to $\mathbf{y}$

Table 3.4 – PLS step of the OPLS algorithm for a single response.

C.	PLS of y on X_p	Compute a PLS with one component on X_p
(1)	$w_p^{nipals} = X_p' y (y' y)^{-1}$	Calculate weight vector w_p from deflated X
(2)	$w_p^{nipals} = w_p (w_p' w_p)^{-1/2}$	Normalize w_p
(3)	$t_p = X_p w_p$	Calculate prediction score vector t_p
(4)	$p_p = X_p' t_p (t_p' t_p)^{-1}$	Calculate corresponding loading vector
(5)	$w_p = w_p^{nipals} (p_p' w_p^{nipals})^{-1}$	The “undeflated” form of the predictive weight vector
(6)	$c = y' t_p (t_p' t_p)^{-1}$	Regress y on t_p
(7)	$b_p^{opls} = w_p c$	Derive the regression coefficients
(8)	$b^{opls} = (I_m - W_o \sum_p') b_p^{opls}$	Derive the corrected regression coefficients to be applied to the unfiltered mean-centred X matrix with the “undeflated” orthogonal weights W_o

$$\text{corr}(\mathbf{t}_p, \mathbf{X}_i) = \frac{\text{cov}(\mathbf{t}_p, \mathbf{X}_i)}{s_{\mathbf{t}_p} s_{\mathbf{X}_i}} ; \quad (3.16)$$

where $s_{\mathbf{t}_p}$ and $s_{\mathbf{X}_i}$ are the estimated standard deviations of *resp.* $\mathbf{t}_p$ and $\mathbf{X}_i$.

It is meant to take into account the reliability (correlation) and contribution (covariance) of the spectral variables to the model. Hence, it helps to identify effective descriptors in the class separation problem that are considered as potential biomarkers. Indeed, only focusing on descriptors with high absolute covariance will lead to the selection of biomarkers with a high spectral intensity while selecting only on the basis of correlations will lead to biomarkers with very low concentration and increase the risk of false positives *i.e.* Type I error in the identification of descriptors as potential biomarkers (Wiklund *et al.* , 2008).

Biomarkers identification

As in PLS, the absolute value of regression coefficients can be used to rank the variables and select the most related ones to variations in the response vector: $\mathbf{B}^{opls-coef} = |\mathbf{b}^{opls}|$ and m^* is set arbitrarily and $\mathcal{A}^{opls-coef} = \{i : |b^{opls}|_{(i)} \geq |b^{opls}|_{(m^*)}\}$. The VIP value is also considered for variable ranking in OPLS. It is now measured based on the predictive X -scores, X -weights and y -loadings. the feature scores are $\mathbf{B}^{opls-vip} = \mathbf{B}^{opls-vip1} = \mathbf{v}^{opls}$ and the corresponding decision rules from PLS are transposed here in the context of OPLS: $\mathcal{A}^{opls-vip} = \{i : |v^{opls}|_{(i)} \geq |v^{opls}|_{(m^*)}\}$ and $\mathcal{A}^{opls-vip1} = \{i : |v^{opls}| \geq 1\}$.

3.3.4 Sparse PLS

Description

Given its potential to select discriminant variables, strong arguments are in favour of a LASSO-type penalisation (Tibshirani, 1996) to affect the PLS regression coefficients.

The usual purpose of variable selection is to pick a restricted set of strongly meaningful potential biomarkers. It enables a better interpretability of the biological study results and displays biomarkers that are really important to take into consideration. The method is considered as an embedded FS method with the advantage of the number of selected features which is no more dependent of an arbitrary threshold parameter, as usually in (O)PLS.

Another rationale for sparse regression is argued in Chun & Keleş (2010) for high dimensional data. The consistency of PLS estimators have been proven in Naik & Tsai (2000) under normality assumptions on $\mathbf{X}$ and $\mathbf{y}$, consistency of covariances $\text{cov}(\mathbf{X}, \mathbf{y})$ and $\text{cov}(\mathbf{X}, \mathbf{X})$ and an additional condition (see Chun & Keleş (2010)

for a detailed description). The latter requires m to be fixed and much smaller than n . When dealing with multivariate metabolomics data in practice, this prerequisite is, however, rarely achieved. Chun & Keleş (2010) extended their result to more common usage situations where the sample size is small compared to the number of variables. They investigated estimators consistency for the case where m is allowed to grow at an appropriate rate. Their finding is that additional conditions are necessary for consistent sample covariance estimates: m should grow much slower than n . However, it cannot be met in the common paradigm of large m / small n datasets. Therefore, and despite their ability to resume data information by a few LV, the PLS estimators cannot fully handle the curse of dimensionality of such omics data.

In this context, two Sparse Partial Least Squares (SPLS) algorithms have been developed independently and in parallel by Chun & Keleş (2010) and Lê Cao *et al.* (2008), the latter being integrated into the broader MixOmics collaborative project for multi-omics (Rohart *et al.* , 2017). Both techniques are based on a LASSO regularisation technique on the direction vectors (*i.e.* w_j , the X -weights).

We decided to apply the algorithm from Chun & Keleş (2010), which is directly based on a penalised version of the original SIMPLS optimisation problem. Besides, their method has been proven to ensure interesting theoretical properties, such as the one of Krylov subspace that ensures the convergence of the algorithm.

Statistical methodology

Chun & Keleş (2010) observed that noise variables altered the direction vectors w_j , creating a form of inconsistency. Therefore, they decided to impose sparsity on those vectors, rather than directly on the regression coefficients.

The SPLS method is based on the SIMPLS algorithm. The constrained optimisation problem from PLS-SIMPLS (Eqs. (3.7) and (3.9)) is modified accordingly by imposing an additional L_1 sparsity constraint as in LASSO regression (James *et al.* , 2013).

$$w_j = \operatorname{argmax}_w w' X' y y' X w \quad \text{s.t.} \quad w_j' w_j = 1, \quad t_j' t_k = w_j' X' X w_k = 0 \\ \text{and } |w_j|_1 \leq \lambda_1 ; \quad (3.17)$$

for $j = 1, \dots, a$, $k = 1, \dots, j - 1$, and λ_1 is the thresholding parameter that fine-tunes the amount of sparsity.

Based on a generalisation of the SPCA formulation from Zou *et al.* (2006), this criterion can be reformulated with penalties imposed onto a surrogate direction

vector $\mathbf{w}_j^*$ rather than on $\mathbf{w}_j$ directly, while keeping them close to each other. Under this formulation, it ensures the LASSO property of shrinkage to exact zero:

$$\begin{aligned} \operatorname{argmin}_{\mathbf{w}, \mathbf{w}^*} & -\kappa \mathbf{w}' \mathbf{M} \mathbf{w} + (1 - \kappa)(\mathbf{w}^* - \mathbf{w})' \mathbf{M}(\mathbf{w}^* - \mathbf{w}) + \lambda_1 |\mathbf{w}^*|_1 + \lambda_2 |\mathbf{w}^*|_2 \\ \text{s.t. } & \mathbf{w}' \mathbf{w} = 1 ; \end{aligned} \quad (3.18)$$

The first term maximises the covariance between $\mathbf{y}$ and $\mathbf{X}$, whereas the second term ensures that $\mathbf{w}$ and $\mathbf{w}^*$ are close enough to each other. The L_2 sparsity constraint (with parameter λ_2) takes care of potential singularities in $\mathbf{M}$, as in Ridge regression (James *et al.*, 2013). $\kappa \in [0, 0.5]$ is a fixed parameter set to avoid local minima issues. In SPLS, λ_2 is set to ∞ and the solution to this optimisation problem has the form of a soft thresholding estimator.

SPLS algorithm

For the sake of simplicity, the algorithm is explained here for a single $\mathbf{y}$ response. In such case, the solution to Eq. (3.18) is obtained by soft thresholding of $\mathbf{w}$:

$$\mathbf{w} = (\tilde{\mathbf{s}} - \lambda_1/2) \mathbf{I}(\tilde{\mathbf{s}} \geq \lambda_1/2) \operatorname{sign}(\tilde{\mathbf{s}}) = (\tilde{\mathbf{s}} - \lambda_1/2)_+ \operatorname{sign}(\tilde{\mathbf{s}}) ; \quad (3.19)$$

where

$$(x)_+ = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise;} \end{cases}$$

and $\tilde{\mathbf{s}} = \frac{\mathbf{X}'\mathbf{y}}{\|\mathbf{X}'\mathbf{y}\|}$, which corresponds to the first direction vector of SIMPLS. This solution is now independent from the meta-parameter κ .

A form of soft thresholded estimators to derive $\mathbf{w}$ that only retains components which are greater than a predefined fraction of the maximum component is proposed by the authors:

$$\mathbf{w} = (|\tilde{\mathbf{s}}| - \eta \cdot \max_{1 \leq i \leq m} |\tilde{\mathbf{s}}_i|)_+ \operatorname{sign}(\tilde{\mathbf{s}}) ; \quad (3.20)$$

where $\eta \in [0, 1]$ is equivalent to the sparsity parameter λ_1 from Eq. (3.19).

Note that in the case of a multivariate response matrix $\mathbf{Y}$, The solution to Eq. (3.18) is obtained by alternatively iterating between solving for $\mathbf{w}$ for fixed $\mathbf{w}^*$ using Lagrange multipliers, and inversely, solving for $\mathbf{w}^*$ for fixed $\mathbf{w}$ using the LARS algorithm (Efron *et al.*, 2004).

The described objective function to derive the first direction vector could in theory be used to derive the subsequent ones. However, this straightforward approach does not guaranties the Krylov subsequences structure of vectors in $\mathbf{W}$ that will ensure

the algorithmic convergence of the sparse solution. To achieve this property, the SPLS fits a PLS regression with the non-zero w_j variables (called active variables) $\mathbf{X}_{\mathcal{V}}$ with the covariates contained in subspace $\mathcal{V}$. Hence, $\mathbf{W}$ will form a Krylov subsequence on the subspace of active variables.

The full algorithm is described in Table 3.5.

Table 3.5 – The SPLS algorithm for a single response.

A. Compute SPLS components		
(1)	Initialisation: $\mathbf{y}_0 = \mathbf{y}_1 = \mathbf{y}$ and $\mathbf{b}_0^{spls} = \mathbf{0}$	Define initial mean-centred $\mathbf{X}$ matrix. Recall that $\mathbf{y}$ is mean-centred as well here. Set to 0 the original vector of regression coefficients
(2)	Iterate for $j = 1, \dots, a$	
(2.1)	$ \tilde{\mathbf{s}}_j = \mathbf{X}'\mathbf{y}_j / \ \mathbf{X}'\mathbf{y}_j\ $	Calculate and norm the cross-product of $\mathbf{X}$ and $\mathbf{y}_j$
(2.2)	$\mathbf{w}_j = \frac{(\tilde{\mathbf{s}}_j - \eta \cdot \max_{1 \leq i \leq m} \tilde{s}_{ij}) \mathbf{I}(\tilde{\mathbf{s}}_j \geq \eta \cdot \max_{1 \leq i \leq m} \tilde{s}_i)}{ \tilde{\mathbf{s}}_j } \text{sign}(\tilde{\mathbf{s}}_j)$	Find the direction vector $\mathbf{w}_j$
(2.3)	$\mathcal{V} = \{i : w_{ij} \neq 0\} \cup \{i : b_{i(j-1)} \neq 0\}$	Update the active variables set with $i = 1, \dots, m$
(2.4)	Fit the PLS model $\mathbf{y}_j = \mathbf{X}_{\mathcal{V}} \boldsymbol{\beta}^{pls} + \mathbf{e}$	With PLS-SIMPLS with j LV on the selected variables: $\mathbf{X}_{\mathcal{V}} = \{\mathbf{X}_i, i \in \mathcal{V}\}$. Estimated regression coefficients $\mathbf{b}^{pls}$ are used in the next steps
(2.5)	$b_{ij}^{spls} = \begin{cases} b_i^{pls} & \text{if } i \in \mathcal{V} \\ 0 & \text{otherwise;} \end{cases}$	Update $\mathbf{b}_j^{spls}$: unselected variables are zeroed, selected variables are given by previous PLS regression
(2.6)	$\mathbf{y}_{j+1} = \mathbf{y}_0 - \mathbf{X} \mathbf{b}_j^{spls}$	Deflation step over the response, the new response $\mathbf{y}_{j+1}$ is the residual response

SPLS results interpretation

The interpretation of the outputs produced by SPLS is equivalent to the PLS one since SIMPLS is applied on the set of the active variables to fit the model. The main difference in the results with PLS is the presence of regression coefficients with exact zero values.

Biomarkers identification

In SPLS, the regression coefficients are commonly used for FS, and here $B^{spls-coef} = |b^{spls}|$. The decision rule to keep variable i is $A^{spls-coef} = \{i : |b^{spls}| \neq 0\}$ and the number of features is automatically fixed accordingly.

3.3.5 Light Sparse OPLS

Description

A recent article from Féraud *et al.* (2017) suggested to combine the previously described OPLS and SPLS into the Light Sparse OPLS (LSOPLS). The idea was to take advantage of these two approaches for biomarkers discovery: sparsity and orthogonal variation filtering. The authors showed that this new approach could lead to improved predictive performances by retaining only a small number of predictors.

Statistical methodology

Similarly to OPLS, LSOPLS is composed of two subsequent stages: an orthogonalisation and a modelling step where only predictive variability is retained to explain the dependent variable. The orthogonalisation step is derived from some of the original OPLS algorithm and the modelling step consists in applying SPLS on the previously filtered out predictive matrix.

Note that for predictive purposes, the coefficients β_p^{lsopls} need to be corrected into β^{lsopls} in order to be applied directly to the original predictive matrix X , as in OPLS. The uncorrected coefficients β_p^{lsopls} are still used for variable selection, since those coefficients are the sparse ones, contrarily to β_p^{lsopls} .

LSOPLS results interpretation

The interpretation results are similar to the ones of the previous approaches, since the produced outputs are from an SPLS model that is described previously.

Biomarkers identification

The uncorrected regression coefficients constitute the LSOPLS feature scores $B^{lsopls-coef} = |b_p^{lsopls}|$. As for SPLS, the subset of selected variables is: $A^{lsopls-coef} = \{i : |b^{lsopls}| \neq 0\}$.

Table 3.6 – The LSOPLS algorithm for a single response.

A. Orthogonalisation step	Apply steps (1) and (2) from OPLS algorithm with $(a - 1)$ orthogonal components and mean-centred $\mathbf{X}$ and $\mathbf{y}$.
B. Build global matrices	Build the following global matrices: $\mathbf{P}_o$, $\mathbf{W}_o$ and $\mathbf{X}_p$ derived from Step A
C. SPLS	Apply SPLS on the filtered matrix $\mathbf{X}_p$
(1) Fit the SPLS model $\mathbf{y} = \mathbf{X}_p \boldsymbol{\beta}_p^{spls} + \mathbf{e}$	With SPLS-SIMPLS where $\mathbf{X} = \mathbf{X}_p$ and for 1 PrC.
(2) $\mathbf{b}_p^{lsopls} = \mathbf{b}_p^{spls}$	Define the sparse LSOPLS coefficients as the ones of the SPLS model above
(2) $\mathbf{b}^{lsopls} = (\mathbf{I}_m - \mathbf{W}_o \mathbf{P}_o') \mathbf{b}_p^{lsopls}$	Derive the non sparse regression coefficients to be applied to unfiltered $\mathbf{X}$

3.3.6 Summary of the feature selection procedures

In Table 3.7 is summarised $\mathbf{B}$'s content and the decision rule to selected variables as discriminant features.

The decision rule depends either on an arbitrary threshold or on a predefined one. When arbitrary, m^* is set to the number of true features, *i.e.* $m^* = 46$.

3.4 Meta-parameters tuning

Except for multiple t-tests, all supervised methods have meta-parameters (MP) to be tuned in order to fit – without overfitting – a parsimonious model to the data at hand, to ensure predictive power and a good model generalisation. For PLS and SPLS, it corresponds to the number of PrC. Whereas, for OPLS and LSOPLS, it corresponds to the number of orthogonal LV, since only one predictive LV is considered in their case. In addition, sparse methods (SPLS and LSOPLS) have a second meta-parameter to adjust: the degree of sparsity η .

In practice, due to the usually restricted number of data samples, their tuning is achieved by cross-validation (CV) on a *training set*. This training set represents a certain percentage of the total dataset (*e.g.* 80 %), the remaining part being defined as the *test set*. Note that in practice, with a limited number of observations,

Table 3.7 – Summary of the elements in $\mathcal{B}$ (Index) for $i = 1, \dots, m$; the decision rule to include the variable in $\mathcal{A}$ and the methods abbreviation in this chapter.

Underlying model	Feature element	score	Decision rule for $i \in \mathcal{A}$	FS method abbrev.
t -test	$ \delta _{(i)}$		$i \leq m^*$	t-stat
	$ \delta _i$		$\rho_{i, fdr} \leq 0.05$	t-fdr
PLS	$ b^{pls} _{(i)}$		$i \leq m^*$	pls-coef
	$v_{(i)}$		$i \leq m^*$	pls-vip
	v_i		$v_i \geq 1$	pls-vip1
OPLS	$ b^{opls} _{(i)}$		$i \leq m^*$	opls-coef
	$v_{(i)}$		$i \leq m^*$	opls-vip
	v_i		$v_i \geq 1$	opls-vip1
SPLS	$ b^{spls} _i$		$b_i^{spls} \neq 0$	spls-coef
LSOPLS	$ b^{lsopls} _i$		$b_i^{lsopls} \neq 0$	lsopls-coef

the training set often represents 100% of the total dataset and no test set is defined. The training set is then divided into k subgroups of data and a model is built based on samples from $k - 1$ groups. The model then predicts the remaining samples that were not used in the model construction and evaluates the prediction performances based on a predefined criterion. The most frequent CV segmentation is the k -folds CV with $k = 10$. It divides at random the training dataset into 10 different folds and is usually preferred to others to prevent a possible overfit. Note that if $k = n$, the technique is called Leave-One-Out CV.

The final predictive performances are then evaluated on the basis of the test set that was left aside, whose response is now predicted by the adjusted model on the training set with the chosen meta-parameters.

In the regression context where PLS-derived methods have been developed, the Root Mean Square Error (RMSE) that estimates the error of prediction is an usual tuning criterion to evaluate predictive performances (Wehrens, 2011). Its name becomes RMSECV if the criterion is used in CV or RMSEP if it is used to estimate the prediction errors of an independent test set.

$$RMSE = \sqrt{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / n}; \quad (3.21)$$

where $\hat{y}_i$ is the predicted response of observation i .

Other well known derived criteria are the Predictive Residual Sum of Squares $PRESS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ and $Q^2 = 1 - \frac{PRESS}{SST}$ where SST stands for Total Sum of Squares and $SST = \sum_{i=1}^n y_i^2$, with mean-centred y . They can be both used in CV or validation as well and are all equivalent.

In the context of classification problems, other criteria such as the discriminant Q2 (Westerhuis *et al.*, 2008) or the Area Under the ROC Curve (AUC) may seem more appropriate to evaluate the classification performances. According to our own previous trials, meta-parameter tuning was slightly better handled by the AUC (which will be described latter in this chapter), which, in a sense, is the more natural way to evaluate the classification fit. Note that an exception is made to decide on the number of orthogonal components. In this case, the RMSE is used as suggested by Trygg & Wold (2002b). Though this issue is beyond the scope of this chapter, finding an appropriate criterion in the classification context remains an open question and would deserve to be investigated further.

3.5 Application of the methods to a sub-sample

In order to complete the description of the methods, they have been applied to one sub-sample (with $n = 200$) from the analysed semi-artificial dataset and the results are presented here. 10-fold CV has been applied to decide on the MP that needed optimisation. The chosen parameters are presented in Table 3.8. One can see that the total number of components is in the same order of magnitude for PLS and OPLS. SPLS and OPLS have much different MPs. In order to discuss the MP selection for sparse methods that have both the number of components and the sparsity, one can inspect the two graphs presented in Figure 3.4. They illustrate the AUC and the number of selected features m^* given fixed MPs for SPLS. The AUC increase through the MPS is not even approaching monotonicity as it is the case for univariate MP selection. Highest values are met at various combinations of $\{a, \eta\}$. Though AUC is an objective and adapted criterion, the wining MPs combinaison is plural and heterogeneous, and so may be the resulting sets of selected features. On the other hand, the impact of the MPs' choice on the size of the feature subset is very clear: a larger sparsity constraint or a lower number of components logically leads to a smaller set of selected variables.

Table 3.8 – Selected meta-parameters for supervised classifiers.

	PLS	OPLS	SPLS	LSOPLS
# PrC	7	1	11	1
# OC	-	6	-	12
Sparsity (η)	-	-	0.8	0.25

Scores and loadings plots are traditionally interpreted to visually assess the links between: the samples, the original variables, and each other, in a reduced space of a few components (PrC or OC). Scores plot enables to see if structured information or clusters emerge from those samples *i.e.* if they are well separated according

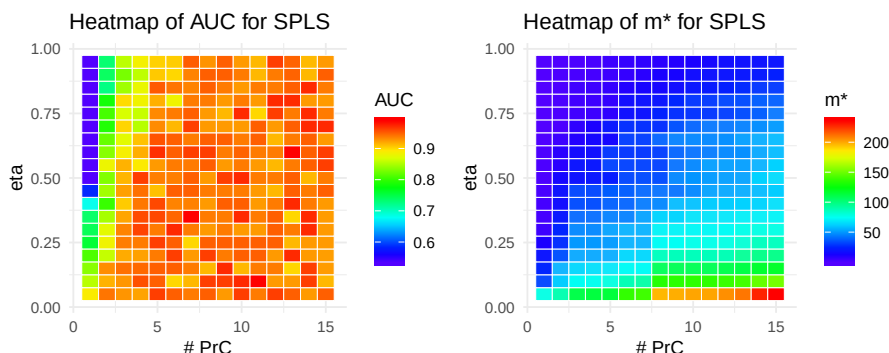


Figure 3.4 – Heatmaps of AUC and m^* obtained by 10-fold CV for SPLS, and given the values of the MPs: the number of LV a and sparsity η .

to their class label, or to detect another structure in the data unrelated to the response of interest. It is also particularly useful to detect outliers. These plots are displayed in Figure 3.5 for all the studied multivariate methods, plus PCA. The latter was not able to discriminate the two classes of spectra along PC1-2. The four supervised methods were mainly able to recover the group structure from the data along the first component (LV1). However, the observations are only well separated by the combination of LV1 and LV2 for (S)PLS methods. Indeed, methods with orthogonal filters ((LS)OPLS) only need one component to discriminate between the two classes. Moreover, the visualisation of the orthogonal components in (LS)OPLS is informative: it helps to identify other structured information than the class membership, to understand if the samples were appropriately prepared, acquired and pre-processed before their actual data analysis.

Since discrimination between the spectra is mainly driven by the first PrC, the main loading of interest will also be the first one, and we will now concentrate on this one. Figure 3.6 represents the loadings from the corresponding first component for each method. Their magnitude represents the importance of each original variable in the construction of the latent variables. Only the amplitude is important here, as well as the relative direction within a loading, and not the overall direction of that loading. The sign of the latter is set arbitrarily during the components' construction. For all methods, higher loading values are present on the right side of the spectral window where half of the true biomarkers are located. The first two axes in PCA is not built on the discriminant variables. PLS and OPLS have a few high values, representing the true biomarkers, and a lot of smaller intensities, representing false discoveries ; whereas SPLS and LSOPLS have fewer low values.

Scores plots

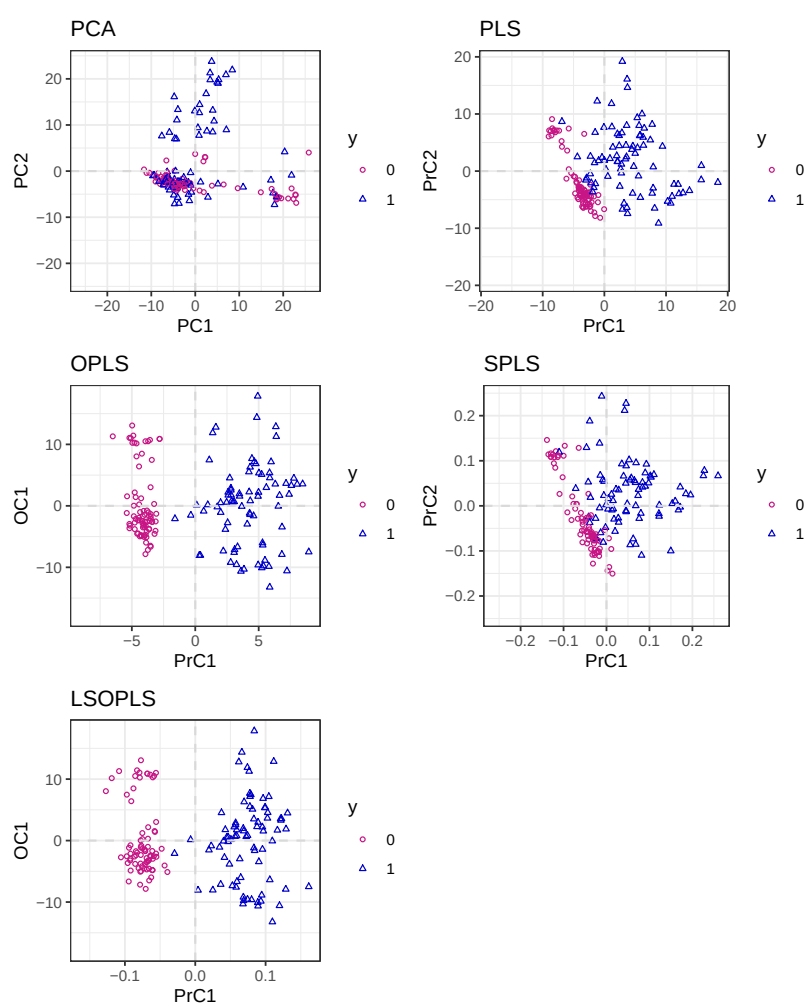


Figure 3.5 – Scores plots. The first two scores are represented for each method. For OPLS and LSOPLS, since only one predictive component is retained, the second axis represents the first orthogonal component.

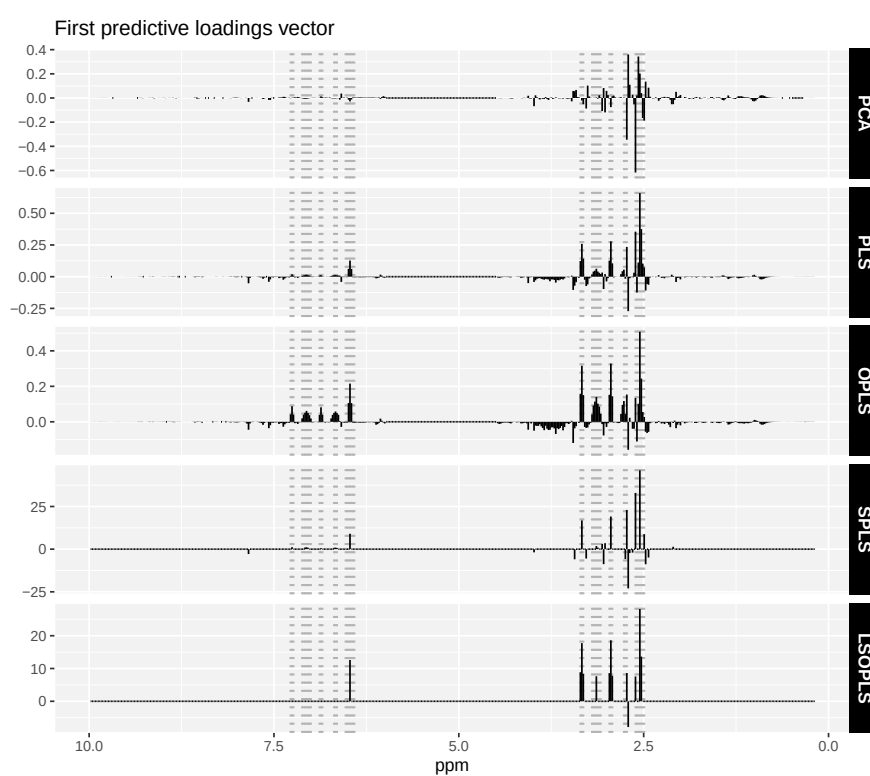


Figure 3.6 – Loadings plots. The first loadings vector is represented for each method.

Values of the difference in means per class and the t-statistic per variables are compared in Figure 3.7. While the difference in means is clearly pointing at the true biomarkers, the t-statistic has in addition a series of high values for all the uninformative and noise without signal variables. This is due to the very low standard deviation of those descriptors, directly associated to their very low mean values. This result highlight the main limitation of using t-tests for feature selection with wide and short data tables, as it definitely leads to a large amount of false discoveries on such data.

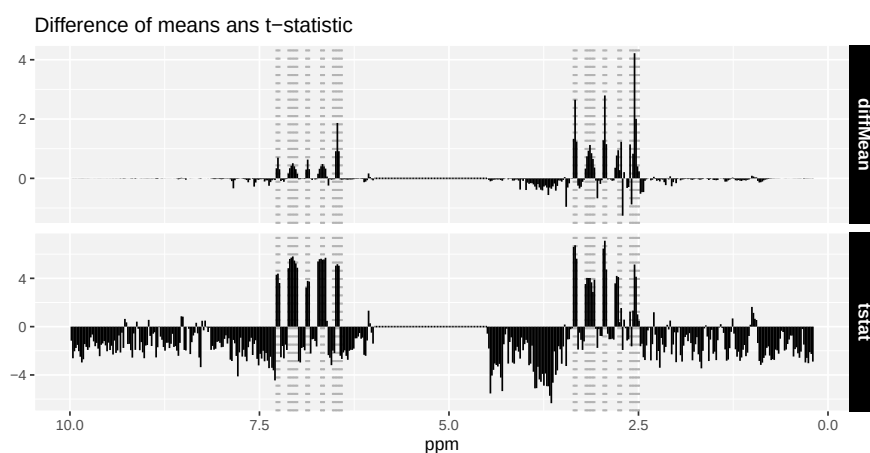


Figure 3.7 – Plots of the difference of the descriptors mean values per class and the t-statistic value. The vertical dotted lines represent the location of the true features.

The regression coefficients and VIP values of the supervised classifiers are of direct interest here since they summarise by themselves the global importance of variables to discriminate between the two classes of spectra and are used later on in the generic evaluation study for feature selection.

The regression coefficients are presented on the same ordinate axis in Figure 3.8. Recall that, based on the regression coefficients, SPLS and LSOPLS select the non-null values as potential features of interest whereas for PLS and OPLS, an arbitrary threshold is set and descriptors with the higher absolute values than that threshold are selected. For LSOPLS, the regression coefficients are the sparse uncorrected ones b_p^{lsopls} that will be used for feature selection. Based on a binary response vector, positive coefficients indicate which descriptors have a higher intensity in the presence of the alteration ($y = 1$), while negative coefficients indicate that those descriptors tend to have a lower intensity in that case. The following comments can be made after analysing this figure. Logically, sparse

methods have a restricted amount of non-null regression coefficients. As expected from the loadings plot, SPLS and LSOPLS are leading to sparser solutions. PLS and OPLS have very similar regression coefficients, as a consequence to the X -weights equality (*cf.* Section 3.3.3) in the case of a single response vector. They have good performances in the identification of features from the noise-free region since true features coefficients are higher in this area. Compared to LSOPLS, SPLS is better at recovering correlated features from a same biomarker. Finally, LSOPLS has the sparser solution but could not find all the peaks' locations.

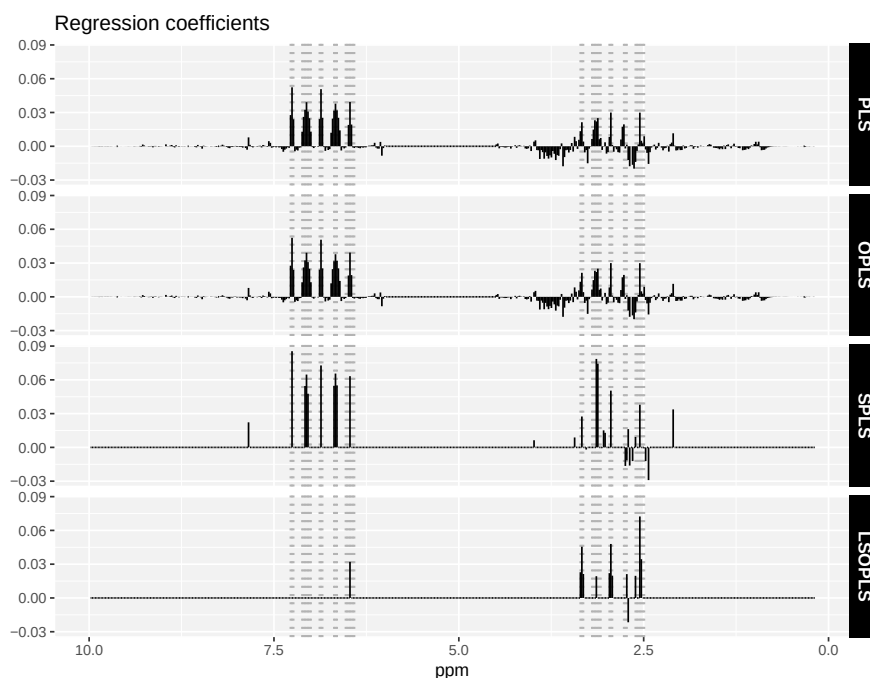


Figure 3.8 – Plots of the regression coefficients for all the methods. The vertical dotted lines represent the location of the true features.

Figure 3.9 represents the coefficient path plot for sparse methods along the iterative recruitment of a new model components (PrC or OC). For each new PrC in SPLS, there are new variables selected. The already enrolled variables tend to have a higher regression coefficient value as well. In LSOPLS, the effect of subtracting an OC to X has almost no influence on the selection subset, though a slight increase in the coefficients value is observed, especially for the first OCs.

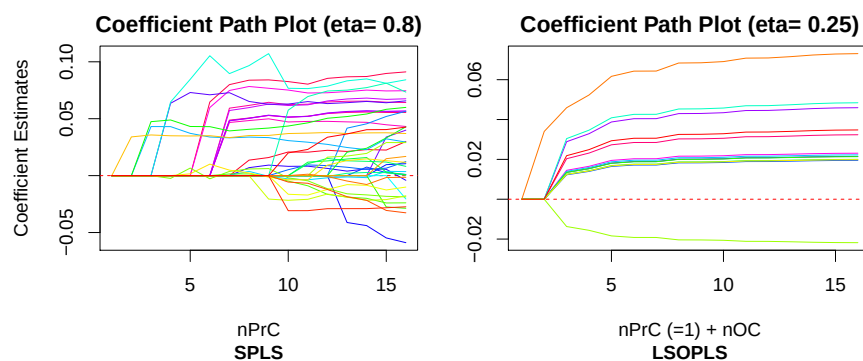


Figure 3.9 – Coefficient path plot of SPLS and LSOPLS. The plot illustrates the variation in the estimated coefficients as a function of the number of LV, with η fixed at its optimised level by CV.

Finally, VIPs are also considered for feature selection and the threshold of 1 is illustrated in Figure 3.10. Note that VIP values are always positive. They indicate which descriptor is important in the class discrimination but tell nothing about the direction of the varying intensities in the presence of the alteration. Based on this threshold, one can see that PLS will select all the peaks, which is not the case for OPLS. This threshold of 1 discards most of the false discoveries but also some of the true features, especially for OPLS. Compared to regression coefficients, VIPs seem to lead to a better feature recovery in the noisy part of the spectra.

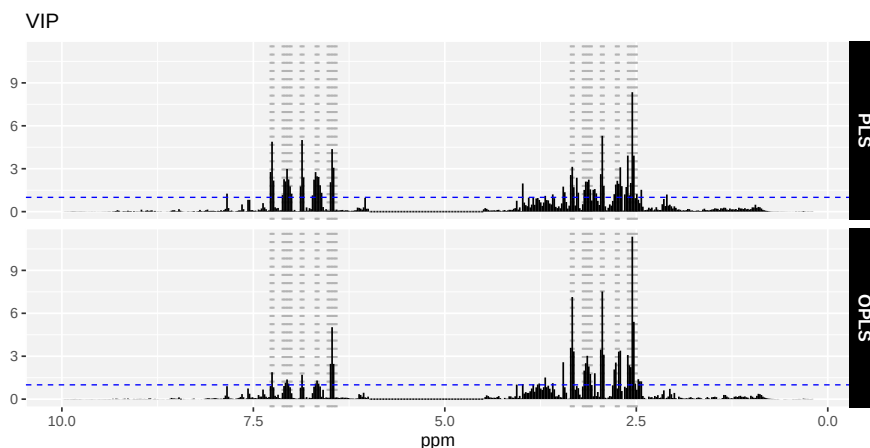


Figure 3.10 – Plots of the VIPs for PLS and OPLS. The vertical dotted lines represent the location of the true features and the horizontal dashed line represents the threshold of 1 for VIP values.

3.6 Evaluation of the classifiers performances

3.6.1 Introduction

The classifiers performances are now systematically evaluated on a large bunch of resampled sub-datasets from the semi-artificial database described above. The 5 biomarker selection methods are then applied to these sub-datasets. This approach is close to the spirit of the ensemble feature selection techniques (Abeel *et al.*, 2009) that consists in the combination of a large set of different models, gathered into a global and more robust one. Those sub-models are based on resampled sub-samples of the original dataset.

The main aims of this study are to:

- avoid misleading results due to a particular draw of one sample.
- study which method more likely discovers all true features – *i.e.* potentially correlated features.
- analyse the sensitivity and the FDR of features selection for each method.
- test the robustness of each method to different realistic sample sizes, in particular to a small sample size for $n = 50$.
- measure selected features and meta-parameters stabilities across resampling.

Note that methods comparison will be principally conducted on the $n = 200$ sample size dataset. Nevertheless, specific variations for a lower sample size ($n = 60$) will be highlighted and all the corresponding results are available in Appendix 3.A.

To achieve these above-mentioned goals, a double loop evaluation scheme was implemented for all methods but the t-test, with resampling and k -fold CV. This validation pattern is illustrated in Figure 3.11. The outer loop consists in resampling⁵ the raw spectral matrix R times (here, $R = 50$). An inherent incentive for resampling is the ability to measure the variability due to data sampling. Each resampled sample of size $n = n_{train} + n_{test}$ is split into a training and a test sets in the *outer loop* and each training set is again split in a series of training and test sets within an *inner loop* during k -fold CV. The outer loop is used to assess the predictive models performances while the *inner loop* decides on the optimal MPs for trained models.

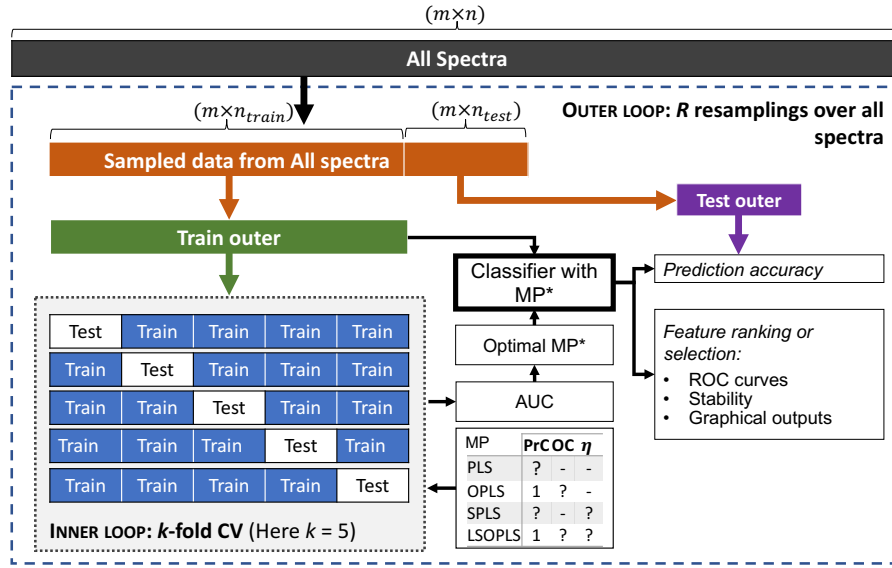


Figure 3.11 – Double loop validation scheme applied to all the methods. The outer loop consists in $R = 50$ resampled subsamples and the inner loop in a 10-fold CV (*rem.*: no inner loop for t-tests).

The results are produced and discussed in the following order or appearance: MPs stability, classification accuracy, global assessment of feature selection accuracy, visual inspection of the selected features on the spectral scale, and feature stability.

⁵Resampling consists here in a random draw of the spectral samples and their associated class with replacements.

Some measurements are based on the known alterations' true location, but others are "unsupervised" and could be used as *proxi* to evaluate the true features recovery in experimental datasets, where true location is unknown (e.g. by measuring the predictive performances or the selected features stability).

3.6.2 MPs stability

Deciding on the meta-parameters is inherent to the training of multivariate classifiers. They should be set wisely to avoid an overfit of the model to the data and ensure a parsimonious classifier. The variability in the choice of these MPs give insights into the algorithmic stability of the methods. Indeed, spectra sampled from a same population should lead to similar MPs, otherwise the algorithm is highly influenced by small changes in the data and is therefore considered unstable. The MPs variability will also influence the stability of the features selection process, and especially for procedures with automatic threshold, where the number of retained features directly depends on those MPs. Note that t-tests do not have MPs to tune as they are not classifiers *per se*. Recall that MP are on the one side, the number of predictive (PLS and SPLS) or orthogonal (OPLS and LSOPLS) components and, on the other side, the degree of sparsity η for SPLS and LSOPLS.

From Figure 3.12, it can be seen that the distribution of those MPs linked to the number of components is equivalent between the classifiers, with median values around 8 LVs. LSOPLS is the less stable algorithm to determine the number of LVs.

With regards to the sparsity parameter, SPLS has a larger median sparsity degree than LSOPLS. This could be explained by the prior orthogonal variation filtering in LSOPLS that already removes the uninformative variation from the data matrix, hence the filtering of uninformative descriptors is less important. This parameter is also less stable for SPLS. As illustrated for SPLS in Section 3.5 (Figure 3.4), a wide variety of possible combinations of LVs and sparsity is indeed leading to high AUC values for sparse classifiers. Hence, their MPs' stability will be affected accordingly. Since the size of the selection set is directly influenced by the amount of sparsity, this can be seen as a weakness of such sparse methods having multiple parameters to be optimised.

With a lower sample size (*cf.* Figure 3.23), MPs tend to be more spread and less stable across the sub-samples. The median number of OC is lowered, which is compensated by a higher degree of sparsity for LSOPLS. For SPLS, a decrease in its sparsity median value can also be observed.

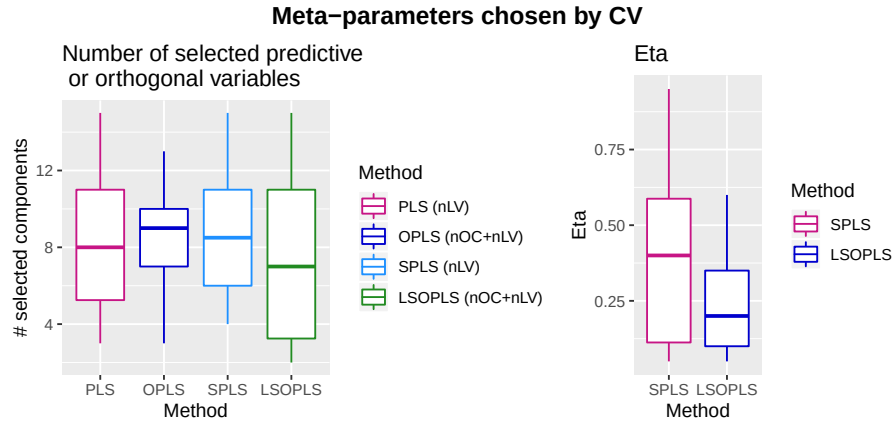


Figure 3.12 – Distribution of MPs chosen by the inner-loop CV ($n = 200$).

3.6.3 Size of the selection sets for automatic threshold methods

Graphical result in Figure 3.13 represents the size of the actual selection sets retained for automatic threshold methods. T-fdr has the highest median value around 100 of selected features, namely the double of the true features set size. Compared to PLS methods, OPLS ones tend to select less features. LSOPLS is the sparsest method with the lowest size of the selection set among all those methods. SPLS has a median value of 50 features selected but this amount is highly variable across the resampled sub-samples, as expected from the MPs variability results. In contrast, Vip1 methods have the lowest variability in the number of selected features. In the case of a reduced sample size (Figure 3.24), and except for t-fdr that is becoming much more conservative, results are approximately the same.

3.6.4 ROC curves and classification accuracies

Features and classification accuracies are evaluated in this section on the basis of indices derived from a binary confusion table.

Confusion matrix indices

In binary classification, a series of performance indices can be derived from a confusion matrix, based on *true* and *predicted* states *e.g.* of a patient (see Table 3.9).

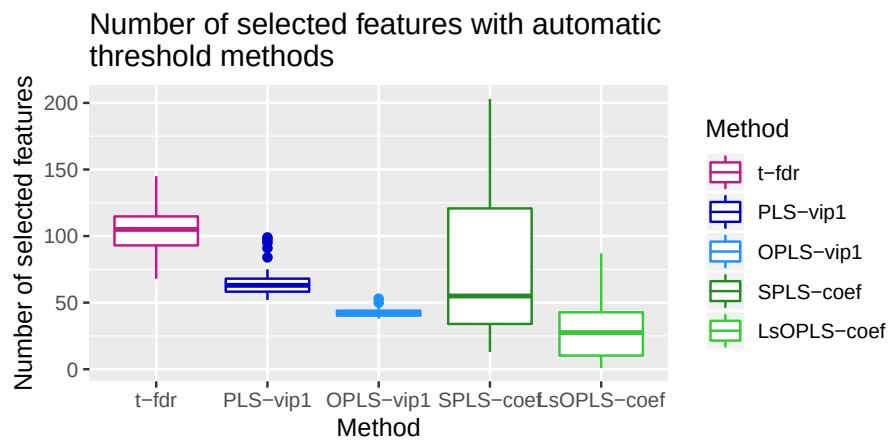


Figure 3.13 – Number of selected features for automatic threshold methods ($n = 200$).

Table 3.9 – Confusion matrix for binary classification.

		True class	
		Positive	Negative
Predicted class	Positive	True Positives (TP)	False Positives (FP)
	Negative	False Negatives (FN)	True Negatives (TN)

From this table, the *confidence in classification* is measured with the *sensitivity* – or number of actual positives that are correctly identified – $TP/(TP + FN)$ and the *specificity* – or number of actual negatives that are correctly identified – $TN/(TN + FP)$. The *predictive ability* of the classifier is measured with the *Positive Predictive Value* (PPV or *precision*): $TP/(TP + FP)$ and the *Negative Predictive Value* (NPV): $TN/(TN + FN)$. Another measure of interest is the opposite of PPV, the *False Discovery Rate* (FDR): $1 - PPV$.

Different aggregated measures that are independent of the group size *i.e.* can be used even in the case of unbalanced group sizes are:

- The Average predictive ability: $(PPV + NPV)/2$
- The Average confidence in classification: $(sensitivity + specificity)/2$
- The F1 score (harmonic mean of PPV and sensitivity) focusing on the correct classification of true positive features: $2 \cdot \frac{PPV \times sensitivity}{PPV + sensitivity} = \frac{2TP}{2TP + FP + FN}$

Application to classification accuracy

The ROC curves (Fawcett, 2006) are an appropriate tool to visualise and evaluate the behaviour of classifiers with a varying parameter. They are based on two specific confusion matrix indices: the sensitivity and the FDR. Generally, ROC curves are designed to evaluate a binary classifier with the discrimination threshold that varies along the curve or to compare several classifiers between each other.

The space of a ROC curve (illustrated in Figure 3.14) extends from (0,0) with no positive classification to (1,1) representing an unconditional positive classification. The perfect classification occurs at D (0,1). Overall, a point located in a Northwest position compared to another one is better (B compared to C). Furthermore, classifiers appearing at the lower left-hand side of the graph are conservative: they make few false discoveries errors and need strong evidence to classify as positive - whereas classifiers at the upper right-hand side of the graph tend to have high false discoveries by classifying as positive even if only little evidence exists. Besides, the diagonal line where *sensitivity* = FDR represents a random class allocation, and for a classifier appearing in the lower triangle (E), the class allocation should be inverted to perform better than a random guessing.

Another aggregated scalar, the Area Under the Curve (AUC) can be used for model comparison (Fawcett, 2006). Its value is always between 0 and 1 and a classifier should have at least an AUC of more than 0.5.

Note that all these indices are averaged over the outer loop.

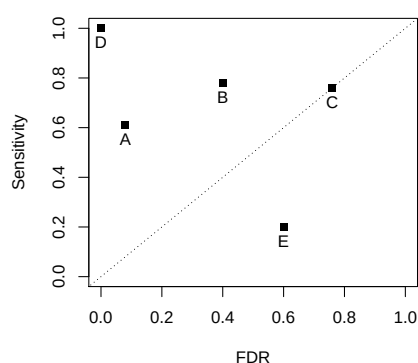


Figure 3.14 – The ROC space.

Classification accuracy results

Classification performances depicted in Figure 3.15 (for $n = 200$ and $n = 60$) and in Table 3.10 are measured at the outer loop level. They are calculated by applying the classifier with the optimal MPs on an independent test set in order to avoid an over-optimistic fit (Varmuza & Filzmoser, 2016). Classifiers perform quite well as expected by dataset construction with clear and distinct artificial alterations. Therefore, classification performances here are of little help to rate the classifying methods between themselves. It should be noted that, although very good, orthogonal methods performances are reduced when compared to the other classifiers. For all the methods, performances are lowered when the sample size decreases (*cf.* Figure 3.25) since less samples are available to train the models.

Table 3.10 – Classification performances: mean values of sensitivity, FDR, average confidence in classification, average predictive ability and F1 score at best threshold ($n = 200$).

	AUC	Sensitivity	FDR	Average confidence in classification	Average predictive ability	F1 score
PLS	0.97	0.97	0.02	0.98	0.98	0.98
OPLS	0.92	0.88	0.08	0.9	0.9	0.89
SPLS	0.98	0.98	0.03	0.97	0.98	0.97
LsOPLS	0.89	0.82	0.13	0.85	0.87	0.84

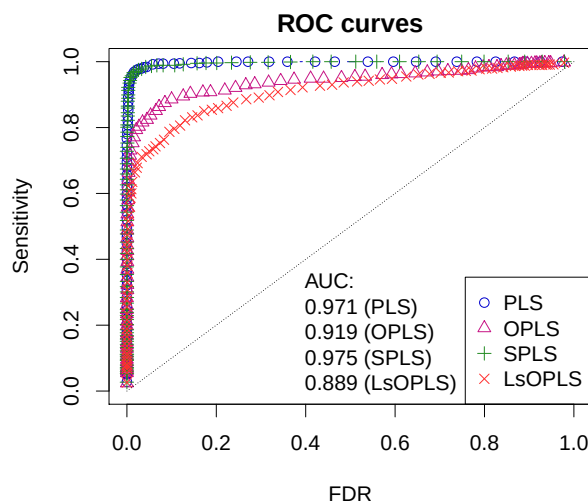


Figure 3.15 – ROC curves for classification accuracy ($n = 200$).

3.6.5 General assessment of features accuracy

A global assessment of features accuracy is presented in this section, before inspecting in more details the recovery of biomarkers by spectral area in the next one.

Use of ROC curves to evaluate features accuracy

Since the exact location of the true features is known in advance, the ROC curves principles can also be applied to evaluate each method's ability to discover the true features, as in Rousseau (2011) or Christin *et al.* (2013).

Each ROC curve is assigned to a classifier. Along these curves, two different parameters impacting features accuracy are varied: either the number of selected features $m^* = 1, \dots, m$ when an arbitrary threshold is to be decided or the sparsity level $\eta = 0.05, \dots, 0.95$ for sparse classifiers. Note that in this case, the curves can be non-monotonic, as should be expected for their regular use. here, it is important to decipher how the *true* features are classified, rather than how are classified the non-discriminant ones. To this end, we focus on indices where TP and FN are present such as in sensitivity, FDR, PPV or F1. Because the number of potential biomarkers detected will affect the range of subsequent research experiments (e.g. purification processes of those biomarkers), classifiers with a lower number of false discoveries (or higher precision) are preferred over classifiers with a higher sensitivity (Christin *et al.* , 2013). However, given the fact that different spectral descriptors

represent a single metabolite and that metabolites are correlated, one may also be interested to keep all the discriminant features, even if they are highly related to each other. In this case all the relevant information is preserved and one would like to have a higher sensitivity. In ROC curves, the sensitivity and the FDR are illustrated.

The indices are as previously averaged over the outer resampling loop.

ROC curves for feature accuracy results

Figure 3.16 represents the ROC curves with a different varying parameter given the type of method applied, and for $n = 200$.

With regards methods with a varying threshold m^* , all curves are nearly monotonic, except for the t-stat. The more interesting part of the graph is located on the left side of it, where only a restricted set of potential features are selected.

At the beginning of the curves, coefficient methods outperform vip methods. They further have the higher AUC among all the arbitrary threshold methods. T-stat is poorly performing at the start of the curve as its ratio sensitivity/FDR is very low, but outperforms vip methods at some point. OPLS-coef has the highest sensitivity and lowest FDR when selecting a lower number of potential features. For a short threshold value window, t-stat has the best sensitivity/specificity trade-off. However, PLS-vip is quickly outperforming the other methods at some point. OPLS-vip performs less well than the other (O)PLS derived methods.

The number of selected variables at best thresholds ranges from 43 to 57, corresponding to approximately 10% (8.6 - 11.4) of the total number of spectral variables. At their best threshold, t-stat has surprisingly the best sensitivity/FDR trade-off. It can also be observed that the average number of selected features or threshold across R with the VIP threshold of 1 (blue dot on the graph) is either above or below the best measured mean threshold respectively for PLS-vip and OPLS-vip.

Among those methods, no one is outperforming the others over the entire curve. Nevertheless, coefficient methods outrun others for a restricted amount of selected features, which is often the case in the omics sciences.

In the case of a lower sample size (*cf.* Figure 3.26), t-stat is performing extremely poorly, and PLS-vip do also especially suffer from the lowered sample size.

With regards to sparse methods with their sparsity level that varies along the curve, LSOPLS has a lower sensitivity than SPLS at equal η , and optimal η is lower for LSOPLS ($\eta = 0.05$) than for SPLS ($\eta = 0.3$), as expected from the

previous results. For higher levels of sparsity, LSOPLS outperforms SPLS: for a same sensitivity rate, it will recover less false discoveries. SPLS has the advantage to be able to select more true features, but at the expense of more false discoveries as well. Beware that these averaged performances are very poor for LSOPLS with $\eta > 0.35$, thus LSOPLS should not be applied in general with a large amount of sparsity to ensure an acceptable features accuracy. SPLS performances are also affected by a too large η (here $\eta > 0.7$). Although performances are globally lowered in the case of lower sample sizes (*cf.* Figure 3.26), methods comparison leads to the same observations as above.

Number of true features selected

Figure 3.17 presents the number of correct features selected across the 50 resampled sub-samples with values ranging from 0 (no correct feature selected) to 46 (all the true features were recovered). Note that all the methods cannot be directly compared to each other since the threshold for a part of them is arbitrarily chosen.

Methods with an arbitrary threshold have their median value varying between 30 (OPLS-vip) and 35 (t-stat). T-stat has the highest median number of correct identifications. However, from previous results, it has been seen that it is at the cost of much more false discoveries. For some of the resampled samples, PLS had a lower rate of recovery. Otherwise, performances do not vary a lot across the sub-samples.

For methods with an automatic threshold, t-fdr is also recovering a high amount of true features (but still with a large amount of false discoveries), closely followed by PLS-vip1. OPLS-vip1 has a less important recovery rate. Finally, SPLS and LSOPLS have a much more scattered number of TP and their sensitivity can be above or below the other methods' ones. This is partly due to those sparse classifiers' construction itself that will directly decide on the size of the selection set, which that could be much lower than 46.

Table 3.11 complements the results from the previous graph for the automatic threshold classifiers. In this table are represented ROC indices derived from the selection sets of those classifiers that were trained with optimal MPs, and averaged across the 50 sub-samples. T-fdr leads to a higher recovery rate of true features but its FDR is highly inflated as well. The F1 score, which summarises the trade-off between these two indices, is best for PLS-vip1. Both vip1 methods outperform sparse ones. It can be noted that LSOPLS has the lowest FDR but is not able to recover more than half of the true features, which is due to its lower number of selected ones. In the lower sample size case (*cf.* Table 3.14), t-fdr only recovers a very small amount of features. Based on the F1 score, LSOPLS is more affected by the size drop whereas other methods are moderately affected.

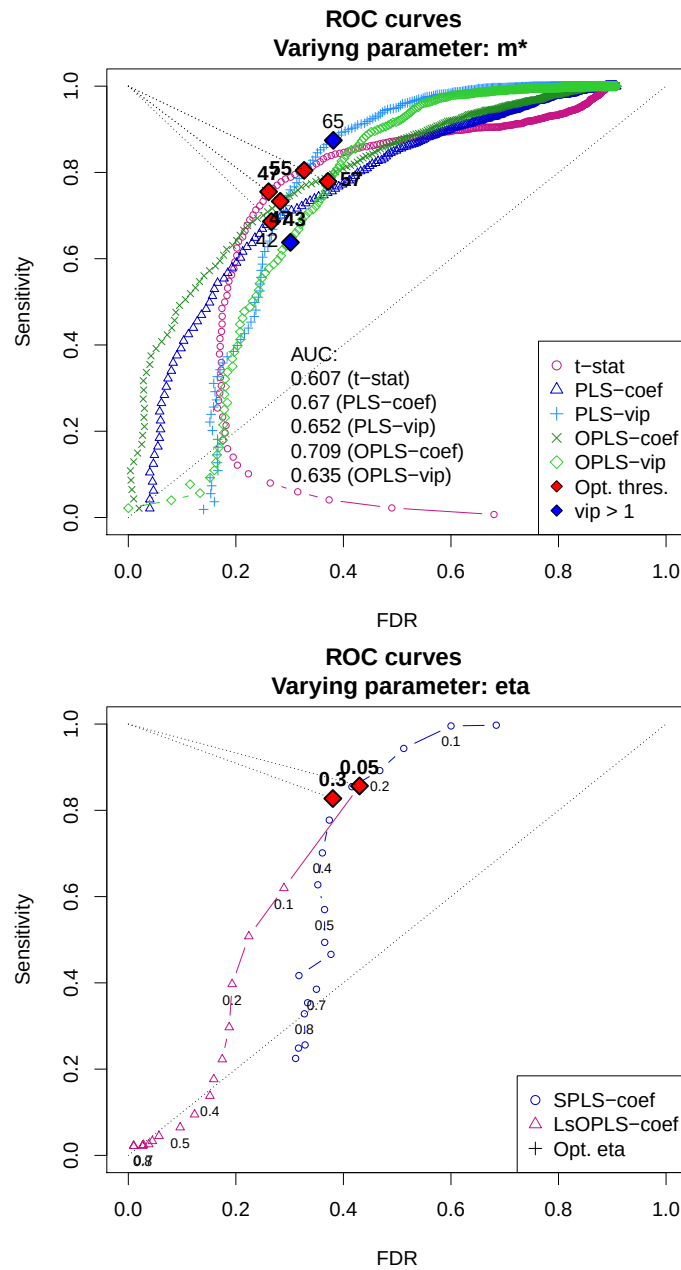


Figure 3.16 – ROC curves for feature accuracy (methods with an arbitrary threshold or sparse methods, $n = 200$). Varying parameters are either the selection threshold $m^* = 1, \dots, m$ or the sparsity parameter $\eta = 0.05, \dots, 0.95$ depending on the classifier. Optimal thresholds are represented in red.

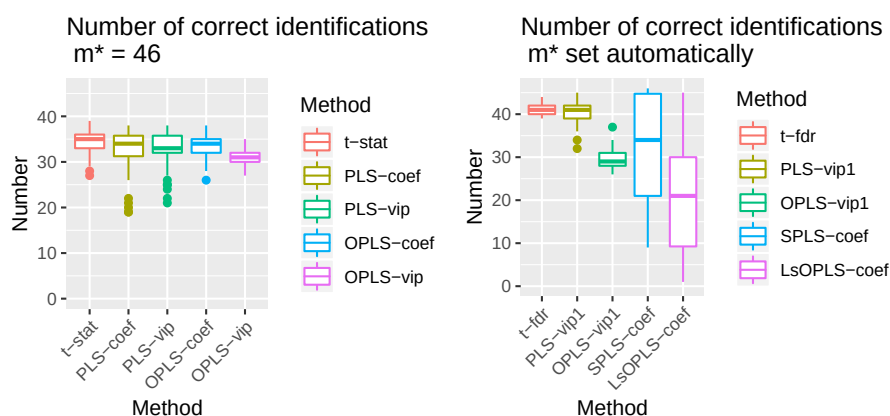


Figure 3.17 – Number of correct features selection. Note that for arbitrary threshold methods, it corresponds to the number of selected features for the 1:46th ranked features ($n = 200$).

Table 3.11 – Mean ROC indices from the actual selection sets based on the optimal MP (automatic threshold methods) for features accuracy ($n = 200$).

	Sensitivity	FDR	F1
t-fdr	89.35	59.97	55.28
PLS-vip1	87.65	36.92	73.36
OPLS-vip1	64.65	30.23	67.11
SPLS-coef	70.91	45.5	61.63
LsOPLS-coef	44.22	23.55	56.03

3.6.6 Selected features inspection on the spectral scale

This section proposes several graphical results to illustrate, on the entire spectral scale or by independent biomarker, the feature selection achieved by the different methods. Once more, for methods needing to set a threshold m^* , one arbitrarily decides to keep the total number of generated true features (46) as the number of variables to select based on their features scores.

In this first graph (Figure 3.18), features scores B are averaged over the 50 data sub-samples ($n = 200$). Here, the coefficient intensities can be directly compared to each other. The main interest of this graph lies in the representation of the selection threshold for each method. This threshold is either: values different from 0 (t-fdr, SPLS-coef and LSOPLS-coef) or superior to the horizontal dotted lines for the remaining ones.

T-stat has high negative mean intensities for almost all non-discriminant descriptors. As a consequence, it leads to a larger amount of false discoveries. With the threshold fixed at $m^* = 46$ (blue line), few false discoveries are still observed, although much less than for t-fdr.

Among the coefficient methods, PLS and OPLS share approximatively the same profile and have much more small intensities, corresponding to noise only, across the spectrum than their sparse counterparts. They have higher mean feature scores in the noiseless area than in the noisy part, to the contrary of vip methods. SPLS is also able to well recover biomarkers from the noise-free side of the spectrum but has much lower scores on the noisy part of it. It still has a lot of small non-null intensities, especially on the right side. LSOPLS has the lowest mean scores for the noise-free features across all coefficient methods. Finally, the threshold of 1 applied to vip methods is too high for OPLS-vip to detect biomarkers in the noise-free area.

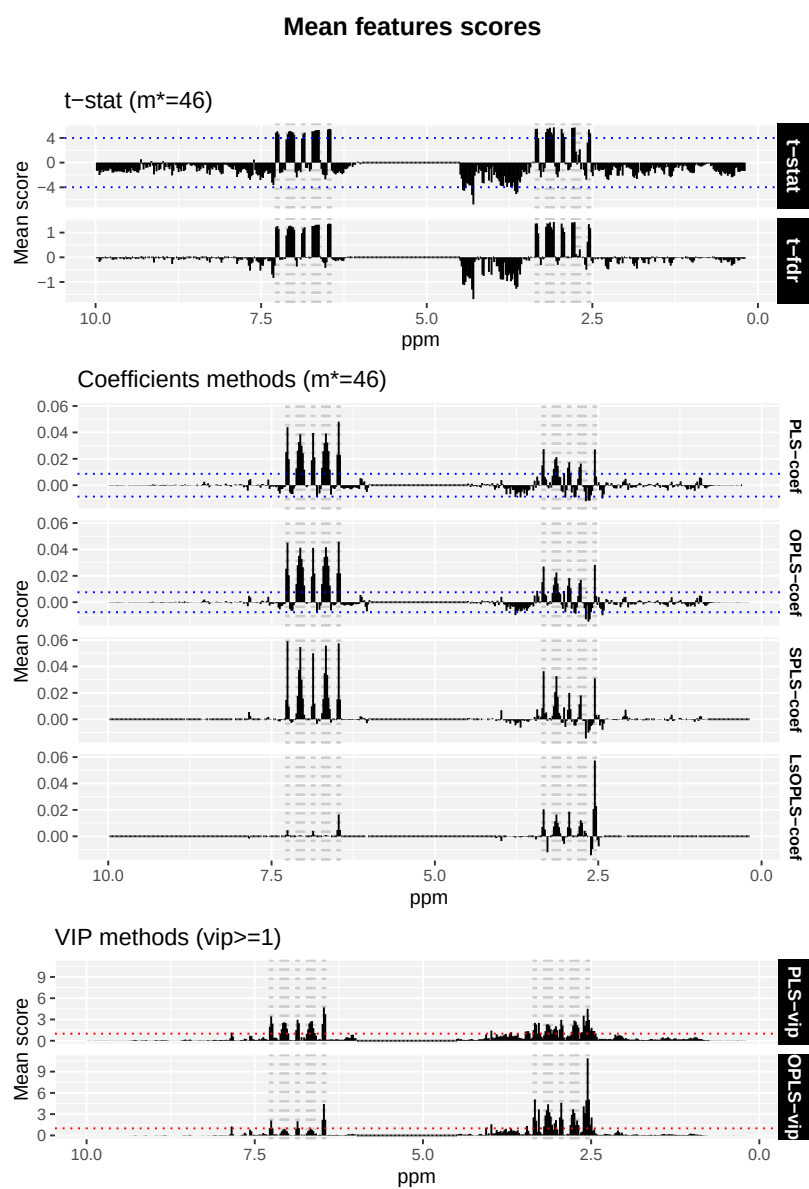


Figure 3.18 – Mean features scores **B**. Blue lines set the arbitrary threshold, it corresponds to the selection of 46 features ($n = 200$) and red lines represent the $VIP \geq 1$. For sparse methods, recall that all coefficients $\neq 0$ are selected.

An interesting way to present the results is to draw the frequencies of TP as positive intensities and FP as negative ones (Figure 3.19, $n = 200$). Note that in this second graph, all the methods cannot be directly compared to each other since the threshold for a part of them is arbitrarily chosen. As observed previously, both approaches based on the t-statistic have a high amount of FP but do have the advantage to detect all the independent biomarkers. Vip methods are better at recovering all relevant features in the noisy part of the spectra, where alteration intensities are also higher, but have more false discoveries than (O)PLS coefficients. On the contrary, the latter have a higher sensitivity in the left side of the spectra, *i.e.* in the noise-free area, where spectral intensities are lower.

With regards to the automatic threshold techniques, vip1 methods do, as previously noted, better recover true features in the noisy part of the spectra. PLS-vip1 has the highest TP frequencies among those methods, with a limited amount of false discoveries. SPLS and LSOPLS are surprisingly leading to a lot of false discoveries. Combined with the figure of the mean scores above, this can be explained by the fact that those algorithms are still selecting variables with very low – still non-null – intensities, and the sparsity parameter is somewhat under-estimated. Furthermore, the frequencies of TP are not identical among the descriptors of a same biomarker. This means that those sparse techniques will generally select only a subset of features from a same biomarker as they represent redundant information. In all logic such classifiers are better suited for minimal-optimal problems. When the sample size is lowered (*cf.* Figure 3.29), t-fdr is affected with a clearly reduced TP rate while other methods' results are mainly unaffected.

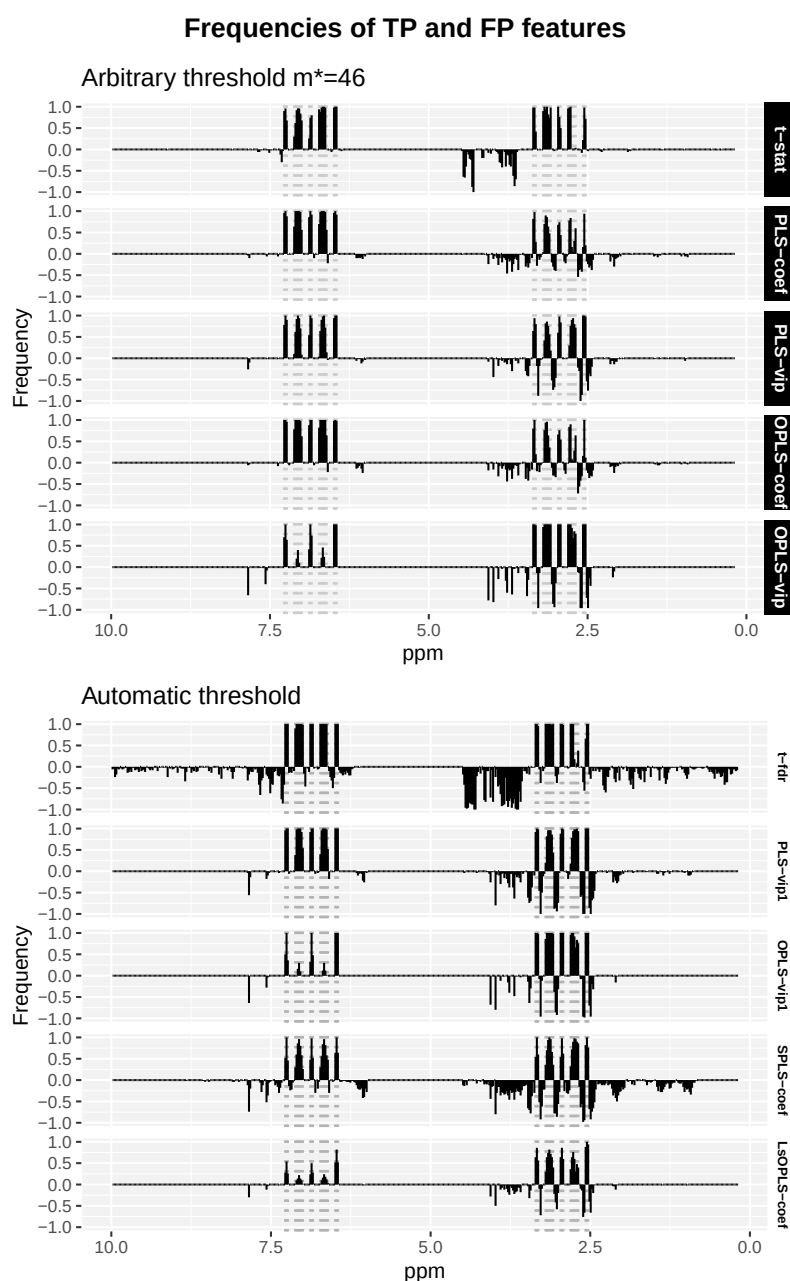


Figure 3.19 – Frequencies of TP and FP. Note that for methods with an arbitrary threshold, it corresponds to frequencies of selected features for the 1:46th ranked features ($n = 200$).

The last graphs for this section (Figures 3.20 and 3.21 , $n = 200$) are a summarised representation of the correct identification percentage by independent biomarkers for each method. It is measured by biomarker as either (A) the number of TP divided by the number of total features in this biomarker or (B) as the mean value of a binary-coded vector with 1 = at least one feature from this biomarker is found and 0 = none of the features from this biomarker is selected. Therefore, methods that are able to recover all the correlated features from a same biomarker will have a higher % of correct identification and are in favour of the *all-relevant problem*, whereas being able to recover at least one feature per independent biomarker represents the solution to the *minimal-optimal problem*. Recall that Biomarkers 1-3 are located in the noise-free area while Biomarkers 4-6 are situated in the noisy area. Here again, methods with arbitrary and automatic thresholds cannot be compared to each other.

In the A-scenario (Figure 3.20) and for methods based on an arbitrary threshold, no one method is outperforming the others. Regression coefficients are better at recovering true features located in the noise-free area but are much less efficient regarding biomarkers in the noisy area. The opposite is true for vip methods though they still maintain a high true discovery rate for Biom 3. For both biomarker locations, OPLS methods have a better percentage of identification than PLS ones. T-stat has mid-way performances across all biomarkers.

With regards to the automatic threshold methods, t-fdr can recover the largest amount of total discriminant features, but again at the expense of a high FDR. It is closely followed by PLS-vip1 that shows a satisfying recovery rate of at least 75% for every biomarker. In comparison, OPLS-vip1 could recover biomarkers in the noisy area with higher intensities but is affected by lower intensities as it was not able to find all features in the noise-free area. Sparse methods logically selected less features by biomarker. As observed for other methods, SPLS better recovered features in the noise-free area. For smaller sample sizes (*cf.* Figure 3.30), the average rate of identification is lowered for all methods and they have a similar behaviour, except for methods derived from the t-statistic.

In the B-scenario depicted in Figure 3.21, (the *minimal-optimal problem*), all arbitrary threshold methods (except for OPLS-vip and Biom 3) were able to recover at least one feature from each independent biomarker across sub-samples. For methods with an automatic threshold, vip1 methods and SPLS also recovered $\simeq 100\%$ of biomarkers. LSOPLS for its part was not able to discover at least one discriminant feature for all resampled sub-samples, and especially in the noise-free area. This is probably linked to the small amount of selected variables by this

method. Again, when the sample size drops, performances are lowered but their behaviour remains equivalent (except for t-stat and t-fdr).

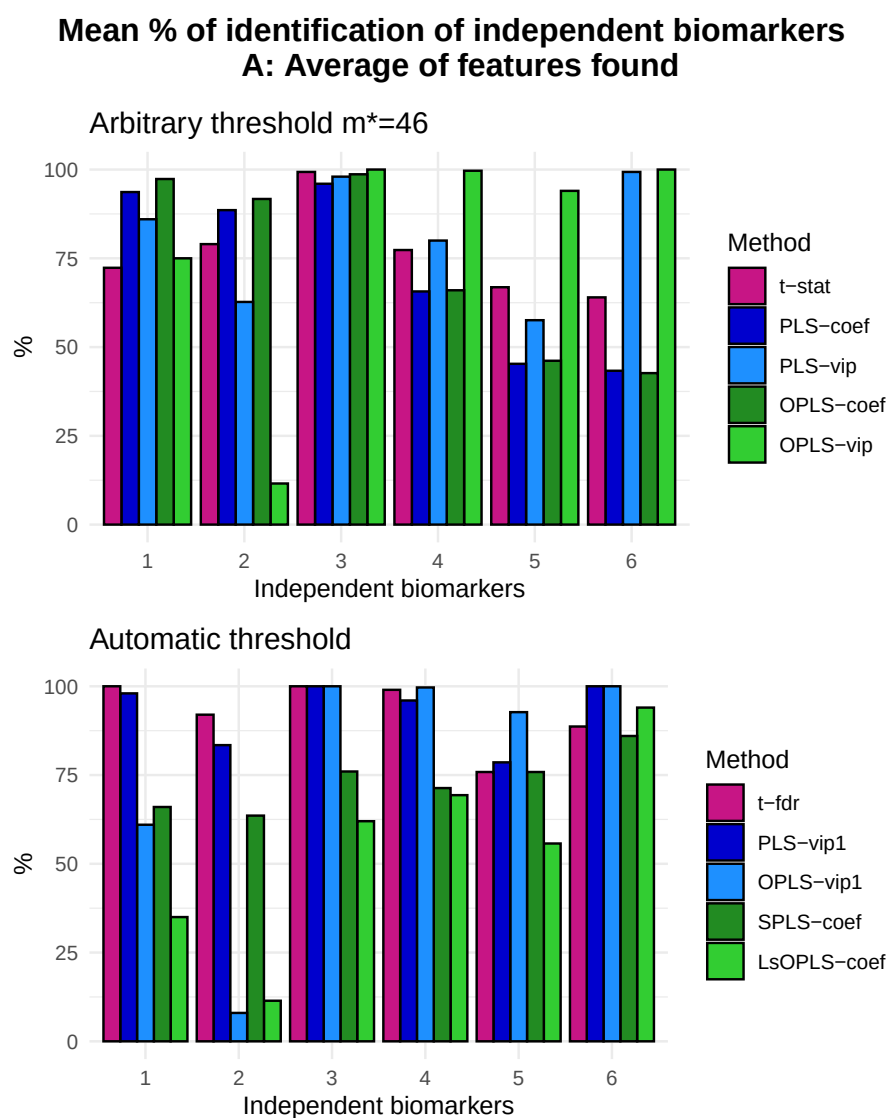


Figure 3.20 – Mean number of identification of independent biomarkers per method calculated as the average of features found for each biomarker. Note that for arbitrary threshold methods, it corresponds to frequencies of selected features for the 1:46th ranked features ($n = 200$).

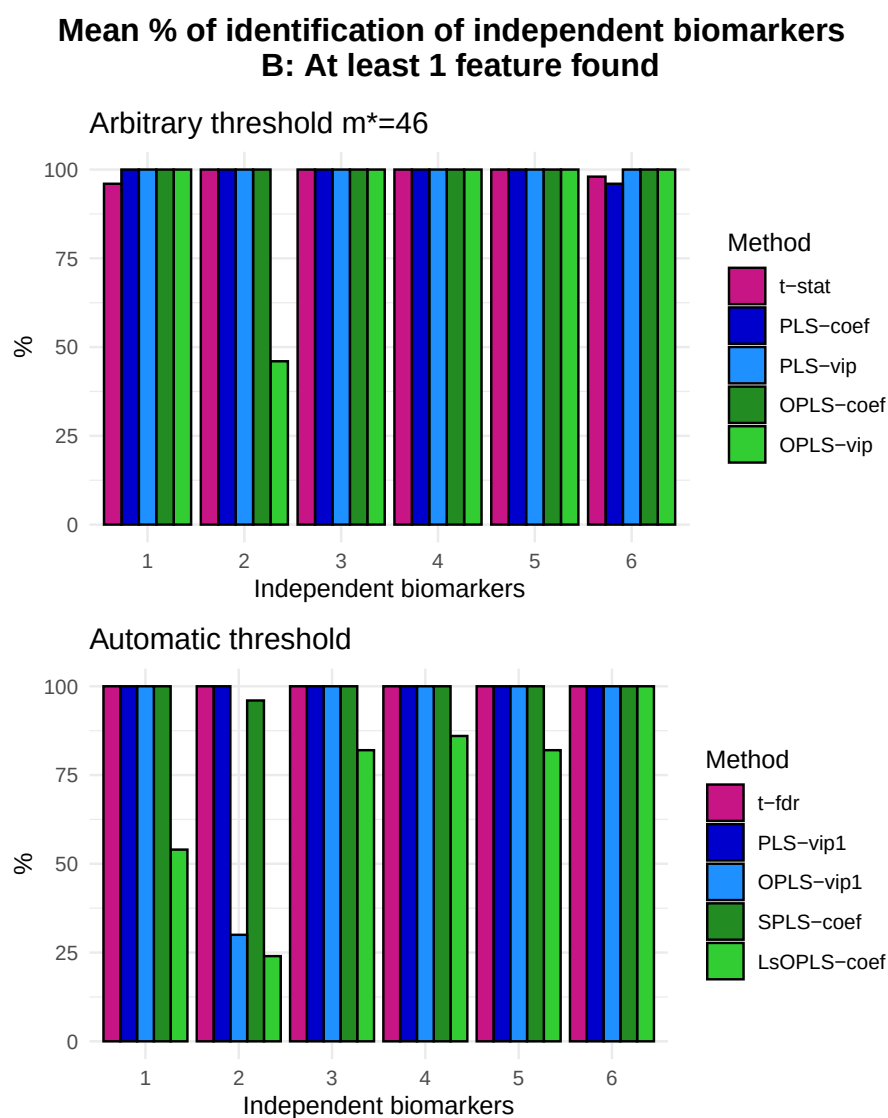


Figure 3.21 – Mean number of identification of independent biomarkers per method where at least 1 feature found for each biomarker. Note that for arbitrary threshold methods, it corresponds to frequencies of selected features for the 1:46th ranked features ($n = 200$).

3.6.7 Feature selection stability

Feature selection stability is essential to achieve reproducible research: having a stable set of features is as much important as identifying relevant ones (Nogueira *et al.*, 2017). The features ranking/selection reproducibility across subsets of data is also a major concern for the classifiers evaluation and this topic starts to gain more importance in the feature selection context (read for example Abeel *et al.* (2009), and also (Christin *et al.*, 2013; Rinaudo *et al.*, 2016) for metabolomics applications).

Indeed it is also important for the rankings or selection sets to be similar across the resampled sub-samples for every feature selection technique. In a way, it evaluates the robustness of the classification algorithms, given the data characteristics (noise, dimensionality, etc.). A procedure is considered unstable if a small change in the dataset leads to a very different ranking or subset selection. If the true features are unknown, which is the case for experimental datasets, feature stability could be a good indicator of the results reliability. Although not used to that purpose, such stability criterion could also presumably be used to select the sparsity level of sparse algorithms, such as LASSO-type regressions (Meinshausen & Bühlmann, 2010).

The selected stability measures are based on the work of Nogueira *et al.* (2017), where the authors compare existing measures to a new one, based on a statistical point of view. Indeed, there is plenty of existing robustness criteria. Their measurement is not trivial and desired properties must be met to adequately quantify the features stability.

The criterion suggested by Nogueira *et al.* (2017) meets the expected properties of such estimator: it is fully defined, is strictly monotonic, is bounded reaches its maximum value *iff* all features across the features subsets are identical, it has correction for chance. It further has the ability to compare subsets of features with different sizes.

Consider a binary matrix $\mathbf{D}$ of size $(m \times R)$ where each element d_{ij} corresponds to the selection of a given feature $i = 1, \dots, m$ in the resampled samples $r = 1, \dots, R$, where

$$d_{ir} = \begin{cases} 1 & \text{if } i \in \mathcal{A}_r \\ 0 & \text{otherwise.} \end{cases}$$

Let $\Phi(\mathbf{D})$ be the stability function with input the binary matrix $\mathbf{D}$ and as output a percentage between 0 and 100%:

$$\Phi(\mathbf{D}) = 1 - \frac{\frac{1}{m} \sum_{i=1}^m s_i^2}{\mathbb{E}[\frac{1}{m} \sum_{i=1}^m s_i^2 | H_0]} = 1 - \frac{\frac{1}{m} \sum_{i=1}^m s_i^2}{\frac{\bar{m}^*}{m} (1 - \frac{\bar{m}^*}{m})}; \quad (3.22)$$

where $s_i^2 = \frac{R}{R-1} f_i(1 - f_i)$ is the unbiased sample variance of the selection of the i^{th} feature with f_i the observed frequency of selection of the i^{th} feature: $f_i = \frac{1}{R} \sum_{r=1}^R d_{ir}$. Note that this formula is derived from a Bernoulli distribution. $\bar{m}^* = \frac{1}{R} \sum_{i=1}^m \sum_{r=1}^R d_{ir} = \sum_{i=1}^m f_i$ is the mean number of selected features across all selection sets $\mathcal{A}$.

Stability results are presented in Figure 3.22. The first graph concerns methods with an arbitrary threshold where m^* varies along the x-axis. The second graph shows the feature selection stability for sparse methods, with their degree of sparsity (η) varying along the x-axis.

For methods with an arbitrary threshold, this stability is at its lowest at the start of the curve. It then greatly increases when the selection sets are including the first 1:50 features. Then, they decrease to reach an approximatively constant rate. Vip methods achieve the best stability rate among all those methods and OPLS-vip has the best stability at lower values of the threshold. Coefficient methods have equivalent performances. The feature selection stability is greatly varying for t-stat. After reaching its best value around $m^* = 50$, it considerably drops to become the less stable method. From this graph, we can say that feature selection stability is greatly depending on the number of features selected. Those results can be analysed in parallel with the ROC curves from Figure 3.16. For vip and t-stat methods, the observed peak corresponds to the optimal trade-off on the ROC curve. However, for coefficient methods the correspondance is less clear.

With regards to the sparse methods, feature selection stability is slightly decreasing with an increase in the degree of sparsity for SPLS while it has a V-shape for LSOPLS. The latter has his feature selection stability more affected by the level of sparsity. Finally, note that stability is in average higher for vip methods than for sparse methods. In Table 3.12 are given the actual selection sets from other methods with an automatic threshold. OPLS-vip1 has the highest stability, a result that corroborates with the previous graph.

In the case of a reduced sample size (*cf.* Figure 3.32), curves keep the same shapes. Except for t-stat that is greatly affected, their stability is slightly diminished. Their relative performances are maintained as well.

Table 3.12 – Feature selection stability for the other automatic threshold methods ($n = 200$).

t-fdr	PLS-vip1	OPLS-vip1
60.98	75.77	82.32

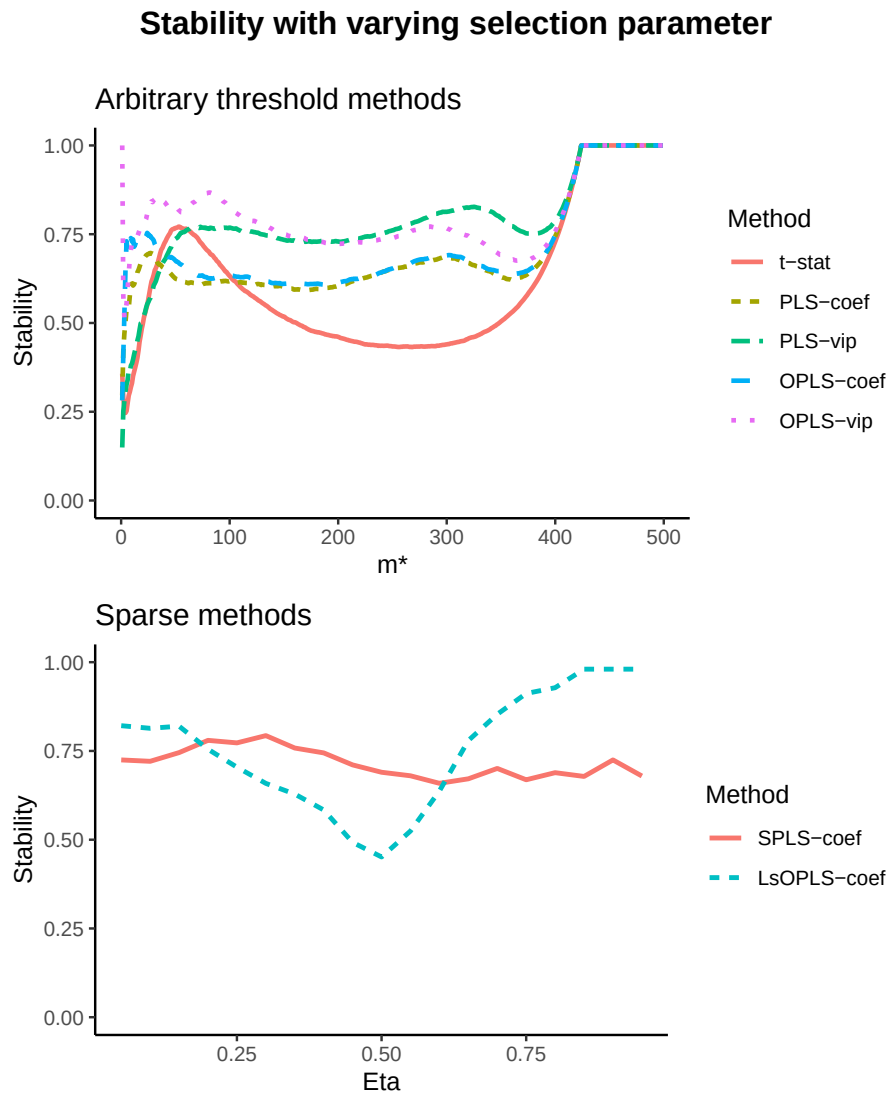


Figure 3.22 – Features Stability across the number of selected features for methods with an arbitrary threshold and across η for sparse methods ($n = 200$).

3.7 Conclusions

The success of biomarker discovery in the omics sciences is heavily dependent on the multivariate method(s) used to extract the features but also on the data characteristics. In this chapter, **classic supervised approaches** – t-test, (O)PLS and their **sparse** counterparts – have been compared based on the Semi-Artificial database, in order to mimic real experimental conditions and with **known in advance biomarkers location**. In this study, emphasis has been placed to (1) use numerous resampled sub-datasets to **average the methods performances** (2) **train** appropriately the supervised models to avoid their overfit to the data and (3) compare the feature selection techniques with several **evaluation measurements**, which, for part of them, takes into account the prior information from the true features location.

Both approaches derived from the t-statistic led to a **high amount of false discoveries**. By reducing the arbitrary threshold in t-stat, most of these could be avoided. In contrast, t-fdr selected a large amount of FP due to the low standard deviation of noise descriptors, even in the case of a small FDR rate that was fixed at 0.05. Still, the advantage of this approach is that it recovers all correlated features. It could therefore be used in the **all discriminant approach**, provided that the FDR can truly be controlled or with a low threshold. However, we found that performances of both methods were also highly **sensitive to a reduced sample size**, which is a more realistic experimental condition.

With regards to methods based on (O)PLS, coefficient and vip methods have the disadvantage that **no clear threshold** is defined, though it indeed affects the feature selection stability and accuracy. In this study, the ratio TP/FDR and feature selection stability were maximum around 50 features selected, namely around 10% of the total number of variables in the data. Selecting a lower number leads to a much lower sensibility and stability whereas, selecting more features decreases specificity. The fixed threshold of 1 for vip methods enabled to have a good feature accuracy (F1 score) for PLS but this threshold was too high for OPLS to recover discriminant features with lower intensities. Nevertheless, both **vip1 methods outperformed sparse classifiers** to detect all the discriminant features. These non-sparse multivariate methods do have each their advantages and drawbacks. Vip methods were better at recovering biomarkers in the noisy area with higher features intensities, whereas coefficients-based approaches were in general better at recovering biomarkers having lower intensities in the noise-free area. Models with an **orthogonal filter** (OPLS and LSOPLS) seemed indeed to be more **sensible to the peaks intensity**: they could not well recover lower intensity biomarkers in the noise-free area.

We saw that the **degree of sparsity** also affected the stability and accuracy of feature selection for sparse methods. Here can be concluded to the importance of a **fine tuning of the MPs** for such embedded approaches. Those methods should perhaps rely on another criterion to decide on the optimal combination of the number of LV and the sparsity degree since multiple solutions were equivalent in terms of the AUC. Further investigations on the interaction between the MPs should be led, as they directly impact features accuracy and stability. Indeed, it seems that sparse methods have the **potential to achieve better performances** but would need a better training criterion, based e.g. on the desired features subset size or its stability. Due to its prior removal of orthogonal variation, LSOPLS has an optimal sparsity parameter that tends to be lower than SPLS. Results also showed that applying SPLS, and even more importantly LSOPLS, with a **large degree of sparsity is counterproductive** as their feature selection sensibility and specificity will drop considerably. Surprisingly, still a lot of **FP** could be observed for such sparse classifiers. Since the degree of sparsity should not be increased excessively, one solution could be to slightly increase the 0 threshold applied to the regression coefficients for feature selection. Furthermore, the frequencies of selected TP were not identical among the descriptors of a same biomarker. This means that those sparse techniques will generally select only a subset of those features since they represent redundant information. In all logic such classifiers are better suited for **minimal-optimal problems**.

Results were mainly conducted on the larger sample size. When only a **reduced sample size** is available, overall performance is logically reduced, but relative performance was maintained. Nevertheless, some of the studied methods were more robust while others did suffer more: t-stat and t-fdr had their performances affected with a significant drop of true features recovery. LSOPLS was also less robust to the sample size decrease than the other related methods.

In the case of our resampled dataset, PLS-vip1 had the best overall performances regarding all aspects of this study, and even in the low sample size case. But other methods were sometimes better with regard to the various investigated evaluation measurements. Yet, this result is truly **dependent on the dataset at hand and its characteristics** (level of noise, number of true discriminant features, number of samples, intensities of those features, etc.). Therefore PLS-vip1 may not be always the best method to detect potential biomarkers.

Besides these compared methods, there is a myriad of other possible feature selection algorithms based on PLS, (see for example Mehmood *et al.* (2012) for a review of such methods) and another myriad of other classifiers derived from PLS, in particular the logistic PLS suggested by Meyer *et al.* (2010), which is a more

logical choice in terms of interpretation in the case of a binary response. Other multivariate methods that already proved their usefulness in metabolomics should deserve to be considered as serious competitors as well (Gromski *et al.* , 2015).

Methodological aspects that were not investigated are the combination of feature selection methods to take advantage of their different assets. This could be adequate since here, no method was always superior to the rest. Other sparse classifiers, such as the (sparse) group lasso (Friedman *et al.* , 2010), could also take into account the feature membership to each particular biomarker. With regards to the problem of not selecting all the correlated biomarkers, a solution could be to apply the bagging principle that originates from Random Forest, by resampling in the feature space (T. K. Ho, 1998). Finally, more realistic datasets (*e.g.* a spiked experimental dataset), could be applied with our methodology to inspect the replicability of these results.

3.A Other results with $n = 60$

3.A.1 MPs stability

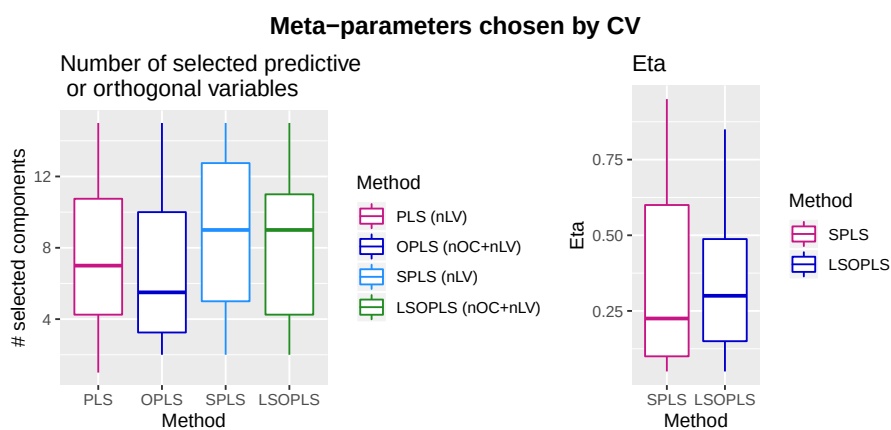


Figure 3.23 – Distribution of MPs chosen by the inner-loop CV ($n = 200$).

3.A.2 Size of the selection sets for automatic threshold methods

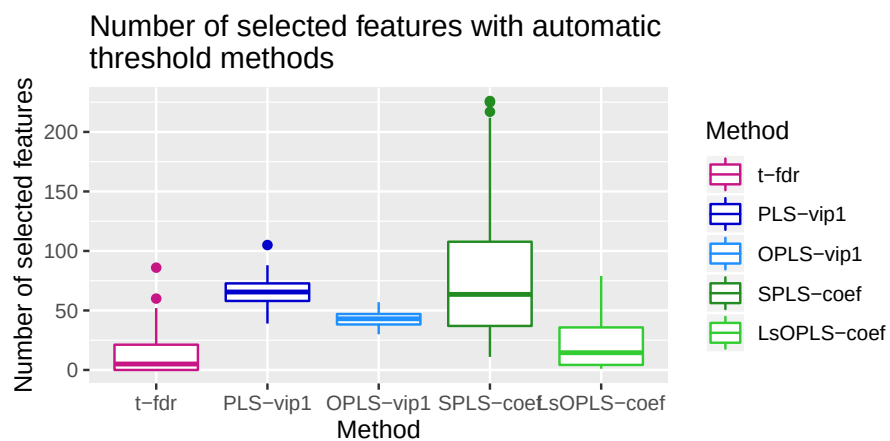


Figure 3.24 – Number of selected features for automatic threshold methods ($n = 60$).

3.A.3 Classification accuracy

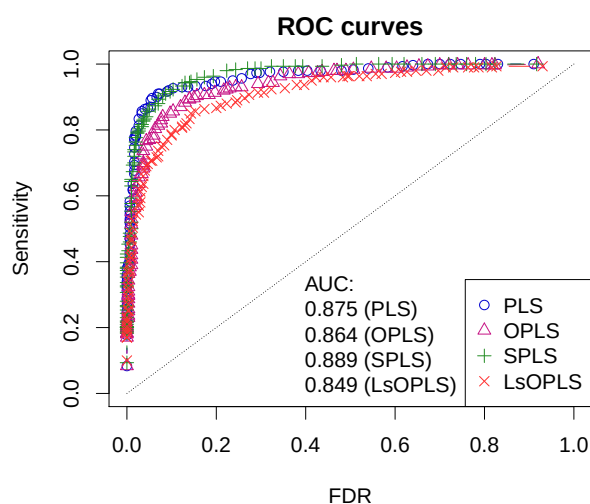


Figure 3.25 – ROC curves for classification accuracy ($n = 60$).

Table 3.13 – Classification performances: mean values of sensitivity, FDR, average confidence in classification, average predictive ability and F1 score at best threshold ($n = 60$).

	AUC	Sensitivity	FDR	Average confidence in classification	Average predictive ability	F1 score
PLS	0.88	0.91	0.07	0.92	0.93	0.92
OPLS	0.86	0.88	0.13	0.88	0.9	0.88
SPLS	0.89	0.91	0.09	0.91	0.93	0.91
LsOPLS	0.85	0.85	0.14	0.85	0.87	0.85

3.A.4 General assessment of features accuracy

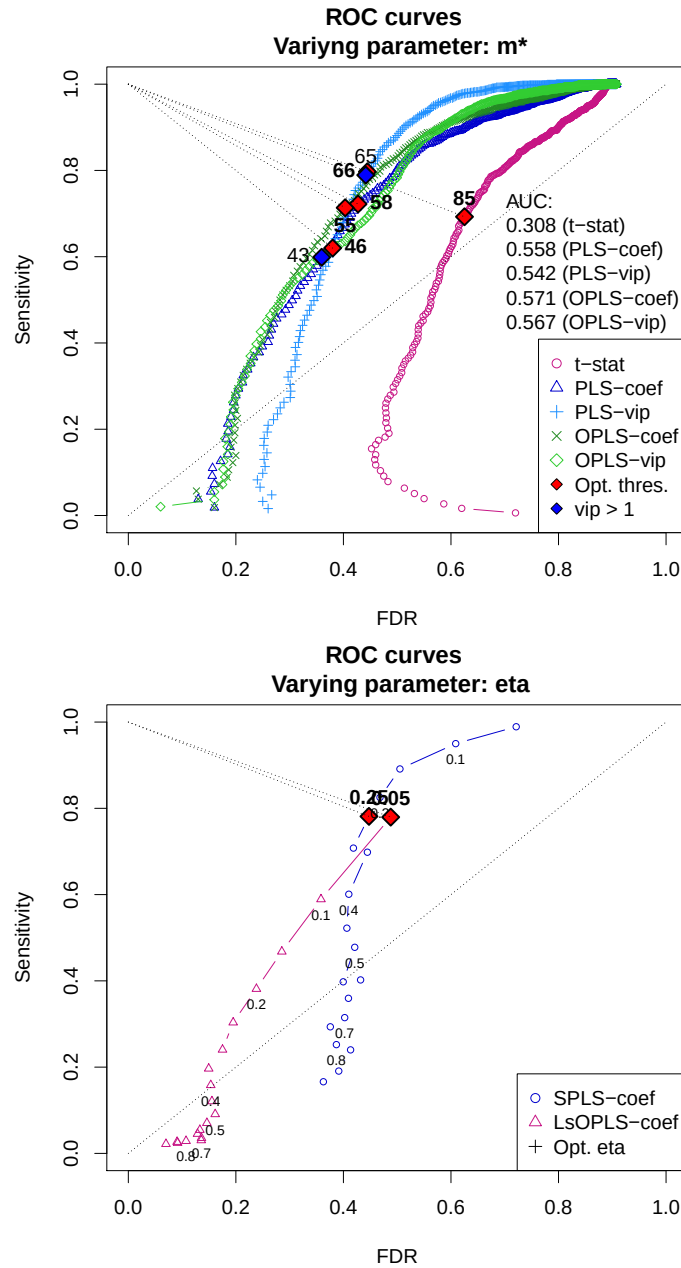


Figure 3.26 – ROC curves for feature accuracy (methods with an arbitrary threshold or sparse methods, $n = 60$). Varying parameters are either the selection threshold $m^* = 1, \dots, m$ or the sparsity parameter $\eta = 0.05, \dots, 0.95$ depending on the classifier. Optimal thresholds are represented in red.

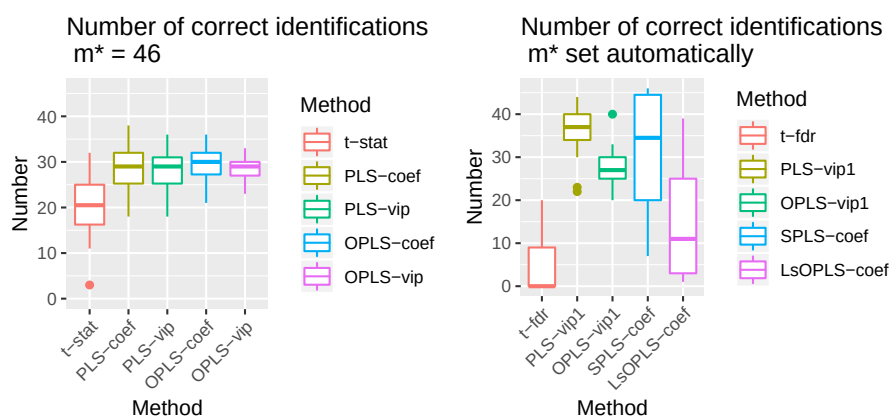


Figure 3.27 – Number of correct features selection. Note that for arbitrary threshold methods, it corresponds to the number of selected features for the 1:46th ranked features ($n = 60$).

Table 3.14 – Mean ROC indices from the actual selection sets based on the optimal MP (automatic threshold methods) for features accuracy ($n = 60$).

	Sensitivity	FDR	F1
t-fdr	10.22	72.4	14.91
PLS-vip1	79.13	43.42	65.98
OPLS-vip1	59.65	36.35	61.59
SPLS-coef	69.26	52.29	56.5
LsOPLS-coef	31.87	23.92	44.92

3.A.5 Selected features inspection on the spectral scale

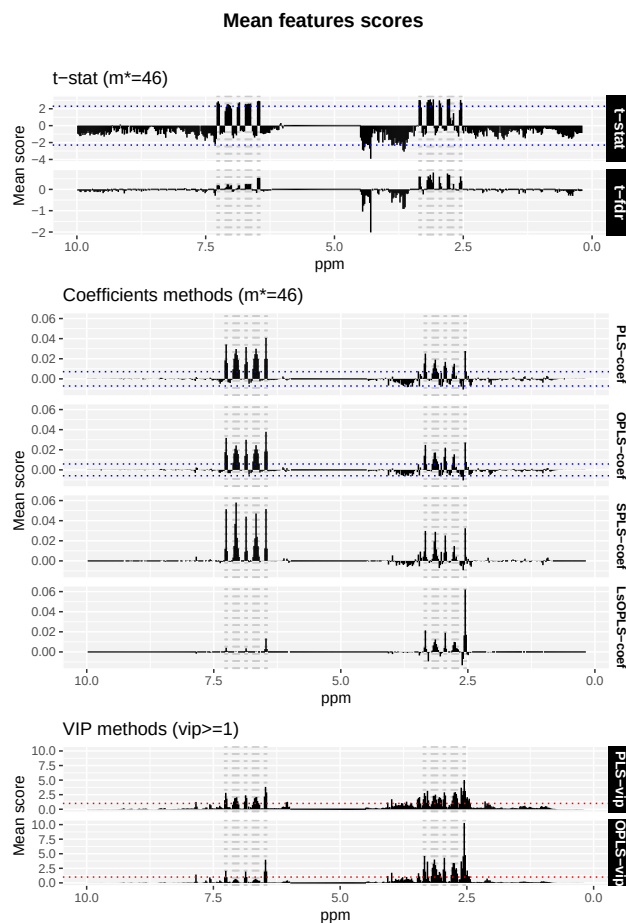


Figure 3.28 – Mean features scores *B*. Blue lines set the arbitrary threshold, it corresponds to the selection of 46 features ($n = 60$) and red lines represent the $VIP \geq 1$. For sparse methods, recall that all coefficients $\neq 0$ are selected.

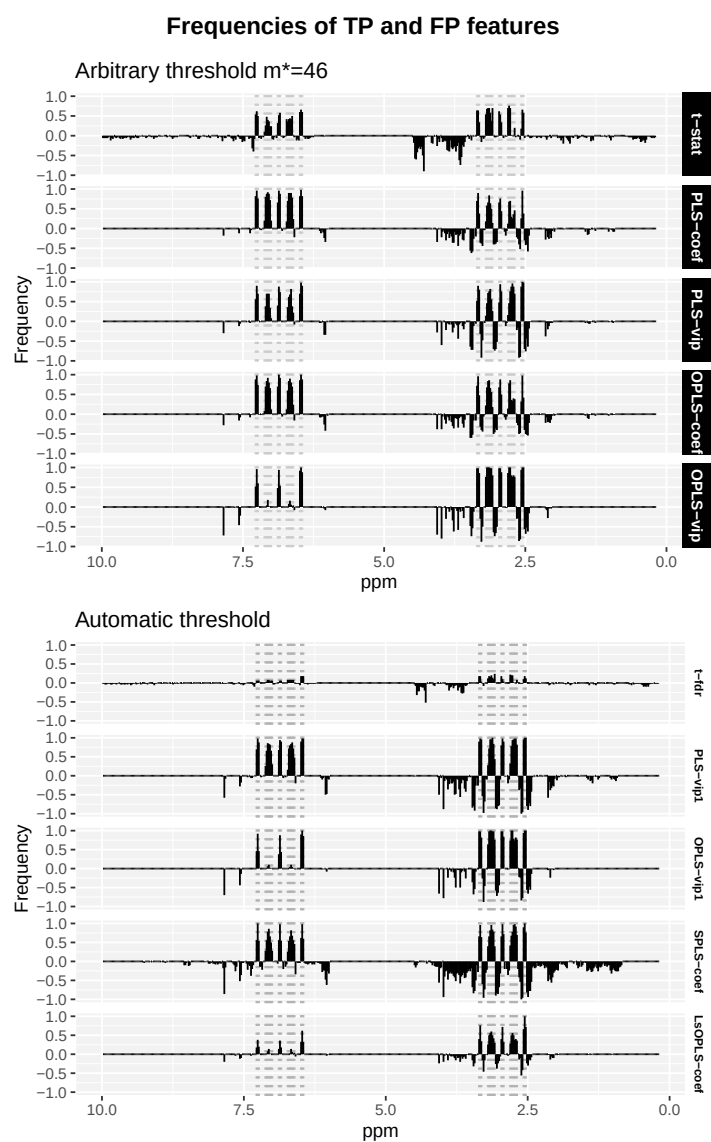


Figure 3.29 – Frequencies of TP and FP. Note that for methods with arbitrary threshold, it corresponds to frequencies of selected features for the 1:46th ranked features ($n = 60$).

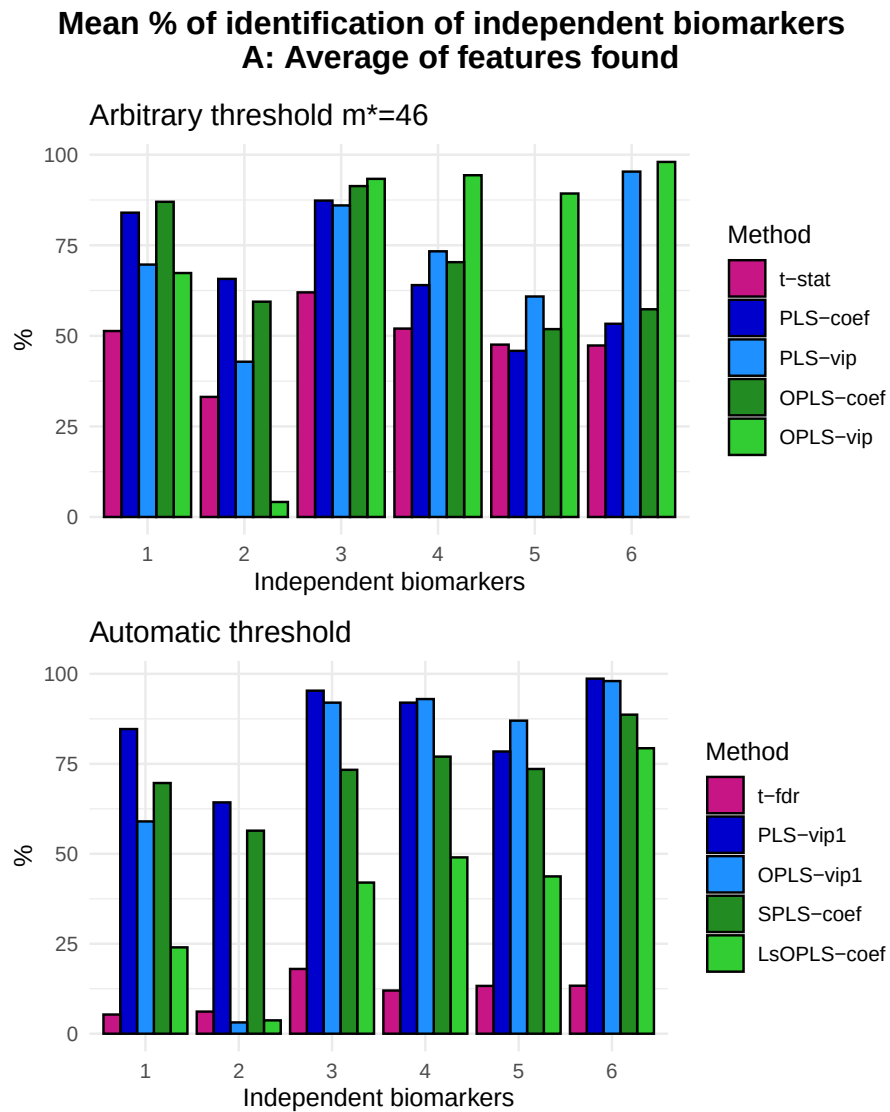


Figure 3.30 – Mean number of identification of independent biomarkers per method calculated as the average of features found for each biomarker. Note that for arbitrary threshold methods, it corresponds to frequencies of selected features for the 1:46th ranked features ($n = 60$).

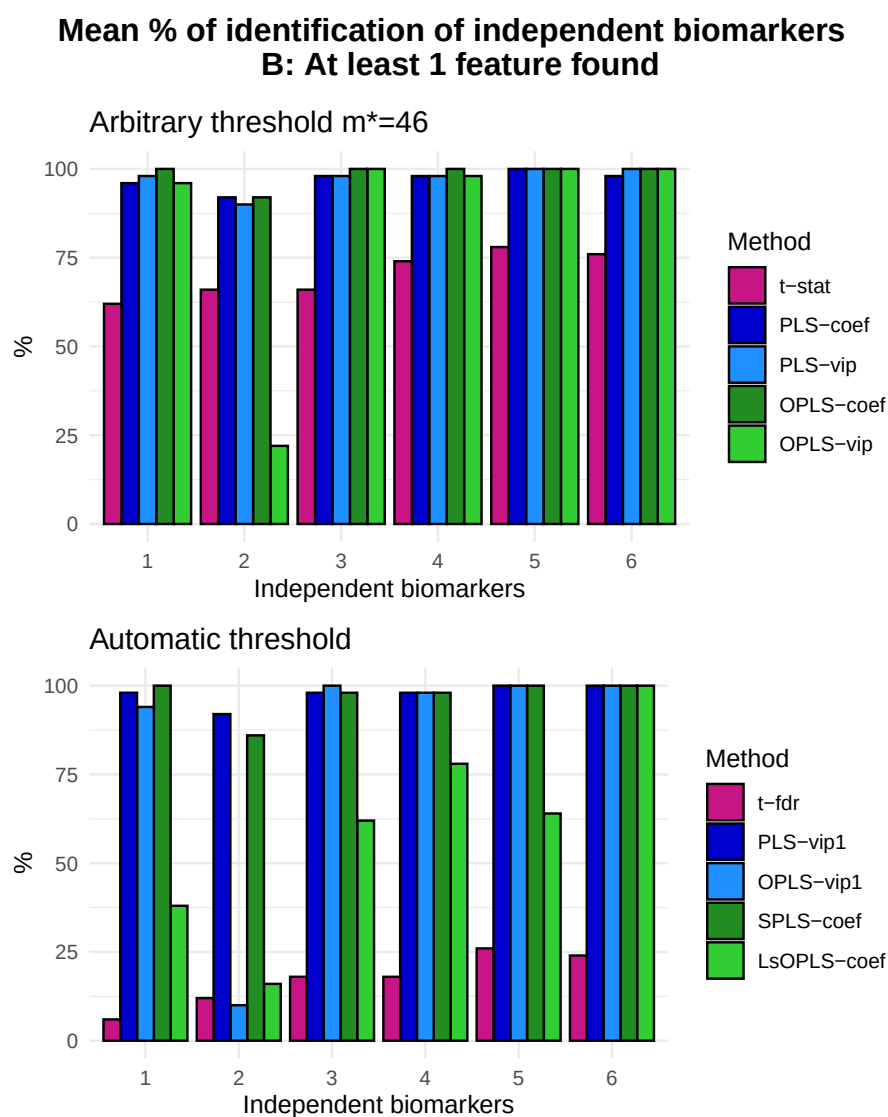


Figure 3.31 – Mean number of identification of independent biomarkers per method where at least 1 feature found for each biomarker. Note that for arbitrary threshold methods, it corresponds to frequencies of selected features for the 1:46th ranked features ($n = 60$).

3.A.6 Feature stability

Table 3.15 – Feature selection stability for the other automatic threshold methods ($n = 60$).

t-fdr	PLS-vip1	OPLS-vip1
9.142	67.94	72.42

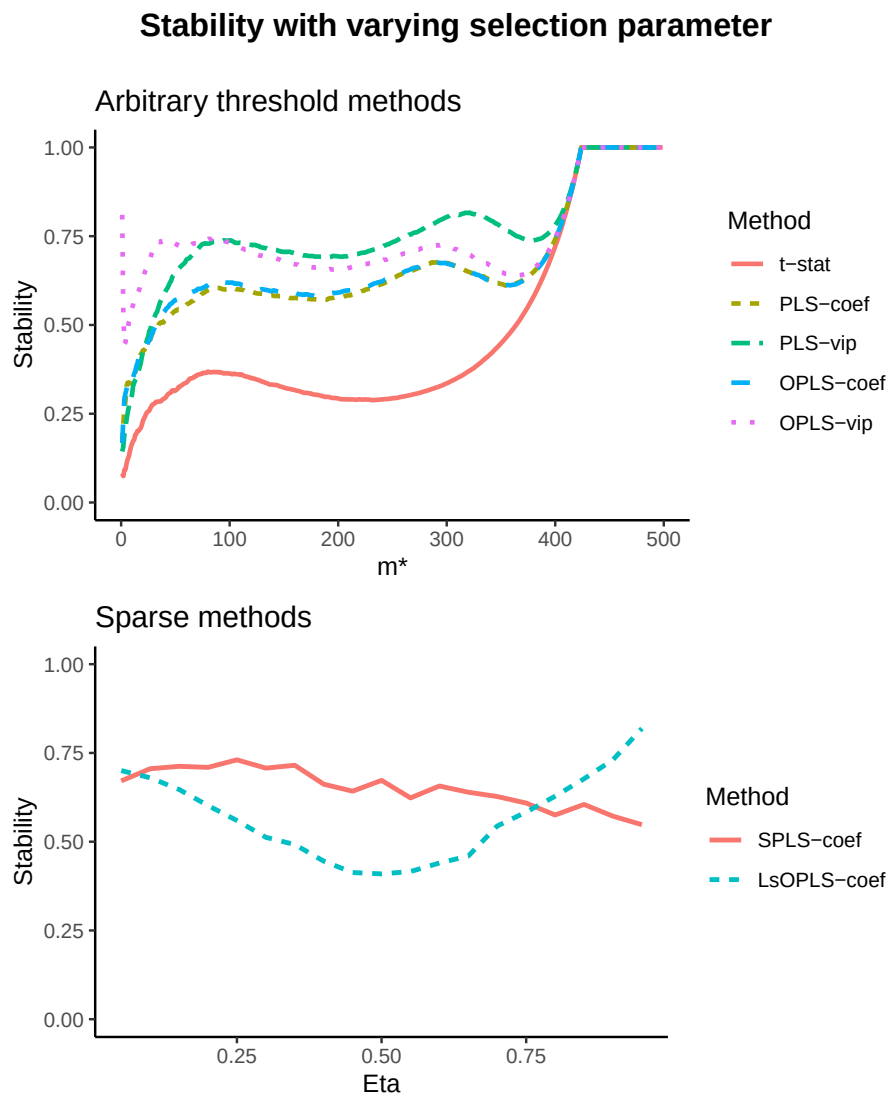


Figure 3.32 – Features Stability across the number of selected features for methods with an arbitrary threshold and across η for sparse methods ($n = 60$).

CHAPTER 4

ASCA-GLM and other GLM-multivariate techniques for advanced designs

4.1 Introduction

4.1.1 Analysis of designed experiments with high dimensional responses

In the field of experimental studies in life sciences, and especially for the “omics” sciences (e.g. genomics, transcriptomics, metabolomics), a large amount of information is commonly produced and results in multivariate quantitative data tables with an explosive number of highly correlated variables, often much larger than the number of samples available. Besides, in a large portion of these studies, the data are acquired on the basis of a multifactorial experimental design, where there is a need to decipher all the factors’ impact on the generated multivariate dataset. The inclusion of possible advanced data structures (such as nesting, crossing or unbalance), either related to the objectives or to the constraints of the study, ensures to model the data as accurately and completely as possible. Both aspects – high dimensional correlated variables and their advanced experimental design structure – are important features that need to be taken into account during their analysis.

Principal Component Analysis (PCA) and Partial Least Squares regression (PLS) are two popular multivariate dimension reduction techniques that are particularly

suited to analyse such correlated and high dimensional data: they simultaneously model the relationship between the variables and reduce the dataset dimensionality. Both methods provide powerful visualisation tools to ease the interpretation of the results. However, PCA is unsupervised, hence there is no insurance that it will give a clear insight on the influence of the factors from the design if unwanted nuisance factors are the main sources of variability. By contrast PLS (Wold *et al.* , 2001) is a supervised method and is, in addition, built on the basis of the correlation between the two blocks of data. But it needs to be tuned for model complexity to prevent overfitting and is essentially meant for very simple experimental designs, often including only one factor of interest. Furthermore, for both these multivariate methods, statistical properties and inferential possibilities are limited, apart from the jack-knife approach by Martens *et al.* (2001).

Regarding the advanced design issue, Analysis of Variance (ANOVA) (Neter *et al.* , 1996; Gelman, 2005) is a particularly well established and reliable tool for the analysis of a univariate quantitative response. MANOVA (Mardia *et al.* , 1979) is its direct multidimensional generalisation that tests the significance of the factors on a multivariate response matrix. However, the latter cannot handle data tables with more variables than observations because of the covariance matrix singularity. Moreover, it provides no further interpretation tools and its different related test statistics can give conflicting results (d. B. Harrington *et al.* , 2005).

Given the lack of tools to deal with such *wide collinear multi-response experimental design data*, several methods, that are typically based on a combination of statistical modelling and multivariate analysis, have emerged from the literature and are currently receiving the attention they deserve. Figure 4.1 presents the family of multivariate techniques suitable for the analysis of such data that will be discussed in this chapter. All techniques include a decomposition of the initial data matrix (or response matrix) into model effect matrices, followed by a multivariate dimension reduction step which allows exploring the relationship between experimental design factors and the observations or variables in the data set. Visualisations of observations and variables are available in all techniques, while some include the computation of importance measures for the factor effects and related significance tests.

Originally applied to metabolomic data, ANOVA-Simultaneous Components Analysis (ASCA) (Smilde *et al.* , 2005; Jansen *et al.* , 2005a) was developed as another multivariate generalisation of ANOVA to study the impact of a combination of fixed categorical predictors on a multivariate response. It combines the advantages of statistical modelling (ANOVA) and multivariate analysis (PCA). The first step of ASCA applies ANOVA in parallel on each response vector to decompose the response matrix into a sum of effect matrices. Then, the influence of each model effect on the multivariate response is analysed by

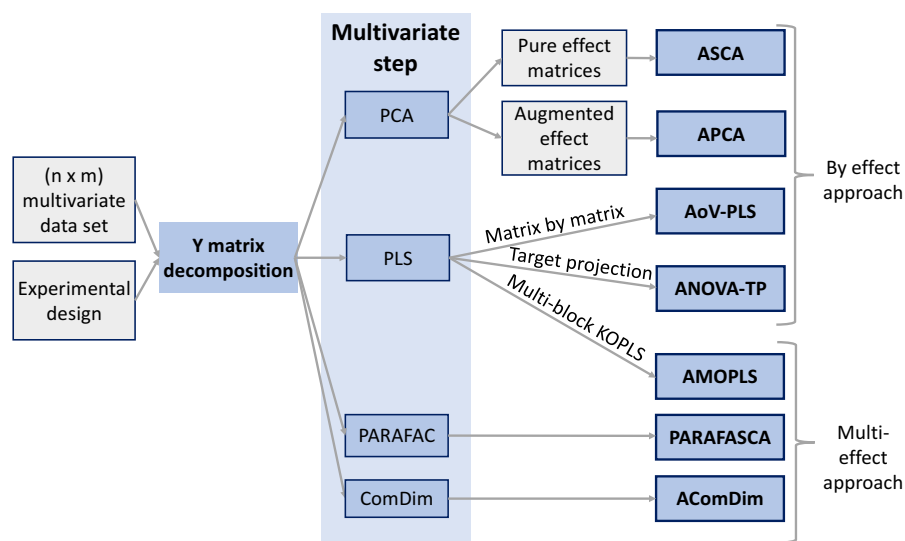


Figure 4.1 – Overview of techniques combining ANOVA modelling and multivariate data analysis

applying PCA on the effect matrices. Concurrently developed on proteomic data, ANOVA-PCA (APCA) (d. B. Harrington *et al.*, 2005) differs from ASCA in the multivariate step, when adding the residuals to the effect matrices prior to the PCA.

Several variants of ASCA and APCA have since been suggested in the literature. Analysis of Variance Partial Least Squares (AoV-PLS) applies PLS to the matrices resulting from the ANOVA decomposition (El Ghaziri *et al.*, 2015). In this approach, an individual PLS regression is implemented on each effect matrix individually. The PLS response matrix is the pure effect matrix and the predictor matrix is the residual-augmented effect matrix. If the PLS regression identifies the factor levels in the residual-augmented effect matrix, the effect of interest can be considered more important than the noise in the data.

ANOVA with target projection (ANOVA-TP) consists in applying the ANOVA decomposition followed by a PLS with target projection (Marini *et al.*, 2015). The PLS model is applied separately for each effect, with the corresponding residual-augmented effect matrix as the predictor matrix, and a matrix or vector of factor levels as the response. Once the model has been estimated, a target projection rotation is applied to ease interpretation. This consists in projecting the data matrix onto the regression vector(s) of the optimal PLS model in order to identify one (or several) latent variable(s), the target-projected component(s). These four techniques allow analysing wide and short data tables with an

advanced design, but are all based on an individual study of each model effect, which can be tedious when the model is complex. In contrast, this article will focus on three other approaches - PARallel Factor - Simultaneous Component Analysis (PARAFASCA, PARAFAC combined to ANOVA), ANOVA Common Dimensions (AComDim, ComDim combined to ANOVA), and ANOVA Multi-block Orthogonal Partial Least Squares (AMOPLS Multibloc-PLS combined to ANOVA) - where the multivariate step allows to analyse and interpret all or several model effects with an integrated approach. These three global methods have already been applied in the chemometrics and/or metabolomics literature. ComDim and MOPLS are two similar multiblock techniques designed to describe a series of data blocks observed for the same samples while PARAFAC is originally meant for multi-way data. Even though they differ in their construction, they have the common goal to analyse the effects of variables from the design on the measured response matrix.

4.1.2 ASCA⁺, an extension to unbalanced experimental designs

The ANOVA decomposition presented in the ASCA/APCA literature is not suitable for unbalanced designs as it is based on effect estimators that become biased when the design is not balanced. Extensions of ASCA to unbalanced designs have been developed by Stanimirova *et al.* (2011) based on type III Sum of Squares (SS) and, more generally, by Thiel *et al.* (2017) who suggests the use of the General Linear Model (GLM) notation and subsequent adaptations as an alternative to ANOVA based effect matrices estimation. This GLM framework, called ASCA⁺, also has the advantage (in comparison with the ANOVA approach) to be very generic since it is not based any more on specific variance decomposition equations linked to each possible model. This allows for an easier computation of the effect matrices. GLM also offers a potential extension to models including quantitative or random effects.

4.1.3 Objectives

The overall aim of this chapter is to present and compare possible ASCA/APCA-based techniques that are applicable to wide and short data tables with balanced or unbalanced advanced designs, and allow a multivariate analysis of all or several model effects in one global approach.

More precisely, this chapter presents within one single framework and with common notations the three global techniques listed in Figure 4.1: PARallel Factor - Simultaneous Component Analysis (PARAFASCA), ANOVA Common Dimensions (AComDim), and ANOVA Multi-block Orthogonal Partial Least

Squares (AMOPLS). They are combined with the ASCA/APCA⁺ GLM approach in order to allow their extension to unbalanced designs. This article shows that each of these methods has specific valuable or problematic properties and offers several tools to order, interpret and represent graphically model effects results. The article also suggests some enhancements of the methods (e.g. testing procedures and new types of graphical outputs), clarifies some aspects that were not detailed in the original articles and suggests simplifications where possible. The GLM extension of ASCA/APCA to unbalanced experimental designs and a review of ASCA/APCA are presented before outlining the three other techniques of interest. These are then applied to a ¹H NMR unbalanced data set, compared to each other and to ASCA/APCA approaches.

Note that in the next Chapter 5, the ASCA⁺ methodology will be extended to Linear Mixed Model (LMM) in order to integrate random effects to the modelling step.

4.2 ASCA⁺ and related methods

4.2.1 Overview

This section introduces a general framework for ASCA/APCA-like methods based on an effect matrix decomposition followed by a dimension reduction multivariate step.

Figure 4.2 shows the steps of ASCA/APCA approaches. The general scheme is illustrated here for a 2-Factor ANOVA. Its background and steps are summarised in broad terms below. In a first step, the original response data matrix $\mathbf{Y}$ is decomposed in effect matrices $\hat{\mathbf{M}}_f$ according to a model linked to the experimental design. In the standard ASCA/APCA method, an ANOVA model is used, but this approach can be generalised using a General Linear Model (GLM) framework. The main difference between the ANOVA and GLM approaches is the parameter estimation method. The GLM extends linear regression and ANOVA to cases with several response variables, different types of independent variables and unbalanced designs. In a second step, the decomposed effect matrices are analysed using a dimension reduction multivariate approach. PCA is used for ASCA and APCA, while AComDim, PARAFASCA and AMOPLS are based on other multivariate techniques that will be described later.

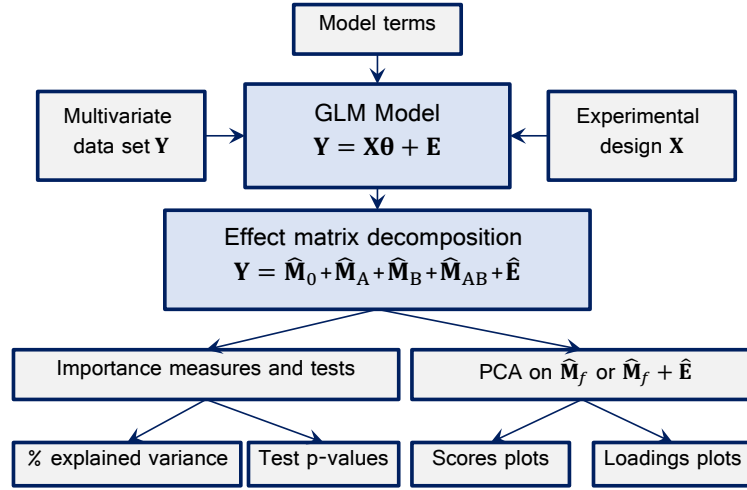


Figure 4.2 – Overview of the GLM-based approach of ASCA/APCA.

4.2.2 Fixed effects ANOVA in the General Linear Model notation

The response matrix decomposition consists in decomposing the response matrix $\mathbf{Y}$ into matrices that each corresponds to the estimation of a given effect in a model adapted to the original experimental design.

In Figure 4.2, the model is an ANOVA 2 model with two crossed factors A and B. Such an ANOVA 2 model with one response variable can be expressed using the “cell means model” (Neter *et al.*, 1996) as: $y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}$ where y_{ijk} is the value of the response variable for the i -th level of factor A, the j -th level of factor B and the k -th observation, with $i=1$ to a , $j=1$ to b and $k=1$ to n . μ is the mean value of the response variable and ϵ_{ijk} is the error term usually supposed to be $\epsilon_{ijk} \sim iN(0, \sigma^2)$. α_i and β_j are the main effects for factors A and B and $\alpha\beta_{ij}$ is the interaction term.

In the standard ANOVA approach used in ASCA/APCA, the estimators for the model effects are based on mean differences: $y_{ijk} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (\bar{y}_{.j} - \bar{y}_{..}) + (\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}) + (y_{ijk} - \bar{y}_{ij}) = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j + \hat{\alpha}\beta_{ij} + e_{ijk}$. The following constraints are verified for these estimators when the design is balanced: $\sum_{i=1}^a \hat{\alpha}_i = 0$, $\sum_{j=1}^b \hat{\beta}_j = 0$, $\sum_{i=1}^a \hat{\alpha}\beta_{ij} = 0 \forall j$ and $\sum_{j=1}^b \hat{\alpha}\beta_{ij} = 0 \forall i$.

These estimators can then be gathered together column by column (one column per response) in effect matrices $\hat{\mathbf{M}}_f$ and the response matrix decomposed as $\mathbf{Y} = \hat{\mathbf{M}}_0 + \hat{\mathbf{M}}_A + \hat{\mathbf{M}}_B + \hat{\mathbf{M}}_{AB} + \mathbf{E}$ where $\hat{\mathbf{M}}_0$ contains the overall means, $\hat{\mathbf{M}}_A$

corresponds to the main effects of factor A, $\hat{M}_B$ to the main effects of factor B, $\hat{M}_{AB}$ to the interactions between A and B, and E to the model residuals.

Since fixed effects ANOVA can be seen as a special case of linear regression, the GLM extension of ASCA, called ASCA⁺ (Thiel *et al.*, 2017), will serve here as a basis for further generalisations. The main motivation is that ANOVA estimators are biased in the case of unbalance, and the total variance decomposition does not anymore produce orthogonal SS summing to Total SS (SS_{Total}). GLM thus provides a general solution to analyse unbalanced and designed data with fixed effects and is equivalent to ANOVA in balanced designs.

The fixed effects ANOVA can be rewritten in the form of a GLM matrix notation as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad (4.1)$$

where $\mathbf{y}$ is the $(n \times 1)$ response vector, $\mathbf{X}$ is the $(n \times (p + 1))$ model matrix, $\boldsymbol{\beta}$ is the $((p + 1) \times 1)$ vector of parameters and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$ is the $(n \times m)$ vector of residuals. When the model contains F effects of interest (main factor effects, interactions, ...), $\mathbf{X}$ will be decomposed into $F + 1$ blocs that includes the constant term, and one for each model effect: $\mathbf{X} = (\mathbf{X}_0 \mid \mathbf{X}_1 \mid \dots \mid \mathbf{X}_F)$. In order to take into account the linear constraints between parameters linked to each model effect, Thiel *et al.* (2017) recommend, in the ASCA context, to use a *sum* or *deviation* coding of the categorical variables in the model matrix. It guarantees the orthogonality between the F effects blocs of the model matrix in the balanced case and equivalence between GLM and ANOVA results (see Govaerts *et al.* (2019) more details). With this coding, a main effect of factor A with a levels, the (ij) th element of $\mathbf{X}_A$ is defined as:

$$\mathbf{X}_{A(ij)} = \begin{cases} 1 & \text{if level of } A = j \text{ for experiment/observation } i, \\ -1 & \text{if level of } A = a \text{ for experiment/observation } i, \\ 0 & \text{otherwise.} \end{cases} \quad (4.2)$$

The model matrices for the interaction terms are accordingly built by a column-wise multiplication of the corresponding main factors model matrices.

4.2.3 Effect matrix decomposition

For a multivariate response matrix $\mathbf{Y}$ of size $(n \times m)$, the GLM can be written in matrix notation as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Theta} + \mathbf{E}; \quad (4.3)$$

where $\boldsymbol{\Theta}$ is the $(p \times m)$ parameter matrix and $\mathbf{E}$ is the $(n \times m)$ error matrix.

By splitting the ordinary least squares estimator of Θ : $\hat{\Theta} = (X'X)^{-1}X'Y$ with respect to each model effect: $\hat{\Theta}' = (\hat{\Theta}'_0 \mid \hat{\Theta}'_1 \mid \dots \mid \hat{\Theta}'_F)$, the ASCA⁺ decomposition of the response matrix into effect matrices is:

$$\hat{Y} = \hat{M}_0 + \sum_{f=1}^F \hat{M}_f + \hat{E}; \quad (4.4)$$

where $\hat{M}_f = X_f \hat{\Theta}_f$ and $\hat{E} = Y - X \hat{\Theta}$.

The estimated effect matrices are recovered from the estimated model parameters and the design matrix with the following formula: $\hat{M}_f = X_f \hat{\Theta}_f$ for an effect f .

4.2.4 Outputs

A range of outputs are available in ASCA/APCA⁺. They are explained in this subsection and will be implemented on the case study in Section 4.4. Two classes of outputs are usually derived from the ASCA-related methods: (1) diagnostic measures to evaluate the importance and significance of model terms and (2) graphical outputs to interpret and visualise the model effects in terms of observations and variables of the original dataset.

For balanced designs, the importance of a given effect f can be quantified by the percentage of the total variance explained explained by that effect: $\%Var\ f = \frac{\|\hat{M}_f\|^2}{\|Y - \hat{M}_0\|^2} \times 100$ where $\|\cdot\|$ is the Frobenius norm of a matrix. For unbalanced cases, the analysis of variance decomposition equation being not more valid, (Thiel *et al.*, 2017) proposed an adapted solution based on the type III Sum of Squares commonly used in GLM.

To evaluate the statistical significance of a model effect, the principle of the general F test for embedded models serves as a basis of all approaches suggested in the ASCA literature. This test compares the importance of model errors when the effect is or is not included in the model, or equivalently, the importance of the factor effect compared to the residual errors. For univariate fixed GLM, the test statistic is commonly calculated as $F_{obs} = MS_f / MS_E$ where MS_f and MS_E are, respectively, the Mean Square of effect f and of the Residual Mean Square. F_{obs} is expected to be close to 1 if the effect is not useful in the model. To determine the significance of the statistic, a p -value is calculated from an $F(p_f, n - p, 1 - \alpha)$ distribution where p_f is the number of parameters linked to an effect f , the degrees of freedom are linked to both the numerator and denominator of the test statistic, and $1 - \alpha$ is the $(1 - \alpha)^{th}$ (e.g. 95th) percentile of the $F_{(p_f, n-p)}$ distribution. In the multivariate balanced case, mean squares are replaced by squared Frobenius norms of the effect and the error matrices. p -values can then be derived with permutation or bootstrap procedures. Several permutation methods are described

in the literature and must be adapted to the type of effect and model of interest (see (Anderson & ter Braak , 2003; Vis *et al.* , 2007) and Section 4.4.4).

The subsequent step is the visual inspection of the results, one of the main purposes of ASCA. The loadings and scores obtained from the PCA decomposition of the pure or residual-augmented effect matrices provide rich information to interpret the result of the ANOVA/GLM modelling. Three types of scores plots can be obtained and are illustrated in Figure 4.3 with example plots on the effect of the three-level Hippurate factor in the case study of this article (further described in Section 4.4).

In ASCA, to interpret each model effect f , a Principal Component Analysis (PCA) is performed on the “pure” effect matrix : $\hat{M}_f = T_f P'_f$. Plotting the first loadings vectors of P_f directly highlights which responses are most affected by a model effect f . The scores are usually not interpreted directly because the corresponding plots lack interesting information: they only show how effect levels are located to each other, without providing an indication about the significance of the distances between data points. In the ASCA scores plot presented in Figure 4.3, the data points corresponding to the three levels of the Hippurate factor are mostly spread along the first PC, which explains 97.7% of the variation of the Hippurate effect matrix, while the second component only explains 2.3% of the variation. This means that the effect of Hippurate on the response matrix is mostly linear over these three levels, with a small second order effect (see details in Section 4.4).

In APCA, PCA is applied to a residual-augmented effect matrix. The corresponding scores plots show how the factor levels and the n observations are situated with respect to each other when only the corresponding effect and the residuals are kept in the model and all the information related to the other model effects has been removed from the data. In Figure 4.3, the APCA scores plot shows that 88.7% of the variance is explained by the first component and separates the three Hippurate levels. The second PC explains 4.5% of the variance and is linked to model residuals. More residual information is included in further components. No other specific structure appears in these data.

A third type of scores plot, suggested by (Zwanenburg *et al.* , 2011) and referred to as “ASCA-E” in (Thiel *et al.* , 2017), projects rather the augmented effect matrix $\hat{M}_f + \hat{E}$ on the principal components (PCs) obtained from PCA on the pure effect matrix, to include individual variations between observations that is not explained by the model. It leads to a better visualisation of the size of each model effect with respect to the variability of the residuals. The “augmented”

scores matrix is calculated as $\hat{T}_f^a = (\hat{M}_f + \hat{E}) \times \hat{P}_f$. The ASCA-E approach can be advantageous when the percentage of variation due to the residuals in the model is high because, in this case, the APCA scores plots may show a dispersion of the samples along PC1 due to noise while the separation of the factor levels is along PC2 or PC3. In Figure 4.3, the ASCA-E scores include individual variation on both axes as expected but caution must be taken in its interpretation: the horizontal variation observed on the first PCA component is of interest since this component explains 97.7% of the model effect variation. However, the second axis is of limited interest since it is linked to a very small second order effect of Hippurate and does not provide the best representation of the variation of the model residuals.

The loading plots extracted from the PCA decomposition of the pure or the residual-augmented effect matrices are important to identify which of the m original variables are linked to each effect in the model. In Figure 4.3, the loading plots corresponding to the first PC in ASCA and APCA clearly display the four expected Hippurate peaks (see Section 4.3 for more details). In general, ASCA loadings should be preferred to APCA ones because they are less affected by residuals and will show a purer image of the model effect for each of the m variables.

In this thesis, loadings plots will all be calculated with the ASCA approach and scores plots with the ASCA-E approach.

The three methods are presented in chronological order of their publication in the literature.

4.2.5 PARAFASCA

PARAFASCA is described in Jansen *et al.* (2008). In this article, the technique presented consists in implementing a PARAFAC following the ANOVA decomposition instead of applying one PCA to each effect matrix as done in ASCA/APCA. PARAFASCA can be seen as complementary to ASCA: it can help visualise the complex model effects in a more interpretable and parsimonious way. The original article does not provide the detailed steps of a general technique as PARAFAC was applied to a three-factor repeated measure design where only the main effect of one factor and the interaction effect of this factor with another are of interest in the full three-way ANOVA model. The corresponding model coefficients are folded into a three-way array with the three dimensions (or modes) corresponding to the two factors of interest and the vector of m response variables. PARAFAC is then applied to this array in order to provide a suitable way to visualise and interpret the set of effects of interest.

The approach presented in the current chapter is more generic: it shows how PARAFASCA can be applied to a full multi-factor crossed ANOVA model using

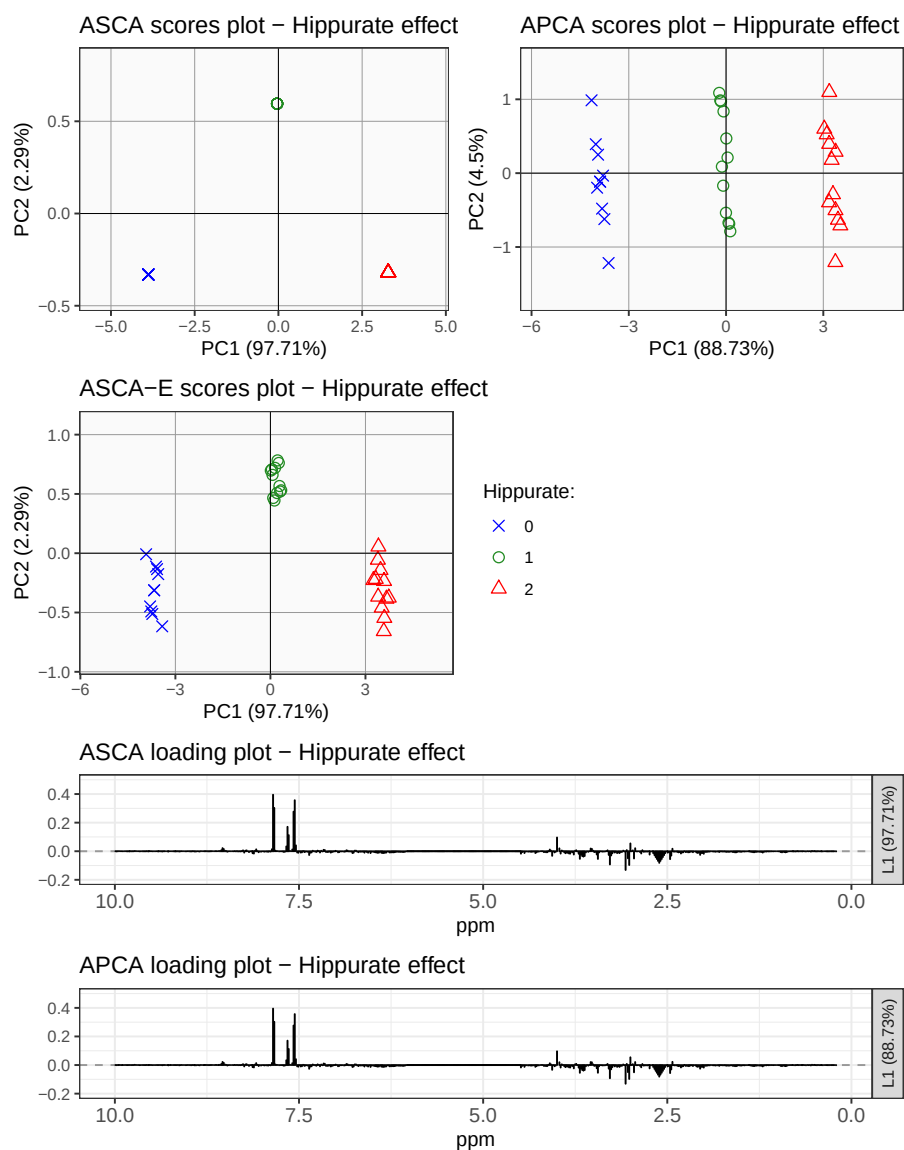


Figure 4.3 – Examples of scores and loadings for ASCA, APCA and ASCA-E for the effect of the three-level Hippurate factor in the case study of this article (see Section 4.4)

the GLM approach in keeping the spirit of Jansen *et al.* (2008). It also illustrate, on the case study, the interest of applying PARAFASCA on a partial model.

Approach

PARAFAC is a multivariate technique to analyse D -way arrays. First suggested in Hitchcock (1927), the technique was again put forward by Carroll & Chang (1970) under the name Canonical Decomposition (also called CANDECOMP) and Harshman (1970) under the name PARAFAC. The technique is sometimes referred to as CP, which stands for CANDECOMP/ PARAFAC (Favier & de Almeida, 2014).

PARAFAC is conceptually a generalisation of PCA on a $(n \times m)$ matrix with 2 “modes” (observations and variables) to arrays with D modes. PARAFAC provides a multilinear decomposition of the array into C sets of D vectors, where C is the number of components extracted. The vectors can be interpreted as scores and loadings as in PCA but since they are treated the same numerically, the word used to name them is unimportant and will depend on the context (Bro, 1997). The corresponding loadings/scores matrices can be computed by Alternating Least Squares (ALS) or other algorithms which are discussed in more details in Faber *et al.* (2003) and Tomasi & Bro (2006).

PARAFAC is used in a range of scientific fields among which chemistry, biology, medicine, environmental and computer sciences. In particular, it is widely used in chemometrics to analyse higher dimension data obtained by spectroscopy, chromatography or NMR (Lenhardt & Dramićanin, 2016; Faber *et al.* , 2003).

PARAFASCA has been applied to metabolomic data obtained by NMR spectroscopy in the original article (Jansen *et al.* , 2008), while approaches close to PARAFASCA and based on ANOVA and PARAFAC have been used to analyse multi-way fluorescent probe spectroscopy anaesthetics data in (Horochowska *et al.* , 2016) and three-way descriptive sensory analysis data in (Romano *et al.* , 2015).

Model

The first step of PARAFASCA is based on the same principle as the matrix decomposition in ASCA⁺. It consists in estimating the ANOVA or GLM model parameters for each of the m response variables but without building the complete effect matrices $\mathbf{M}_f$. The next step of PARAFASCA applies a PARAFAC decomposition to a selected combination of model terms according to the questions of interest in the study and the structure of the design (crossed, nested etc...). Such decomposition is especially suited to models containing complex interaction terms. The existing articles related to PARAFASCA do not present a generic methodology but illustrate the principles on case studies. On this basis, we present below a generic way to apply PARAFASCA to the ANOVA 2 model discussed in Section 2 when no prior knowledge is available on the way to

combine model effects.

After the GLM model estimation, the idea is to combine and fold in a three way array $\mathbf{Y}_G$ all the parameters estimated in the ANOVA/GLM model except for the constant term (see Figure 4.4). $\mathbf{Y}_G$ is of size $(a \times b \times m)$ and the element y_{ijh} of $\mathbf{Y}_G$ is calculated as follows:

$$y_{ijh} = \hat{\alpha}_{ih} + \hat{\beta}_{jh} + \hat{\alpha}\hat{\beta}_{ijh}$$

where $i = 1, \dots, a$, $j = 1, \dots, b$ are the factor levels and h is the response of interest with $h = 1, \dots, m$.

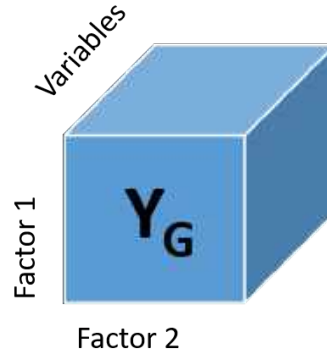


Figure 4.4 – Example of a PARAFASCA array

The PARAFAC decomposition can then be written as (Bro, 1997; Smilde *et al.*, 2004):

$$y_{ijh} = \sum_{c=1}^C u_{ic}v_{jc}w_{hc} + e_{ijh} \quad (4.5)$$

or

$$\mathbf{Y}_G = \sum_{c=1}^C \mathbf{u}_c \Delta \mathbf{v}_c \Delta \mathbf{w}_c + \mathbf{E}_G \quad (4.6)$$

where C is the number of computed PARAFAC components with $c=1$ to C . $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ are the three loadings/scores matrices of the PARAFAC decomposition with elements u_{ic} , v_{jc} and w_{hc} , and dimensions $(a \times C)$, $(b \times C)$ and $(m \times C)$ respectively. $\mathbf{u}_c$, $\mathbf{v}_c$ and $\mathbf{w}_c$ are vectors and the c -th columns of $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ respectively. $\mathbf{E}_G$, with elements e_{ijh} , is the residual array of the PARAFAC model, of the same dimension as $\mathbf{Y}_G$. Δ denotes a tensor product.

Figure 4.5 presents a schematic visualisation of the three-way PARAFAC decomposition of matrix $\mathbf{Y}_G$ with C components and a residual array $\mathbf{E}_G$.

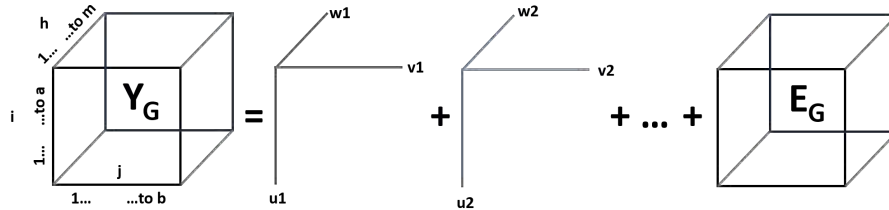


Figure 4.5 – Visualisation of a three-way PARAFAC decomposition of matrix $\mathbf{Y}_G$ with C components and a residual array $\mathbf{E}_G$

The details of an Alternating Least Squares (ALS) algorithm which can be used to compute PARAFAC components are presented in Appendix 4.A.

Different techniques are available to identify the optimal number of components to extract in a PARAFAC analysis and are discussed in more details in Bro (1997), Harshman & Lundy (1994) and Bro & Kiers (2003). The split-half technique cannot be applied in this context as none of the array dimensions corresponds to the observations. The number of components can however be selected based on a scree plot and confirmed by the interpretation of the results. Please note that a high number of components may lead to unstable algorithm results, with different effects corresponding to different components in each run of the same model. Some ALS algorithms also allow to force the orthogonality of the components on one or several modes. In the PARAFASCA context, a reasonable approach is to impose orthogonality between the loadings of the Dth mode. This prevents model terms mixing and makes their interpretation easier.

Output

The following outputs can be derived from a PARAFASCA analysis. For a chosen number of components C , a (marginal) percentage of explained variance can be computed for each component c of the PARAFAC decomposition. If a given component can be linked to one of the model effects, this percentage provides information on the relative size of the effect in the ANOVA model. Please note that if components are not constrained to be orthogonal in the ALS algorithm, these marginal percentages of variance explained may be difficult to interpret

because they do not sum to the total variance explained by the decomposition.

$$PC_c = \frac{\|u_c \Delta v_c \Delta w_c\|^2}{\|Y_G\|^2} \times 100$$

A scree plot presenting the total percentage of variance explained by models with different numbers of components C can also be used to select the number of components to be included in the model. This approach will be applied in Section 4.4.

The scores/loadings matrices U , V and W can then be represented graphically in order to interpret the parameters of the m ANOVA/GLM models. Mean effect scores available in matrices U and V allow to interpret the main and interaction effect of factors on the response and the contribution of the model terms to each component. The loading matrix W of size $(m \times C)$ may be represented column by column to visualise the link between the response variables and each PARAFAC component. Section 4.4 will show that it is convenient to represent the scores either by mode, to illustrate which mode or combination of modes is linked to each component, and then, to represent for each component the tensor product of scores in parallel to the corresponding loading vector (mode D) to understand which response changes are linked to the possible combinations of factor levels.

4.2.6 AComDim

AComDim is described in Jouan-Rimbaud Bouveresse *et al.* (2011).

Approach

AComDim applies the ComDim approach to the matrices obtained from the ANOVA/GLM matrix decomposition. AComDim is based on a Singular Value Decomposition (SVD) of **sample** "variance-covariance matrices" calculated on the residual-augmented effect matrices M_f^a . It aims at comparing the variation of the response due to the factor levels of the experimental design with the residual variability (Jouan-Rimbaud Bouveresse *et al.* , 2011) and therefore provides information on the importance of specific factor effects and interactions. ComDim finds its origins in the Common Components and Specific Weight Analysis (CCSWA) approach which was first used to analyse sensory and chemometric data linked to food (Qannari *et al.* , 1995, 2000; Mazerolles *et al.* , 2002). The method was then applied to other contexts and renamed Common Dimensions (El Ghaziri *et al.* , 2016).

ComDim aims at analysing simultaneously several data matrices of size $(n \times m_i)$ by finding a common space of representation, with each matrix having a specific

weight (or salience) λ_f^c in each of the orthogonal common dimensions or common components (CC) of this space (Mazerolles *et al.*, 2002). The technique relies on the weighted sum of **sample** variance-covariance matrices or association matrices. This is different from many multivariate approaches where the **variable** variance-covariance matrix is used.

ACoDim has been applied to the analysis of duplicate apple reflectance spectra, NMR relaxation curves of starch-lignin, fluorescence and Mid-InfraRed spectroscopies of wine in the original article (Jouan-Rimbaud Bouveresse *et al.*, 2011). The technique has also been used on near infrared sensors spectral data (Amat *et al.*, 2010), irradiated resins spectral data (Korifi *et al.*, 2015) and metabolic indicators of olive oil (Korifi *et al.*, 2016).

Model

The first step of ACoDim is the ANOVA/GLM matrix decomposition. In a second step, the residual-augmented effect matrices M_f^a ($f = 1, \dots, F$) and the residual matrix E (written $M_{(F+1)}^a$ below) are column-centred - which is already the case for a balanced design - and normalised at the matrix level through a division by their Frobenius norm. This normalisation ensures that the matrices are of similar magnitude and avoids the dominance of one matrix over the others (Jouan-Rimbaud Bouveresse *et al.*, 2011). ComDim is then simultaneously applied to this set of residual-augmented effect matrices M_f^a (including the residual matrix E) to compute common components and related effect saliences. To do this, these centred and normalised matrices M_f^a ($f = 1, \dots, (F+1)$) are used to build related $(n \times n)$ **sample** variance-covariance matrices Ψ_f defined as $M_f^a M_f^{a'}$. This approach ensures that the resulting matrices $\Psi_f = M_f^a M_f^{a'}$ all have the same $(n \times n)$ dimensions and can be summed, even if the matrices M_f^a include a different number of variables m_i (although this is not the case with effect matrices).

ComDim decomposition is then computed dimension by dimension in order to minimise the following final criteria:

$$\sum_{f=1}^{(F+1)} \|\Psi_f - \lambda_f^c q_c q_c'\|^2 \quad (4.7)$$

where each q_c is obtained from an SVD and $\lambda_f^c = q_c \Psi_f q_c'$.

It is then possible to summarise the final results in the following global formulae (Mazerolles *et al.*, 2006; Jouan-Rimbaud Bouveresse *et al.*, 2011). Each sample

variance covariance matrix is firstly decomposed as

$$\Psi_f = M_f^a M_f^{a'} = \sum_{c=1}^C \lambda_f^c q_c q_c' + E_f = \Psi_f^* + E_f \quad (4.8)$$

and, written globally, as

$$\Psi_G = \sum_{f=1}^{(F+1)} \Psi_f^* + E_G = Q \Gamma Q' + E_G \quad (4.9)$$

where Q is the $(n \times C)$ matrix $(q_1, q_2, \dots, q_C)$. Γ is a diagonal matrix with $\Gamma_{cc} = \sum_{f=1}^{(F+1)} \lambda_f^c$ and $E_G = \sum_{f=1}^{(F+1)} E_f$.

The details of the algorithm used to compute common components and saliences are presented in Appendix 4.B.

Output

The percentages of explained variance for each common dimension c show their relative importance. They can be calculated as (Qannari *et al.*, 1995, 2000):

$$\frac{\mu_c}{\sum_{f=1}^F \|\Psi_f\|^2} \text{ where } \mu_c = \sum_{f=1}^{(F+1)} (\lambda_c^f)^2 \quad (4.10)$$

A scree plot presenting the cumulative percentage of variance explained by the common dimensions can be used to select the number of common dimensions to visualise. This approach will be applied in Section 4.4.

For each common dimension c , the saliences λ_c^f help to interpret the link between the dimension and the model effects. A high salience for an effect matrix on a given common dimension indicates that this dimension contains a large amount of information linked to the corresponding model effect. The sum of saliences for a given matrix over all dimensions tends to 1, which means that saliences can be interpreted as proportions. The first common dimension is linked to the largest common source of variance in the data. This is usually the residuals as they were added to each effect matrix to produce the residual-augmented effect matrices. The residual matrix is thus expected to have a high salience on the first common component.

The scores and loadings computed in AComDim can be interpreted in the same way as for other techniques in the same family. The scores for each dimension are simply defined as the singular vectors q_c . They link common dimensions with design factor levels and allow visualising single observations.

Loadings link common components with the original variables. The loadings are not precisely defined in the original AComDim article and different options are available to calculate and visualise them. In Jouan-Rimbaud Bouveresse *et al.* (2011), the authors suggest in their application to compute loading vectors for dimension c as the product $p_c = M_f^{a'} q_c$ where f is the model effect with maximum salience λ_f^c for dimension c . In this chapter, we suggest using a more general definition which takes into account effect matrices for all model terms weighted by their respective saliences: $p_c = \sum_{f=1}^{(F+1)} \sqrt{\lambda_f^c} M_f^{a'} q_c$. These loadings will provide profiles similar to the previous ones when one salience is really dominant. Case studies show that this is often the case for components which are not linked to noise.

Jouan-Rimbaud Bouveresse *et al.* (2011) also suggest an F-test to test the significance of the main effects and interactions in the ANOVA model based on the fact that saliences can be considered proportions of variance and that the first common dimension is linked to the residuals. A ratio of the saliences of an effect matrix M_f^a and the residual matrix on the first common dimension can therefore be computed to assess how much noise the effect matrix contains:

$$\frac{\lambda_1^{residuals}}{\lambda_1^f} = \frac{q_1' E E' q_1}{q_1' \Psi_f q_1} \quad (4.11)$$

This ratio is by definition always greater than or equal to one and, according to Jouan-Rimbaud Bouveresse *et al.* (2011), a F distribution with $((n-1), (n-1))$ degrees of freedom can be used to provide a p-value to test the significance of the corresponding model effect. The null hypothesis of the test is that the model term has no effect on the response and the p-value can be calculated as $P(F_{n-1; n-1} < \lambda_1^{residuals} / \lambda_1^f)$. If the p-value is small, the term is considered significant, which indicates that some of its related parameters are not null.

It is however important to highlight that no mathematical justification is given in Jouan-Rimbaud Bouveresse *et al.* (2011) to justify the fact that the test statistics follows a $F_{n-1; n-1}$ distribution under H_0 . In Section 4.4 of this article, a permutation test will be suggested to compute an alternative p-value.

4.2.7 AMOPLS

AMOPLS is described in Boccard & Rudaz (2016).

Approach

AMOPLS consists in applying a kernel-based multi-block OPLS approach on the pure and residual-augmented matrices issued from the ANOVA decomposition of

the design response matrix and provides, as output, a global view of important effects in the model which allows their interpretation.

PLS is considered the “standard method in supervised linear modelling in the field of chemometrics” (Rantalainen *et al.*, 2007). It models the relationship between a predictor matrix $\mathbf{X}$ and a response matrix $\mathbf{R}$ and is suited to high dimensional data sets. OPLS is an extension of PLS which uses an orthogonal signal correction filter in order to distinguish Z -predictive (PrC) and R -orthogonal (OC) components in the analysis. The resulting predictive components therefore only include variation related to $\mathbf{R}$, which simplifies the interpretation of the results (Trygg & Wold, 2002a; Rantalainen *et al.*, 2007).

AMOPLS is a supervised method where, the response matrix $\mathbf{R}$ is defined in order to guide the choice of OPLS predictive components that are directly linked to significant model effects. Orthogonal components gather the data variability due to the residuals and non significant model effects.

AMOPLS relies on a kernel-based version of OPLS, KOPLS, suggested in Rantalainen *et al.* (2007). In KOPLS, the predictor matrix can be mapped (by a kernel function) to a higher dimensional space in order to model potential non-linear relationships. The resulting kernel matrix $\mathbf{K}$ is of size $(n \times n)$. In AMOPLS, however, the kernel transformation is only used to base the OPLS decomposition on sample variance-covariance matrices, such as the ones used in AComDim. The multi-block characteristic of AMOPLS comes from its ability to analyse several matrices simultaneously (Boccard & Rudaz, 2016).

It will be shown below that AMOPLS presents several similitudes with AComDim even if the general approach seems very different. Finally, it is important to note that the AMOPLS algorithm is applicable to balanced or unbalanced designs, but the structure of its response matrix implies that it is recommended to use it only for balanced or almost balanced designs to produce valuable results.

Model

Figure 4.6 presents the AMOPLS approach. Following the response matrix decomposition by ANOVA/GLM modelling, the effect matrices are used to compute a predictor matrix and a response matrix subsequently used in a KOPLS regression.

The detailed steps of the technique described in Boccard & Rudaz (2016) and Boccard & Rutledge (2013) are as follows:

1. Compute the matrix decomposition by ANOVA/GLM modelling: $\mathbf{Y} = \hat{\mathbf{M}}_0 + \sum_{f=1}^F \mathbf{M}_f + \mathbf{E}$
2. Compute the $(n \times p)$ KOPLS response matrix $\mathbf{R}$ as a concatenation of non-null score matrices extracted from each pure effect matrix.

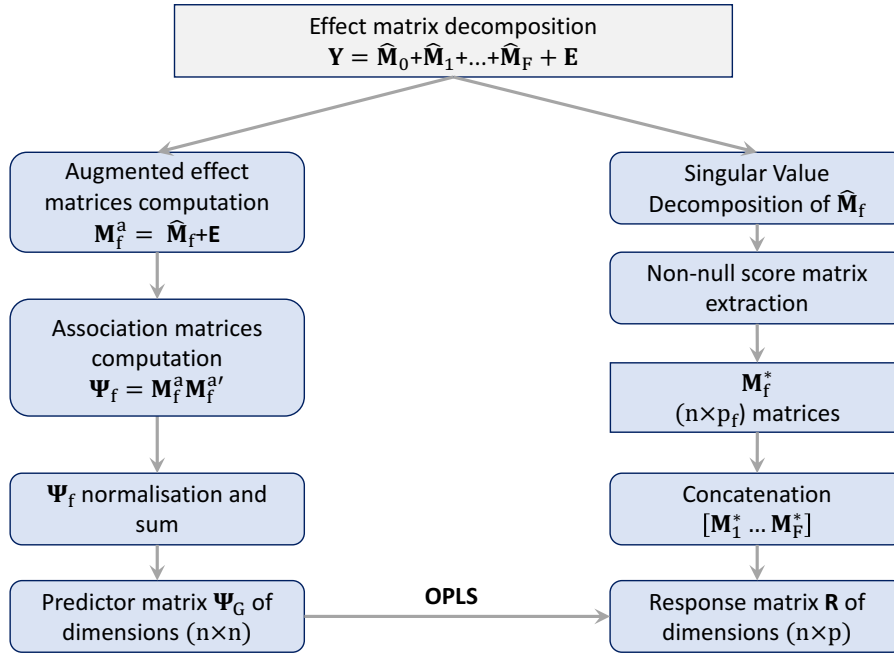


Figure 4.6 – Overview of the AMOPLS approach

- Apply a SVD to each pure effect matrix $\hat{M}_f$ for $f=1$ to F : $\hat{M}_f = Q_f \Gamma_f V_f'$.
- Compute the score matrix for each effect $Q_f \Gamma_f$ and gather the first non null columns in M_f^* . M_f^* is of size $(n \times p_f)$ where p_f is the number of model parameters linked to effect f in the GLM model.
- Concatenate the resulting matrices to create the KOPLS response matrix R :

$$R = (M_1^* \quad M_2^* \quad \dots \quad M_F^*)$$

The final number of variables of the response matrix R is equal to p , the number of model parameters minus 1. R is an orthogonal matrix when the design is balanced. When the design is unbalanced, some or all blocs M_i^* of R are not orthogonal to each other anymore which may decrease the interpretability of the final results.

3. Compute the $(n \times n)$ KOPLS predictor matrix Ψ_f on the basis of the augmented effect matrices aggregated in a similar way to AComDim:
 - Compute the residual-augmented effect matrices $M_f^a = \hat{M}_f + E$ for $f=1$ to F .

- Transform the F residual-augmented effect matrices in sample variance covariance matrices or “association” matrices $\Psi_f = \mathbf{M}_f^a \mathbf{M}_f^{a'}$ of size $(n \times n)$. This transformation is equivalent to a polynomial kernel transformation of order 1 in KOPLS.
- Divide each Ψ_f matrix by its Frobenius norm.
- Define the KOPLS predictor matrix Ψ_G as the sum $\Psi_G = \sum_{f=1}^{(F+1)} \Psi_f$ of these normalised matrices.

4. Fit the KOPLS model:

- Identify the optimal number of orthogonal components for the OPLS of $\mathbf{R}$ on Ψ_G in order to remove most of the non-significant information from the predictor matrix. The selection of the number of orthogonal components is discussed in Subsection 4.4.2.
- Set the number of predictive components as the number of columns of the response matrix $\mathbf{R}$, i.e. p .
- Run a KOPLS regression explaining the response matrix $\mathbf{R}$ by the predictor matrix Ψ_G . The KOPLS algorithm in (Boccard & Rudaz, 2016) is simply an OPLS where the predictive matrix is the kernel transformation Ψ_f of the initial matrix.

Output

For a given number of components C , the percentage of explained variance can be computed in AMOPLS via the classic OPLS goodness-of-fit indicators R^2Y and R^2X . R^2X gives a measure of the percentage of variance of Ψ_G explained by C OPLS components and R^2Y of the percentage of variance of $\mathbf{R}$ explained by Ψ_G through these components.

A scree plot presenting the total percentage of variance explained for models with different numbers of predictive components can be used to select the number of components of interest in the final model. This approach will be applied in Section 4.4.

The contribution of each model effect f to component c can be computed as $\delta_f^c = \frac{\mathbf{t}_c' \Psi_f \mathbf{t}_c}{\sum_{f=1}^F \mathbf{t}_c' \Psi_f \mathbf{t}_c}$ where $\mathbf{t}_c$ is the orthogonal or predictive score vector and $\sum_f \delta_f^c = 1$.

These contributions can also be used to build a measure of model term importance, the RSR (for *Residual Structure Ratio* (Boccard & Rudaz, 2016)). For effect matrix f the RSR is defined as: $RSR_f = \delta_E^{o1} / \delta_f^{o1}$, the ratio between the contribution of the error to the first orthogonal component and the contribution of the effect f to this component. This ratio follows the same spirit as the F

ratio defined in AComDim: since the first orthogonal component in AMOPLS represents the noise in the data which is orthogonal to the model effects, it is expected to be very similar to the first common component of AComDim. The contribution of an effect f to t_{o1} , δ_f^{o1} will then be close to δ_E^{o1} if f has no effect in the model and smaller than δ_E^{o1} when f has an effect. As in AComDim, a permutation test based on the RSR can be set up to assess the effect significances.

Scores and loadings can also be computed for each OPLS component and visualised and interpreted as for the other techniques. In (Boccard & Rudaz, 2016), loadings are computed using the same approach as in the original AComDim article (Jouan-Rimbaud Bouveresse *et al.*, 2011), but this chapter uses the general form suggested in Section 4.2.6: $p_c = \sum_{f=1}^{(F+1)} \sqrt{\delta_f^c} \mathbf{M}_f^{a'} t_c$.

4.3 Data

The data set used to illustrate the methods in the next section is a subset of the Urine-Citrate-Hippurate data set described in Appendix C. It has specifically been chosen in order to highlight and compare the properties of the different approaches.

The data contain metabolomic profiles obtained by ^1H NMR spectroscopy of a pool of rat urine samples. These samples were prepared by spiking the pool with known concentrations of chemicals (Hippurate and Citrate) in order to obtain data based on an experimental design. Classic NMR data pretreatments steps were applied to the original FID's prior to statistical analysis (see details in Appendix C). The spectra have been reduced to 600 descriptors and normalised with a constant sum normalisation. This traditional approach in metabolomics is suitable here since an appropriate quantity of buffer has been added to the samples in order to avoid a decrease of urine signal in the presence of the spiked molecules. No variable scaling was applied as, in such NMR data, scaling can artificially inflate noisy spectral areas. Please note that scaling in the ASCA context must be applied with caution, taking into account the considerations raised in Timmerman *et al.* (2015).

In the subset of this data set used here, and further referred to as the UCH data set, each of these factors has three levels: no added chemical (a concentration of 0), a medium concentration (2 mM for the Citrate factor (Qc/4) and 1 mM for the Hippurate factor (Qh/4)) and a high concentration (4 mM for the Citrate factor (Qc/2) and 2 mM for the Hippurate factor (Qh/2)). This leads to an experimental design with 9 possible combinations of concentrations illustrated in Figure 4.7(a). It is important to highlight than a linear increase in concentration

usually leads to a linear increase in spectral intensity when using NMR. This means that the Hippurate and Citrate effects can be expected to be almost one-dimensional in the ANOVA model in spite of having three levels each.

Four samples of each of the nine combinations were prepared, then frozen and analysed (by sets of 18) over two days. Each day, the analysis of the two samples with the same Hippurate and Citrate proportions was done under two conditions: just after defreezing and one hour later in order to evaluate stability of samples over time. The UCH data selected for the chapter originally contained 36 spectra: 9 combinations of concentrations $\times$ 2 time points $\times$ 2 replicates. However, two outliers were removed, resulting in 34 observations and a slightly unbalanced experimental design (see Figure 4.7(b)).

The resulting (34×600) spectral matrix is defined as the response matrix $\mathbf{Y}$. In these spectra, 4 Hippurate peaks are located around 3.96, 7.54, 7.62 and 7.82 ppm and the aggregated Citrate peak is located around 2.6ppm (Wishart *et al.*, 2009) (see Figure 4.7(c)). Figure 4.7(d) presents the scores plot of the first two PCs for the response matrix. It shows that the information on the Citrate and Hippurate factor levels from the experimental design can be recovered, but only to some extent.

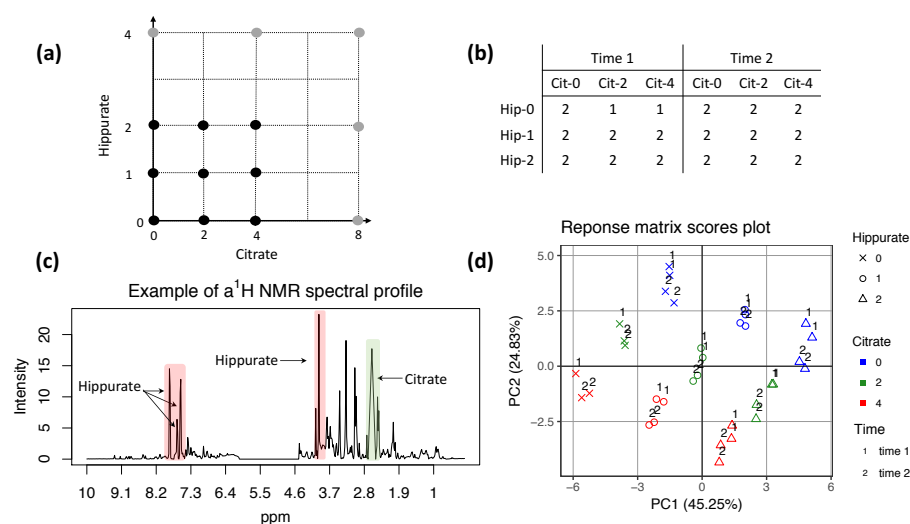


Figure 4.7 – Presentation of the UCH data set (a) Experimental design (b) Sample sizes for each of the 9 combinations of Citrate, Hippurate and Time levels (c) Example of a ^1H NMR spectral profile showing the Hippurate and Citrate peaks (d) Response matrix scores plots for the first two PCs of the UCH data set

4.4 Results

ASCA(-E), PARAFASCA, AComDim and AMOPLS were applied to the UCH data set. The results for each method are presented in parallel below with a particular effort to facilitate their comparison.

4.4.1 GLM decomposition

For all approaches, the following 3-way crossed ANOVA model was first fitted to the UCH data set. It includes three main factors, the concentrations of Hippurate (α) and Citrate (β) and the Time (γ), and four interactions, Hippurate-Citrate ($H \times C$), Hippurate-Time ($H \times T$), Citrate-Time ($C \times T$), and Hippurate-Citrate-Time ($H \times C \times T$).

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} + \epsilon_{ijkl}$$

The model parameters and effect matrices were estimated by the GLM approach in order to take into account the unbalanced property of the design. For each response l ($l = 1, \dots, m$) this model contains 17 independent parameters (constant term excluded). The first useful result is displayed in Table 4.1: based on the type III Sum of Squares, it represents the percentage of variation (squared Frobenius norm) of the response matrix $\mathbf{Y}$ explained by each model effect in the GLM decomposition. It shows that the three main effects correspond to a meaningful percentage of variation in the data, the $H \times T$ interaction to a small but notable percentage and the three other interactions to a very small percentage of variation. Please note that, because the design is unbalanced, the effect matrices are not orthogonal to each other and the total variation of $\mathbf{Y}$ cannot be exactly decomposed into percentages of variation summing to 100% (Thiel *et al.*, 2017).

Table 4.1 – Percentage of type III Sum of Squares for each effect matrix from the GLM decomposition

Hippurate	Citrate	Time	H:C	H:T	C:T	H:C:T	Error	Total
39.31	29.91	16.24	1.54	6.23	0.54	1.68	4.30	99.74

4.4.2 Components data decomposition and measure of component importance

All four approaches include factorial decompositions of combinations of the GLM effect matrices. This section presents, for each approach, the percentage of variation of the data explained by the calculated components and show that scree

plots are a good complement to model term significance tests to decide how many components are informative in each model. It must be stressed that these percentages of variance cannot directly be compared across approaches because the initial decomposed matrices and decomposition algorithms are not identical. For ASCA, a PCA decomposition was applied separately to each pure effect matrix $\hat{M}_f$ and to the residual matrix E . Table 4.2(a) shows that, for the three main effects, one component explains most or all of the variation. This was expected for the Time because it only has two levels (i.e. one degree of freedom). For the three-level Hippurate and Citrate factors, this is an indication that the main effect is probably linear and also has only one degree of freedom. The PCA decomposition of the $H \times T$ interaction is also almost one-dimensional and since Table 4.1 showed that other interaction terms are not very important in the model, the PCA decompositions of the corresponding interaction matrices are therefore not expected to contain substantial information. Keeping one component for each model effect can thus be considered enough in further steps. Finally, the residuals' decomposition shows that at least two PCs are necessary to study their behaviour. In conclusion, for ASCA, keeping 8 components, each directly linked to one model effect or to the two first residuals PCs, should allow understanding most of the information available in the data. Please note that these 8 components are not all orthogonal to each other because of the unbalanced structure of the design.

PARAFASCA was applied in the spirit of the methodology described in Section 4.2.5. The centred model predictions for all factors levels $\hat{y}_{ijkl} - \hat{\mu}$ were gathered in a $(3 \times 3 \times 2 \times m)$ Y_G matrix and a 4-mode PARAFAC decomposition was applied to Y_G with different numbers of components R .

Figure 4.8(a) shows, on a scree plot, the percentage of variance of Y_G explained by models with 1 to 17 components (17 being the number of parameters in the model). It shows that a model with four components is best suited to these data. Table 4.2(b) shows the percentage of variation explained by each component. Please note that, as suggested in Section 4.2.5, an orthogonality constraint has been imposed on Mode 4 in the decomposition.

For AComDim, a screeplot was also used to visualise the information provided by each additional dimension. It shows that 6 components are very informative and explain together 91% of the information of the aggregated effect matrix Ψ_G . The choice of the number of components in AMOPLS was done in two steps. First, a number of orthogonal components needed to be selected. Boccard & Rudaz (2016) suggest applying a permutation test to decide how many orthogonal components to include but this is time-consuming for complex designs. It was therefore decided to reflect the dimensionality of the data while applying the principle of parsimony. The ASCA results show that the error is

Table 4.2 – Percentage of variance of the decomposed matrices explained by each dimension for ASCA, PARAFASCA, AComDim and AMOPLS. The total percentages of variance cannot directly be compared across approaches because the initial decomposed matrices are not identical.

	Hippurate	Citrate	Time	H:C	H:T	C:T	H:C:T	Error
PC 1	97.71	98.22	100	44.01	93.92	90.76	47.23	48.54
PC 2	2.29	1.78	0	38.51	6.08	9.24	27.49	16.90
Total of $\hat{M}_f$	100	100	100	82.52	100	100	74.72	65.44

(a) ASCA-E effect matrix variance for PC 1 and PC 2.

comp. 1	comp. 2	comp. 3	comp. 4	Total of Y_G
38.03	28.27	16.64	11.63	94.58

(b) PARAFASCA variance decomposition for components 1 to 4.

CC 1	CC 2	CC 3	CC 4	CC 5	CC 6	Total of Ψ_G
20.44	21.54	20.39	17.63	2.435	8.543	90.98

(c) AComDim variance decomposition for common components 1 to 6.

R2X	R2X	R2X	R2X	R2X	R2X	R2X	R2Y
OC 1	OC 2	PrC 1	PrC 2	PrC 3	PrC 4	Total of Ψ_G	Total of Z
26.84	12.31	8.13	7.79	7.85	6.62	69.54	94.09

(d) AMOPLS X and Y variance decomposition for orthogonal and predictive components.

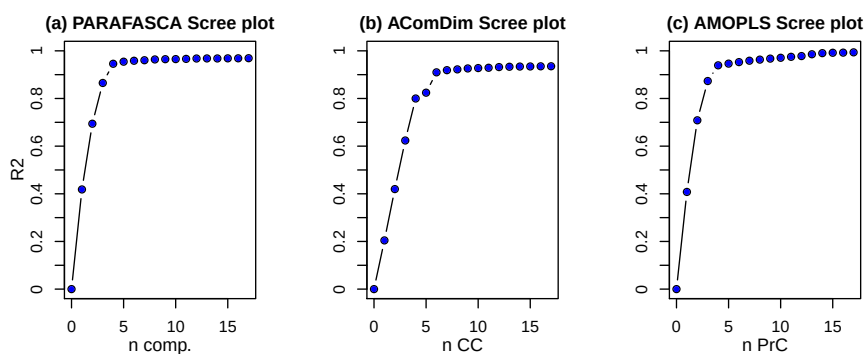


Figure 4.8 – Scree plots based on R2 to select the optimal number of components for PARAFASCA (4 components), AComDim (6 Common Components) and AMOPLS (4 Predictive Components)

bi-dimensional and the authors of Boccard & Rudaz (2016) advise selecting a small number of orthogonal components. A model with two orthogonal components was therefore chosen. Next, the maximal number of predictive components was set to 17, the number of parameters in the model, but their interpretation in the next sections will focus on the first four on the basis of the scree plot shown in Figure 4.8(c).

4.4.3 Components interpretation

In ASCA, the principal components calculated on each effect matrix are directly linked to the model effects and can then further be interpreted via their scores and loadings (see Sections 4.4.5 and 4.4.6). This section shows how this link can be visualised and quantified for the three approaches of interest.

Figure 4.9 presents the PARAFASCA (mean) scores for the three components and three modes A, B and C, corresponding respectively by construction to Hippurate, Citrate and Time factors. These results clearly show the link between scores for component 1 and the Hippurate effect, scores for component 2 and the Citrate effect and scores for component 3 and the Time effect. The PARAFAC decomposition recovers then perfectly the three main effects in the model and the response loadings for the first three components will therefore be easily interpretable. These scores also allow highlighting the linearity of the Hippurate and Citrate effects. The results for component 4 are less straightforward to interpret but point towards an $H \times T$ interaction effect. This is expected because this interaction is the 4th more important term in the GLM model (see Table 4.1) and because, in Figure 4.9, the Hippurate and Time effects are steep and corresponding scores change sign, while the Citrate effect is smaller and the corresponding scores remain positive. Its interpretation will be complemented by the scores and loadings plots presented in subsections 4.4.5 and 4.4.6.

Both AComDim and AMOPLS provide an index to quantify the contribution of the model effects to each dimension of the data space decomposition. These indices allow interpreting directly the loadings and scores in terms of their link with the model effects. They are represented in Figure 4.10 for the two methods.

In AComDim, the contributions are based on the saliences and in AMOPLS the contributions are obtained with a formula similar to the AComDim saliences (see Section 4.2.7). For these two methods, a careful interpretation of the contributions in terms of proportions is possible: in AComDim, saliences sum to 100% for each model effect (over all dimensions) whereas the contributions are normed to one within a component in AMOPLS.

In AComDim, the first component is related to the residuals. This result could be expected because, in this method, the error matrix E appears eight times in the decomposed data matrix - one time in each residual-augmented effect matrix

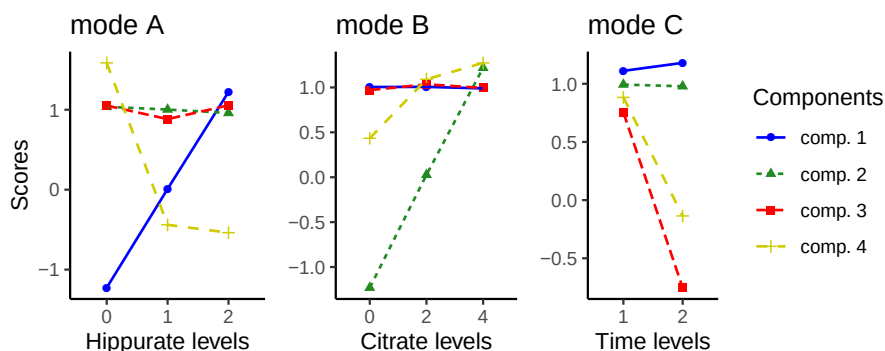


Figure 4.9 – PARAFASCA scores for modes A, B and C.

and one time alone - and therefore has a high weight. E thus represents “common information” in each data block used in the analysis.

The next three components clearly correspond to the Hippurate, Citrate and Time main effects. This ordering is consistent with the percentage of variance explained by these terms in Table 4.1. Component 5 seems to contain more information concerning model residuals, while component 6 clearly corresponds to the $H \times T$ interaction effect.

In AMOPLS, the two orthogonal components are linked to the residuals and their effect contribution profiles are very similar to what is observed in AComDim. This is expected since the residual matrix E is orthogonal to the KOPLS response matrix based on pure effect matrices only. The effect contributions for the three predictive components and the $H \times T$ interaction are similar to the ones obtained in AComDim but, in AMOPLS, small contributions of other model effects appear and are linked to the fact that AMOPLS is very sensitive to unbalanced designs.

These contribution plots can finally be linked to Table 4.2(c) and (d) to show that variance explained by components linked to the main model effects are, in AComDim and AMOPLS, generally smaller than the real proportion of the variance explained by the main effects in the data (see Table 4.1). This is particularly true for AMOPLS. This decrease is clearly due to the overweighting of the error term in the decomposed data matrices and can be deduced mathematically.

4.4.4 Significance tests of model effects

Significance tests of the model effects (Table 4.3) were performed for ASCA, AComDim and AMOPLS. The test statistic is based on the type III Sum of Squares of the effects for ASCA, on a salience ratio for AComDim and on a

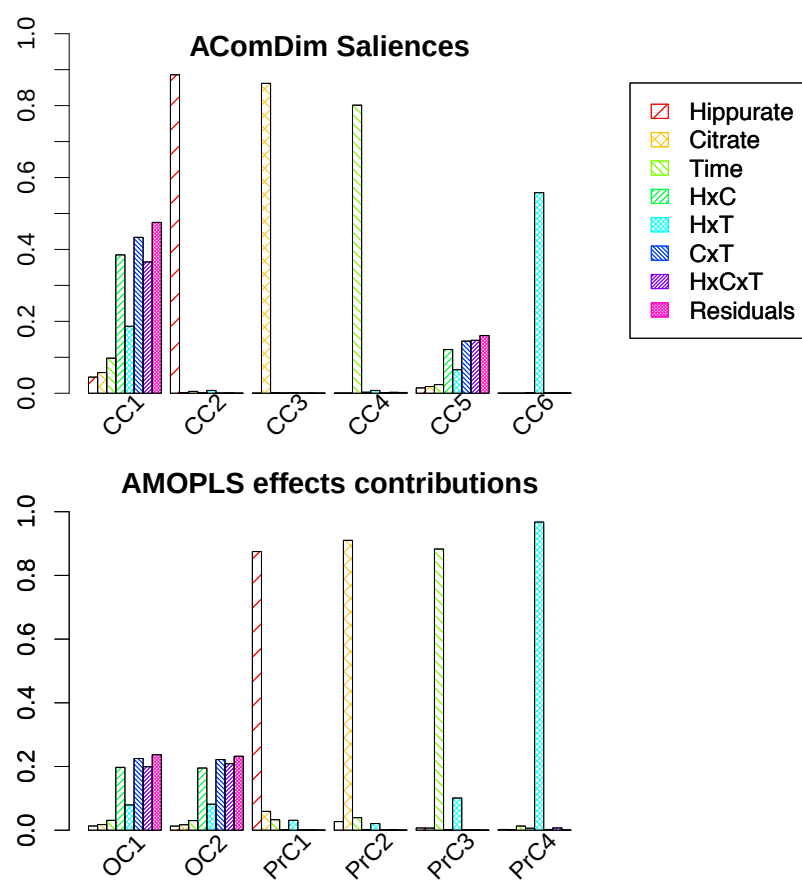


Figure 4.10 – Model effects and residuals contributions for each AComDim and AMOPLS predictive components.

contribution ratio (RSR) for AMOPLS (see Section 3.3). Permutations ($n=1000$) were implemented to assess the statistical significance of the test statistics according to the approach suggested in Anderson & ter Braak (2003) : for main effects , responses were permuted between factor levels keeping other factor levels constant and, for interaction effects, Manly's approach was used. The p-values were estimated as the proportion of test statistics obtained from the random permutation response matrices which are larger than the test statistic obtained for the original response matrix. For AComDim, the p-values based on the $((n - 1), (n - 1))$ F-distribution recommended in Jouan-Rimbaud Bouveresse *et al.* (2011) were also computed.

Table 4.3 shows that results from all approaches seem consistent: the Hippurate, Citrate, Time main effects and the H×T interaction are significant at the $\alpha = 0.05$ level. The other interactions are not significant across all methods. The interpretation of the significant effect will be done on the basis of scores and loadings in the next sections. Note also that for AComDim, the same test statistic is used for both the F and the permutation approaches but the p-values are calculated in different ways. In this example, the resulting p-values are consistent but one must note that, in spite of this, the densities of the test statistic under the null hypothesis in both cases (the $F((n - 1), (n - 1))$ density and the density obtained from permutations) are quite different. As an illustration, the density of the empirical distribution of the F -statistic under H_0 for the Hippurate effect is right-skewed, having minimum values of 1.01 and quantiles $q_{0.05} = 1.03$ and $q_{0.95} = 1.26$ whereas the F distribution with $(33, 33)$ df is more symmetric, centered around 1 with quantiles $q_{0.05} = 0.56$ and $q_{0.95} = 1.79$.

The results of these significance tests also confirm the adequacy of the number of components selected in Section 4.4.2 on the basis of the scree plots.

Table 4.3 – Model effects significance tests for ASCA, AComDim & AMOPLS

	ASCA		AComDim			AMOPLS	
	Test statistic	p-val. permut.	Test statistic	p-val. F-test	p-val. permut.	Test statistic	p-val. permut.
Hippurate	9.15	< 0.001	10.59	< 0.001	< 0.001	17.97	< 0.001
Citrate	6.96	< 0.001	8.26	< 0.001	< 0.001	13.39	< 0.001
Time	3.78	< 0.001	4.87	< 0.001	< 0.001	7.68	< 0.001
H:C	0.36	0.146	1.24	0.274	0.366	1.20	0.499
H:T	1.45	<0.001	2.55	0.004	< 0.001	2.99	< 0.001
C:T	0.13	0.448	1.10	0.397	0.499	1.05	0.680
H:C:T	0.39	0.104	1.30	0.227	0.179	1.19	0.548

4.4.5 Scores

In the context of this chapter, scores plots help visualise the amplitude of model effects directly on the data, identify outlying or specific observations, and detect potential patterns. The scores plots are presented as scatter plot matrices for ASCA-E, AComDim and AMOPLS in Figures 4.11, 4.14 and 4.15. For each component, the model effect with the highest contribution is indicated and the observations are identified with respect to different factors in the upper and lower part of the scatter plot matrix. In the upper part, colours and markers differentiate Hippurate et Citrate levels and in the lower part, the observations are identified with respect to the Time. For PARAFASCA, score plot are different and their construction is explained below.

For ASCA-E (Figure 4.11), 9 components have been kept to visualise the data and factor effects but the 3 components related to the non-significant interaction terms are less relevant. The PC1 scores for the three main effects Hippurate, Citrate and Time show a clear separation between the factors levels which is a confirmation of their significance and are consistent with the initial shape of the $3 \times 3 \times 2$ factorial design. On line 5 and column 1, the ($H \times T$, Hippurate) scores plot shows a distinctive pattern which helps to interpret the interaction between Hippurate and Time. The observations for the two Time levels cross each other when Hippurate increases, which shows that the effect of Time is different for level 0 of Hippurate compared to levels 1 and 2. This will be much clear in the PARAFASCA score plot below. Finally, PC2 of residuals allows identifying two moderate outliers which will be confirmed by other methods' scores plots.

Score plots for the full PARAFASCA are built in this section in a different way than in Figure 4.9. For each component c , the tensor product $u_c \Delta v_c \Delta w_c$ between the scores of the three first modes have been calculated and all numbers in the resulting three-way array represented on one graphic with respect to the most important factor highlighted in Figure 4.9. These graphs (see Figure 4.12) show much more clearly that components 1, 2 and 3 are respectively linked to H, C and T alone and that component 4 is linked to the $H \times T$ interaction. Putting these graphics next to the loadings plots provided in the next section will allow understanding very clearly how the NMR spectra are affected by factor changes. To obtain a clearer view of the interaction effect, a three-way PARAFASCA was then applied to the $H \times T$ effect alone, which provides the interaction plot of Figure 4.13. This graph is very similar to the graph (5,1) obtained in the ASCA-E score plot matrix.

AComDim scores plots are very similar to those of ASCA-E but present a more parsimonious image of this information. In Figure 4.14, the main effects of Hippurate, Citrate and Time are recovered very clearly by the common components 2, 3 and 4. The ($H \times T$, Hippurate) scores plot shows the same

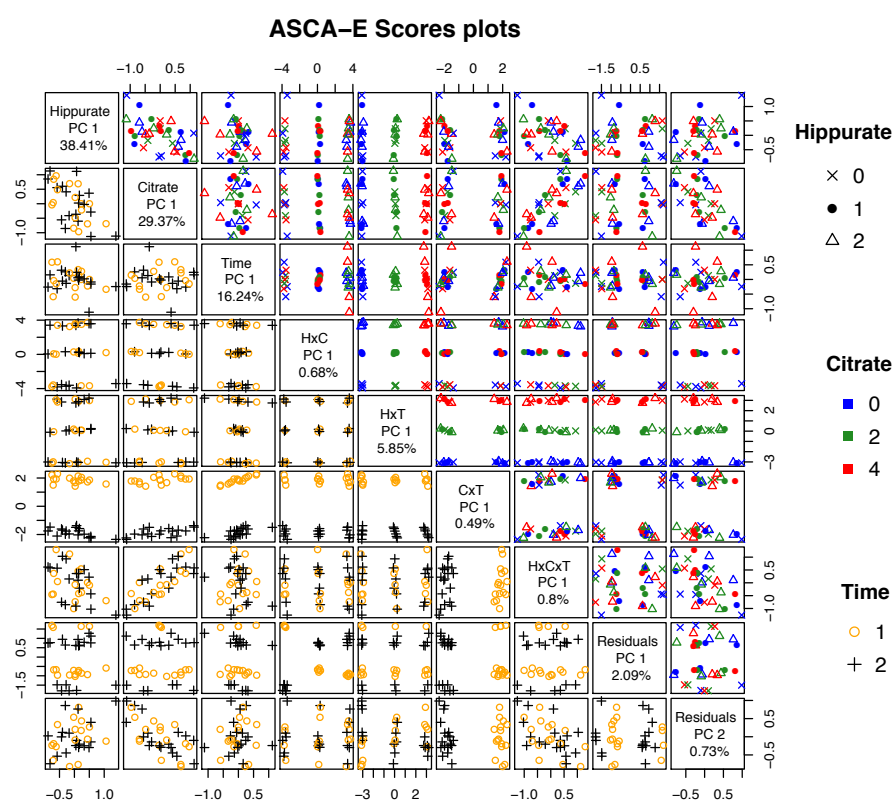


Figure 4.11 – ASCA-E Scores plots.

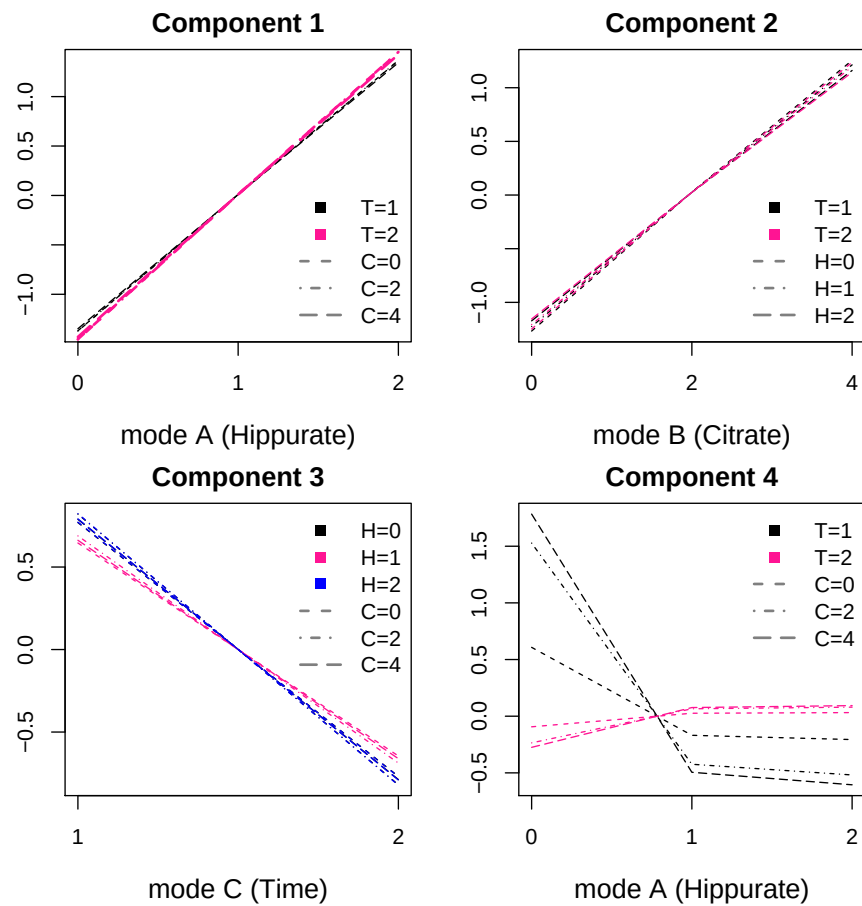


Figure 4.12 – (Mean effect) score plots by component for the full PARAFASCA model

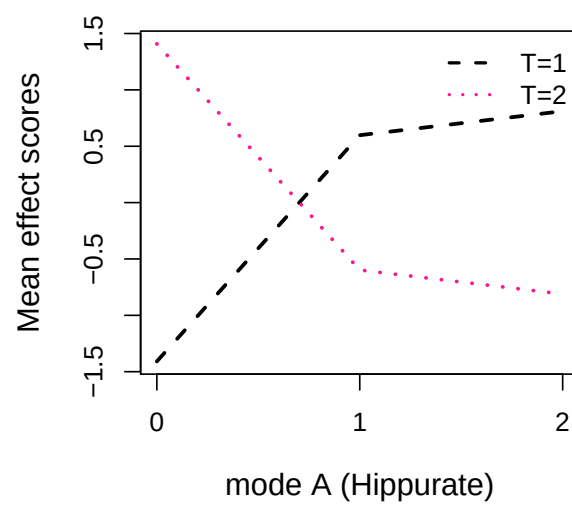


Figure 4.13 – (Mean effect) score plots by component for the partial PARAFASCA model on $H \times T$

pattern as in ASCA-E. Common Dimension 5, linked to residuals, allows visualising the same two outliers as in ASCA-E.

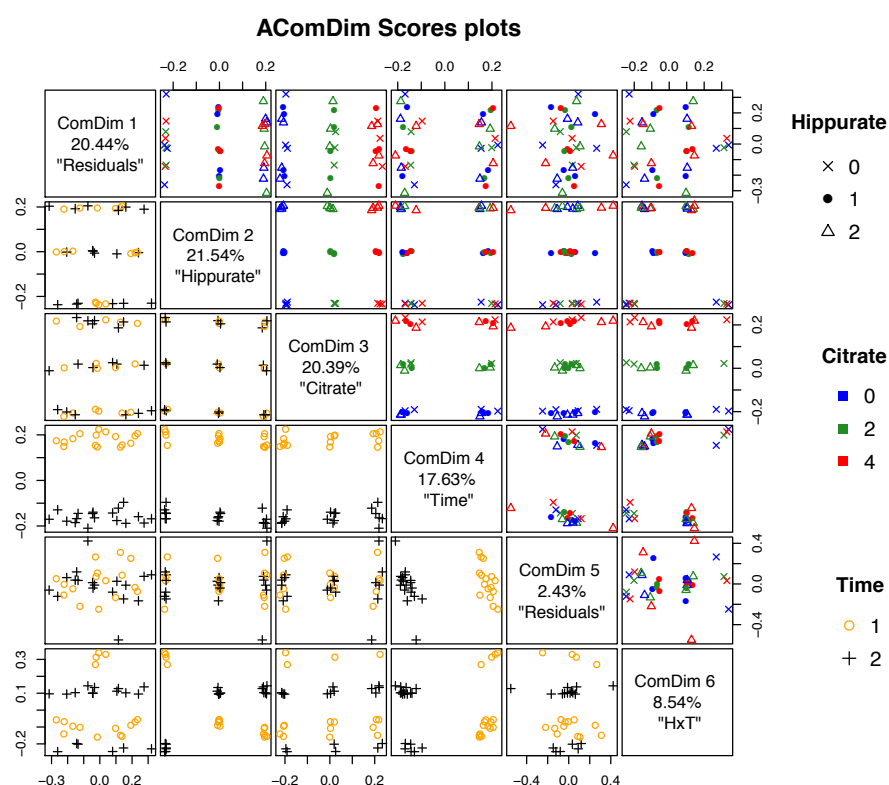


Figure 4.14 – ACoDim scores scatter plot matrix.

For AMOPLS, predictive Components 1 and 2 scores plots clearly correspond to the 9 combinations of concentrations available in the experimental design, but the results are slightly less clearly displayed than in ACoDim and ASCA-E. This is due to the fact that AMOPLS has been designed for balanced design and is thus quite sensitive to unbalanced designs. Predictive components 3 and 4 respectively correspond to the Time effect and the $H \times T$ interaction. The same interaction pattern appears on the ($H \times T$, Hippurate) scores plot. The two Orthogonal Components are mostly noise and OC2 lets appear the two outliers.

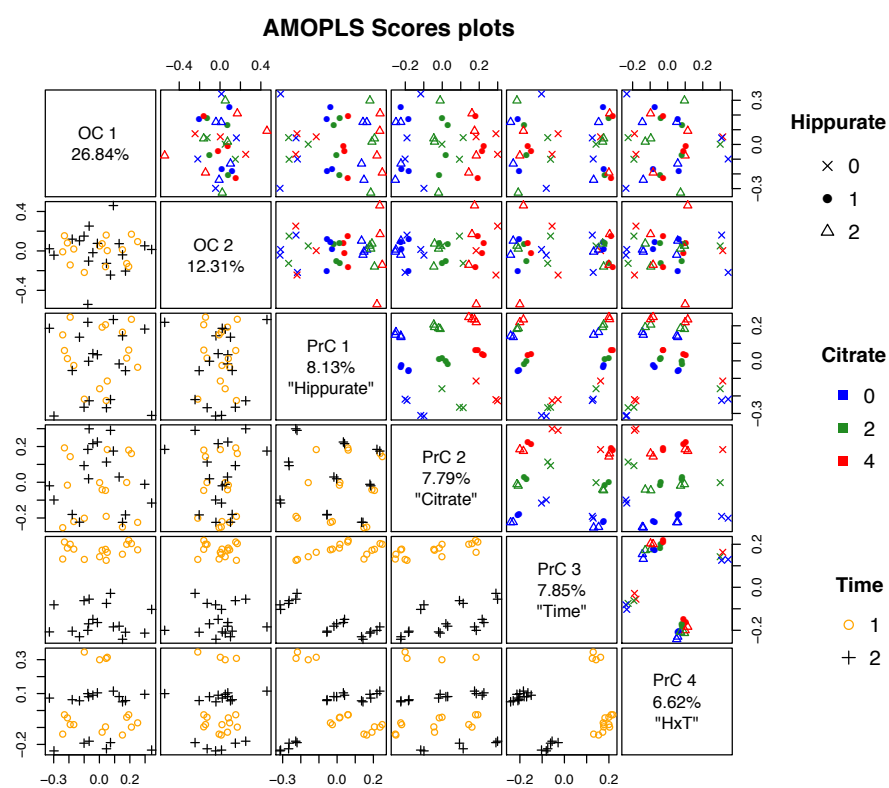


Figure 4.15 – AMOPLS scores scatter plot matrix.

4.4.6 Loadings

Loadings plots provide a way to visualise the link between the model effects and the m responses (the 600 descriptors from the spectral matrix). Figures 4.16 and 4.17 presents, for each method of interest, the loadings for the components corresponding to the model main effects Hippurate, Citrate, Time and H×T interaction and are very similar accros methods. The loadings for the Hippurate and Citrate main effects clearly show the expected molecule peaks: 4 peaks for Hippurate and one for Citrate (this peak has been aggregated in the preprocessing step). The Time effect is located at the border of the water zone and the interaction effect shows alternating peaks around 3 ppm typically due to peak shifting across spectra.

The Figure 4.12 score plot combined with these loading plots provide a very clear way to understand how factor changes affect the spectra in the three-factor design.

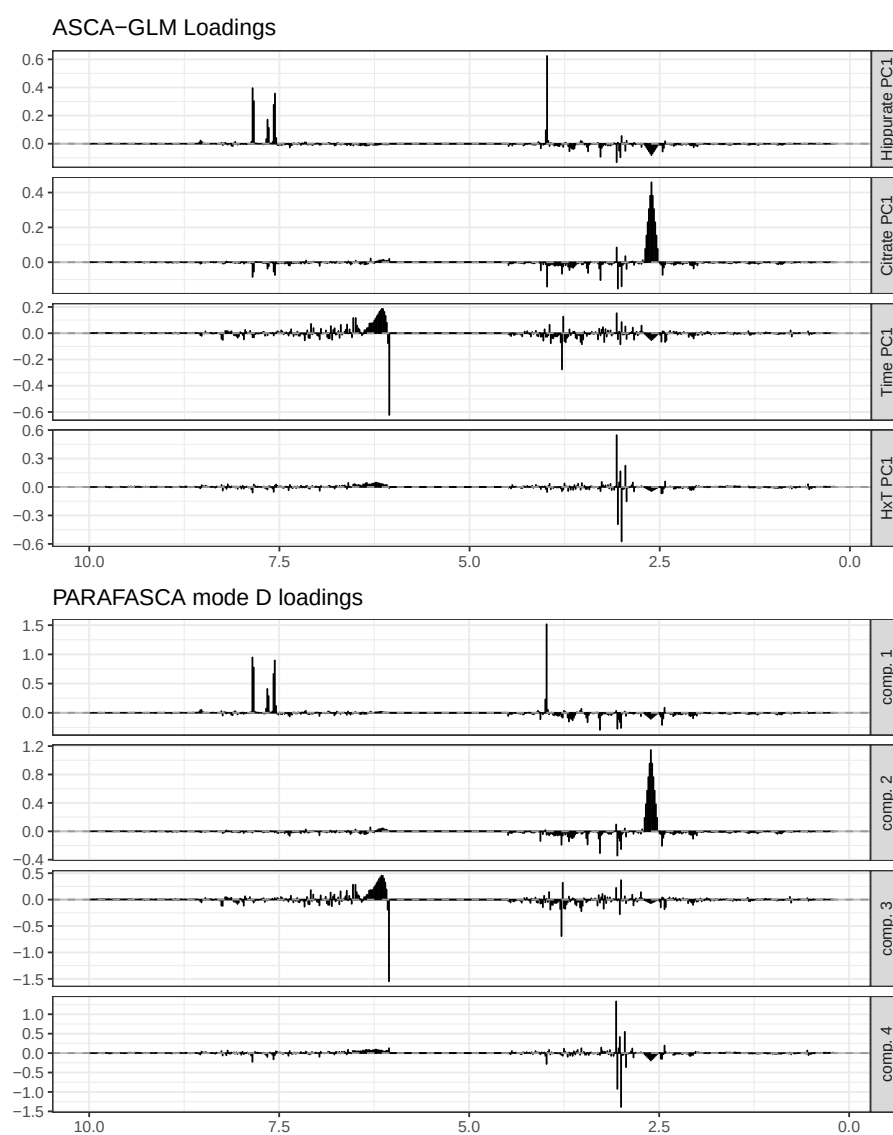


Figure 4.16 – Loadings plots for ASCA-E and PARAFASCA calculated for the components corresponding to the three main effects Hippurate, Dilution and Citrate.

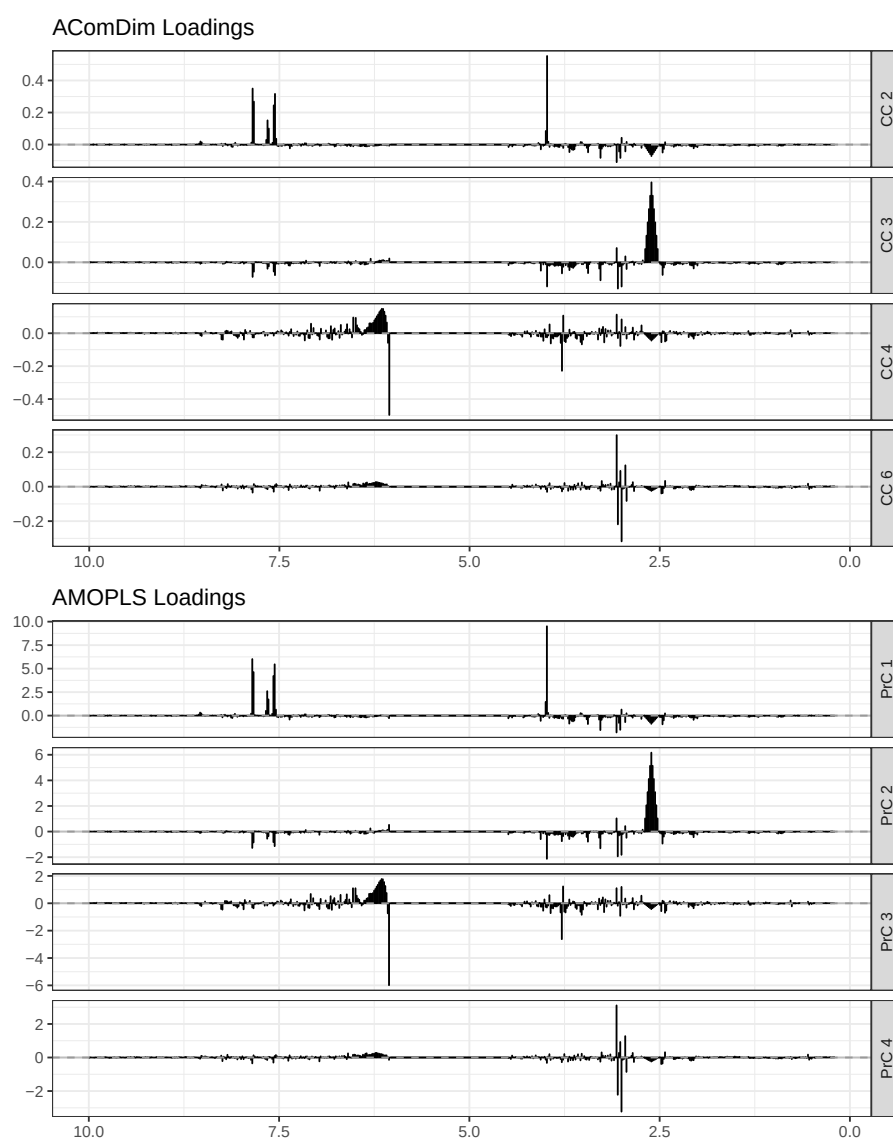


Figure 4.17 – Loadings plots for AComDim and AMOPLS calculated for the components corresponding to the three main effects Hippurate, Dilution and Citrate.

4.5 Conclusions

This chapter aimed at presenting together three “global” methods suitable for the analysis of wide multi-response experimental design data, compare them to the traditional ASCA/APCA approaches and extend them to unbalanced designs thanks to an effect matrices estimation by GLM instead of ANOVA modelling. These techniques first decompose the response matrix with respect to the ANOVA/GLM model before analysing the resulting effect matrices with one multivariate technique to interpret the model results. It has been shown that the main advantages of these methods are their ability to evaluate the effect of all model terms on the responses under a **common framework** with an **automatic ranking of their importance**, and that they provide several interesting tools to **ease the interpretation** of the components generated by the multivariate step.

The case study has illustrated that these methods provide interesting tools to **complement ASCA and APCA**, especially when the multifactorial model has a large number of terms. On the other hand, these global approaches can induce a **remixing** of the model effects because they are recombined in a common object, although the experience of the authors of this chapter and the related literature show that they are quite robust to this problem especially in the balanced case. This will be discussed further below.

The objective of this chapter was not to select one method over the others but to present under a common framework, analyse objectively, and illustrate on a case study similar approaches in parallel. All techniques lead to **useful and interpretable results** and provide a decomposition of the data into components linked to specific model effects, but they **differ in subtle ways**. Table 4.4 presents some of the main differences between techniques.

Table 4.4 – Comparison of the three global techniques

Technique	PARAFASCA		AComDim		AMOPLS		
Data type	D-dimensional array		Multi-block		Multi-block		
Matrices used	Effects estimation related to the question of interest		Residual-augmented effect matrices		Pure effect matrices (response) and residual-augmented effect matrices (predictors)		
Residual matrix	Not included		Included		Included for the predictor matrix, excluded for the response matrix		
Underlying methodology	Parallel Analysis	Factor	Common analysis	Dimension	Kernel Least Squares	Orthogonal	Partial
Unbalanced Design	Applicable		Applicable		Applicable for no or minimally unbalanced data		

PARAFASCA can be seen as an elegant and parsimonious way to complement the results provided by ASCA to visualise all or some factor effects, especially in the presence of **interactions**. PARAFASCA, being a multi-way method, provides results that are closer to the initial structure of the interaction terms in the model and works directly on the effect matrices estimated by the ANOVA/GLM model without including model residuals in the analysis. Thanks to these two properties, it can provide a cleaner and **more readable view of factor effects** than other methods. For complex factor behaviours or if too many model terms are included in the PARAFASCA array, the method could induce **effect remixing**, but the case study of this chapter illustrate at least that, even on a full three-factor cross design, PARAFASCA may provide very elegant results. Note also that the ALS algorithm can be **unstable** when the chosen number of components is too high.

In AComDim, the interpretation is a strength: it is facilitated by the use of **saliences**, which allow easily identifying the links between the model effects and common components. AComDim could in theory be affected by effect remixing but the case study of this chapter, other experiences by its authors and the results on balanced data found in Jouan-Rimbaud Bouveresse *et al.* (2011) always show very pure profiles for common components not interpreted as noise.

Some additional results provided in Appendix 4.C confirm this and also show that, even under high unbalance, the common components remain mainly related to only one model term. The use of residual-augmented effect matrices in AComDim also deserves some comments: residuals are added to each effect matrix and therefore have a big weight in the ComDim decomposition. The percentage of variance explained by the residuals is thus **overestimated** and the cleanliness of the scores and loadings may be slightly impacted but, on the other hand, this extra noise has the important advantage to hide non significant model terms into the noise and automatically extract the important ones in the analysis. Lastly, in AComDim, the F-test suggested in the original article is not suitable in this context, although the alternative permutation test suggested in this chapter can alleviate this problem.

AMOPLS is the only supervised method, using the pure effect matrices to guide the construction of components. In this respect, it is close to AoV-PLS and ANOVA-TP. This technique can therefore potentially lead to better components and avoid effect mixing in the balanced case due to the orthogonal structure of the O-PLS response matrix. On this last aspect, Appendix 4.C show that, unfortunately, in the **highly unbalanced case**, the AMOPLS components may be very affected, unlike what was observed in AComDim. Data imputation could then be considered if one want to apply AMOPLS on unbalanced designs. Note also that the AMOPLS algorithm may seem fairly complex for users who are not familiar with PLS/OPLS, but this chapter shows that some steps of the

procedure can be simplified. On the other aspects, AMOPLS tools, results and properties are very close to those provided by AComDim.

The **choice of the number of components** is also an important matter to consider in AComDim, PARAFASCA and AMOPLS. In AComDim, things are close to the spirit of PCA : they are extracted one by one and one can choose a sufficiently large number of components to be sure to extract all the information. Complementary tools like a scree plot may then be used to decide how many are of interest. In PARAFASCA the components profiles change for different number of components extracted by the ALS algorithm. It is then recommended to compare the results obtained with several model dimensions and chose an adequate number on the basis of a scree plot, the profiles of the obtained score and loading plots and by remaining parsimonious in order to avoid the instability of the results. In AMOPLS, the choice of the number of orthogonal components may be simply based on the dimensionality of the residuals identified via ASCA and empirical results also show that the choice of one or two orthogonal components in AMOPLS often does not lead to major changes in the results. Next, the choice of the number of (useful) predictive components can be guided, as in AComDim, by the use of a scree plot.

The three tested methods rely on the choice of joint components for all the model effect matrices, which is not the case in ASCA when separate PCA are applied to each effect matrix instead of an integrated SCA analysis. This last method was not the aim of this chapter and would probably not have led to different results in the case study, but its robustness to data unbalance and effect mixing would gain to also be tested in further works.

In conclusion, the three methods discussed have shown their usefulness to analyse wide and short data tables with an advanced design, while **each has its own advantages and disadvantages**. They allow for a fast and parsimonious analysis when the number of model terms is large, while remaining within the ASCA/APCA framework in terms of available information. Besides, the GLM approach helps analyse the data on the same basis as the ANOVA approach, but may be extended to unbalanced designs without biasing the estimation. Depending on its data and context, researchers may choose to apply an ASCA/APCA approach or one of the techniques described here in order to reach their objectives.

4.A PARAFASCA algorithm

One of the most common algorithms to compute a PARAFAC decomposition is the alternating least squares algorithm. Its objective is to derive iteratively each loading/score matrix by finding a least square solution to equations relating the matrix of interest and the unfolded array built from the loading/score matrices related to the other models. An important PARAFAC step is the unfolding of the array $\mathbf{Y}_G$ to obtain two-dimensional matrices. Figure 4.18 shows the possible unfolded arrays obtained from $\mathbf{Y}_G$ and used in the computation of the PARAFAC components when the number of modes is 3.

The full algorithm includes the following steps (Bro, 1997; Tomasi & Bro, 2006):

■ Initialisations

- Compute $\mathbf{Y}_G$, the three-dimensional array resulting from concatenating the residual-augmented effect matrices (Figure 4.4).
- Decide on a number of components C to compute.
- Initialise loading/scores matrices $\mathbf{V}_0$ and $\mathbf{W}_0$ of size $(b \times C)$ and $(m \times C)$ respectively. Various options can be used, among which random initialisation or an SVD-based initialisation.

■ START OF THE LOOP with k , the number of iterations until convergence and c , the component, with $c=1$ to C

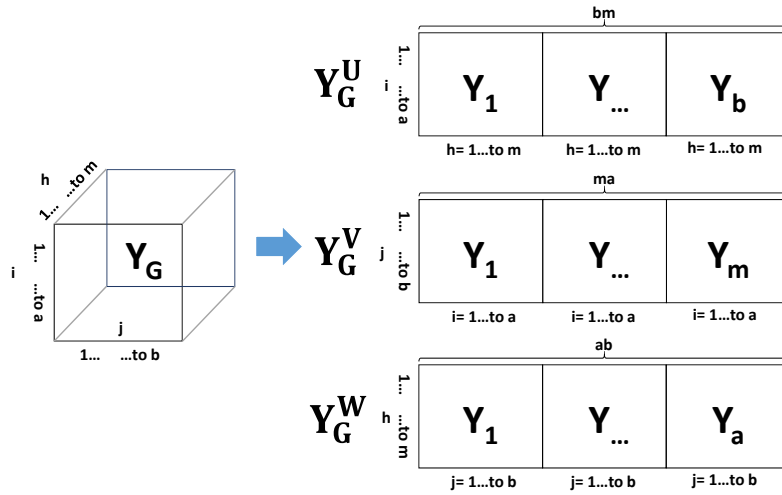


Figure 4.18 – Visualisation of the unfolded PARAFASCA arrays adapted from Bro (1997)

- Unfold $\mathbf{Y}_G$ to an $(a \times (m.b))$ matrix $\mathbf{Y}_G^U$ and define $\mathbf{Z}_{U_k}$ as the $(C \times (m.b))$ matrix with row vectors $\mathbf{v}'_c \otimes \mathbf{w}'_c$ where $\mathbf{v}_c$ and $\mathbf{w}_c$ are respectively the c^{th} columns of $\mathbf{V}_{k-1}$ and $\mathbf{W}_{k-1}$.
- Estimate matrix $\mathbf{U}_k$ (from $\mathbf{V}_{k-1}$ and $\mathbf{W}_{k-1}$) by least square regression: $\mathbf{U}_k = \mathbf{Y}_G^U \mathbf{Z}_{U_k}' (\mathbf{Z}_{U_k} \mathbf{Z}_{U_k}')^{-1}$.
- Unfold $\mathbf{Y}_G$ to an $(b \times (m.a))$ matrix $\mathbf{Y}_G^V$ and define $\mathbf{Z}_{V_k}$ as the $(C \times (m.a))$ matrix with row vectors $\mathbf{u}'_c \otimes \mathbf{w}'_c$ where $\mathbf{u}_c$ and $\mathbf{w}_c$ are respectively the c^{th} columns of $\mathbf{U}_k$ and $\mathbf{W}_{k-1}$.
- Estimate matrix $\mathbf{V}_k$ (from $\mathbf{U}_k$ and $\mathbf{W}_{k-1}$) by least square regression: $\mathbf{V}_k = \mathbf{Y}_G^V \mathbf{Z}_{V_k}' (\mathbf{Z}_{V_k} \mathbf{Z}_{V_k}')^{-1}$.
- Unfold $\mathbf{Y}_G$ to an $(m \times (b.a))$ matrix $\mathbf{Y}_G^W$ and define $\mathbf{Z}_{W_k}$ as the $(C \times (b.a))$ matrix with row vectors $\mathbf{u}'_c \otimes \mathbf{v}'_c$ where $\mathbf{u}_c$ and $\mathbf{v}_c$ are respectively the c^{th} columns of $\mathbf{U}_k$ and $\mathbf{V}_k$.
- Estimate matrix $\mathbf{W}_k$ (from $\mathbf{U}_k$ and $\mathbf{V}_k$) by least square regression based on $\mathbf{W}_k = \mathbf{Y}_G^W \mathbf{Z}_{W_k}' (\mathbf{Z}_{W_k} \mathbf{Z}_{W_k}')^{-1}$.
- Compute the loss function $\sum_{i=1}^a \sum_{j=1}^b \sum_{h=1}^m (y_{ijh} - \sum_{c=1}^C u_{ic} v_{jc} w_{hc})^2$.

- If the loss function does not decrease by more than a specified threshold value, the algorithm is considered as having converged and the loop ends. Otherwise, the algorithm goes back to the start of the loop ($k + 1$) using the new matrices U_k , V_k and W_k .

■ END OF THE LOOP when convergence has taken place

■ The loadings/scores matrices are the matrices U , V and W computed in the last iteration of the loop.

Note that this algorithm can be slow to converge but some enhancements are proposed in the related literature (see (Bro, 1997)).

4.B AComDim algorithm

The iterative algorithm used in ComDim can be detailed as follows (Qannari *et al.*, 2000; Jouan-Rimbaud Bouveresse *et al.*, 2011):

■ Initialisations

- Calculate the residual-augmented effect matrices $M_f^a = \hat{M}_f + E$ with $f = 1$ to F and define $M_{(F+1)}^a = E$
- Center each matrix by column and normalise it by its Frobenius norm.
- Compute initial sample variance-covariance matrices $\Psi_f^1 = M_f^a M_f^{a'}$ for $f=1$ to $(F + 1)$.

■ START OF LOOP 1 : repeat for each common dimension c to be extracted with $c=1$ to C

- Initialise the saliences λ_f^c to 1 for $f=1$ to $(F + 1)$.
- START OF LOOP 2: repeat until convergence
 - Compute the weighted sum of the Ψ_f^c matrices $\Psi_G^c = \sum_{f=1}^F \lambda_f^c \Psi_f^c$.
 - Decompose Ψ_G^c by singular value decomposition into $Q_c \Gamma_c Q_c'$.
 - Assign the value of the first eigenvector of Ψ_G^c (the first column vector of Q_c) to the vector q_c .
 - Update the saliences $\lambda_f^c = q_c' \Psi_f^c q_c$ with $f=1$ to $(F + 1)$.
 - Compute the loss function $\sum_{f=1}^F \|\Psi_f^c - \lambda_f^c q_c q_c'\|^2$.
 - Evaluate the convergence of the loss function. If the loss function does not decrease by more than a specified threshold value (for instance 10^{-20}), the algorithm is considered as having converged.

- If the algorithm has not converged, go back to the beginning of LOOP 2 using the new values of the saliences λ_f^c .
 - END OF LOOP 2: when convergence has taken place
 - The final values of $\mathbf{q}_c$ and λ_f^c are respectively defined as the c -th common dimension and the c -th salience of matrix f with $f=1$ to $(F+1)$.
 - Compute the deflated matrices: $\Psi_f^{r+1} = (\mathbf{I}_n - \mathbf{q}_c \mathbf{q}_c') \Psi_f^c (\mathbf{I}_n - \mathbf{q}_c \mathbf{q}_c')'$ for $f=1$ to $F+1$ and go back to the beginning of LOOP 1 to compute the next common component on the deflated matrices Ψ_f^{r+1} , $f = 1, \dots, (F+1)$.
- END OF LOOP 1 when $c = C$.

The process is iterated until a pre-selected number of common dimensions has been computed. The percentage of variance explained by each common component can guide the choice of the number of dimensions to analyse.

4.C Factor mixing in ACOMDIM and AMOPLS for different degrees of data balanceness

Table 4.5 – Observation frequencies in the balanced and highly unbalanced datasets.

	Balanced design				Highly unbalanced design			
	Hippurate	Citrate			Hippurate	Citrate		
Time = 1	0	2	2	2	0	1	1	1
	1	2	2	2	1	2	2	2
	2	2	2	2	2	1	1	2
Time = 2	0	2	2	2	0	2	2	1
	1	2	2	2	1	1	2	2
	2	2	2	2	2	2	2	2

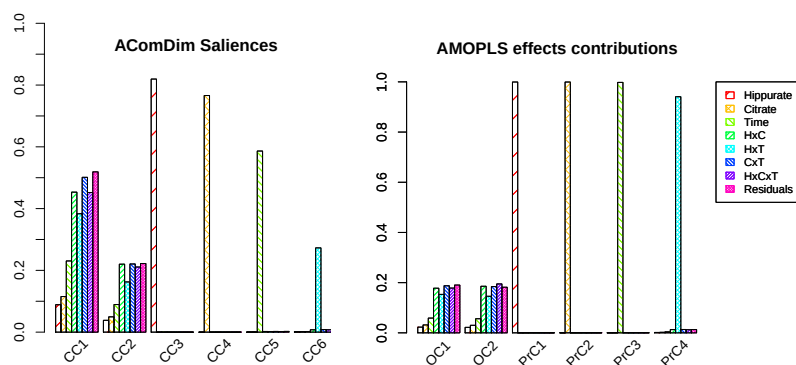


Figure 4.19 – Model effects and residuals contributions for each AComDim and AMOPLS predictive components for a balanced dataset.

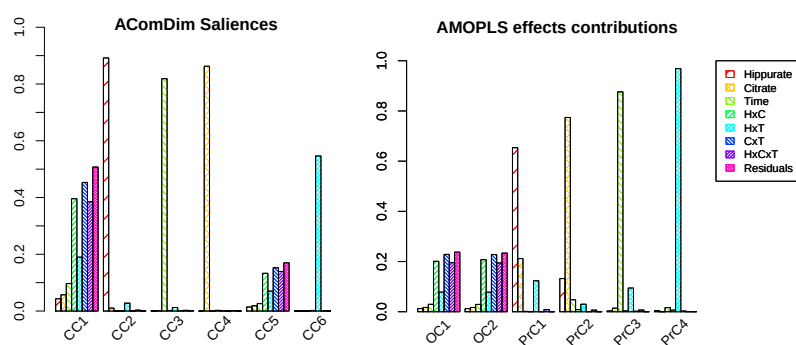


Figure 4.20 – Model effects and residuals contributions for each AComDim and AMOPLS predictive components for a highly unbalanced dataset.

CHAPTER 5

LiMM-PCA : combining ASCA⁺ and linear mixed models to analyse high dimensional designed data

5.1 Introduction

Most of the integrated approaches presented in Chapter 4 were developed to study classic fixed and crossed factor designs. But in the case of particular design structures (nested factors, multilevel, repeated measurements), the researchers should consider the need to rely on more advanced modelling possibilities. Exceptions in the literature include Marini *et al.* (2015) with ANOVA-Target Projection (ANOVA-TP) and in Luciano & Næs (2009) with ASCA and PC-ANOVA that both applied mixed ANOVA with multiple crossed or nested factors in the presence of interactions. ASCA variants for multilevel data has also been developed by Timmerman (2006) with the Multilevel SCA (MSCA) which was applied to a time-resolved metabolomics dataset (Jansen *et al.* , 2005b) and by de Noord & Theobald (2006) with the Multilevel PLS (MLPLS). However, inferential properties of such models (*e.g.* effect tests) and their potential generalisations to more advanced design structures are not dealt in depth within these papers.

As a complement to ANOVA and GLM, the Linear Mixed Model (LMM) (Snijders & Bosker, 1999; McCulloch & Searle, 2001; Brown & Prescott, 2006) has nowadays become the gold standard statistical tool to model advanced balanced or unbalanced designs that include *fixed* but also *random* effects. An

effect is usually considered as fixed when the differences between its levels are of direct interest (*e.g.* treatment effect) and we want to compare their mean effect. In contrast, the values of a random effect are considered as unreproducible random realisations of a larger population (*e.g.* the patient effect). They affect the error structure of the data and are able to model complex dependence structures present in them. This class of modelling is extensively used for designs where observations cannot be considered as independent between each other. In this context, the LMM is therefore an ideal candidate to generalise the modelling step of ASCA-related methodologies to more advanced designs.

5.1.1 Motivations and objectives

Accordingly, this chapter aims to suggest a generic framework to extend the ASCA⁺ methodology to LMM, called LiMM-PCA, taking into account the possible unbalance of the data as suggested in Thiel *et al.* (2017). The entire framework has to be adapted consequently and new methodological aspects developed to enable the comparison of both fixed and random sources of variation (*e.g.* test the significance of the effects, effect matrix augmentation, etc.).

In this chapter, we will focus on an extension of ASCA but the elaborated procedure could be applied to other multivariate methods, such as those mentioned in the background section here-above.

In the field of chemometrics, this methodology has a direct and broad range of potential applications. For example a repeatability/reproducibility study could decipher which of the measured biological or technical sources of variability are prominent and affect the measured signal; one could also be interested in longitudinal data where the dynamics of a system need to be integrated in the modelling process. Other potential typical applications would be cross-over studies *e.g.* where several treatments are administered consecutively to a same group of subjects.

5.1.2 Content

After the contextualisation and rationale of this research in Sections 5.1 and 5.2, Section 5.3 will describe the basics of LMM in this context. Section 5.4 will then describe our suggested LiMM-PCA methodology with emphasis on new developments and adjustments compared to ASCA, Section 5.5 will apply LiMM-PCA to two typical applications before closing by concluding remarks and perspectives in Section 5.6.

5.2 Limitations of ASCA⁺ when random factors & effects are included in the model

A huge amount of literature is available on ANOVA and GLM models that explains extensively the usefulness of random effects in statistical models and the associated methodological difficulties. We suggest to the more applied readers to refer to Neter *et al.* (1996) for a very didactic presentation of many of these models. These texts provide insights into the main reasons why ASCA is not well adapted to models including random factors.

Let's first start with balanced designs. In this case, the ANOVA theory is very well developed for models including only random effects (random models) or including both random and fixed effects (mixed models). The principle of total variance decomposition by an ANOVA equation remains correct regardless of the fixed or random nature of the factors. Consider, for example, the LMM with two factors, one fixed A and one random B . This model will be applied in Case Study 1 (Section 5.5.1) to analyse how a candies (factor A) are rated by b assessors (factor B) through several candies attributes in a sensory experiment, where all ratings are repeated c times. The model equation takes the following form:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \quad \text{with } i = 1, \dots, a; j = 1, \dots, b; k = 1, \dots, c; \quad (5.1)$$

where y_{ijk} is the observed response for the i th level of factor A , j th level of factor B and replicate k . $\mu_{..}$ is the overall mean, α_i is the effect of level i of factor A , β_j is the (random) effect of level j of factor B and $(\alpha\beta)_{ij}$ is the (random) interaction between i and j . In a mixed model, the β_j 's are supposed to be independent normal random variables $\sim iN(0, \sigma_B^2)$, the $(\alpha\beta)_{ij} \sim iN(0, \sigma_{AB}^2)$ and the errors $\epsilon_{ijk} \sim iN(0, \sigma_\epsilon^2)$. In this model, the main objectives are to understand how the mean response is affected when the level of A changes (e.g. the candy) and what amount of variability is produced by the levels of factor B (e.g. the assessors).

The related analysis of variance equation (which is the cornerstone of the ASCA decomposition) is identical to the one of the fixed model:

$$\begin{aligned} SS_T &= \sum_{ijk} (y_{ijk} - \bar{y}_{...})^2 \\ &= \sum_i (\bar{y}_{i..} - \bar{y}_{...})^2 + \sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2 + \sum_{ij} (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2 + \sum_{ijk} (y_{ijk} - \bar{y}_{ij.})^2 \end{aligned} \quad (5.2)$$

$$= SS_A + SS_B + SS_{AB} + SS_E \quad (5.3)$$

Until now, nothing has been altered by the randomness of factor B , but is what follows, adaptations are necessary. First, the sums of squares cannot be directly used to estimate the variances for random effects, they must be first corrected by other sum of squares. The consequences in ASCA are that the effect matrix $\hat{M}_r$ of a random effect r , derived from Equation (5.3) cannot be directly used to interpret this effect.

In the ANOVA framework, the correct estimation is obtained on the basis of the so called *Expected Mean Squares* ($E(MS)$) provided in Table 5.1 for this model. This table shows for example that $E(MS_B) = \sigma_\epsilon^2 + ca\sigma_\beta^2$. The variance linked to factor B should therefore be estimated as follows: $\hat{\sigma}_\beta^2 = (MS_B - MS_E)/ca$.

Table 5.1 – ANOVA table and $E(MS)$ for the crossed 2-Factor mixed ANOVA model. In this table, x represents the effect name (A , B or AB) or the error (E).

Effect	SS	df_x	MS_x	$E(MS_x)$	F-test
A	SS_A	$a - 1$	SS_A/df_A	$\sigma_\epsilon^2 + \frac{cb \sum \alpha_i^2}{(a-1)} + c \frac{a}{a-1} \sigma_{\alpha\beta}^2$	MS_A/MS_{AB}
B	SS_B	$b - 1$	SS_B/df_B	$\sigma_\epsilon^2 + ca\sigma_\beta^2$	MS_B/MS_E
AB	SS_{AB}	$(a - 1)(b - 1)$	SS_{AB}/df_{AB}	$\sigma_\epsilon^2 + c \frac{a}{a-1} \sigma_{\alpha\beta}^2$	MS_{AB}/MS_E
Error	SS_E	$ab(c - 1)$	SS_E/df_E	σ_ϵ^2	

In mixed models, the F -test hypotheses to test for the presence of a fixed factor, say A , are written as $H_0 : \alpha_1, \dots, \alpha_a = 0$ versus $H_1 : \text{not all } \alpha_i = 0$ and alternatively as $H_0 : \sigma_\beta^2 = 0$ versus $H_1 : \sigma_\beta^2 > 0$ when testing for the presence of a random factor, say B . In this setting, the classic F -tests for the significance test of an effect that were based on the ratio MS_f/MS_E cannot be used anymore. One should rather refer to the $E(MS)$ to decide which MS should be assigned to the denominator of each test statistic. For example, to test if fixed factor A is significant, the F -ratio MS_A/MS_{AB} is the one that should be used because the ratio of the corresponding expected mean squares is equal to 1 when A has no effect ($\alpha_i = 0$). In a multivariate setting, Marini *et al.* (2015) also uses this justification to decide which source of random variation to use in the case of mixed effects in order to obtain the reduced model residuals and perform permutation tests in ANOVA-TP. hence, this has an incidence on ASCA since the denominator of these tests should naturally be used to visualise the noise in the ASCA-E effect matrix augmentation, for the computation of confidence bands or in testing procedures.

Finally, this question is becoming more complex in the case of unbalanced designs since, in this case, the ANOVA matrix decomposition equation is no more valid. Behind this issue lies the real motivation of introducing the LMM as a GLM extension. The aim is indeed to derive a generalisation of ASCA for the analysis unbalanced experimental designs including random factors and a large number of responses. The main principles of LMM are presented in the next section before introducing LiMM-PCA, the suggested generalisation of ASCA to these models.

5.3 The Linear Mixed Model

5.3.1 Model terms definition and justification

This chapter gives a succinct description of LMM. LMM is a versatile regression toolkit in statistics when the observations cannot be assumed to be independent between each other (e.g. multiple measurements on the same patient, hierarchical data coming from different labs, repeated measures, etc.). The model is built on the differentiation between the *fixed* and the *random* effects (Searle *et al.*, 1992; Gelman, 2005). Typically, an effect is considered as fixed when the differences between its mean levels effects are of direct interest (e.g. treatment effect); whereas, the values of a random effect are considered as random realisations drawn from a larger population (e.g. the patient effect), the levels are a priori unknown and cannot be reproduced. Key aspects advocating the use of mixed models are that: it leads to more valid inferences by taking into account the complex error structure of the data, the inferential results are generalised at the population of the random effect(s) and not just limited to their sampled levels; the random sources of variation can be compared to each other. Compared to ANOVA, LMM based on Restricted Maximum Likelihood (REML) estimates offers more benefits, especially in the case on unbalanced data: in this context, consistency, asymptotic normality of the estimators, and the asymptotic sampling dispersion matrix of the estimators is known and enables to establish confidence intervals, or test hypothesis about derived parameters. Therefore, for advanced designs and models, LMM offers a flexible and generic framework for parameter estimation and inference compared to ANOVA.

5.3.2 General notations and assumptions

The matrix notation of linear mixed model can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}; \quad (5.4)$$

with $\mathbf{Z}$ the model matrix for random effects of size $(n \times q)$ and $\boldsymbol{\alpha}$ the random parameters vector of size $(q \times 1)$.

As in GLM, linear constraints apply between levels of a same effect for every fixed effect. Hence, the model matrix $\mathbf{X}$ is built with a *factor effects* coding as explained in Section 4.2.2. A dummy coding is used for the random effects design matrix $\mathbf{Z}$: the number of columns in $\mathbf{Z}$ is the sum of the levels' numbers for all the random effects.

The general assumptions for the random effects are $\alpha \sim N(0, \mathbf{G})$ and $\epsilon \sim N(0, \mathbf{\Sigma})$ with $\mathbf{G}$ of size $(q \times q)$ and $\mathbf{\Sigma}$ of size $(n \times n)$ the respective covariance matrices.

Here we will assume the traditional *variance components* model with only categorical fixed and random effects as predictors. Moreover, there is no covariance between each pair of random effects such that: $\text{var}(\alpha) = \mathbf{G} = \{\sigma_r^2 \mathbf{I}_{q_r}\}_{r=1}^R$ for a random effect r having q_r levels, with $\mathbf{I}_{q_r}$ the identity matrix and $\text{var}(\epsilon) = \mathbf{\Sigma} = \sigma_\epsilon^2 \mathbf{I}_n$. The variance components for the random effects are contained in $\gamma = (\sigma_1^2, \dots, \sigma_R^2)$.

The variance-covariance matrix of $\mathbf{y}$ is then: $V = \text{var}(\mathbf{y}) = \text{var}(\mathbf{X}\beta + \mathbf{Z}\alpha + \epsilon) = \mathbf{Z}\text{var}(\alpha)\mathbf{Z}' + \text{var}(\epsilon) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{\Sigma} = \sum_{r=1}^R \mathbf{Z}_r \mathbf{Z}_r' \sigma_r^2 + \sigma_\epsilon^2 \mathbf{I}_n$.

Thus, since $\mathbb{E}[\mathbf{y}] = \mathbf{X}\beta$, we have $\mathbf{y} \sim N(\mathbf{X}\beta, \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{\Sigma})$.

5.3.3 Estimation

The model fitting process is separated into 3 distinct parts: estimate the fixed effects β by Maximum Likelihood (ML), the variance parameters by REML, and predict the random effects by ML. More information about the estimation procedure is provided in Appendix 5.A. Final estimators for the fixed effects are: $\hat{\beta} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$, where $\hat{\mathbf{V}}$ is the REML estimator of $\mathbf{V}$. They are both obtained by an iterative numerical optimisation procedure. The ML predictors of the random effects are: $\hat{\alpha} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta)$, where $\hat{\beta}$ is the ML estimator of β and $\hat{\mathbf{V}}$ is the REML estimator of $\mathbf{V}$. Finally, the estimators of conditional errors, called the conditional residuals are simply (Santos Nobre & da Motta Singer, 2007): $\hat{\epsilon} = \mathbf{y} - \mathbf{X}\hat{\beta} - \mathbf{Z}\hat{\alpha}$.

5.3.4 Test of effects significance

In order to test the significance of one or several fixed effects β_f for $f = 1, \dots, F$ or one or several variance parameters σ_r^2 for $r = 1, \dots, R$, the Wald F-statistic and the Log Likelihood Ratio (LLR) are the two most well-known test procedures (Gumedze & Dunne, 2011). The latter compares two nested models and is based on the (R)LLR statistic: $2[\log(L_{H_1}) - \log(L_{H_0})]$, where L_{H_1} and L_{H_0} are the likelihoods of two nested models, with (under H_1) and without (under H_0) the tested effect(s). Given its versatility and the multivariate testing procedure explained thereafter in Section 5.4.5, LiMM-PCA will rely on the (R)LLR approach to test the effects significance.

For fixed effects, maximum likelihood estimates must be used in the test statistic but for variance parameters, the REML likelihood values should be used to compare different variance structures through the LLR statistic (Verbeke & Molenberghs, 2000).

When testing for the fixed effects, the LLR follows asymptotically a χ_k^2 distribution under H_0 where k is the difference in the number of parameters between the two models. This approach is frequently used in this context, but produces anti-conservative p-values for moderate to small sample sizes (Pinheiro & Bates, 2000). For the variance components, other difficulties arise in the derivation of the asymptotic distribution of the RLLR to test for non-null variance parameters. Since H_0 is on the boundary of the parameter space, the standard asymptotic theory does not hold anymore as regularity conditions are not met and the limiting distribution does not follow a χ^2 distribution under H_0 . Self & Liang (1987) and Stram & Lee (1994) demonstrated for a number of special cases that the distribution of the RLLR under standard conditions (independent and identically distributed (i.i.d.) random terms) is indeed a mixture of χ^2 distributions. For example, consider the case of one random effect α , and testing $H_0: \sigma_\alpha^2 = 0$ against $H_1: \sigma_\alpha^2 > 0$, the asymptotic distribution of the RLLR is a 50:50 mixture of χ_0^2 and χ_1^2 . Giampaoli & Singer (2009) further proved that the mixture of χ^2 distributions is the actual asymptotic distribution of the RLLR under nonstandard conditions for models with one or two random effects. Nevertheless, Crainiceanu & Ruppert (2004) demonstrated for simple cases that the 50-50 weights are inappropriate for a moderate to small amount of random levels, and Drikvandi *et al.* (2012) explained that determining the weights of this mixture distribution can be difficult, especially when testing multiple variance components in more advanced designs.

An alternative solution for significance tests, that can be applied both for fixed effects and variance parameters is to determine the full and null distributions of the (R)LLR test statistic by a parametric bootstrap approach (Davison *et al.* , 1997; Halekoh *et al.* , 2014). This approach has been adapted to a multivariate setting in LiMM-PCA.

5.4 LiMM-PCA: combining ASCA⁺ and LMM for advanced designs and modelling

The aim of this chapter is to extend the current ASCA⁺ methodology to more advanced designs and models by the use of a multivariate mixed model that includes both fixed and random effects. The proposed method is called Linear Mixed Models-PCA or LiMM-PCA. A summary of its workflow is illustrated in Figure 5.1 where the yellow boxes highlight new developments compared to ASCA⁺.

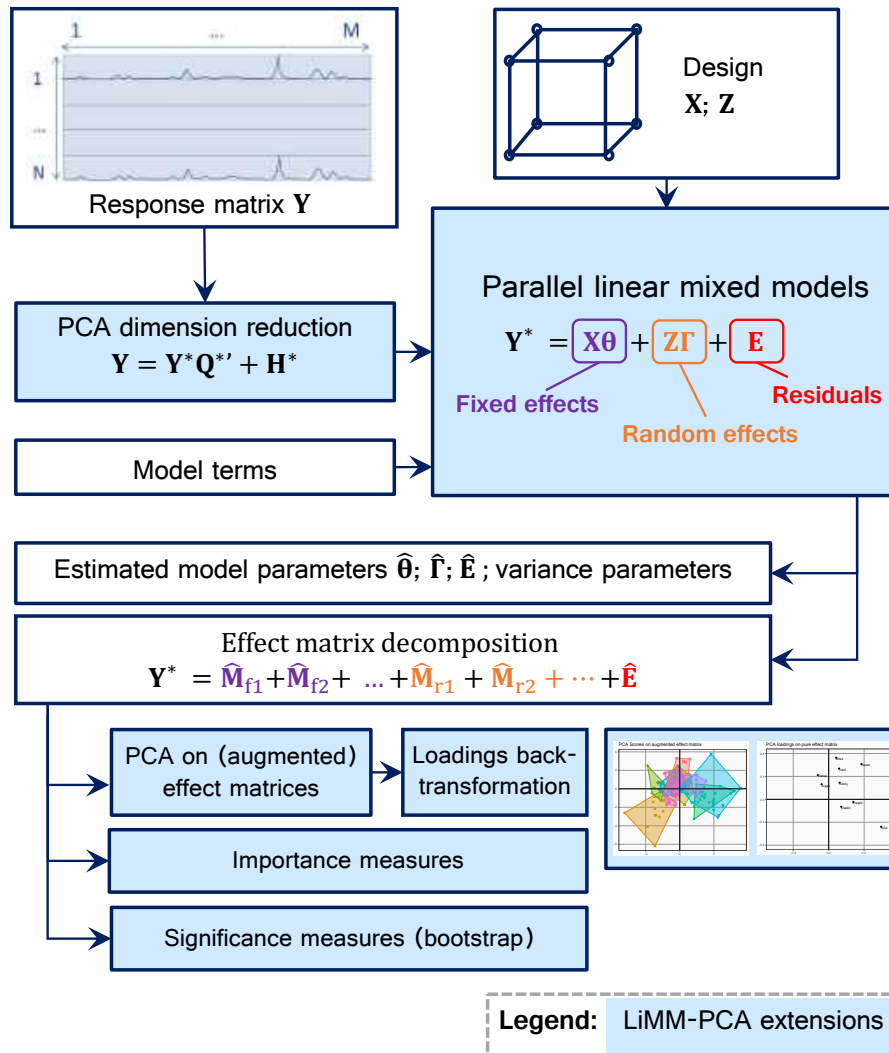


Figure 5.1 – Combination of ASCA⁺ and LMM to analyse high dimensional designed data.

5.4.1 STEP 1: PCA orthogonalisation and dimension reduction of the response matrix

In LiMM-PCA, before assessing links between the multivariate responses and the factors of interest, the response matrix $\mathbf{Y}$ should be prepared in order to comply as much as possible with the statistical properties that are expected in LMM. First, a mathematical transformation can be applied to each response to ensure that its distribution is close to a normal one (e.g. $\log(\mathbf{Y})$, $\sqrt{\mathbf{Y}}$, or any other Box-Cox transformations). Then, a linear transformation should be applied to the data in order to get the more independent responses as possible. This is motivated by the fact that linear mixed modelling of multiple responses benefits from much simpler inferential properties when the responses are independent. This second step can be achieved by PCA orthogonalisation of the response matrix because it produces a set of uncorrelated responses, which are independent if the original responses do follow a multivariate normal distribution. This transformation is motivated by the fact that linear mixed modelling of multiple responses benefits from much simpler inferential properties when the responses are independent, see Section 5.4.2.

After these transformations, the methodology developed subsequently implies that the hypotheses of independence and normality are almost verified for the transformed responses. If these properties cannot be achieved for the data of interest, most of the methodology remains still valid but some inferential results may be affected.

This PCA transformation should be associated with a dimension reduction by dropping the very last components, especially when the number of variables exceeds that of the observations. This will have a positive impact on the computational burden without losing the important information contained in the data. This transformation – PCA and dimension reduction – is not affecting the final results interpretation since a back-transformation, applied to the loadings, will allow to project them onto the original m -dimensional response space.

More formally, the normalised and centred $(n \times m)$ response matrix $\mathbf{Y}$ will be transformed by PCA into a $(n \times m^*)$ matrix $\mathbf{Y}^*$ defined as $\mathbf{Y} = \mathbf{Y}^* \mathbf{Q}^{*'} + \mathbf{H}^*$, where m^* , the number of PC retained, should be high enough to retain most of the information contained in the data. The choice of m^* can be assessed by a scree plot or any other classic tool. The $(m \times m^*)$ loadings matrix $\mathbf{Q}^*$ is kept for later use to revert to the original variables space.

5.4.2 STEP 2: Parallel mixed modelling

As in ASCA – where an ANOVA/GLM model is first fitted to each response variable – a linear mixed model is fitted, by maximum likelihood estimation, to each column

$\mathbf{y}_j^*$ of the new response matrix $\mathbf{Y}^*$:

$$\mathbf{y}_j^* = \mathbf{X}\boldsymbol{\beta}_j + \mathbf{Z}\boldsymbol{\alpha}_j + \boldsymbol{\epsilon}_j \quad \text{for } j = 1, \dots, m^*. \quad (5.5)$$

The model should to be built according to the underlying experimental design and we will assume, as in ASCA, that the model matrix $\mathbf{X}$ and $\mathbf{Z}$ are organised in blocks corresponding to the F fixed and R random effects of the model: $\mathbf{X} = (\mathbf{X}_0 \mid \mathbf{X}_1 \mid \mathbf{X}_2 \mid \dots \mid \mathbf{X}_F)$ and $\mathbf{Z} = (\mathbf{Z}_1 \mid \mathbf{Z}_2 \mid \dots \mid \mathbf{Z}_R)$. The first column of $\mathbf{X}$ is a constant term which is not always necessary if the design is balanced.

For each response j , the fixed parameters $\boldsymbol{\beta}_j$ and random predictors $\boldsymbol{\alpha}_j$ are estimated as described in Section 5.3.3 where: $\hat{\boldsymbol{\beta}}_j = (\mathbf{X}'\hat{\mathbf{V}}_j^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}_j^{-1}\mathbf{y}_j^*$, $\hat{\boldsymbol{\alpha}}_j = \hat{\mathbf{G}}_j\mathbf{Z}'\hat{\mathbf{V}}_j^{-1}(\mathbf{y}_j^* - \mathbf{X}\hat{\boldsymbol{\beta}}_j)$, and $\hat{\mathbf{G}}_j$ is the estimated random effects covariance matrix for response j . These estimators are completed by the (RE)ML estimators of random parameters: a variance $\hat{\sigma}_{rj}^2$ is associated to each random effect r and the variance $\hat{\sigma}_{\epsilon j}^2$ to the residuals.

The estimation of these m^* models results in an estimated multivariate model of the form:

$$\mathbf{Y}^* = \mathbf{X}\hat{\boldsymbol{\Theta}} + \mathbf{Z}\hat{\boldsymbol{\Gamma}} + \hat{\mathbf{E}}; \quad (5.6)$$

where $\hat{\boldsymbol{\Theta}}$ is the $((p+1) \times m^*)$ matrix of estimated fixed parameters and $\hat{\boldsymbol{\Gamma}}$ is the $(q \times m^*)$ matrix of random predictors.

According to the decomposition of matrices $\mathbf{X}$ and $\mathbf{Z}$ in blocks, the $\hat{\boldsymbol{\Theta}}'$ and $\hat{\boldsymbol{\Gamma}}'$ can be decomposed into $\hat{\boldsymbol{\Theta}}' = (\boldsymbol{\Theta}'_0 \mid \boldsymbol{\Theta}'_1 \mid \boldsymbol{\Theta}'_2 \mid \dots \mid \boldsymbol{\Theta}'_F)$ and $\hat{\boldsymbol{\Gamma}}' = (\boldsymbol{\Gamma}'_1 \mid \boldsymbol{\Gamma}'_2 \mid \dots \mid \boldsymbol{\Gamma}'_R)$.

$\hat{\mathbf{E}}$, the $(n \times m^*)$ residual matrix, is deduced from previous calculations as: $\hat{\mathbf{E}} = \mathbf{Y}^* - (\mathbf{X}\hat{\boldsymbol{\Theta}} + \mathbf{Z}\hat{\boldsymbol{\Gamma}})$.

Note that the hypothesis of independence assumed in the first step of the LiMMPCA procedure becomes handy at this stage because it enables to neglect complicated covariance structures between the random effects of the m^* parallel models.

5.4.3 STEP 3: Effect matrix decomposition

Similarly to ASCA⁺, the response matrix can then be decomposed into a sum of fixed and random effect matrices: $\mathbf{Y}^* = \hat{\mathbf{M}}_0 + \sum_{f=1}^F \hat{\mathbf{M}}_f + \sum_{r=1}^R \hat{\mathbf{M}}_r + \hat{\mathbf{E}}$.

For a fixed effect f , $\hat{\mathbf{M}}_f$ is simply calculated as $\mathbf{X}_f\hat{\boldsymbol{\Theta}}_f$, and for a random effect r , $\hat{\mathbf{M}}_r$ is calculated as $\mathbf{Z}_r\hat{\boldsymbol{\Gamma}}_r$.

Note that these effect matrices are orthogonal to each other for balanced cases and when an appropriate coding for the factors is chosen for the model matrices (see Thiel *et al.* (2017) for more details).

5.4.4 STEP 4a : Quantification of effects importance

Quantifying the effects importance is a common output derived in ASCA-like methods applications through the literature. Before the visual interpretation of the effect matrices content, it is indeed important to get an idea of what portion of the total variance of the data is explained by each model term. Another, more formal, approach consists in evaluating which of them have a statistically significant effect on the responses. The graphical interpretation of the effect matrices should therefore be guided by this prior information at hand.

Compared to ANOVA and GLM, the quantification of effect importance is not as straightforward for mixed models since it should be able to compare the fixed and random sources of variation based on a common scale. It should further solve the non-orthogonality issue between the effect matrices when the design is not balanced. Hence, a simple calculation of their Frobenius norm, as recommended in the traditional ASCA approach, is no longer appropriate.

The suggested approach is based on the work of Nakagawa & Schielzeth (2013) that offers a general definition of marginal and conditional R^2 in the univariate LMM and GLMM framework for independent random effects. Their approach provides a direct solution to quantify fixed and random effects importance in the balanced case.

For the random effects, the estimated variance for one response $\mathbf{y}_j^*$ conditional on $\mathbf{X}$ is defined as the sum of the estimated variance components linked to all random effects of the model (residual included): $\widehat{var}(\mathbf{y}_j^*|\mathbf{X}) = \sum_{r=1}^R \hat{\sigma}_{rj}^2 + \hat{\sigma}_{\epsilon j}^2$. Regarding the fixed effects, Nakagawa & Schielzeth (2013) and Rousseau (2011) both suggest to quantify the variance of an effect f over a response j by the population variance of the estimated effect matrices: $\hat{\sigma}_{fj}^2 = var(\hat{\mathbf{M}}_{fj})$ where $\hat{\mathbf{M}}_{fj}$ is the j^{th} column of $\hat{\mathbf{M}}_f$. When the design is balanced and the factors are encoded with a sum coding, the population variance is basically the Frobenius norm of $\hat{\mathbf{M}}_{fj}$ divided by n . The total variance of $\mathbf{y}_j^*$ is later defined as the sum of all variance components linked to fixed and random effects.

At the multivariate stage, and thanks to the orthogonal property between the responses $\mathbf{y}_j^*$, all these individual variances can be aggregated over all responses. That way one can obtain the global variance for the multivariate response and the portion of this variance that is linked to each effect included in the model. The global variance of $\mathbf{Y}^*$ is estimated as:

$$\widehat{var}(\mathbf{Y}^*) = \sum_{j=1}^{m^*} \widehat{var}(\mathbf{y}_j^*) = \sum_{j=1}^{m^*} \left(\sum_{f=1}^F \hat{\sigma}_{fj}^2 + \sum_{r=1}^R \hat{\sigma}_{rj}^2 + \hat{\sigma}_{\epsilon j}^2 \right) ;$$

and the variance linked to one model term/effect $g = \{r, f\}$ is $\widehat{Var} g = \sum_{j=1}^{m^*} \hat{\sigma}_{ij}^2$. The latter can also be expressed as a percentage of the total variance as: $\% \widehat{Var} g = \widehat{Var}(g) * 100 / \widehat{var}(\mathbf{Y}^*)$.

Note that if the design is balanced, and the model contains only fixed effects, this approach is equivalent to what is usually done in ASCA. In contrast, if the design is unbalanced, the formula for the fixed effect variances can be adapted on the basis of type III SS, as suggested in Thiel *et al.* (2017). In the multivariate LMM framework $\widehat{Var} f$ can be calculated as follows : $\hat{\Theta}' \mathbf{L}' (\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{L}')^{-1} \mathbf{L} \hat{\Theta} / n$ where $\mathbf{L}$ is the contrast matrix to select in $\mathbf{X}$ and $\hat{\Theta}$ the columns/lines linked to an effect f .

5.4.5 STEP 4b : Testing effects significance

This section provides a methodology to test for the statistical significance of one or a group of model terms over the response variables to complement the computation of the effects importance. In single response LMM models, the more general testing procedure is the LLR test. It is extended here to a multivariate test procedure able to deal with univariate tests multiplicity issues. Once more, this test relies on the hypothesis of near independence between the individual transformed responses $\mathbf{y}_j^*$ thanks to the initial response matrix $\mathbf{Y}$ orthogonalisation.

The Global Log Likelihood Ratio test statistics for (a group of) effect(s) $\tilde{g} = \{\tilde{f}, \tilde{r}\}$ $\text{GLLR}_{\tilde{g}}$ is defined as the sum of the independent single response LLR statistics:

$$\text{GLLR}_{\tilde{g}} \text{ statistic: } \Lambda_{\tilde{g}}^{obs} = 2 \left[\sum_{j=1}^{m^*} (\log(L_{H_1, j\tilde{g}}) - \log(L_{H_0, j\tilde{g}})) \right] ;$$

where $L_{H_1, j\tilde{g}}$ and $L_{H_0, j\tilde{g}}$ are the likelihood values of two nested models with response j : the one under H_1 , where the tested effect(s) $\tilde{g}$ is (are) included in the model, and the other under H_0 , without the tested effect(s) $\tilde{g}$ of interest.

When testing for the fixed effects, the asymptotic distribution under H_0 is a $\chi_{m^* \times k}^2$ where k corresponds to the number of parameter's restrictions between the two univariate models under H_0 and under H_1 .

Recall that testing for the random effects is not as straightforward when testing the nullity of variance components at the border of the parameter space. In this context and as shown in Section 5.3.4, the distribution of the RLLR under standard conditions (independent and identically distributed (i.i.d.) random terms) is indeed a mixture of χ^2 distributions, and we will illustrate through the case study in Section 5.5.1, that this approximation is providing good results.

However, due to difficulties in the derivation of finite sample distributions of the (R)LLR, the effects significance are rather assessed by a parametric bootstrap procedure using the scheme suggested in Davison *et al.* (1997); Sinha (2009) or

Halekoh *et al.* (2014). It is extended in the multivariate case to evaluate the significance of (a) given model effect(s) $\tilde{g}$ as follows:

- Define the model under H_0 obtained by excluding the effect(s) $\tilde{g}$ from the full model: $\mathbf{Y}^* = \tilde{\mathbf{X}}\tilde{\boldsymbol{\Theta}} + \tilde{\mathbf{Z}}\tilde{\boldsymbol{\Gamma}} + \mathbf{E}$ where the $\sim$ notation represents the model elements without the effect(s) $\tilde{g}$ to be tested.
- Fit the multivariate mixed model under H_1 and H_0 by respectively including and excluding the effect(s) $\tilde{g}$.
- Calculate the observed GLLR statistic $\Lambda_{\tilde{g}}^{\text{obs}}$.
- Keep the following estimated parameters of the model under H_0 to use them in the bootstrap loop: the fixed parameters matrix $\hat{\boldsymbol{\Theta}}$, the variances of the random effects $\hat{\sigma}_{rj}^2$ and the residual variances $\hat{\sigma}_{\epsilon j}^2 \forall j = 1, \dots, m$.
- Repeat a high number of times ($b = 1, \dots, B$):
 - Generate a bootstrap response matrix $\mathbf{Y}_b^*$ as $\mathbf{Y}_b^* = \tilde{\mathbf{X}}\hat{\boldsymbol{\Theta}} + \tilde{\mathbf{Z}}\tilde{\boldsymbol{\Gamma}}_b + \mathbf{E}_b$ where the matrices $\tilde{\boldsymbol{\Gamma}}_b$ and $\mathbf{E}_b$ are randomly generated from independent normal distributions with zero mean and variances issued from the estimated model under H_0 .
 - Estimate the full and restricted models on this new response matrix $\mathbf{Y}_b^*$ and calculate the GLLR statistic $\Lambda_{\tilde{g}}^b$. Note that this statistic should be close to 0 since the responses have been generated based on the restricted model.
- Calculate the p-value for the test statistic as the tail probabilities defined in Davison *et al.* (1997, Chapter 4): $p_{\tilde{g}}^{\text{boot}} = \frac{\sum_{b=1}^B I(\Lambda_{\tilde{g}}^b \geq \Lambda_{\tilde{g}}^{\text{obs}}) + 1}{B+1}$.

This procedure may be applied individually to each effect r or f of the mixed model in order to get a measure of their significance.

5.4.6 STEP 5 : Visual representation of the effect matrices

The last and probably the more informative step of LiMM-PCA aims at offering to the scientist a way to visualise the modelling results through a set of graphics. LiMM-PCA's philosophy is close to what is nowadays used in the ASCA framework.

In ASCA, each pure effect matrix $\mathbf{M}_g$ for an effect g of interest (*i.e.* those which have been shown to be important in Step 4) is first decomposed by PCA. The loadings are plotted to illustrate which variables are affected by the effect g and the pure or augmented scores are then used to visualise the behaviour of the effect on the experimental points. In LiMM-PCA, this PCA decomposition of $\hat{\mathbf{M}}_g$ is written as $\hat{\mathbf{M}}_g = \mathbf{T}_g \mathbf{P}_g^{*'} where c_g is the rank of $\hat{\mathbf{M}}_g$, $\mathbf{T}_g$ is the "pure" scores matrix of size $(n \times c_g)$, $\mathbf{P}_g^*$ is the loadings matrix of size $(c_g \times m^*)$. As the original responses $\mathbf{Y}$ were orthogonalised into $\mathbf{Y}^*$ by PCA before being introduced in the$

mixed models, the P_g^* loadings must be back-transformed in order to be interpreted in their initial variable space. The resulting $(c_g \times m)$ matrix is obtained as follows: $P_g = P_g^* Q^{*'}.$

In ASCA⁺, the effect matrices were augmented by the residuals to allow to visualise graphically the effects significance via a scores plot. However, the scores augmentation with random effects requires particular attention in LMM as more than one source of random variation is now present in the model. Recall from Table 5.1 that if effects B and AB are random in the two-way mixed model described in Equ. (5.1), the MS_{AB} is the denominator in the F-value test statistic when testing for fixed effect A . In LiMM-PCA, we suggest then to use the corresponding effect matrix M_{AB} to augment M_A . In addition, a correction factor is used here to take into account the degrees of freedom (df) and the F-distribution quantile that are used in the formula of the associated F-test.

For the fixed effect A , we propose concretely to calculate the augmented effect matrix as $M_A + C * M_{AB}$ with C calculated as $\sqrt{df_A/df_{AB} * F_{df_A, df_{AB}, 1-\alpha}}$, where $F_{(df_A, df_{AB}, 1-\alpha)}$ is the $(1 - \alpha)^{th}$ (e.g. 95th) percentile of the $F_{(df_A, df_{AB})}$ distribution. Indeed, the correction factor C can be considered as an empirical choice of correction, as the distribution is only following a *pseudo F* distribution. The degrees of freedom in this formula should however be calculated with care because the random effects in LMM are not estimated from the simple ANOVA decomposition given in Equ. (5.3) but with the predictors of the random effects as defined in Section 5.3.3. These estimators are “shrunk” compared to the ANOVA Sum of Squares and actual degrees of freedom must be calculated accordingly. In the statistical literature, the “degrees of freedom” for each model effect are renamed as “effective dimensions” (ED) and can be calculated following a methodology described in Eilers (2018) and in Appendix 5.B. An ED for an effect takes the same value as the degrees of freedom of the ANOVA model (*i.e.* the number of parameters) when the effect is fixed. For a random effect r , the ED can be any value between 0 and the total number of coefficients related to this effect, depending on the value of its variance σ_r^2 estimated from the data.

5.5 Case studies

This section illustrates step by step the results of the application of LiMM-PCA to two datasets, issued from sensory sciences and metabolomics, both already published in the literature.

5.5.1 Candy sensory dataset

Dataset description

The first dataset is issued from a sensory experiment where a panel of 11 judges rated 5 different candies with respect to a series of 9 attributes: transparency, acidity, sweet taste, raspberry flavour, sugar coated texture tested with a spoon, biting strength in the mouth, hardness, elasticity in the mouth and stick to teeth in the mouth. Each candy was evaluated in triplicate by each assessor and the resulting balanced design is presented here in an unfolded data table of dimensions (165×9) where the $(165 = 11 \times 5 \times 3)$ observations are identified by 4 digits: judge-candy-replicate, e.g. 0712 corresponds to judge 7, candy 1 and replicate 2.

This dataset was first published in Luciano & Næs (2009) where they compared the PC-ANOVA and the ASCA approaches to model this dataset with a fixed and crossed 2-Factor ANOVA model. This dataset is also present in Liland *et al.* (2018), where the authors suggest a new analytical solution to estimate the multivariate classification uncertainty based on confidence ellipsoids in the balanced case, and extending their results for multiple testing of effect's levels differences.

Different questions of interest for this dataset will be answered in this section, the main ones being: *What is the mean value of each attribute for the different tested candies? Are these attributes significantly different across the candies? Is there an agreement between the judges? Do some judges appreciate certain candies? Does the dataset contain outlying judges or observations?*

In Figure 5.2, the raw values of the 9 different attributes are displayed and colour-coded according to the candies. Clearly, distinct patterns are appearing: candies $\{5, 1\}$ and $\{2, 3, 4\}$ already seem to share similar attribute rating profiles.

Statistical model

In this chapter, the statistical model chosen to analyse the design results is a typical 2-Factor mixed ANOVA model where the candy is a fixed factor and the subject is considered as random in order to allow the generalisation of the results to the *population* of potential consumers from which the 11 assessors are issued. This sensory profiling modelling is discussed in details in Næs & Langsrud (1998) and, in particular, the consequence of defining an assessor as fixed or random in such studies.

For one attribute (*i.e.* response), the model corresponds to the 2-Factor mixed ANOVA model already presented in Section 5.2. In this model, y_{ijk} represents the value of the attribute of interest for candy i , judge j and replicate k ; and $a = 5$, $b = 11$ and $k = 3$. μ represents the overall mean, α_i is the (fixed) effect of candy i , β_j is the (random) effect of judge j (β_j are independent and normally distributed

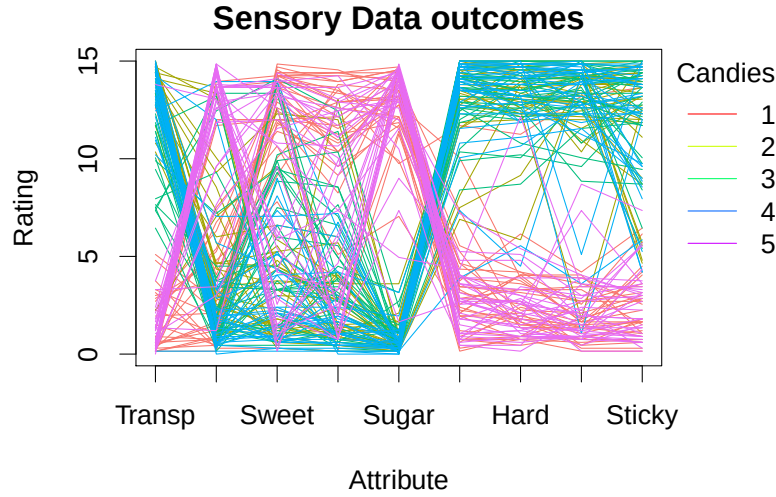


Figure 5.2 – Plot of the sensory attributes matrix.

i.e. $\sim iN(0, \sigma_j^2)$) and $(\alpha\beta)_{ij}$ is the (random) interaction between candy i and j ($(\alpha\beta)_{ij} \sim iN(0, \sigma_{C,J}^2)$). The errors ϵ_{ijk} are $iN(0, \sigma_\epsilon^2)$.

STEP 1: PCA orthogonalisation, data exploration and dimension reduction

The first step of LiMM-PCA is the PCA dimension reduction and orthogonalisation of the response matrix $\mathbf{Y}$. This section also discusses the quality of the data.

The PC1-2 scores and loadings of PCA on the raw attributes matrix are illustrated in Figure 5.3. The first PC captures almost three-quarters of the variance present in the response matrix and clearly discriminates between the two previously observed groups of candies, based on clusters of distinct attributes values. According to PC1, candies $\{2,3,4\}$ ($\{1,5\}$) have a higher (*lower*) attribute value for transparency, biting strength, hardness, elasticity, stick to teeth and a lower (*higher*) one for acidity, sweet taste, sugar coated texture and raspberry flavour. The second PC mainly captures the negative correlation between acidity and {sweet, raspberry} grades: candies $\{2, 5\}$ are mainly rated having more acidity whereas candies $\{1, 3\}$ have a sweeter, more raspberry taste. Note that an outlying observation from candy 1 is observed in the PC1-2 scores plot.

A diagnostic plot measuring Hotelling T^2 vs the Squared residual Q distances for a chosen PCA model (here with 4 PCs retained) (Jackson, 2005) was also built to detect potential outliers in Figure 5.4. T^2 represents the distance of one observation to the center of the points in the PCA model whereas the Q statistic

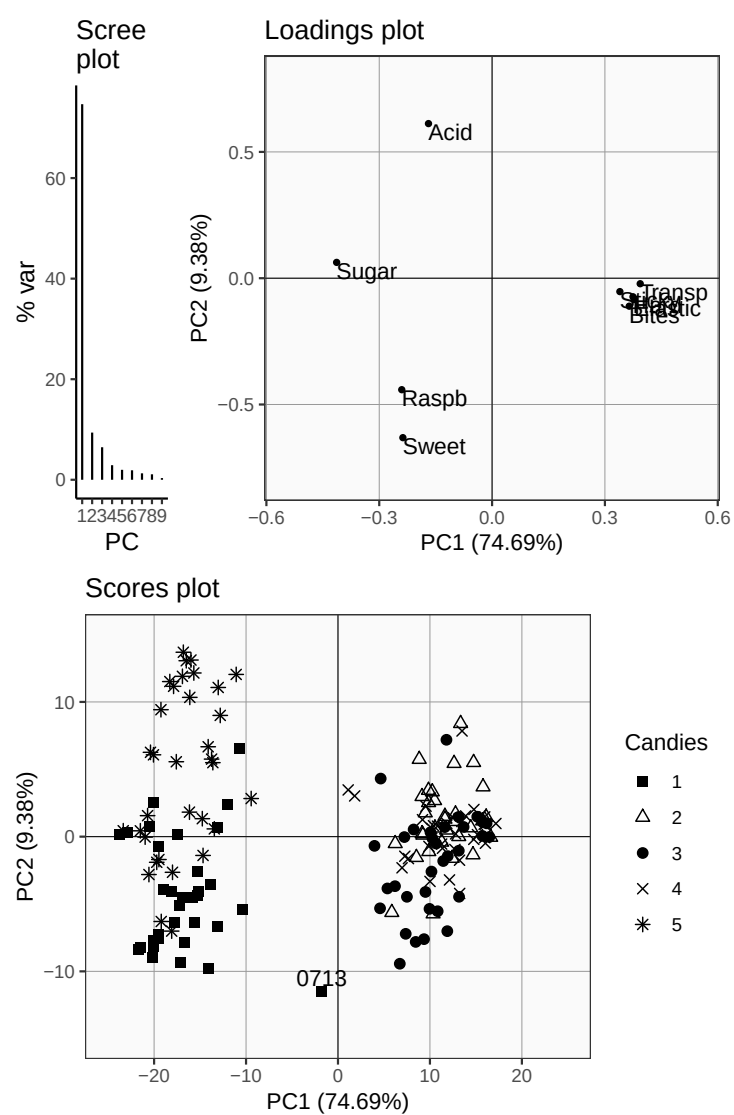


Figure 5.3 – PCA scree plot, plots of PC1-2 scores and loadings for the sensory attributes response matrix.

measures the residual variation not taken into account in the PCA model. At this stage, this graph does not point out important outliers.

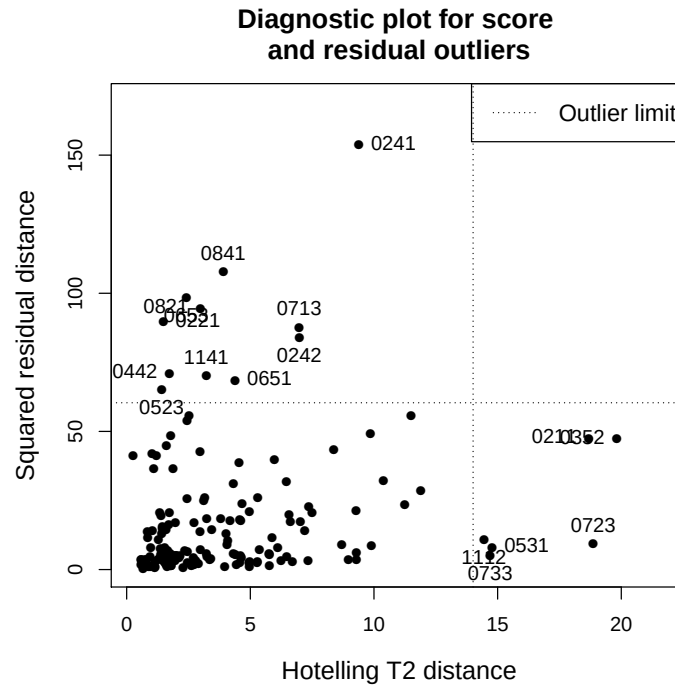


Figure 5.4 – Diagnostic plot with Hotelling T^2 vs the Squared residual distance Q to detect scores and residuals outliers for the sensory dataset.

In Luciano & Næs (2009), PC-ANOVA was applied to these data and their results will serve here as a reference for the subsequent analysis. In their paper, the candy effect was assessed by applying univariate F-tests on each scores vector of the PCA applied to the attributes matrix. However, the authors used the Mean Squares (MS) ratios of fixed effects ANOVA model for their test. For this analysis, we prefer to adopt the more conservative ratios, as advocated in mixed models, along with a Bonferroni correction of the p-values to correct for test multiplicity (p-values were multiplied by $3 \times 8 = 24$, the number of tests performed). The adapted results, given in Table 5.2 show that the effects of the three model terms are deemed significant at least once in the first three PCs. Hence, they can all be considered as significant in the present analysis accordingly. The candy effect is the predominant one in the data since it is significant in the two first PCs, which are capturing most of the attributes variability.

Table 5.2 – p-values of F-tests from the univariate 2-Factor mixed ANOVA with crossed factors performed on retained PCA principal components of the sensory data table.

PC % var. expl.	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Candy	<0.0001	<0.0001	0.0114	1	1	1	1	1
Judge	0.5014	0.3607	<0.0001	1	1	0.0998	0.0008	1
CJ	0.2780	0.0141	1	1	0.781	1	0.0123	1

Only the last PC is discarded in the following steps of LiMM-PCA in order to retain most of the information contained in the data, leaving a total of 8 PCs retained in the new response matrix. The new matrix accounts for an overall of 99.67% of the total variance.

STEPS 2 and 3 : Mixed modelling and effect matrix decomposition

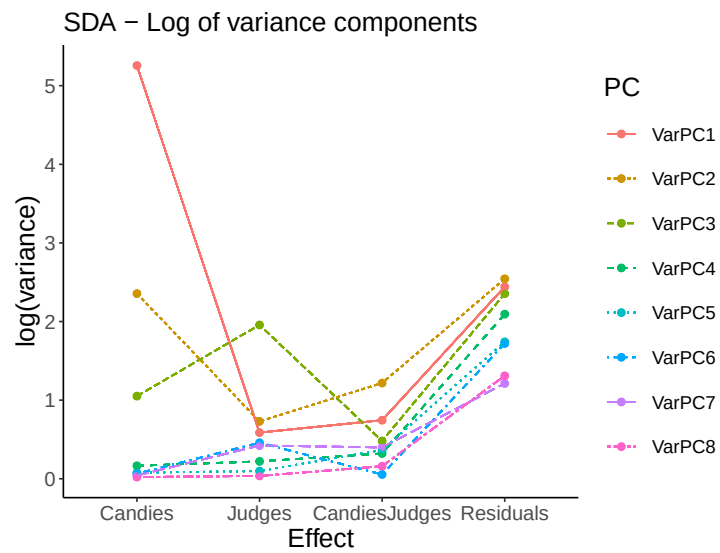
The multivariate LMM for this case study corresponds to the 2-Factor mixed ANOVA model with crossed factors already discussed in Section 5.5.1. It is written as: $\mathbf{Y}^* = \mathbf{X}_0\boldsymbol{\Theta}_0 + \mathbf{X}_C\boldsymbol{\Theta}_C + \mathbf{Z}_J\boldsymbol{\Gamma}_J + \mathbf{Z}_{CJ}\boldsymbol{\Gamma}_{CJ} + \mathbf{E}$, where 0, C, J and CJ respectively account for the fixed intercept, the candy, the judge and the interaction effects. According to Figure 5.20 in Appendix 5.B.2, the fitted model residuals are approximately normal.

After the model estimation by REML, the response matrix can be further decomposed as a sum of effect matrices: $\mathbf{Y}^* = \hat{\mathbf{M}}_0 + \hat{\mathbf{M}}_C + \hat{\mathbf{M}}_J + \hat{\mathbf{M}}_{CJ} + \hat{\mathbf{E}}$.

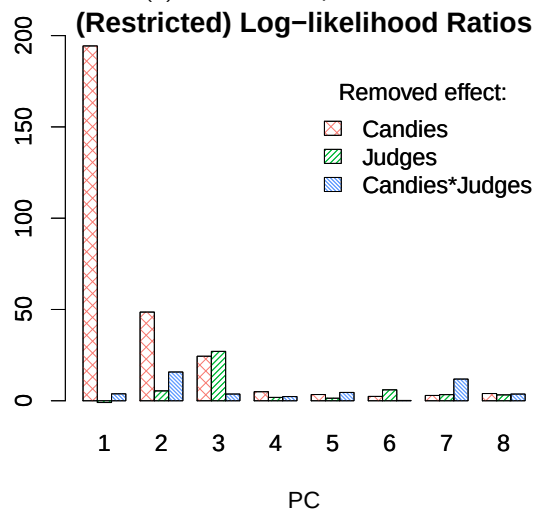
STEP 4 : Effects importance and significance testing

According to the maximum likelihood methodology presented in Section 5.4.4, the percentage of variance for each PC linked to each model effect has been calculated and is illustrated in Figure 5.5a. In addition, aggregated percentages of variance are provided in Table 5.3 along with p-values obtained with the significance testing procedures presented in Section 5.4.5.

Figure 5.5 shows that the 2 first PCs are mainly linked to the candies and the interaction effects whereas the third one is mostly linked to the judge main effect, but also to the candies one. The remaining PCs do not account for a large amount of variability and are not interpreted further. Prior to the effects significance assessment, the graph of the LLR test statistics per PC (Figure 5.5b) already gives a good indication of which effect leads to a larger increase of the global likelihood of the mixed models. Both graphs in Figure 5.5 show that the candy main effect leads to an important increase of the LLR in the first PC, but also in PC2 and 3. The judge and interaction effects are in turn responsible only



(a) Variance components.



(b) (R)LLR.

Figure 5.5 – Variance components and (R)LLR issued from the mixed models applied on the sensory dataset.

for a more moderate increase of some of the first PCs and only a formal test is able to conclude to their true significance.

Table 5.3 – Global effect importance and tests of effect significance for the sensory dataset.

Effect	Global variance (%)	Bootstrapped p-value	χ^2 test
Candy	74.77	< 5e-04	< 5e-04
Judge	3.49	0.0355	-
CJ	2.08	< 5e-04	-
Residuals	19.66	-	-

In Table 5.3, the global variance, accounting for the effects importance, is split into the model effects as expected in sensory profiling *i.e.* the fixed candy effect is by far the main source of variability in the attributes rating, whereas the random effects (the assessor and the interaction) represent a minor part of the variability present in the data. Note that the residual variability (or the intra-assessor variability) is large compared to the other random effects. Table 5.3 further provides the p-values of the GLLR test statistics obtained by bootstrap strategy and by a χ^2 with 32 (4×8) degrees of freedom for the fixed effect. They confirm that all the effects included in the model are highly statistically significant.

Histograms of the bootstrapped G(R)LLR under H_0 are presented for each model effect in Figure 5.6, where they are compared to the observed test statistics. It shows that for the fixed effect, the χ^2_{32} works well to approximate the small sample distribution of the test statistic.

Other derived bootstrap results are depicted in Appendix 5.C where it is shown based on histograms and qqplots that the χ^2_4 distribution matches the LLR under H_0 for the fixed effect as well as mixtures of χ^2 do better approximate RLLR under H_0 for the random effects.

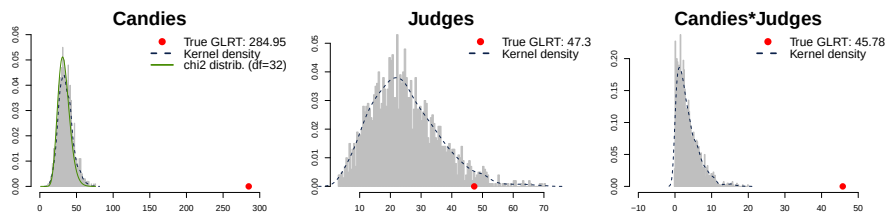


Figure 5.6 – Histogram of bootstrapped G(R)LLR for each effect (fixed candies and random judges and candies*judges effects) and for the sensory dataset.

STEP 5 : Visual representation of the effect matrices

In order to visualise the resulting effect matrices, PCA is now applied on each of them: $\hat{M}_g$ is reconstructed into $T_g P_g^*$. The loadings are then back-transformed to visualise the variables and the augmented effect matrices projected on the space to the first PC's to visualise the observations.

Based on $E(MS)$ and F-tests from the 2-Factor mixed ANOVA model in Section 5.2 (cf. Table 5.1), Table 5.4 presents how the effect matrices should be augmented in the ASCA-E fashion: the interaction effect matrix, corrected with the correction factor as explained in Section 5.4.6, is added to the candy effect matrix, and the assessors and the interaction effect matrices are augmented by the corrected residuals matrix.

Table 5.4 – Effect matrix and correction factor added to augment each effect matrix for the sensory dataset. Type is either fixed (F) or random (R). a is the number of levels of the candy effect (5). F_{ED_1, ED_2} represents the 95% quantile of the F-distribution with ED_1, ED_2 the partial effective dimensions.

Effect	Type	Associated ED	Effect matrix added	Correction factor
Candy	F	$ED_C = a - 1$	$\hat{M}_{CJ}$	$\sqrt{F_{df_C, df_{CJ}} \times \frac{ED_C}{ED_{CJ}}}$
Judge	R	ED_J	$\hat{E}$	$\sqrt{F_{df_J, df_E} \times \frac{ED_J}{ED_E}}$
CJ	R	ED_{CJ}	$\hat{E}$	$\sqrt{F_{df_{CJ}, df_E} \times \frac{ED_{CJ}}{ED_E}}$
Error	R	$ED_E = n - (ED_C + ED_J + ED_{CJ})$	-	-

The resulting scores and loadings plots are provided in Figure 5.7. Note that to enhance the visual interpretation of scores plots, ellipses were added for the candy and judge effects, and points were grouped and linked by the candy-judge centroid value of those scores. They lead to the conclusions that follow.

As already observed in the PCA scores of the raw data (Figure 5.3), the candies (A) are well separated. Differences between the candies $\{1,5\}$ and $\{2,3,4\}$ are mainly driven along PC1 by the previously described attributes clustering in this Figure. Furthermore, the loadings indicate that candies having an important sugar texture are likely to have a lower rating for transparency, biting strength, hardness, elasticity and stick to teeth attributes. Similarly, sweet candies with raspberry flavour will have a lower acidity taste.

In the scores plot for the judge effect (B), the first PC captures most of the variance linked to this effect. Clearly, some judges are less homogeneous in their rating than others (e.g. 2, 5 and 7) and tendencies of under- or over-rate the

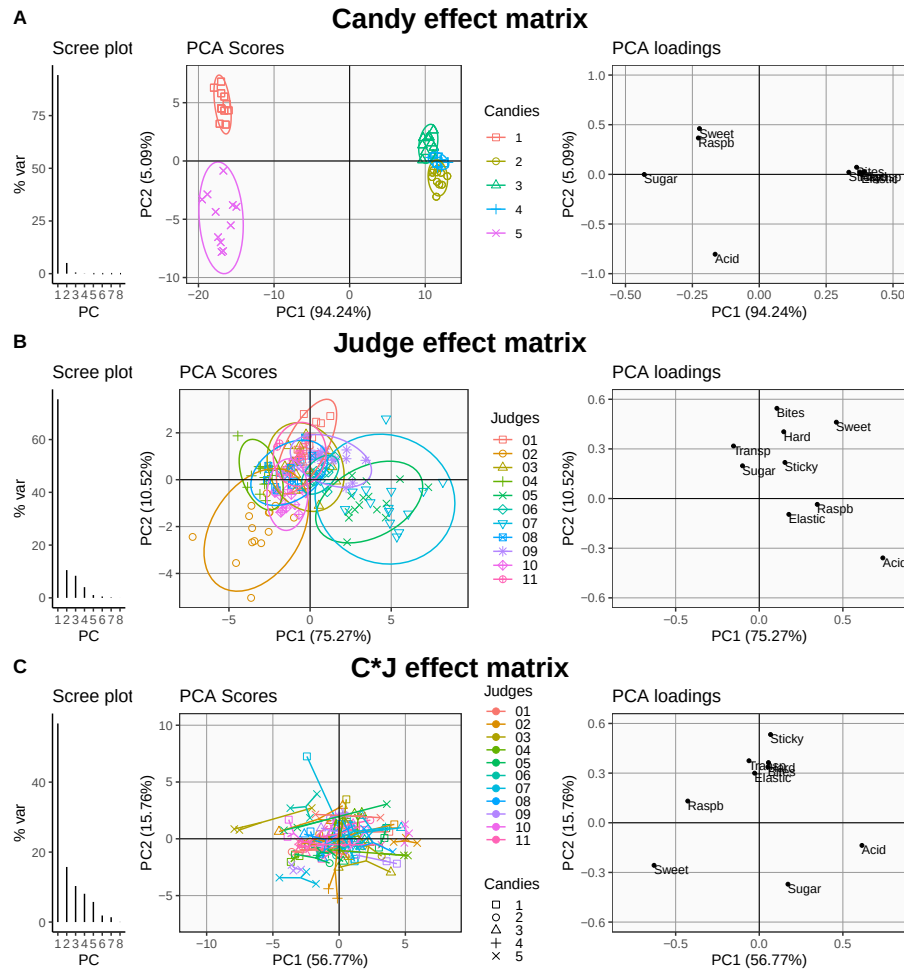


Figure 5.7 – Individual PCA on the effect matrices for the sensory dataset. From left to right: scree plots, scores and loadings of {PC1,PC2}. The scree plots and loadings plot are from the PCA on the pure effect matrices whereas the scores are augmented.

attributes can be observed: e.g. assessors 7 and 5 {resp. 2 and 4} tend to overrate {resp. underrate} the attributes, and mainly the acidity.

For the interaction effect (C), the first PC is driven by the distinction between the acidity and {sweet, raspberry taste}, already spotted by the PCA on the candy main effect matrix. From the scores plot, the main judges responsible for the interaction effect are 2 and 7. It means that they are the one differing the most from the average profile of attributes rating. The main candy responsible for the interaction is candy 5. It can be interpreted as a disagreement over the ratings of candy 5 attributes by the judges, which is also visible from the greater scattering of scores linked to this candy in the scores plot of the candies main effect matrix.

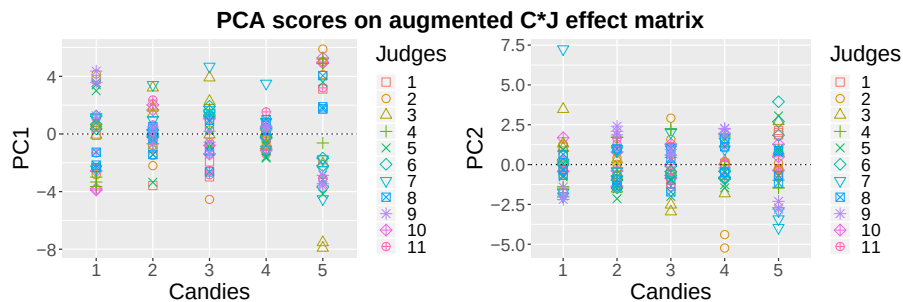


Figure 5.8 – Representation of PC1-2 scores for the augmented candy*judge interaction effect matrix for the sensory dataset.

A comparison of these LiMM-PCA results to the ASCA results in Luciano & Næs (2009) shows clearly the same results for the loadings of candy effect but only a similar pattern for the scores, that are augmented in our case.

Scores and loadings for the judge effect are comparable along PC1 to these ASCA results except that here PC1 captures a larger part of the effect matrix variability. Nevertheless, results differ along PC2. For the interaction, PC1-2 loadings are similar, although more informative than ASCA. However, the scores augmentation enables to spot the main observations responsible for this interaction.

Finally, Figure 5.9 provides PCA scores and loadings on the residuals matrix, representing the variability of the 3 replicates for each volunteer and candy. What stands out first is that there is no emerging structure from those residuals. The variability in the PC1-2 space is mainly driven by 2 outlying observations on candy 1: one from assessor 7 and the other from assessor 2. Note that removal of those outliers would lead to a slight transfer of the global variance distribution between the residuals and all the model effects. Compared to Table 5.3, global variances would respectively be: 76.39% 3.65% 2.72% and 17.24%. The loadings plot shows that, along PC1, the variability is driven by the outlying observation from assessor 2,

mainly regarding the attributes sweet and raspberry taste and transparency, which are clearly misattributed (*cf.* Table 5.5). The outlying observation from assessor 7 is the main source of variability that built PC2 with the attributes acidity, biting strength, hardness, elasticity and stick to teeth being erroneously rated given the other replicates from that judge. This is confirmed when looking to the row data gathered in Table 5.5. Clearly, the two judges provided abnormal scores for these two evaluations. Note that scores and loadings of the interaction effect and the residuals identify the principal outliers already highlighted in the Hotelling T^2 vs squared residuals distance Q in Figure 5.4. However, the LiMM-PCA analysis is richer in terms of interpretation as these outlying observations observed in the scores plot can be directly related to the attributes rating represented in the corresponding loadings plot.

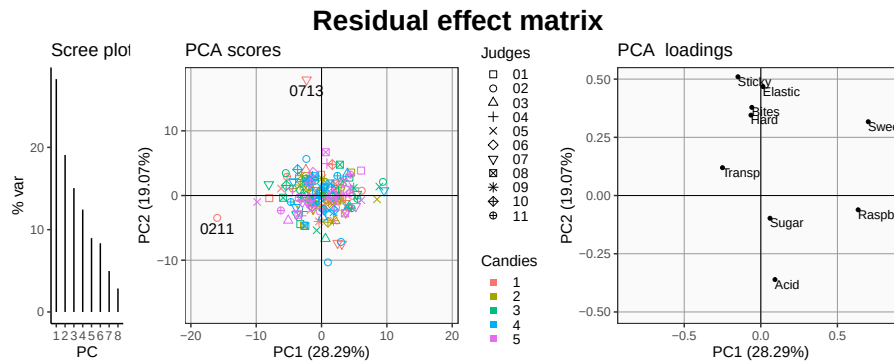


Figure 5.9 – Residuals matrix PCA scores and loadings for the sensory dataset.

Table 5.5 – Attributes rating for candy 1 and assessors 2 and 7. The outlying observations are marked as bold italic.

Obs. ID	Transp	Acid	Sweet	Raspb	Sugar	Bites	Hard	Elastic	Sticky
2_1_1	12.75	0.3	0.3	0.15	14.7	0.3	0.75	0.15	0.15
2_1_2	0.15	0.75	14.85	14.55	13.35	2.85	2.7	0.45	0.6
2_1_3	0.6	0.75	14.25	12.9	13.2	1.95	2.1	0.3	0.3
7_1_1	0.45	13.95	14.25	13.95	14.55	0.75	0.9	0.75	0.75
7_1_2	0.3	13.65	13.95	13.35	12.9	0.75	0.75	0.9	0.6
7_1_3	2.25	1.05	13.95	13.65	9.75	11.7	11.25	13.2	13.65

5.5.2 Variance components analysis of metabolomics data

Data set description

The metabolomics Human Serum dataset (Rousseau, 2011) is meant to study how serum NMR metabolomics spectra are influenced by various sources of variability linked to either biological (volunteer, sampling) or analytical (tube, time replication) factors. This dataset was mainly designed to quantify and compare: (1) the variability between and within volunteers and (2) the biological and analytical sources of variability. It also enables to test if these sources of variability must be considered as statistically significant with respect to pure repeatability caused by analytical noise.

In this study, 12 healthy volunteers were enrolled. For each volunteer, 3 blood samplings were performed on non-consecutive days and under similar conditions. For each sampling, 2 tubes were collected. The serum was then extracted from each sample and frozen. Each serum sample was then analysed twice by ¹H NMR spectroscopy the same day with a couple of hours between the two measurements in order to evaluate whether the samples can be conserved after defrosting. Each spectrum is available at a resolution of 750 buckets (*i.e.* variables) and observations are named with a series of 5 digits as: volunteer-sampling-tube-time; *e.g.* 01312 corresponds to volunteer 01, sampling 3, tube 1 and time 2. For more information on data preparation and pre-processing, please refer to Rousseau (2011).

4 observations had an acquisition problem and were removed beforehand (04221, 08222, 11121 and 11122), rendering the design unbalanced with a total of 140 observations instead on 144, as represented in Figure 5.10.

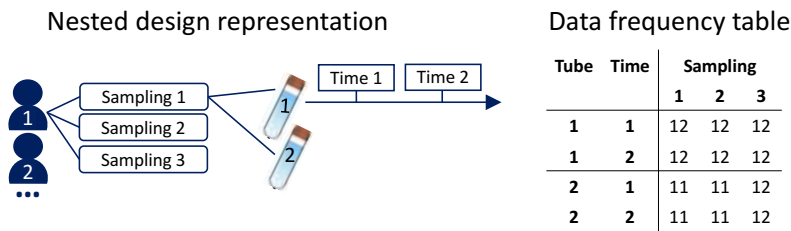


Figure 5.10 – Nested design representation and data frequency table for the metabolomics Human Serum dataset.

A stacked representation of the spectra from this dataset is given in Figure 5.11.

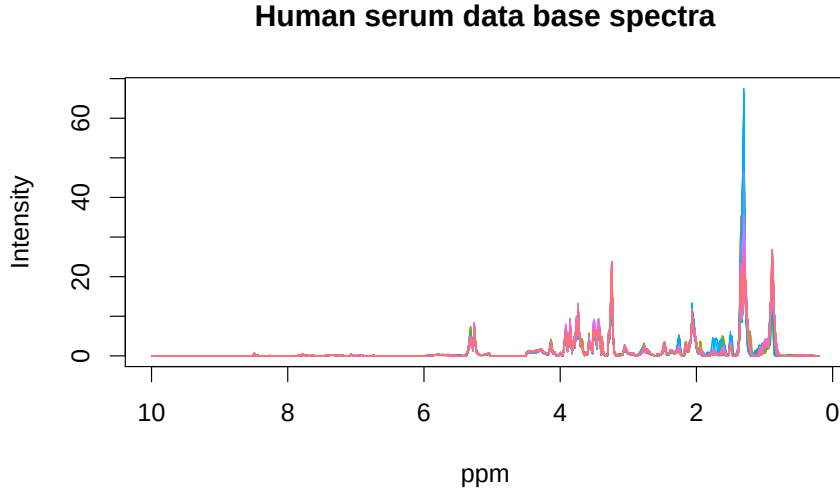


Figure 5.11 – Plot of the metabolomics dataset

Statistical model

In Rousseau (2011), the 750 responses of this four factors experiment were already analysed in parallel by univariate LMM, including the main effect of the three random effects (volunteer (V), sampling (S) and tube (U)) and the fixed factor of this study (time (T)). Note that the time effect should indeed be introduced as fixed in the model since the same time laps was applied to all samples. All other factors were considered as random because their levels represent specific instances among the population of all possible factor's levels. All the interactions were considered as negligible in the model.

In this chapter, we will proceed with the same model but rather consider the tube as a fixed factor because the data shows an unexpected but clear shift between the first and second tubes collected for each patient.

The univariate mixed ANOVA model can then be written as:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_{l(k)} + \epsilon_{ijkl} \quad (5.7)$$

with $i = 1, \dots, a; j = 1, \dots, b; k = 1, \dots, c; l = 1, \dots, d$;

where y_{ijkl} is the value of the response of interest for time i , tube j , volunteer k and sample l ; and $a = 2$, $b = 2$, $c = 12$ and $d = 3$. μ is the overall mean, α_i is the fixed effect of time i , β_j is the fixed effect of tube j , γ_k is the random effect

of volunteer k ($iN(0, \sigma_V^2)$) and $\delta_{l(k)}$ is the random effect of sampling l nested in volunteer k ($iN(0, \sigma_S^2)$). The errors ϵ_{ijkl} are supposed to be $iN(0, \sigma_\epsilon^2)$.

In this quite intricate model, it can be shown that the following ANOVA ratios should be used to test the significance of each effect: MS_T/MS_E for time effect, MS_U/MS_E for tube effect, MS_V/MS_S for volunteer effect and MS_S/MS_E for sample effect. A detailed explanation for effect matrix augmentation in the multivariate case is provided in Section 5.5.2 below.

STEP 1: PCA orthogonalisation, data exploration and dimension reduction

The prior PCA applied to the raw spectral matrix shows that 15 PCs of the 140 available cumulate 99.12% of the total variance present in the spectral profiles and are retained in what follows. PCA scores and loadings of the first 2 PCs are illustrated in Figure 5.12. They gather more than 85% of the total variability. The scores plot shows already a clear volunteer effect among the four factors of interest.

A graph diagnostic for a PCA model with 4 PCs retained in Figure 5.13 shows potential outliers in this dataset. Some of them will be more clearly identified during the LiMM-PCA modelling.

STEPS 2 and 3 : Mixed modelling and effect matrix decomposition

The multivariate mixed model corresponding to the ANOVA model defined in Section 5.5.2 can be written as: $\mathbf{Y}^* = \mathbf{X}_0\boldsymbol{\Theta}_0 + \mathbf{X}_T\boldsymbol{\Theta}_T + \mathbf{X}_U\boldsymbol{\Theta}_U + \mathbf{Z}_V\boldsymbol{\Gamma}_V + \mathbf{Z}_{S[V]}\boldsymbol{\Gamma}_{S[V]} + \mathbf{E}$; where the 0 index accounts for the fixed intercept, and "[V]" describes the nested model structure. According to Figure 5.21 in Appendix 5.B.2, the fitted model residuals are approximately normal.

After the parallel (RE)ML estimation of the model for the 15 PCs, the effect matrix decomposition equation is the following: $\mathbf{Y}^* = \hat{\mathbf{M}}_0 + \hat{\mathbf{M}}_T + \hat{\mathbf{M}}_U + \hat{\mathbf{M}}_V + \hat{\mathbf{M}}_{S[V]} + \hat{\mathbf{E}}$. Each effect matrix is of size (140×15) . Note that since the design is slightly unbalanced, these matrices are not anymore perfectly orthogonal.

STEP 4: Effects importance and significance testing

For each retained PC y_j^* , the total variance is decomposed by effect in Figure 5.14a, and aggregated over all PCs in Table 5.6. The volunteer effect captures more than two thirds of the total dataset variability, followed by the sampling effect. The remaining effects (tube and time) only account for a marginal portion of the total information extracted from the dataset. The global residual variance is also negligible. For all the effects but time, their variability contribution is widespread across multiple PCs. The time effect is only caught by PC1. Note that the 3 first

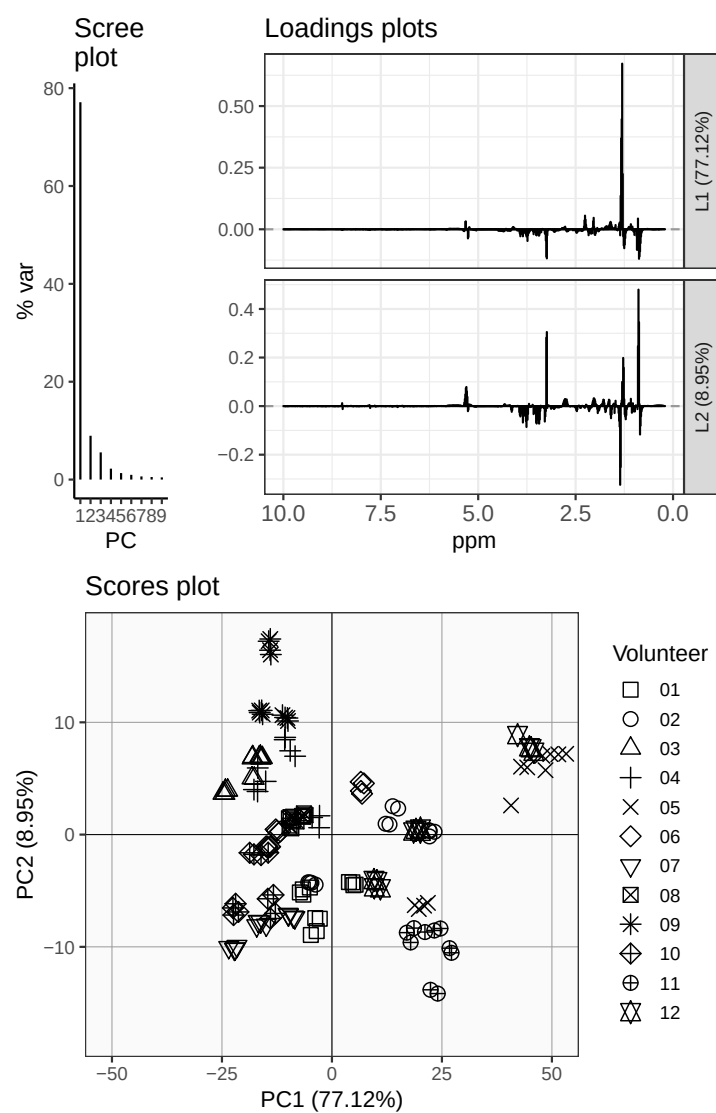


Figure 5.12 – PCA scores and loadings plots of the raw metabolomics dataset.

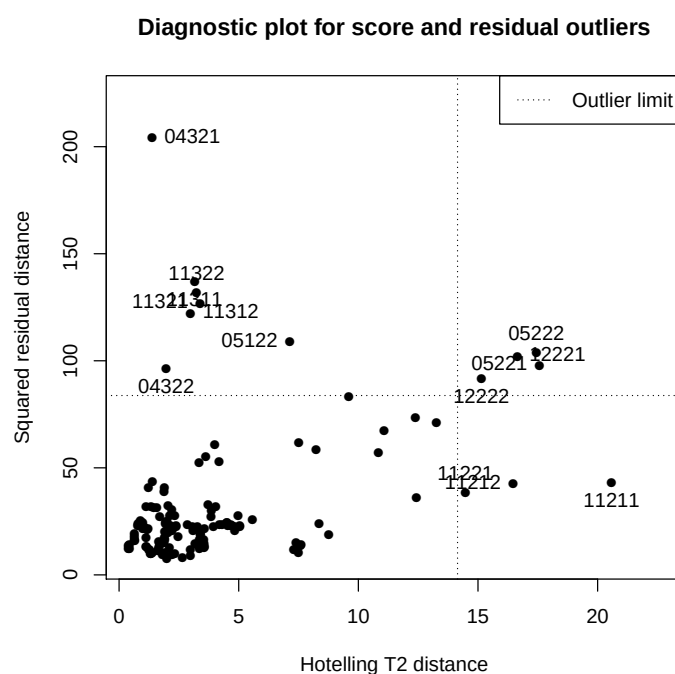


Figure 5.13 – Diagnostic plot with Hotelling T^2 vs the Squared residual distance Q to detect scores and residuals outliers for the metabolomics dataset.

PCs capture most of the volunteer's and sampling's variabilities and the tube effect is mainly contained in PC4 and 6.

For the effects statistical significance, the (R)LLR values in Figure 5.14b indicate to what extent the addition of an effect leads to an increase in the (restricted) log-likelihoods: unsurprisingly, the volunteer and the sampling effects appear to be the prominent ones. As shown for variances components, the effects are spanned across all the PCs except for the time effect, that only affects PC1. The bootstrap global test results in Table 5.6 confirms that all the fixed and random effects are deemed significant in the multivariate mixed model, although the importance of the time and tube effects is rather small.

Table 5.6 – Global effect importance and tests of effect significance for the metabolomics dataset.

Effect	Global variance (%)	Bootstrapped p-value	χ^2 test
Time	0.05	0.0015	< 5e-04
Tube	0.07	< 5e-04	< 5e-04
Volunteer	70.80	< 5e-04	-
Sampling	27.69	< 5e-04	-
Residuals	1.38	-	-

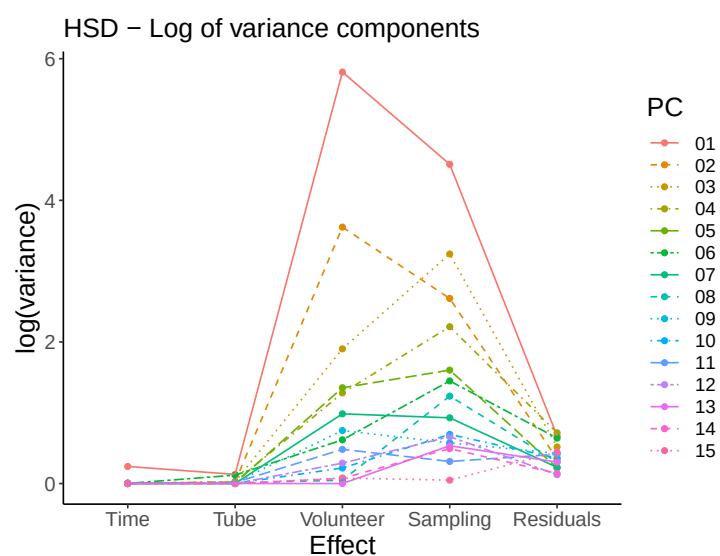
Histograms of the bootstrapped (R)LLR under H_0 are presented for each model effect in Figure 5.15, where they are compared to the observed test statistics. It confirms that for the fixed effects, the χ^2_{15} (15 PCs $\times$ 1 degree of freedom) works well to approximate the small sample size distribution of the test statistic.

STEP 5 : Visual representation of the effect matrices

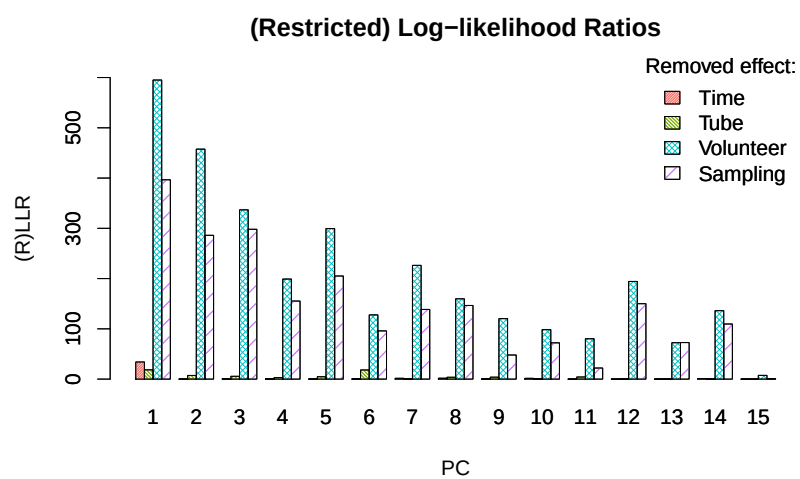
Based on the E(MS) and F-test calculations from Table 5.10 (Appendix 5.D), Table 5.7 tells which of the effect matrices must be added, corrected with the correction factor as explained in Section 5.4.6, to the effect matrix of interest in order to compare this effect against the appropriate random source of variation.

Results of PCA applied to the effect matrices with the augmented scores are given in Figure 5.16. Without going into the metabolites identification details here, a general remark concerning the loadings is that they reveal which of the spectral descriptors are most likely responsible for the variability observed in the scores between the groups of effect's levels. They can also tell if a shift between peaks of the same metabolites across spectra has been detected.

With regards to the time and tube effects (A and B), only one PC is built because these terms only have 1 degree of freedom. Therefore, PC1 captures 100% of the corresponding effect matrix variability. These two graphs show a very clear



(a) Variance components.



(b) (R)LLR.

Figure 5.14 – Variance components and (R)LLR from the multivariate mixed model applied to the metabolomics dataset.

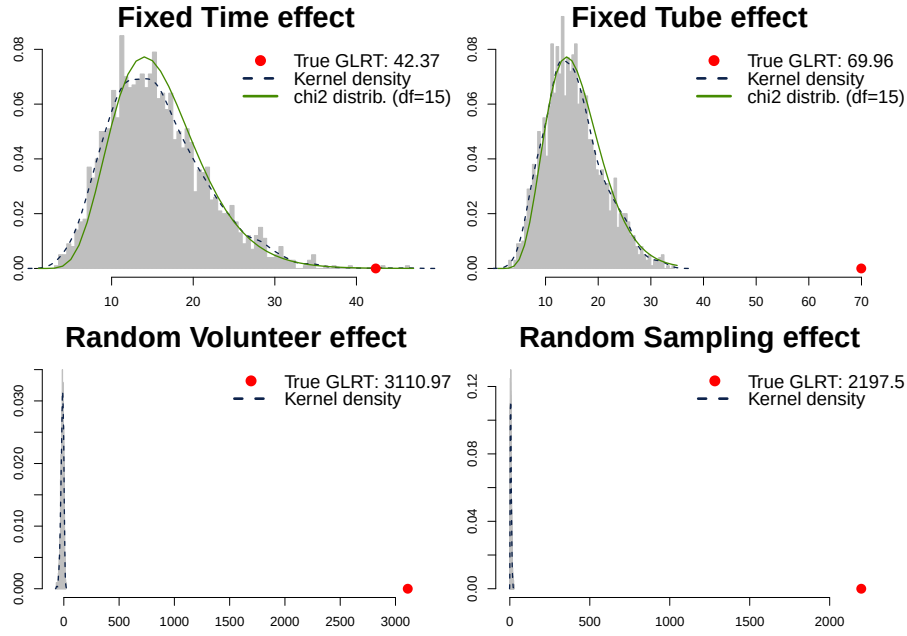


Figure 5.15 – Histogram of bootstrapped G(R)LLR for each effect for the metabolomics dataset.

Table 5.7 – Effect matrix and correction factor added to augment each effect matrix for the metabolomics dataset. Type is either fixed (F) or random (R). a and b are the number of levels respectively of the time (2) and tube (2), and F_{ED_1, ED_2} represents the 95% quantile of the F-distribution with ED_1, ED_2 the partial effective dimensions.

Effect	Type	Associated df	Effect matrix added	Correction factor
Time	F	$ED_T = a - 1$	$\hat{E}$	$\sqrt{F_{ED_T, ED_E} \times \frac{ED_T}{ED_E}}$
Tube	F	$ED_U = b - 1$	$\hat{E}$	$\sqrt{F_{ED_U, ED_E} \times \frac{ED_U}{ED_E}}$
Volunteer	R	ED_V	$\hat{M}_S$	$\sqrt{F_{ED_V, ED_S} \times \frac{ED_V}{ED_S}}$
Sampling	R	ED_S	$\hat{E}$	$\sqrt{F_{ED_S, ED_E} \times \frac{ED_S}{ED_E}}$
Error	R	$ED_E = n - (ED_T + ED_U + ED_S + ED_V)$	-	-

separation of the observations between times 1 and 2 and tubes 1 and 2. These graphs highlight also several outliers for spectra coming from volunteers 4, 5 and 11.

The sampling scores plot (C) depicts observations coming from the same volunteer and sampling. All clusters of points are well separated, confirming the significant effect of the sample within each volunteer regarding model residuals. The differences between the 3 samplings of a same volunteer are particularly noticeable for volunteers 5 and 11. The volunteer scores plot (D), shows, as expected, a very clear separation in the PC1-2 space. These two scores plots also illustrate the random nature of the volunteer and sample factors. Even if they are highly significant and show a clear separation between their factor levels, they later behave randomly around (0,0) as expected from the underlying normality assumption.

A look into the PCA scores and loadings of the residuals (Figure 5.17) – which account for the excluded higher order interaction terms and other data randomness – spots mainly the “outliers” from volunteers 4, 5 and 11, already spotted in the diagnostic plot (Figure 5.13) and the scores from the time and tube effect matrices. It seems that samples 3 of volunteer 4, 1 of volunteer 5 and 2 of volunteer 11 are especially concerned across all these graphs. The residual loadings indicate that the variability along these first two PCs is mainly driven by peak shifts in the spectra.

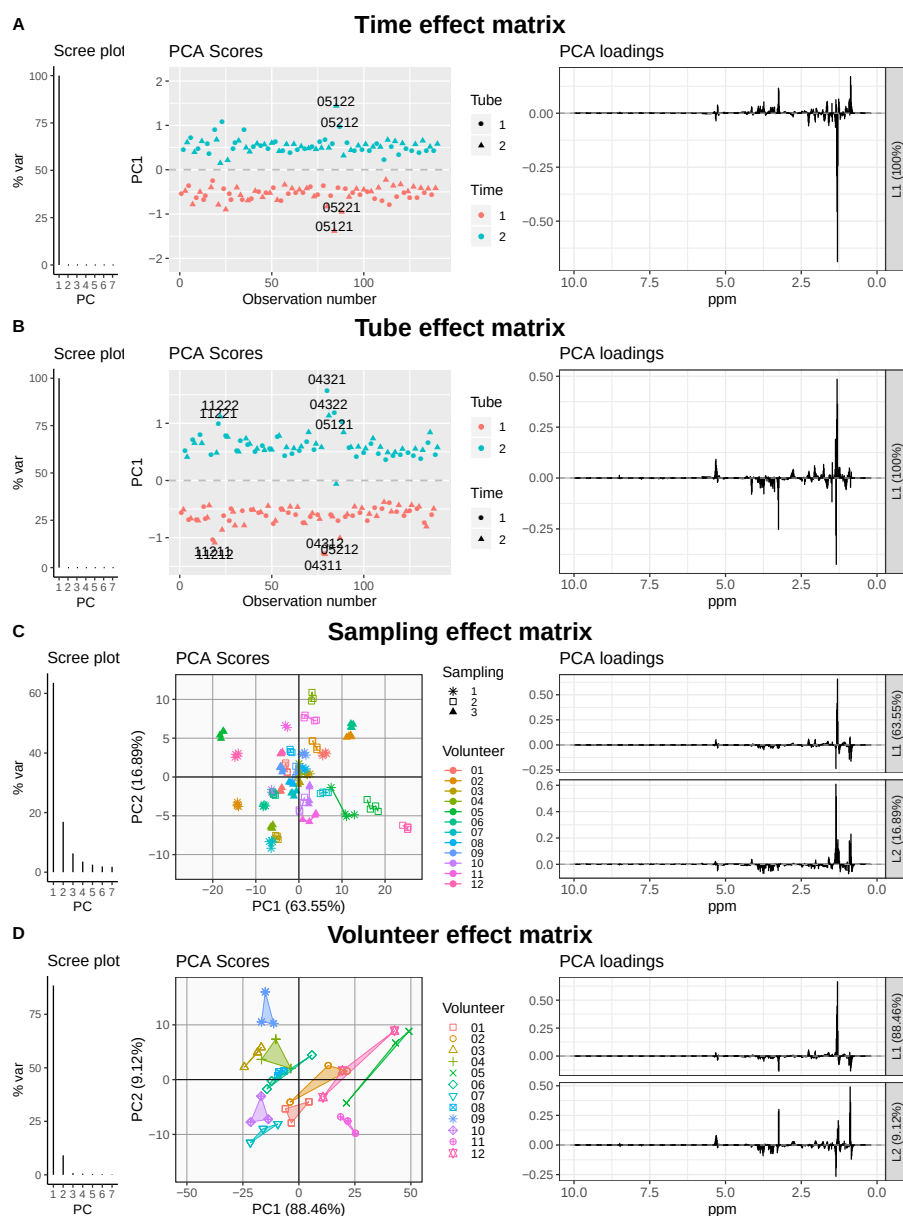


Figure 5.16 – Individual PCA on the effect matrices for the metabolomics dataset. From left to right: scree plots, scores and loadings of {PC1,PC2}. The scree plots and loadings are from the PCA on the pure effect matrices whereas the scores are augmented.

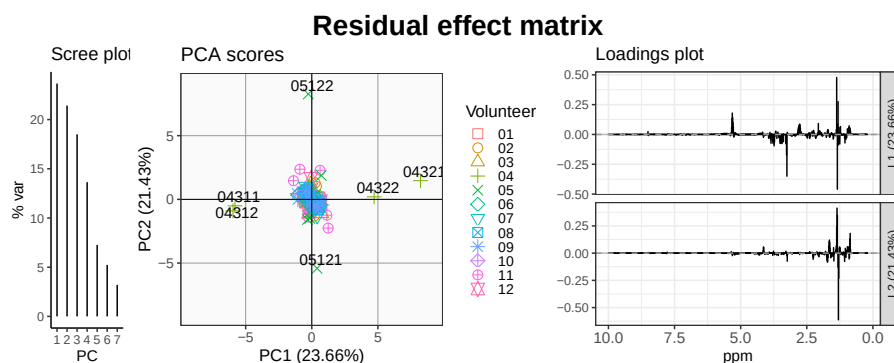


Figure 5.17 – Residuals matrix PCA scores and loadings for the metabolomics dataset.

5.6 Conclusions and perspectives

In summary, the suggested methodology, called LiMM-PCA, manages to model advanced error structures and thus **generalises ASCA** to a broader framework, by using LMM. In this respect, new methodological developments have been suggested to appropriately augment the effect matrices, as well as being able to compare the contributions of the mixed effects and test their statistical significance. The property to include random effects is indeed highly recommended to have a correct inference for specific designs such as in repeatability/reproducibility studies or for longitudinal data.

Two case studies, based on already published real experimental datasets, showed that it successfully leads to a comprehensive and fine-tuned analysis of such high dimensional and designed data. On the one side, it keeps the **advantages of ASCA⁺** by combining dimension reduction techniques and classic modelling. Effect matrix projections are undoubtedly a powerful visualisation tool of the multivariate structures in the space of each effect of the statistical model linked to the experimental design. This approach further enables to model without bias an unbalanced dataset as in ASCA⁺. On the other side, new proposed developments lead to correct inference when considering the **inclusion of random effects**, and new methodological aspects are established to **quantify and compare the mixed variability sources** as well as to provide a **global test procedure of effects significance**.

Some obstacles or unconsidered aspects are still remaining, such as the influence of the initial PCA on the generalisation of the results. One solution would be to apply a resampling scheme on the whole procedure. Also, when applying PCA in

the visualisation step, augmenting the matrices here only serves for visual purposes as the general scope of LiMM-PCA is rather to model unbalanced data. As a matter of fact, the augmented scores cannot be used to formally assess the significance of the effects. Further research should lead to a modification of the methodology to enable a visual assessment of effects significance, such as in Liland *et al.* (2018), extended to unbalanced data. One intricate aspect of ASCA-related methods that has not been debated here is the data pre-processing, and principally the scaling issue, already discussed in Timmerman *et al.* (2015). The authors demonstrated that scaling of such methods influences the data analysis results and subsequent interpretation. In the future, more emphasis and research should be conducted in this non-trivial topic.

Other perspectives arising from this work are both methodological and practical: beyond the two presented applications, LiMM-PCA could be directly applied to other kinds of chemometric applications of interest. In addition, its theoretical aspects could especially be extended to analyse data from longitudinal studies. Also, the methodology could be extended to include continuous variables in the modelling.

5.A The model fitting process in LMM

This section explains, for an audience of chemometricians which may not be familiar with LMM, the essential aspects of LMM parameters estimation.

The classic Maximum Likelihood (ML) approach to fit a LMM is separated into 3 distinct parts: estimate the fixed effects, the variance parameters, and predict the random effects. A major problem is that to estimate the fixed effects β by ML, V should be known but in practice, it is estimated. Therefore, the estimation procedure works as follows: (1) estimate the variance parameters by Restricted Maximum Likelihood (REML), treating the fixed effects as known, next (2) estimate the fixed effects (by considering the variance parameters fixed to their estimated values in (1)), and then (3) predict the random effects.

The log-likelihood function based on the multivariate normal density function is:

$$l = \log(L) = K - \frac{1}{2}[\log |V| + (y - X\beta)'V^{-1}(y - X\beta)]; \quad (5.8)$$

where K is a constant term.

Differentiating (5.8) with respect to β gives the ML estimators for the fixed effects:

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y; \quad (5.9)$$

where one can observe that $\hat{\beta}$ depends indeed on V .

The REML log-likelihood, which is based on the residual terms $(y - X\hat{\beta})$ and only defined in terms of the variance parameters, is used to calculate unbiased estimates of those variance parameters σ_ϵ^2 and γ (see Section 5.3.2 for their definition):

$$l^r(\gamma, \sigma_\epsilon^2; \mathbf{y} - \mathbf{X}\hat{\beta}) = \log(L^r) = K - \frac{1}{2}[\log |\mathbf{V}| + \log |\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|^{-1} + (\mathbf{y} - \mathbf{X}\hat{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta})]; \quad (5.10)$$

Although $\hat{\beta}$ still appears in Equ. 5.10, it does as a function of the parameters when replaced by its estimator $\hat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$. Therefore, the restricted log-likelihood will depend only on variance parameters, and no more on $\hat{\beta}$. Since the derivatives of this REML log-likelihood with respect to the variance parameters are non-linear, these variance estimators cannot be obtained analytically, as for the fixed and random effects. They need iterative numerical optimisation techniques to be derived (e.g. Newton-Raphson algorithm).

The fixed effects are thereafter estimated based on (5.9), where $\mathbf{V}$ is now replaced by its REML estimator $\hat{\mathbf{V}}$. $\hat{\beta}$ are the Empirical Best Linear Unbiased Estimators (EBLUE) of the fixed effects, where E stands for the fact that $\mathbf{G}$ and $\mathbf{V}$ are unknown and need to be estimated.

Assuming that $\mathbf{y}$ and α follow a joint normal distribution, we finally find the ML predictors of the random effects α :

$$\hat{\alpha} = (\mathbf{Z}'\Sigma^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}'\Sigma^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\beta}) \quad (5.11)$$

When replacing β by its ML estimator $\hat{\beta}$ and $\mathbf{V}$, Σ and $\mathbf{G}$ by their REML estimators $\hat{\mathbf{V}}$, $\hat{\Sigma}$ and $\hat{\mathbf{G}}$; $\hat{\alpha}$ becomes the Empirical Best Linear Unbiased Predictor (EBLUP).

Because of the $\mathbf{G}^{-1}$ term added in Equ. (5.11), estimators of the random effects are “shrunk” compared to their estimators if they were considered as fixed.

Finally, the estimators of conditional errors, called the conditional residuals are simply (Santos Nobre & da Motta Singer, 2007): $\hat{\epsilon} = \mathbf{y} - \mathbf{X}\hat{\beta} - \mathbf{Z}\hat{\alpha}$.

Note that in our case, the *lme4* R package has been used (version 1.1-21) to estimate the LMM models. Its fitting process is based on a penalised least squares representation¹ of the mixed model and a non-linear optimiser used to estimate the model parameters via an iterative process (see Bates *et al.* (2015)), which is an alternative and more efficient way to estimate the parameters.

¹The method is based on sparse matrix methods and a penalised least squares representation of the conditional model from Bates & DebRoy (2004). The conditional estimates of the fixed effects and predictors of the random effects are determined simultaneously as the solution to this penalised least squares problem representation and the variances of the random effects are estimated via a profiled log-(restricted)-likelihood.

5.B Effective model dimension in the presence of random effects

5.B.1 Theory

In linear regression models, the computation of the effective model dimension of a given model effect is straightforward: it simply coincides to the corresponding number of coefficients in the model. For example, for a fixed categorical factor with q levels, $q-1$ coefficients are included in the model. In LMM, because the estimated random predictors are *shrunk* (cf. Equ. 5.11), this rule cannot be transposed to calculate the effective dimension of a random effect (Eilers, 2018). In this case, it can be any value between 0 and the total number of random coefficients linked to that effect, depending on its estimated variance component. Inspired by Ye and Harville, Eilers (2018) suggests to rely on the so called hat matrix $\mathbf{H}$, defined as: $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ to derive those partial effective dimensions. The global effective dimension of a model corresponds to the trace of $\mathbf{H}$. In order to explain how this global effective model dimension is calculated, consider the following system of equations in a matrix form from Henderson's mixed model equations that can also be used to obtain the estimates for β and α in the univariate LMM considered in Equ. (5.4):

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\alpha} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}, \quad (5.12)$$

$\mathbf{H}$ can be decomposed as:

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y} = \mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\alpha} = \underbrace{[\mathbf{X} \quad \mathbf{Z}] \mathbf{Q} \begin{bmatrix} \mathbf{X}' \\ \mathbf{Z}' \end{bmatrix}}_{\mathbf{H}} \mathbf{R}^{-1} \mathbf{y} \quad (5.13)$$

$$\text{with } \mathbf{Q} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix}^{-1}.$$

The trace of $\mathbf{H}$ can then be written as follows:

$$\text{tr}(\mathbf{H}) = \text{tr} \left([\mathbf{X} \quad \mathbf{Z}] \mathbf{Q} \begin{bmatrix} \mathbf{X}' \\ \mathbf{Z}' \end{bmatrix} \mathbf{R}^{-1} \right) = \text{tr} \left(\mathbf{Q} \begin{bmatrix} \mathbf{X}' \\ \mathbf{Z}' \end{bmatrix} \mathbf{R}^{-1} [\mathbf{X} \quad \mathbf{Z}] \right). \quad (5.14)$$

Let define

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}^{ff} & \mathbf{K}^{fr} \\ \mathbf{K}^{rf} & \mathbf{K}^{rr} \end{bmatrix} = \mathbf{Q} \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \end{bmatrix}. \quad (5.15)$$

Since only the trace of the hat matrix is important, the global effective model dimension, $tr(\mathbf{H})$, can be decomposed in $tr(\mathbf{H}) = tr(\mathbf{K}) = tr(\mathbf{K}^{ff}) + tr(\mathbf{K}^{rr})$, where $\mathbf{K}^{ff}$ is the diagonal block of $\mathbf{K}$ related to the fixed effects and $\mathbf{K}^{rr}$ is the diagonal block related to the random effects. For each fixed or random effect, their partial effective dimension corresponds to the sum of the diagonal elements in $\mathbf{K}^{ff}$ or $\mathbf{K}^{rr}$ with the indexes of those diagonal elements corresponding to the column indexes of the related effect either in the $\mathbf{X}$ or $\mathbf{Z}$ model matrix. For example, given a fixed effect f , its effective dimension is: $ED_f = tr(\mathbf{K}_{[f,f]}^{ff})$ where $[f, f]$ are row/column indexes in $\mathbf{K}^{ff}$ corresponding to the column index of the effect f in $\mathbf{X}$. Finally, the effective dimension of the residuals is calculated as: $n - tr(\mathbf{K})$.

5.B.2 Results for the two case studies

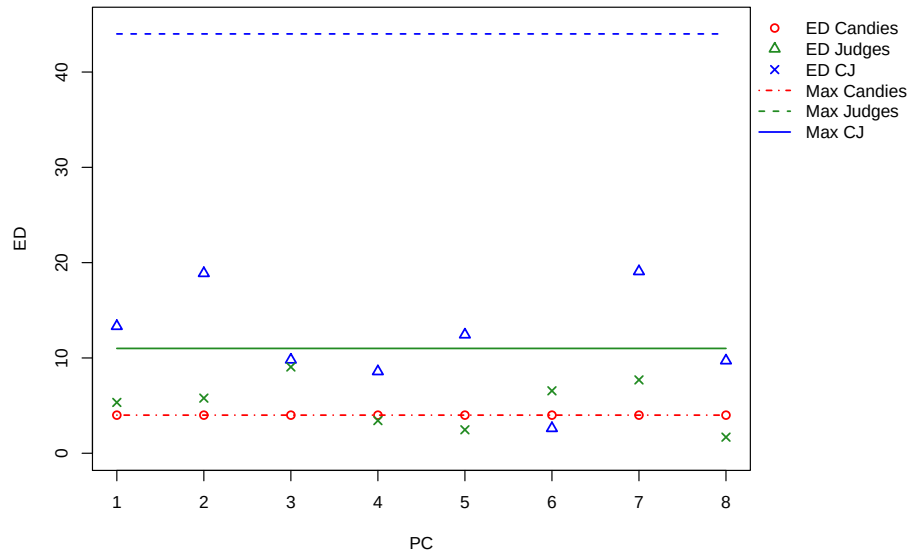


Figure 5.18 – Graphical representation of the effective dimensions of the linear mixed models for the sensory dataset.

	PC1	PC2	PC3	PC4	PC5
Intercept	1	1	1	1	1
Candies	4	4	4	4	4
Judges	5.3	5.8	9.1	3.4	2.5
CJ	13.3	18.9	9.8	8.6	12.4
Residuals	141.3	135.3	141.1	148	145.1
Total	165	165	165	165	165

	PC6	PC7	PC8	ED max.	Maximum possible dimension
Intercept	1	1	1	1	1
Candies	4	4	4	4	4
Judges	6.6	7.7	1.7	9.1	11
CJ	2.6	19.1	9.7	19.1	44
Residuals	150.8	133.2	148.6	150.8	160
Total	165	165	165	165	

Table 5.8 – Effective dimensions of the linear mixed models for the sensory dataset. Note that the maximum residuals dimension is calculated by taking into account only the effective dimension of the fixed effects.

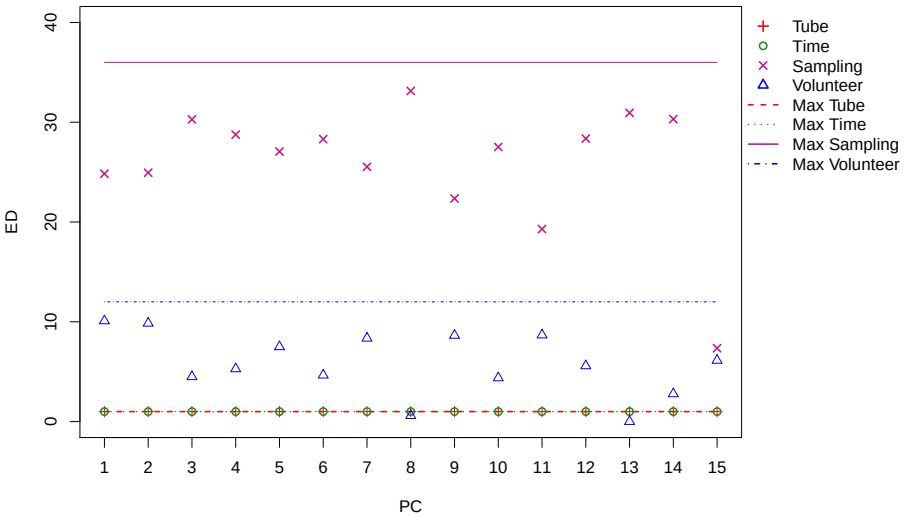


Figure 5.19 – Graphical representation of the effective dimensions of the linear mixed models for the metabolomics dataset.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Intercept	1	1	1	1	1	1	1	1	1
Tube	1	1	1	1	1	1	1	1	1
Time	1	1	1	1	1	1	1	1	1
Sampling	24.8	24.9	30.3	28.8	27.1	28.3	25.5	33.1	22.4
Volunteer	10.1	9.8	4.5	5.3	7.5	4.7	8.3	0.6	8.6
Residuals	102.1	102.2	102.2	103	102.4	104	103.1	103.3	106
Total	140	140	140	140	140	140	140	140	140

	PC10	PC11	PC12	PC13	PC14	PC15	ED max.	Maximum possible dimension
Intercept	1	1	1	1	1	1	1	1
Tube	1	1	1	1	1	1	1	1
Time	1	1	1	1	1	1	1	1
Sampling	27.5	19.3	28.4	30.9	30.3	7.4	33.1	36
Volunteer	4.4	8.7	5.6	0	2.8	6.1	10.1	12
Residuals	105.1	109	103.1	106.1	103.9	123.5	123.5	137
Total	140	140	140	140	140	140		

Table 5.9 – Effective dimensions of the linear mixed models for the metabolomics dataset. Note that the maximum residuals dimension is calculated by taking into account only the effective dimension of the fixed effects.

section Distribution of the model residuals

Figures 5.20 and 5.21 indicate that the model residuals do not present strong deviation from a normal distribution for both case studies: they do not show an accentuated U-shape in their Q-Q-plot, which is characteristic of an asymmetric distribution. Their imperfect kurtosis is mainly due to the outlying observations that were not removed in the two models.

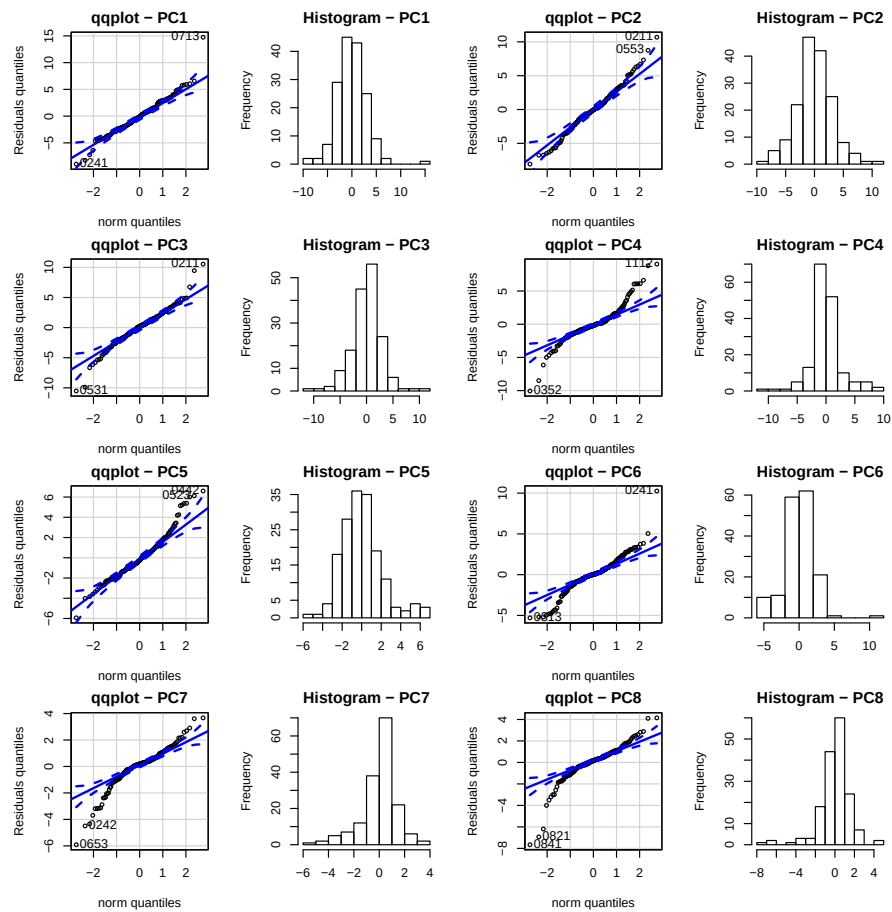
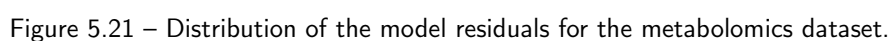


Figure 5.20 – Distribution of the model residuals for the sensory dataset.



5.C Histograms and quantiles of true and bootstrapped (R)LLR statistics for the sensory dataset

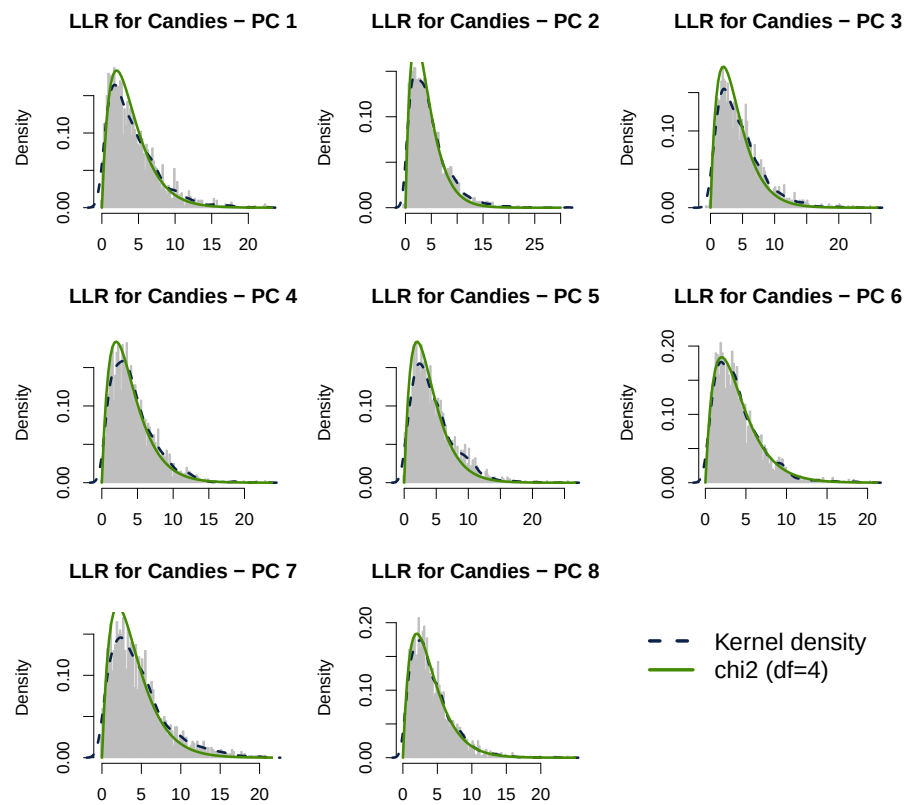


Figure 5.22 – Histogram of bootstrapped LLR for the candies effect per PC for the sensory dataset.

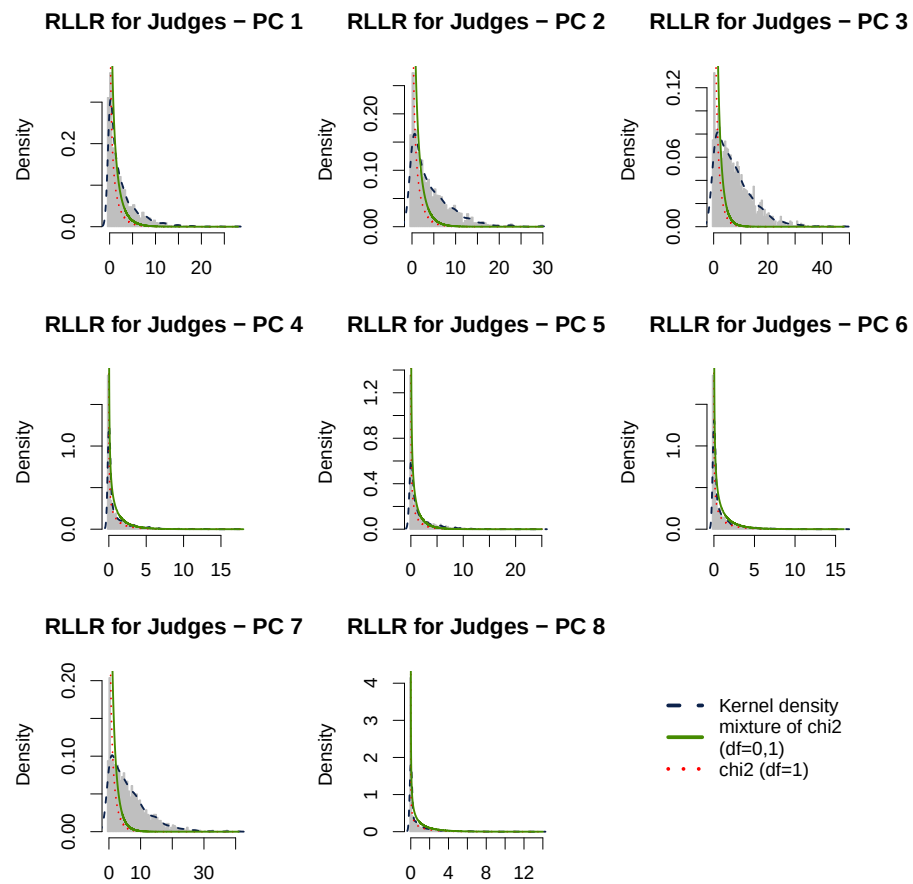


Figure 5.23 – Histogram of bootstrapped LLR for the assessors effect per PC for the sensory dataset.

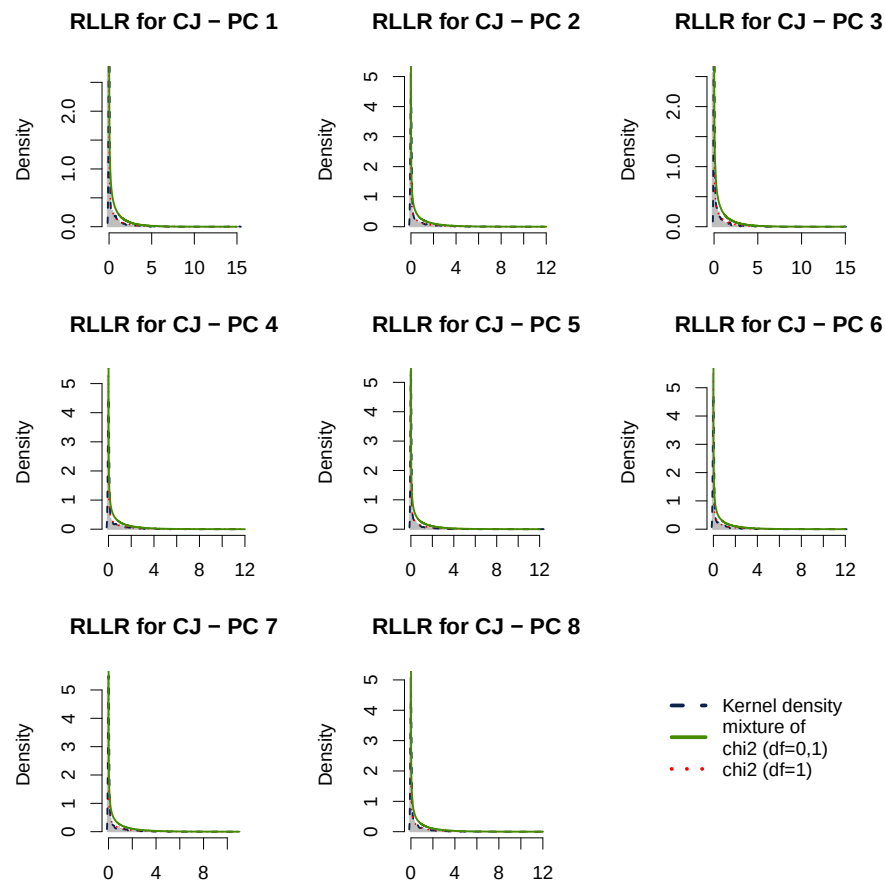


Figure 5.24 – Histogram of bootstrapped LLR for the interaction effect per PC for the sensory dataset.

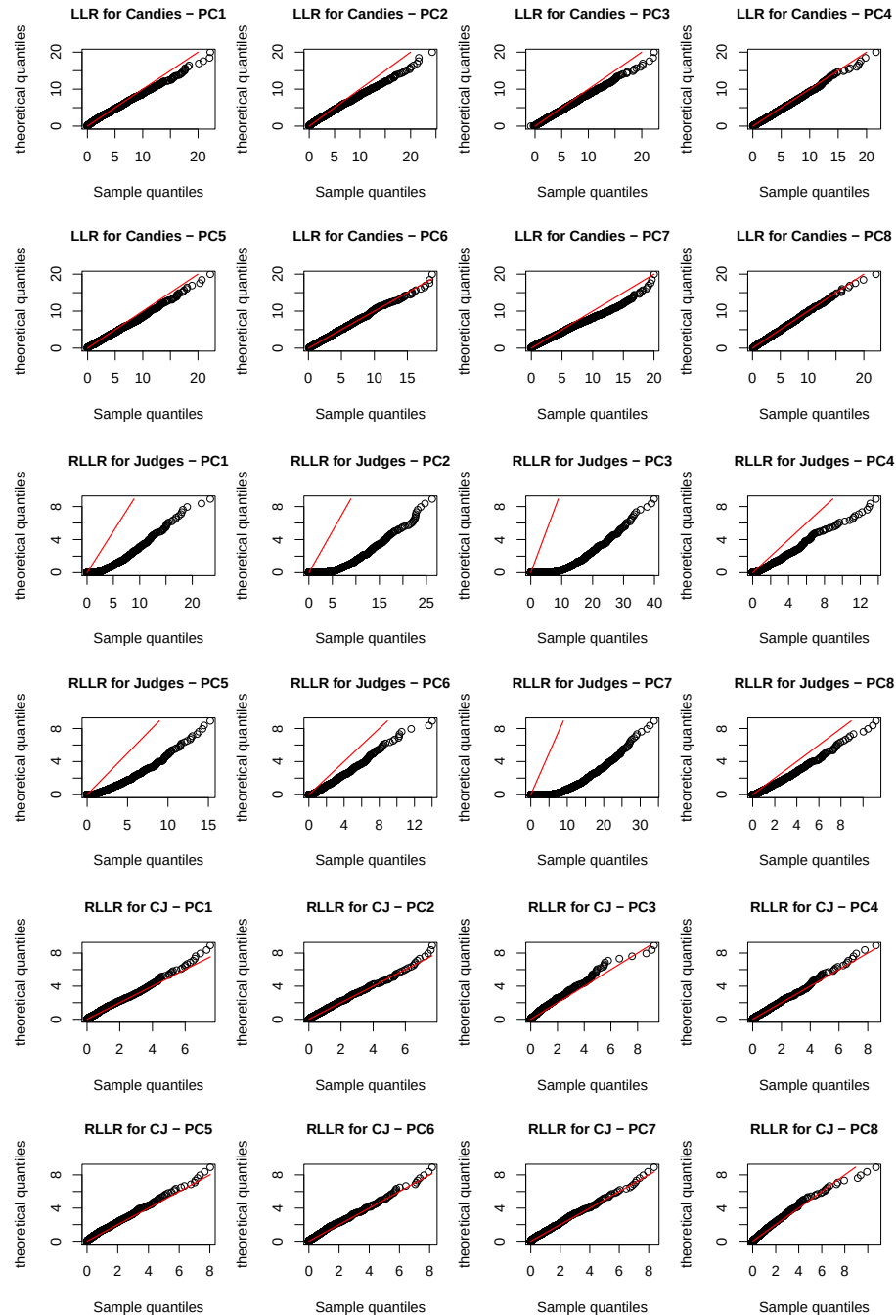


Figure 5.25 – True and theoretical quantiles for the bootstrapped RLLR for the candies, assessors and interaction effects per PC for the sensory dataset.

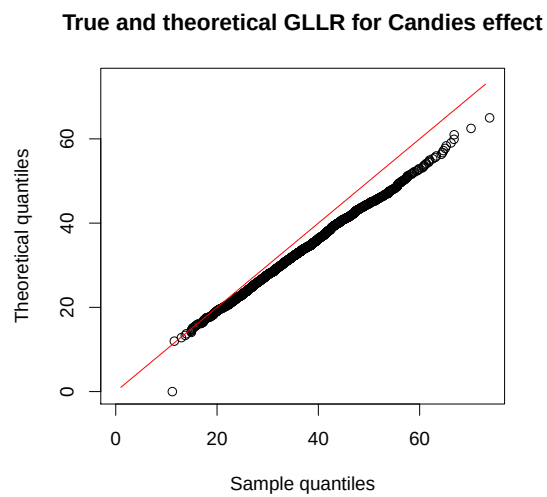


Figure 5.26 – True and theoretical quantiles for the bootstrapped GLLR for the candy effect.

5.D ANOVA $E(MS)$ and F-test table for the metabolomics dataset

Table 5.10 – ANOVA table and $E(MS)$ for the mixed ANOVA model linked to the metabolomics dataset. In this table, x represents the effect name (T, U, V or S) or the error (E). a, b, c and d , are the number of levels respectively of the time (2), tube (2), volunteer (12) and sampling effects (3) in a balanced setting, and F_{df_1, df_2} represents the 95% quantile of the F-distribution with df_1, df_2 degrees of freedom.

Effect	SS	df_x	MS_x	$E(MS_x)$	F-test
Time	SS_T	$a - 1$	SS_T/df_T	$\sigma_\epsilon^2 + \frac{nbcd \sum \alpha_i^2}{a-1}$	MS_T/MS_E
Tube	SS_U	$b - 1$	SS_U/df_U	$\sigma_\epsilon^2 + \frac{nacd \sum \alpha_i^2}{b-1}$	MS_U/MS_E
Volunteer	SS_V	$c - 1$	SS_V/df_V	$\sigma_\epsilon^2 + nbcd\sigma_V^2 + nab\sigma_S^2$	MS_V/MS_S
Sampling	SS_S	$c(d - 1)$	SS_S/df_S	$\sigma_\epsilon^2 + nab\sigma_S^2$	MS_S/MS_E
Error	SS_E	$abcd - a - b - cd + 2$	SS_E/df_E	-	-

CHAPTER 6

General conclusions

Metabolomics, as the other “omics”, generates a very large amount of correlated variables that requires a heavy pre-treatment, and for which classic statistical methods cannot be directly applied.

As for all the omics, the data (pre-)processing chain is long, complex, but also mainly technology- and data-driven. This is why it is important for statisticians performing their data analysis to deeply understand how the data were generated beforehand, in order to grasp all their specificities and extract meaningful information from it. In fact, they should take part of the whole study, from design construction, to molecular biomarkers identification. Additionally, and due to the whole processing chain complexity, collaboration is key in this context to ensure the quality of the realised studies.

This work is in line with the current and general trend in this field to develop strong bases of statistical and software tools in metabolomics, in order to provide practical solutions to non-statistician researchers in the most appropriate manner.

The purposes of this thesis were to publish a proven pre-processing software from raw FID to interpretable spectra, and investigate feature selection for biomarker discovery in simple and more advanced designs. For the latter, classic feature selection techniques for metabolomics data were compared. ASCA-like methods were applied and related new developments were suggested as well. Through these themes, cross-cutting issues arose such as the measurement of spectral quality and repeatability, and more generally the measurement and visualisation of the importance of mixed effects on the spectral response.

The developed methodologies have been successfully applied to experimental datasets, in order to prove their direct practical interest. Another important aspect of this work was to make available all the developed tools of this thesis. To that extent, all the R codes from this work are available as supplementary material, and for a part of them under the form of R packages. The datasets that were used are available as well for reuse.

More precisely, four contributions were made through this thesis. They can be generalised either to other metabolomics ^1H NMR data (pre-processing), or more generally to other high dimensional datasets based on simple or advanced designs:

- The writing and implementation of PepsNMR for the metabolomics ^1H NMR data pre-processing. An adequate pre-processing of such data is of primary importance to translate the data to an interpretable format, remove instrumental artefacts and other biological sources of variation that are of no interest. The already existing Matlab library has been converted to an R package that covers the entirety of the pre-processing pipeline, from raw data importation to their normalisation. Through its implementation on numerous databases, several improvements and new methodological aspects have been suggested. Currently, this R package is published on the Bioconductor repository and is part of the collaborative W4M web platform aimed at a full NMR and MS metabolomics data treatment pipeline. Compared to classic pre-processing based on Topspin and Amix, we were also able to demonstrate that this software had a better spectral repeatability recovery through the MIC criteria. Perspectives arising from this work are to create a design of experiment protocol to investigate more precisely the impact of each pre-processing step on the final data quality, and their potential interactions. Other quantitative quality criteria would also be beneficial to be developed, especially to ensure that the spectral profiles are truly biologically meaningful – *i.e.* they represent adequately the metabolites.
- In a second contribution, classic supervised chemometrics tools to find potential biomarkers in metabolomics data (t-test, PLS, OPLS, SPLS and LSOPLS), were investigated and compared. First, all the statistical models, algorithms and feature selection procedures were depicted. Then, the models were applied to a metabolomics semi-artificial urine database with known in advance true biomarkers locations, and under different conditions (sample size, level of noise and signal intensity). Care has been taken to avoid drawing conclusions from a particular sub-dataset and an overfit of the supervised multivariate models to the data, by training the classifiers in a double loop scheme with resampling and CV. Several evaluation criteria

have been proposed and measured, such as the true features accuracy and feature selection stability. The results pointed out that methods related to the t-test had a high amount of FP, and were very sensitive to lower sample sizes. On the other hand, methods derived from PLS and OPLS had each pros and cons, particularly with regards to the level of the signal-to-noise ratio. Nevertheless, PLS-vip with a decision threshold of 1 was the overall best performing strategy to recover all discriminant features. For their part, sparse classifiers have the potential to achieve “concise” feature selection. However, their MPs combinations should be more rigorously chosen than by conventional AUC-CV, depending *e.g.* on the size of the desired selected features or their feature stability, since they will heavily impact the selection sets size. The main perspectives related to this chapter are the combination of different feature selection methods to overlay their selection sets, the use of other machine learning algorithms that were developed in other research fields with high dimensional data or to use the bagging in the feature space principle for such correlated variables. The selection of optimal MPs for sparse embedded classifiers would also improve their performances.

- A third contribution has been to implement and compare ASCA⁺ to three other global approaches (PARAFASCA, AComDim and AMOPLS) to deal with advanced designs and wide and short multivariate responses. The algorithms were rewritten, sometimes from an unclear description in the literature, and an effort has been made to harmonise the notations and simplify the methodology, if possible. They were then applied to an experimental dataset with spiked metabolites and results of each method were compared to each other. The global techniques first decompose the response matrix with respect to the ASCA⁺ model before analysing the resulting effect matrices with one multivariate technique to interpret the model results.

It has been shown that the main advantages of these global approaches are their ability to evaluate the effect of all model terms on the responses under a common framework and with an automatic ranking of their importance. They also provide parsimonious analysis with several interesting tools to ease the interpretation of the components generated by the multivariate step. Subtle differences between these approaches were observed on the basis on an experimental dataset. PARAFASCA algorithm can be unstable and can induce effects remixing but it also offers a neat and efficient representation of the model terms, and especially the interactions. Saliences in AComDim provided additional information to recover the links between the model effects and common components, but it can also, in theory, be affected by effects remixing. AMOPLS results were close to the

ones of AComDim and it furthermore avoids potential effects remixing thanks to the orthogonal structure of the O-PLS response matrix. However, this approach may be very affected in the case of unbalanced designs. A transversal concern is the number of components' choice, especially for PARAFASCA, whose components profiles are all changing depending on this number.

- The last contribution went deeper into the modelling of advanced designs by combining ASCA⁺ and LMM to include random effects in the multiple response model. This is especially relevant in the case of repeatability/reproducibility or longitudinal studies. In this work, the adapted methodology has been applied to two experimental datasets and it has proven the interest to deal with such advanced design. While it preserves the visual representation advantages of ASCA, it leads to a correct inference of such advanced designs and models in order to compare the mixed sources of variability. Direct perspectives linked to the chapter are to either modify its multivariate step (e.g. use ICA instead of PCA) or its modelling part (include continuous effects to the model, derive necessary methodological aspects to deal with longitudinal studies). The scaling of such data in the ASCA framework is also an important aspect that was not investigated here.

Metabolomics is a recent technology- and data-led science that is likely to have a bright future given its multiple ways of development, from new applications to advances in the analytical technologies, together with new data analysis methods and bioinformatics tools.

To adequately meet the researchers' needs, methodological developments in the data pre-processing and analysis techniques are of vital importance, and much remains to be done in this omics field. As quoted in Buydens (2013a), "*chemometric theory must excel to survive the data tsunami to find adequate and logical solutions to complex data arising from the technology*". Such tools should also be somehow homogenised and able to be used in combination to expand their use so that researchers can keep the best parts in each of them.

In recent years, several topical issues in metabolomics are globally emerging, such as:

Spectral reproducibility/repeatability: The statistical perspective of spectral reproducibility/repeatability is still an open and complex issue of interest. This is especially true with the current fast moving technologies where the necessary critical distance to evaluate them is lacking. Once again, collaboration between experts from these multiple data sources is essential,

in addition to data standardisation, and methodological and informatics tools, to combine these data in an appropriate manner.

Data fusion: A metabolomics dataset in itself can be not sufficient enough to acquire a comprehensive view and better understanding of the biological processes occurring. Integrated approaches with different and complementary analytical platforms and/or data samples and/or multilevel integrations with other omics (proteomics, transcriptomics, etc.) can be integrated together to enrich data analysis. Several levels and techniques of integration are possible (*cf. e.g.* Boccard & Rudaz (2014) and Pinu *et al.* (2019) for more details) but this integrated multi-omics research is still currently a challenging task.

Longitudinal studies: Another aspect of metabolomics is the response evolution in time of a living organism of the metabolic disruption that is caused by an external stimuli. To answer such research question, time-related studies are especially relevant to be explored. In this kind of studies, dynamic data modelling is clearly dependent on the experimental design (number of time points, spacing in time, etc.), among other things (Smilde *et al.* , 2010). Future research should unleash methodological solutions from statistics toolbox to answer this recurrent request from researchers in metabolomics.

While omics data offer a new and exciting playground for statisticians and bioinformaticians to play with high dimensional data, new methods, algorithms and powerful softwares, these data nevertheless remain very complex to handle. In this context, collaboration between researchers remains a key priority.

Bibliography

- Abdi, H. 2010. Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdisciplinary Reviews: Computational Statistics*, **2**(1), 97–106.
- Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., & Saeys, Y. 2009. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, **26**(3), 392–398.
- Alonso, A., Marsal, S., & Julià, A. 2015. Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers in bioengineering and biotechnology*, **3**, 23.
- Amat, S., Dupuy, N., Kister, J., & Rutledge, D.N. 2010. Development of near infrared sensors: Detection of influential factors by the AComDim method. *Analytica Chimica Acta*, **675**(1), 16–23.
- Anderson, M., & ter Braak, C. 2003. Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation*, **73**(2), 85–113.
- Andersson, M. 2009. A comparison of nine PLS1 algorithms. *Journal of Chemometrics*, **23**(10), 518–529.
- Antti, H., Ebbels, T., Keun, H. C., Bollard, M. E., Beckonert, O., Lindon, J. C., Nicholson, J. K., & Holmes, E. 2004. Statistical experimental design and partial least squares regression analysis of biofluid metabonomic NMR and clinical chemistry data for screening of adverse drug effects. *Chemometrics and intelligent laboratory systems*, **73**(1), 139–149.

- Baek, S.-J., Park, A., Ahn, Y.-J., & Choo, J. 2015. Baseline correction using asymmetrically reweighted penalized least squares smoothing. *Analyst*, **140**, 250–257.
- Barker, M., & Rayens, W. 2003. Partial least squares for discrimination. *Journal of chemometrics*, **17**(3), 166–173.
- Bates, D., Mächler, M., Bolker, B., & Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, Articles*, **67**(1), 1–48.
- Bates, D. M., & DebRoy, S. 2004. Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*, **91**(1), 1 – 17. Special Issue on Semiparametric and Nonparametric Mixed Models.
- Benjamini, Y., & Yekutieli, D. 2001. The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, **29**(4), 1165–1188.
- Beyoğlu, D., & Idle, J. R. 2013. Metabolomics and its potential in drug development. *Biochemical pharmacology*, **85**(1), 12–20.
- Bloemberg, T. G., Gerretzen, J., Lunshof, A., Wehrens, R., & Buydens, L. M. C. 2013. Warping methods for spectroscopic and chromatographic signal alignment: A tutorial. *Analytica Chimica Acta*, **781**, 14 – 32.
- Boccard, J., & Rudaz, S. 2014. Harnessing the complexity of metabolomic data with chemometrics. *Journal of Chemometrics*, **28**(1), 1–9.
- Boccard, J., & Rudaz, S. 2016. Exploring Omics data from designed experiments using analysis of variance multiblock Orthogonal Partial Least Squares. *Analytica Chimica Acta*, **920**, 18 – 28.
- Boccard, J., & Rutledge, D. N. 2013. A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion. *Analytica Chimica Acta*, **769**(Mar.), 30–39.
- Boulesteix, A.-L., & Strimmer, K. 2006. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, **8**(1), 32–44.
- Bourne, M. 2015. *The Trapezoidal Rule*. online book, accessed 29/07/2015.
- Bro, R. 1997. PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, **38**(2), 149–171.
- Bro, R., & Kiers, H. A. L. 2003. A new efficient method for determining the number of components in PARAFAC models. *Journal of Chemometrics*, **17**(5), 274–286.

- Brown, H., & Prescott, R. 2006. *Applied mixed models in medicine*. John Wiley & Sons.
- Buydens, L. 2013a. Towards tsunami-resistant chemometrics. *Anal. Sci.*, **813**, 24–29.
- Buydens, L. M. C. 2013b. Towards tsunami-resistant chemometrics. *The analytical Scientist*, **813**, 24–30.
- Bylesjö, M., Rantalainen, M., Cloarec, O., Nicholson, J. K., Holmes, E., & Trygg, J. 2006. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics*, **20**(8-10), 341–351.
- Cambiaghi, A., Ferrario, M., & Masseroli, M. 2016. Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. *Briefings in Bioinformatics*, **18**(3), 498–510.
- Carroll, J. D., & Chang, J.-J. 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, **35**(3), 283–319.
- Christin, C., Hoefsloot, H. C. J., Smilde, A. K., Hoekman, B., Suits, F., Bischoff, R., & Horvatovich, P. 2013. A Critical Assessment of Feature Selection Methods for Biomarker Discovery in Clinical Proteomics. *Molecular & Cellular Proteomics*, **12**(1), 263–276.
- Chun, H., & Keleş, S. 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(1), 3–25.
- Claridge, Timothy D.W. 2016. Chapter 3 - Practical Aspects of High-Resolution {NMR}. Pages 61 – 132 of: Claridge, T. D.W. (ed), *High-Resolution {NMR} Techniques in Organic Chemistry (Third Edition)*, third edition edn. Boston: Elsevier.
- Craig, A., Cloarec, O., Holmes, E., Nicholson, J. K., & Lindon, J. C. 2006. Scaling and Normalization Effects in NMR Spectroscopic Metabonomic Data Sets. *Analytical Chemistry*, **78**(7), 2262–2267. PMID: 16579606.
- Crainiceanu, C. M., & Ruppert, D. 2004. Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**(1), 165–185.
- d. B. Harrington, P., Vieira, N. E., Espinoza, J., Nien, J. K., Romero, R., & Yergey, A. L. 2005. Analysis of variance-principal component analysis: A soft

- tool for proteomic discovery. *Analytica Chimica Acta*, **544**(1), 118 – 127. Papers Presented at the 9th International Conference on Chemometrics in Analytical Chemistry.
- Davison, A. C., Hinkley, D. V., *et al.* . 1997. *Bootstrap methods and their application*. Vol. 1. Cambridge university press.
- de Jong, S. 1993. SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and intelligent laboratory systems*, **18**(3), 251–263.
- de Noord, O. E., & Theobald, E. H. 2006. Multilevel component analysis and multilevel PLS of chemical process data. *Journal of Chemometrics*, **19**(5-7), 301–307.
- de Rooi, J.J., & Eilers, P. H.C. 2012. Mixture models for baseline estimation. *Chemometrics and Intelligent Laboratory Systems*, **117**, 56 – 60. Special Issue Section: Selected Papers from the 1st African-European Conference on Chemometrics, Rabat, Morocco, September 2010 Special Issue Section: Preprocessing methods Special Issue Section: Spectroscopic imaging.
- Determan Jr., C. E. 2015. Optimal algorithm for metabolomics classification and feature selection varies by dataset. *International journal of biology*, **7**(1), 100.
- Dieterle, F., Ross, A., Schlotterbeck, G., & Senn, H. 2006. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics. *Analytical Chemistry*, **78**(13), 4281–4290. PMID: 16808434.
- Domingos, P. M. 2012. A few useful things to know about machine learning. *Commun. acm*, **55**(10), 78–87.
- Drikvandi, R., Verbeke, G., Khodadadi, A., & Partovi Nia, V. 2012. Testing multiple variance components in linear mixed-effects models. *Biostatistics*, **14**(1), 144–159.
- Dumas, M.-E., Maibaum, E. C., Teague, C., Ueshima, H., Zhou, B., Lindon, J. C., Nicholson, J. K., Stamler, J., Elliott, P., Chan, Q., & Holmes, E. 2006. Assessment of Analytical Reproducibility of 1H NMR Spectroscopy Based Metabonomics for Large-Scale Epidemiological Research: the INTERMAP Study. *Analytical Chemistry*, **78**(7), 2199–2208. PMID: 16579598.
- Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., *et al.* . 2004. Least angle regression. *The Annals of statistics*, **32**(2), 407–499.
- Eilers, P. H. C. 2003. A Perfect Smoother. *Analytical Chemistry*, **75**(14), 3631–3636. PMID: 14570219.

- Eilers, P. H. C. 2004. Parametric Time Warping. *Analytical Chemistry*, **76**(2), 404–411. PMID: 14719890.
- Eilers, P. H. C. 2018. The truth about the effective dimension. *Statistica Neerlandica*, **72**(3), 201–209.
- Eilers, P. H. C., & Boelens, H. F. M. 2005. *Baseline correction with asymmetric least squares smoothing*. Medical Centre Report (unpublished results), Leiden University.
- Eilers, P. H. C., & Marx, B. D. 1996. Flexible Smoothing with B-splines and Penalties. *Statistical Science*, **11**(2), 89–102.
- El Ghaziri, A., Qannari, E. M., Moyon, T., & Alexandre-Gouabau, M.-C. 2015. AoV-PLS: a new method for the analysis of multivariate data depending on several factors. *Electronic Journal of Applied Statistical Analysis*, **8**(2), 214–235.
- El Ghaziri, A., Cariou, V., Rutledge, D. N., & Qannari, E. M. 2016. *Extension of ComDim for the analysis of (K+ 1) datasets; Application in Sensometrics*.
- Emwas, A.-H., Saccenti, E., Gao, X., McKay, R. T., dos Santos, V. A. P. M., Roy, R., & Wishart, D. S. 2018. Recommended strategies for spectral processing and post-processing of 1D 1H-NMR data of biofluids with a particular focus on urine. *Metabolomics*, **14**(3), 31.
- Emwas, A.-H., Roy, R., McKay, R. T., Tenori, L., Saccenti, E., Gowda, G. A. N., Raftery, D., Alahmari, F., Jaremko, L., Jaremko, M., & Wishart, D. S. 2019. NMR Spectroscopy for Metabolomics Research. *Metabolites*, **9**(7).
- Emwas, A.-H. M., Salek, R. M., Griffin, J. L., & Merzaban, J. 2013. NMR-based metabolomics in human disease diagnosis: applications, limitations, and recommendations. *Metabolomics*, **9**(5), 1048–1072.
- Engel, J., Gerretzen, J., Szymańska, E., Jansen, J. J., Downey, G., Blanchet, L., & Buydens, L. M.C. 2013. Breaking with trends in pre-processing? *TrAC Trends in Analytical Chemistry*, **50**, 96 – 106.
- Euceda, L. R., Giskeødegård, G. F., & Bathen, T. F. 2015. Preprocessing of NMR metabolomics data. *Scandinavian Journal of Clinical and Laboratory Investigation*, **75**(3), 193–203. PMID: 25738209.
- Faber, N. M., Bro, R., & Hopke, P. K. 2003. Recent developments in CANDECOMP/PARAFAC algorithms: a critical review. *Chemometrics and Intelligent Laboratory Systems*, **65**(1), 119–137.

- Farrés, Mi., Platikanov, S., Tsakovski, S., & Tauler, R. 2015. Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. *Journal of Chemometrics*, **29**(10), 528–536.
- Favier, G., & de Almeida, A. L. F. 2014. Overview of constrained PARAFAC models. *EURASIP Journal on Advances in Signal Processing*, **2014**(1), 142.
- Fawcett, T. 2006. An introduction to ROC analysis. *Pattern recognition letters*, **27**(8), 861–874.
- Féraud, B., Govaerts, B., Verleysen, M., & de Tullio, P. 2015. Statistical treatment of 2D NMR COSY spectra in metabolomics: data preparation, clustering-based evaluation of the Metabolomic Informative Content and comparison with 1H-NMR. *Metabolomics*, **11**(6), 1756–1768.
- Féraud, B., Munaut, C., Martin, M., Verleysen, M., & Govaerts, B. 2017. Combining strong sparsity and competitive predictive power with the L-sOPLS approach for biomarker discovery in metabolomics. *Metabolomics*, **13**(11), 130.
- Frasso, G., & Eilers, P. H. C. 2015. L- and V-curves for optimal smoothing. *Statistical Modelling*, **15**(1), 91–111.
- Friedman, J., Hastie, T., & Tibshirani, R. 2010. *A note on the group lasso and a sparse group lasso*.
- Gebregiorgis, T., & Powers, R. 2012. Application of NMR Metabolomics to Search for Human Disease Biomarkers. *Combinatorial Chemistry & High Throughput Screening*, **15**(8), 595–610.
- Geladi, P. 1988. Notes on the history and nature of partial least squares (PLS) modelling. *Journal of Chemometrics*, **2**(4), 231–246.
- Geladi, P., & Kowalski, B. R. 1986. Partial least-squares regression: a tutorial. *Analytica chimica acta*, **185**, 1–17.
- Gelman, A. 2005. Analysis of Variance: Why It Is More Important than Ever. *The Annals of Statistics*, **33**(1), 1–31.
- Giampaoli, V., & Singer, J. M. 2009. Likelihood ratio tests for variance components in linear mixed models. *Journal of Statistical Planning and Inference*, **139**(4), 1435 – 1448.
- Goodacre, R., Broadhurst, D., Smilde, A. K., Kristal, B. S., Baker, J. D., Beger, R., Bessant, C., Connor, S., Capuani, G., Craig, A., Ebbels, T., Kell, D. B., Manetti, C., Newton, J., Paternostro, G., Somorjai, R., Sjöström, M., Trygg, J.,

- & Wulfert, F. 2007. Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics*, **3**(3), 231–241.
- Govaerts, B., Martin, M., & Francq, B. 2019. *The family of linear regression, ANOVA, general linear and linear mixed models for the chemist: a survival guide*. Forthcoming Discussion Paper. Institute of Statistics, Biostatistics and Actuarial Sciences, UCLouvain, Belgium.
- Grissa, D., Pétéra, M., Brandolini, M., Napoli, A., Comte, B., & Pujos-Guillot, E. 2016. Feature Selection Methods for Early Predictive Biomarker Discovery Using Untargeted Metabolomic Data. *Frontiers in Molecular Biosciences*, **3**, 30.
- Gromski, P. S., Muhamadali, H., Ellis, D. I., Xu, Y., Correa, E., Turner, M. L., & Goodacre, R. 2015. A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta*, **879**, 10 – 23.
- Guenneq, A. Le, & Giraudeau, P. and Caldarelli, S. 2014. Evaluation of Fast 2D NMR for Metabolomics. *Analytical Chemistry*, **86**(12), 5946–5954. PMID: 24856256.
- Guisset, S., Martin, M., & Govaerts, B. 2019. Comparison of PARAFASCA, AComDim, and AMOPLS approaches in the multivariate GLM modelling of multi-factorial designs. *Chemometrics and Intelligent Laboratory Systems*, **184**, 44 – 63.
- Guillon, Y., Tremblay-Franco, M., Corguillé, G. Le, Martin, J.-F., M. Pétéra, Roger-Mele, P., Delabrière, A., Goulitquer, S., Monsoor, M., Duperier, C., Canlet, C., Servien, R., Tardivel, P., Caron, C., Giacomoni, F., & Thévenot, E. A. 2017. Create, run, share, publish, and reference your LC-MS, FIA-MS, GC-MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics. *The International Journal of Biochemistry & Cell Biology*, **93**, 89 – 101.
- Gumedze, F. N., & Dunne, T. T. 2011. Parameter estimation and inference in the linear mixed model. *Linear Algebra and its Applications*, **435**(8), 1920 – 1944.
- Guyon, I., & Elisseeff, A. 2003. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.*, **3**(Mar.), 1157–1182.
- Halekoh, U., Højsgaard, S., et al. . 2014. A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models—the R package pbkrtest. *Journal of Statistical Software*, **59**(9), 1–30.

- Härdle, W., & Simar, L. 2007. *Applied multivariate statistical analysis*. Vol. 22007. Springer.
- Harshman, R. A. 1970. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, **16**(1), 84.
- Harshman, R. A., & Lundy, M. E. 1994. PARAFAC: Parallel factor analysis. *Computational Statistics & Data Analysis*, **18**(1), 39–72.
- Hitchcock, F. L. 1927. The Expression of a Tensor or a Polyadic as a Sum of Products. *Journal of Mathematics and Physics*, **6**(1-4), 164–189.
- Horochowska, M., Stanimirova, I., & Czarnik-Matusiewicz, B. 2016. Studying the influence of enflurane, isoflurane, and sevoflurane on the DPPC lipid bilayer using the analysis of variance and parallel factor analysis. *Chemometrics and Intelligent Laboratory Systems*, **153**(Apr.), 146–152.
- Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Ole's, A. K., Pag'es, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D., Waldron, L., & Morgan, M. 2015. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods*, **12**(2), 115–121.
- Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. 2005. ROBPCA: a new approach to robust principal component analysis. *Technometrics*, **47**(1), 64–79.
- Hyvärinen, A., & Oja, E. 2000. Independent component analysis: algorithms and applications. *Neural networks*, **13**(4), 411–430.
- Indahl, U. G., Liland, K. H., & Næs, T. 2009. Canonical partial least squares - a unified PLS approach to classification and regression problems. *Journal of Chemometrics*, **23**(9), 495–504.
- Ipsen, I. C. F., & Meyer, C. D. 1998. The Idea Behind Krylov Methods. *The American Mathematical Monthly*, **105**(10), 889–899.
- Jackson, J. E. 2005. *A user's guide to principal components*. Vol. 587. John Wiley & Sons.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. 2013. *An introduction to statistical learning*. Vol. 112. Springer.
- Jansen, J. J., Hoefsloot, H. C. J., van der Greef, J., Timmerman, M. E., Westerhuis, J. A., & Smilde, A. K. 2005a. ASCA: analysis of multivariate data obtained from an experimental design. *Journal of Chemometrics*, **19**(9), 469–481.

- Jansen, J. J., Hoefsloot, H. C.J., van der Greef, J., Timmerman, M. E., & Smilde, A. K. 2005b. Multilevel component analysis of time-resolved metabolic fingerprinting data. *Analytica Chimica Acta*, **530**(2), 173 – 183.
- Jansen, J. J., Bro, R., Hoefsloot, H. C. J., van den Berg, F. W. J., Westerhuis, J. A., & Smilde, A. K. 2008. PARAFASCA: ASCA combined with PARAFAC for the analysis of metabolic fingerprinting data. *Journal of Chemometrics*, **22**(2), 114–121.
- Jeanmougin, M., de Reynies, A., Marisa, L., Paccard, C., Nuel, G., & Guedj, M. 2010. Should We Abandon the t-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies. *PLOS ONE*, **5**(9), 1–9.
- Johnson, C. H., Ivanisevic, J., & Siuzdak, G. 2016. Metabolomics: beyond biomarkers and towards mechanisms. *Nature reviews Molecular cell biology*, **17**(7), 451–459.
- Johnson, R. A., Wichern, D. W., et al. . 2002. *Applied multivariate statistical analysis*. Vol. 5. Prentice hall Upper Saddle River, NJ.
- Jolliffe, I. 2005. *Principal component analysis*. Wiley Online Library.
- Jouan-Rimbaud Bouveresse, D., Climaco Pinto, R., Schmidtke, L.M., Locquet, N., & Rutledge, D.N. 2011. Identification of significant factors by an extension of ANOVA-PCA based on multi-block analysis. *Chemometrics and Intelligent Laboratory Systems*, **106**(2), 173–182.
- Keeler., J. 2002. 4 Fourier transformation and data processing. *Pages 48–65 of: Understanding NMR Spectroscopy*. John Wiley & Sons.
- Kell, D. B. 2007. Metabolomic biomarkers: search, discovery and validation. *Expert Review of Molecular Diagnostics*, **7**(4), 329–333.
- Keogh, E., & Mueen, A. 2017. *Curse of Dimensionality*. Boston, MA: Springer US. Pages 314–315.
- Khalheim, O. M. 1985. Scaling of analytical data. *Analytica chimica acta*, **177**, 71–79.
- Korifi, R., Amat, S., Rebufa, C., Labed, V., Rutledge, D.N., & Dupuy, N. 2015. AComDim as a multivariate tool to analyse experimental design application to gamma-irradiated and leached ion exchange resins. *Chemometrics and Intelligent Laboratory Systems*, **141**(Feb.), 12–23.

- Korifi, R., Plard, J., Le Dreau, Y., Rebufa, C., Rutledge, D.N., & Dupuy, N. 2016. Highlighting metabolic indicators of olive oil during storage by the AComDim method. *Food Chemistry*, **203**(July), 104–116.
- Kosmides, A. K., Kamisoglu, K., Calvano, S. E., Corbett, S. A., & Androulakis, I. P. 2013. Metabolomic Fingerprinting: Challenges and Opportunities. *Critical Reviews in Biomedical Engineering*, **41**(3), 205–221.
- Krämer, N. 2007. An overview on the shrinkage properties of partial least squares regression. *Computational Statistics*, **22**(2), 249–273.
- Kuyuk, S., & Ercan, I. 2017. Commonly used statistical methods for detecting differential gene expression in microarray experiments. *Biostatistics Epidemiol Int J*, **1**(1), 00001.
- Lê Cao, K.-A., Rossouw, D., Robert-Granié, C., & Besse, P. 2008. A sparse PLS for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, **7**(1).
- Lenhardt, L., & Dramićanin, M. D. 2016. PARAFAC: A tool for the analysis of phosphor mixture luminescence. *Journal of Luminescence*, **170**(Feb.), 136–140.
- Lenz, E. M., & Wilson, I. D. 2007. Analytical Strategies in Metabonomics. *Journal of Proteome Research*, **6**(2), 443–458. PMID: 17269702.
- Liland, K. H. 2011. Multivariate methods in metabolomics — from pre-processing to dimension reduction and statistical analysis. *TrAC Trends in Analytical Chemistry*, **30**(6), 827 – 841.
- Liland, K. H., Almøy, T., & Mevik, B.-H. 2010. Optimal Choice of Baseline Correction for Multivariate Calibration of Spectra. *Applied Spectroscopy*, **64**(9), 1007–1016. PMID: 20828437.
- Liland, K. H., Smilde, A., Marini, F., & Næs, T. 2018. Confidence ellipsoids for ASCA models based on multivariate regression theory. *Journal of Chemometrics*, **32**(5), e2990.
- Lindgren, F., Geladi, P., & Wold, S. 1993. The kernel algorithm for PLS. *Journal of Chemometrics*, **7**(1), 45–59.
- Lindon, J. C., & Nicholson, J. K. 2008. Analytical technologies for metabonomics and metabolomics, and multi-omic information recovery. *TrAC Trends in Analytical Chemistry*, **27**(3), 194–204.

- Lindon, J. C., Nicholson, J. K., Holmes, E., Antti, H., Bollard, M. E., Keun, H., Beckonert, O., Ebbels, T. M., Reily, M. D., Robertson, D., *et al.* . 2003. 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.
- Luciano, G., & Næs, T. 2009. Interpreting sensory data by combining principal component analysis and analysis of variance. *Food Quality and Preference*, **20**(3), 167 – 175.
- Maher, A. D, Zirah, S. F. M., Holmes, E., & Nicholson, J. K. 2007. Experimental and Analytical Variation in Human Urine in ¹H NMR Spectroscopy-Based Metabolic Phenotyping Studies. *Analytical Chemistry*, **79**(14), 5204–5211. PMID: 17555297.
- Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., & Ferrari, R. 2016. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in Bioinformatics*, **19**(2), 286–302.
- Mardia, K., Bibby, J. M., & Kent, J. 1979. *Multivariate analysis*. Academic Press, London.
- Marini, F., de Beer, D., Joubert, E., & Walczak, B. 2015. Analysis of variance of designed chromatographic data sets: The analysis of variance-target projection approach. *Journal of Chromatography A*, **1405**, 94 – 102.
- Marion, R. 2016. *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, Université catholique de Louvain.
- Martens, H., Høy, M., Westad, F., Folkenberg, D., & Martens, M. 2001. Analysis of designed experiments by stabilised PLS Regression and jack-knifing. *Chemometrics and Intelligent Laboratory Systems*, **58**(2), 151 – 170. PLS Methods.
- Martin, M., & Govaerts, B. 2019a. *Feature Selection in metabolomics with PLS-derived methods*. Discussion Paper DP 2019/20. Institute of Statistics, Biostatistics and Actuarial Sciences, UCLouvain, Belgium.
- Martin, M., & Govaerts, B. 2019b. *LiMM-PCA : combining ASCA⁺ and linear mixed models to analyse high dimensional designed data*. Discussion Paper DP 2019/21. Institute of Statistics, Biostatistics and Actuarial Sciences, UCLouvain, Belgium.

- Martin, M., Legat, B., Leenders, J., Vanwinsberghe, J., Rousseau, R., Boulanger, B., Eilers, P. H.C., Tullio, P. De, & Govaerts, B. 2018. PepsNMR for ^1H NMR metabolomic data pre-processing. *Analytica Chimica Acta*, **1019**, 1 – 13.
- Mazerolles, G., Devaux, M. F., Dufour, E., Qannari, E. M., & Courcoux, P. 2002. Chemometric methods for the coupling of spectroscopic techniques and for the extraction of the relevant information contained in the spectral data tables. *Chemometrics and Intelligent Laboratory Systems*, **63**(1), 57–68.
- Mazerolles, G., Hanafi, M., Dufour, E., Bertrand, D., & Qannari, E.M. 2006. Common components and specific weights analysis: A chemometric method for dealing with complexity of food products. *Chemometrics and Intelligent Laboratory Systems*, **81**(1), 41–49.
- McCulloch, C. E., & Searle, S. R. 2001. *Generalized, Linear and Mixed Models*. Wiley Series in Probability and Statistics. John Wiley & Sons.
- Mehmood, T., & Ahmed, B. 2016. The diversity in the applications of partial least squares: an overview. *Journal of Chemometrics*, **30**(1), 4–17.
- Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. 2012. A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, **118**, 62 – 69.
- Meinshausen, N., & Bühlmann, P. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(4), 417–473.
- Meyer, N., Maumy-Bertrand, M., & Bertrand, F. 2010. Comparaison de variantes de régressions logistiques PLS et de régression PLS sur variables qualitatives: application aux données d'allélotypage. *Journal de la Société Française de Statistique*, **151**(2), 1–18.
- Næs, T., & Langsrud, Ø. 1998. Fixed or random assessors in sensory profiling? *Food Quality and Preference*, **9**(3), 145 – 152. Sensometric Workshop.
- Naik, P., & Tsai, C.-L. 2000. Partial Least Squares Estimator for Single-Index Models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **62**(4), 763–771.
- Nakagawa, S., & Schielzeth, H. 2013. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, **4**(2), 133–142.
- Neter, J., Kutner, M.I H., & Nachtsheim, C. J. 1996. *Applied linear statistical models*. 4th ed edn. The Irwin series in production operations management. Chicago: Irwin.

- Nicholson, Jeremy K., Holmes, Elaine, & Lindon, John C. 2007. Chapter 1 - Metabonomics and Metabolomics Techniques and Their Applications in Mammalian Systems. *Pages 1 – 33 of: Lindon, J. C., Nicholson, J. K., & Holmes, E. (eds), The Handbook of Metabonomics and Metabolomics.* Amsterdam: Elsevier Science B.V.
- Nogueira, S., Sechidis, K., & Brown, G. 2017. On the Stability of Feature Selection Algorithms. *Journal of Machine Learning Research*, **18**, 174–1.
- Parker, L. A., Chilet-Rosell, E., Hernández-Aguado, I., Pastor-Valero, M., Gea, S., & Lumbreras, B. 2018. Diagnostic Biomarkers: Are We Moving from Discovery to Clinical Application? *Clinical Chemistry*, **64**(11), 1657–1667.
- Parsons, H. M., Ekman, D. R., Collette, T. W., & Viant, M. R. 2009. Spectral relative standard deviation: a practical benchmark in metabolomics. *Analyst*, **134**, 478–485.
- Patti, G. J., Yanes, O., & Siuzdak, G. 2012. Innovation: Metabolomics: the apogee of the omics trilogy. *Nature reviews Molecular cell biology*, **13**(4), 263.
- Pinheiro, J. C., & Bates, D. M. 2000. Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*, 3–56.
- Pinu, F. R., Beale, D. J., Paten, A. M., Kouremenos, K., Swarup, S., Schirra, H. J., & Wishart, D. 2019. Systems Biology and Multi-Omics Integration: Viewpoints from the Metabolomics Research Community. *Metabolites*, **9**(4), 76.
- Poirion, O. B., Zhu, X., Ching, T., & Garmire, L. 2016. Single-Cell Transcriptomics Bioinformatics and Computational Challenges. *Frontiers in Genetics*, **7**, 163.
- Poste, G. 2011. Bring on the biomarkers. *Nature*, **469**(7329), 156.
- Qannari, E. M., Wakeling, I., & MacFie, H. J. H. 1995. A hierarchy of models for analysing sensory data. *Food Quality and Preference*, **6**(4), 309–314.
- Qannari, E. M., Wakeling, I., Courcoux, P., & MacFie, H. J. H. 2000. Defining the underlying sensory dimensions. *Food Quality and Preference*, **11**(1-2), 151–154.
- Rantalainen, M., Bylesjo, M., Cloarec, O., Nicholson, J. K., Holmes, E., & Trygg, J. 2007. Kernel-based orthogonal projections to latent structures (K-OPLS). *Journal of Chemometrics*, **21**(7-9), 376–385.
- Ren, S., Hinzman, A. A., Kang, E. L., Szczesniak, R. D., & Lu, L. J. 2015a. Computational and statistical analysis of metabolomics data. *Metabolomics*, **11**(6), 1492–1513.

- Ren, S., Hinzman, A. A., Kang, E. L., Szczesniak, R. D., & Lu, L. J. 2015b. Computational and statistical analysis of metabolomics data. *Metabolomics*, **11**(6), 1492–1513.
- Rinaudo, P., Boudah, S., Junot, C., & Thévenot, E. A. 2016. biosigner: A New Method for the Discovery of Significant Molecular Signatures from Omics Data. *Frontiers in Molecular Biosciences*, **3**, 26.
- Rohart, F., Gautier, B., Singh, A., & Lê Cao, K.-A. 2017. mixOmics: an R package for 'omics feature selection and multiple data integration. *bioRxiv*.
- Romano, R., Næs, T., & Brockhoff, P. B. 2015. Combining analysis of variance and three-way factor analysis methods for studying additive and multiplicative effects in sensory panel data. *Journal of Chemometrics*, **29**(1), 29–37. CEM-14-0017.R1.
- Rosipal, R., & Krämer, N. 2006. Overview and Recent Advances in Partial Least Squares. *Pages 34–51 of: Saunders, C., Grobelnik, M., Gunn, S., & Shawe-Taylor, J. (eds), Subspace, Latent Structure and Feature Selection*. Berlin, Heidelberg: Springer Berlin Heidelberg.
- Rosipal, R., & Trejo, L. J. 2001. Kernel partial least squares regression in reproducing kernel hilbert space. *Journal of machine learning research*, **2**(Dec), 97–123.
- Rousseau, R. 2011. *Statistical contribution to the analysis of metabonomics data in ¹H NMR spectroscopy*. Ph.D. thesis, Institute of Statistics, Biostatistics and Actuarial Sciences, Université catholique de Louvain, Belgium.
- Rousseau, R., Govaerts, B., Verleysen, M., & Boulanger, B. 2008. Comparison of some chemometric tools for metabonomics biomarker identification. *Chemometrics and Intelligent Laboratory Systems*, **91**(1), 54 – 66. Selected papers presented at the Chemometrics Congress “ CHIMIOMETRIE 2006” Paris, France, 30 November - 1 December 2006.
- Santos Nobre, J., & da Motta Singer, J. 2007. Residual Analysis for Linear Mixed Models. *Biometrical Journal*, **49**(6), 863–875.
- Savorani, F., Tomasi, G., & Engelsen, S. B. 2010. icoshift: A versatile tool for the rapid alignment of 1D NMR spectra. *Journal of Magnetic Resonance*, **202**(2), 190 – 202.
- Schorn, C., Taylor, B. F., & Taylor, B. J. 2004. *NMR-spectroscopy: Data acquisition*. John Wiley & Sons.

- Searle, S. R., Casella, G., & McCulloch, C. 1992. *Variance components: Wiley series in probability and mathematical statistics*. John Wiley & Sons, New York.
- Self, S. G., & Liang, K.-Y. 1987. Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *Journal of the American Statistical Association*, **82**(398), 605–610.
- Shi, L., Westerhuis, J. A., Rosén, J., Landberg, R., & Brunius, C. 2018. Variable selection and validation in multivariate modelling. *Bioinformatics*, **35**(6), 972–980.
- Siebert, W. McC. 1986. *Circuits, signals, and systems*. Vol. 2. MIT press.
- Simó, C., Cifuentes, A., & García-Cañas, V. 2014. *Fundamentals of advanced omics technologies: from genes to metabolites*. Vol. 63. Newnes.
- Sinha, S. K. 2009. Bootstrap tests for variance components in generalized linear mixed models. *Canadian Journal of Statistics*, **37**(2), 219–234.
- Sjöström, M., Wold, S., Lindberg, W., Persson, J.-Å., & Martens, H. 1983. A multivariate calibration problem in analytical chemistry solved by partial least-squares models in latent variables. *Analytica Chimica Acta*, **150**, 61 – 70.
- Smilde, A. K., Bro, R., & Geladi, P. 2004. *Multi-way analysis with applications in the chemical sciences*. Chichester, West Sussex, England ; Hoboken, NJ: J. Wiley.
- Smilde, A. K., Jansen, J. J., Hoefsloot, H. C. J., Lamers, R.-J. A. N., van der Greef, J., & Timmerman, M. E. 2005. ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. *Bioinformatics*, **21**(13), 3043–3048.
- Smilde, A. K., Westerhuis, J. A., Hoefsloot, H. C. J., Bijlsma, S., Rubingh, C. M., Vis, D. J., Jellema, R. H., Pijl, H., Roelfsema, F., & van der Greef, J. 2010. Dynamic metabolomic data analysis: a tutorial review. *Metabolomics*, **6**(1), 3–17.
- Smolinska, A., Blanchet, L., Buydens, L. M.C., & Wijmenga, S. S. 2012. NMR and pattern recognition methods in metabolomics: From data acquisition to biomarker discovery: A review. *Analytica Chimica Acta*, **750**, 82 – 97. 750th Anniversary Volume.
- Snijders, T. A. B., & Bosker, R. J. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. SAGE Publications.

- Spruyt, V. 2014. *The Curse of Dimensionality in classification*. <https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification>. Accessed: 2019-08-17.
- Stanimirova, I., Michalik, K., Drzazga, Z., Trzeciak, H., Wentzell, P. D., & Walczak, B. 2011. Interpretation of analysis of variance models using principal component analysis to assess the effect of a maternal anticancer treatment on the mineralization of rat bones. *Analytica Chimica Acta*, **689**(1), 1 – 7.
- Stram, D. O., & Lee, J. W. 1994. Variance Components Testing in the Longitudinal Mixed Effects Model. *Biometrics*, **50**(4), 1171–1177.
- Svensson, O., Kourti, T., & MacGregor, J. F. 2002. An investigation of orthogonal signal correction algorithms and their characteristics. *Journal of Chemometrics*, **16**(4), 176–188.
- T. K. Ho. 1998. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**(8), 832–844.
- Tang, K. W. A., Toh, Q. C., & Teo, B. W. 2015. Normalisation of urinary biomarkers to creatinine for clinical practice and research—when and why. *Singapore medical journal*, **56**(1), 7.
- Teahan, O., Gamble, S., Holmes, E., Waxman, J., Nicholson, J. K., Bevan, C., & Keun, H. C. 2006. Impact of Analytical Bias in Metabonomic Studies of Human Blood Serum and Plasma. *Analytical Chemistry*, **78**(13), 4307–4318. PMID: 16808437.
- ter Braak, C. J. F., & de Jong, S. 1998. The objective function of partial least squares regression. *Journal of Chemometrics*, **12**(1), 41–54.
- Thiel, M. 2015. *Utilisation des methodes multivariees ASCA et APCA dans l'analyse des composantes de variance en metabolomique*. Memoire de Master, Université catholique de Louvain, Louvain-la-Neuve, Belgique.
- Thiel, M., Féraud, B., & Govaerts, B. 2017. ASCA+ and APCA+: Extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs. *Journal of Chemometrics*, **31**(6), e2895–n/a. e2895 cem.2895.
- Tibshirani, R. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.
- Timmerman, M. E. 2006. Multilevel component analysis. *British Journal of Mathematical and Statistical Psychology*, **59**(2), 301–320.

- Timmerman, M. E., Hoefsloot, H. C. J., Smilde, A. K., & Ceulemans, E. 2015. Scaling in ANOVA-simultaneous component analysis. *Metabolomics*, **11**(5), 1265–1276.
- Tomasi, G., & Bro, R. 2006. A comparison of algorithms for fitting the PARAFAC model. *Computational Statistics & Data Analysis*, **50**(7), 1700–1734.
- Trygg, J., & Wold, S. 2002a. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, **16**(3), 119–128.
- Trygg, J., & Wold, S. 2002b. Orthogonal projections to latent structures (O-PLS). *Journal of Chemometrics*, **16**(3), 119–128.
- van den Berg, R. A., Hoefsloot, H. C. J., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. 2006. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC genomics*, **7**(1), 142.
- van der Greef, J., & Smilde, A. K. 2005. Symbiosis of chemometrics and metabolomics: past, present, and future. *Journal of Chemometrics*, **19**(5-7), 376–386.
- van Nederkassel, A.M., Daszykowski, M., Eilers, P. H. C., & Heyden, Y. Vander. 2006. A comparison of three algorithms for chromatograms alignment. *Journal of Chromatography A*, **1118**(2), 199 – 210.
- Vanwinsberghe, J. 2005. *Bubble: development of a matlab tool for automated ¹H-NMR data processing in metabonomics*. Traineeship report (unpublished results), Strasbourg University.
- Varmuza, K., & Filzmoser, P. 2016. *Introduction to multivariate statistical analysis in chemometrics*. CRC press.
- Verbeke, G., & Molenberghs, G. 2000. Springer series in statistics. *Linear Mixed Models for Longitudinal Data*.
- Vis, D. J., Westerhuis, J. A., Smilde, A. K., & van der Greef, J. 2007. Statistical validation of megavariate effects in ASCA. *BMC Bioinformatics*, **8**(1), 322.
- Vu, T. N., & Laukens, K. 2013. Getting Your Peaks in Line: A Review of Alignment Methods for NMR Spectral Data. *Metabolites*, **3**(2), 259.
- Vu, T. N., Valkenburg, D., Smets, K., Verwaest, K. A., Dommissie, R., Lemiere, F., Verschoren, A., Goethals, B., & Laukens, K. 2011. An integrated workflow for robust alignment and simplified quantitative analysis of NMR spectrometry data. *BMC bioinformatics*, **12**(1), 405.

- Wehrens, R. 2011. *Chemometrics with R: multivariate data analysis in the natural sciences and life sciences*. Springer Science & Business Media.
- Wentzell, P. D., & Vega Montoto, L. 2003. Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures. *Chemometrics and Intelligent Laboratory Systems*, **65**(2), 257 – 279.
- Westerhuis, J. A., van Velzen, E. J. J., Hoefsloot, H. C. J., & Smilde, A. K. 2008. Discriminant Q2 (DQ2) for improved discrimination in PLS-DA models. *Metabolomics*, **4**(4), 293–296.
- Whittaker, E. T. 1923. On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, **41**, 63–75.
- Wiklund, S., Johansson, E., Sjöström, L., Mellerowicz, E. J., Edlund, U., Shockcor, J. P., Gottfries, J., Moritz, T., & Trygg, J. 2008. Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Analytical Chemistry*, **80**(1), 115–122.
- Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., Hau, D. D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J. A., Lim, E., Sobsey, C. A., Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., De Souza, A., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhtudinov, R., Li, L., Vogel, H. J., & Forsythe, I. 2009. HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Research*, **37**(Database), D603–D610.
- Wold, S., Antti, H., Lindgren, F., & Öhman, J. 1998. Orthogonal signal correction of near-infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, **44**(1), 175–185.
- Wold, S., Sjöström, M., & Eriksson, L. 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, **58**(2), 109–130.
- Worley, B., & Powers, R. 2013. Multivariate analysis in metabolomics. *Current Metabolomics*, **1**(1), 92–107.
- Worley, B., & Powers, R. 2014. MVAPACK: A Complete Data Handling Package for NMR Metabolomics. *ACS Chemical Biology*, **9**(5), 1138–1144. PMID: 24576144.

- Wu, Y., & Li, L. 2016. Sample normalization methods in quantitative metabolomics. *Journal of Chromatography A*, **1430**, 80–95.
- Zhang, Z.-M., Chen, S., & Liang, Y.-Z. 2010. Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst*, **135**, 1138–1146.
- Zou, H., Hastie, T., & Tibshirani, R. 2006. Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, **15**(2), 265–286.
- Zwanenburg, G., Hoefsloot, H. C. J., Westerhuis, J. A., Jansen, J. J., & Smilde, A. K. 2011. ANOVA-principal component analysis and ANOVA-simultaneous component analysis: a comparison. *Journal of Chemometrics*, **25**(10), 561–567.

APPENDIX A

Abbreviations

Table A.1 – **Abbreviations, in alphabetic order.**

Abbreviation	Description
^1H NMR	Proton Nuclear Magnetic Resonance
AComDim	ANOVA Common Dimensions
AMOPLS	ANOVA Multi-block Orthogonal Partial Least Squares
ANOVA	Analysis Of Variance
ANOVA-TP	Analysis Of Variance-Target Projection
ANOVA-TP	ANOVA with Target Projection
AoV-PLS	Analysis of Variance Partial Least Squares
APCA	Analysis Of Variance-Principal Component Analysis
ASCA	Analysis Of Variance-Simultaneous Components Analysis
AsLS	Asymmetric Least Squares Smoothing
AUC	Area Under the ROC Curve
B-Y	Benjamini-Yekutieli approach
CluPA	Cluster-based Peak Alignment
COW	Correlation Optimised Warping
CSN	Constant Sum Normalisation
CV	Cross-Validation
df	Degrees of freedom
DFT	Discrete Fourier Transform
Continued on next page	

Table A.1 – continued from previous page

Abbreviation	Description
DSS	4,4-Dimethyl-4-silapentane-1-sulfonic acid
DTW	Dynamic Time Warping
ED	Effective Dimension
E(MS)	Expected mean squares
FDR	False Discovery Rate
FID	Free Induction Decay
FN	False Negatives
FP	False Positives
FS	Feature Selection
FT	Fourier transform
GC	Gas Chromatography
GLLR	Global Log-Likelihood Ratio
GLM	General Linear Model
ICA	Independent Component Analysis
icoshift	interval correlation shifting
IQR	Interquartile Range
KOPLS	Kernel OPLS
LARS	Least Angle Regression
LASSO	Least Absolute Shrinkage and Selection Operator
LB	Line Broadening
LC	Liquid Chromatography
LiMM-PCA	Linear Mixed Models - Principal Component Analysis
LLR	Log-Likelihood Ratio
LMM	Linear Mixed Models
LSOPLS	Light Sparse Orthogonal Partial Least Squares
LV	Latent Variables
MANOVA	Multivariate Analysis Of Variance
MIC	Metabolomic Informative Content
ML	Maximum Likelihood
MLPLS	Multilevel Partial Least Squares regression
MP	Meta Parameter
MS	Mean Squares
MSCA	Multilevel Simultaneous Component Analysis
NIPALS	Nonlinear estimation by Iterative PArtial Least Squares
NPV	Negative Predictive Value
OC	Orthogonal Components
OPLS	Orthogonal Partial Least Squares regression
OSC	Orthogonal Signal Correction

Continued on next page

Table A.1 – continued from previous page

Abbreviation	Description
PARAFAC	Parallel Factor Analysis
PARAFASCA	Parallel Factor Analysis- Simultaneous Components Analysis
PC	Principal Component
PC-ANOVA	Principal Component Analysis-Analysis Of Variance
PCA	Principal Component Analysis
PCR	Principal Component Regression
PepsNMR	Packaged Extensive Pre-processing Strategy for NMR data
PLS	Partial Least Squares
ppm	parts per million
PPV	Positive Predictive Value
PQN	Probabilistic Quotient Normalisation
PrC	Predictive Component
PRESS	Predictive Residual Sum of Squares
PTW	Parametric Time Warping
REML	Restricted Maximum Likelihood
RLLR	Restricted Log-Likelihood Ratio
RMSE	Root Mean Square Error
ROC	Receiver Operating Characteristic
RSR	Structure Ratio
SIMPLS	Statistically Inspired Modification of PLS
SNR	Signal-to-Noise Ratio
SPCA	Sparse Partial Least Squares
SPLS	Sparse Partial Least Squares
SPTW	Semi-Parametric Time Warping
SS	Sum of Squares
SST	Total Sum of Squares
SSY	Sum of Squares for explained variation in the response
SVD	Singular Value Decomposition
TMSP	Trimethylsilyl-propionate
TN	True Negatives
TP	True Positives
TVD	Total Variance Decomposition
VC	Variance Component
VIP	Variable Importance in Projection

APPENDIX B

Mathematical notations

Table B.1 – The main mathematical notations for Chapter 2: *PepsNMR for ^1H NMR metabolomic data pre-processing*, in the order of appearance.

s	Decaying FID signal
s_x	Real part of the signal s
s_y	Imaginary part of the signal s
ν	Frequency vector
s_0	Signal amplitude
T	Signal transverse relaxation time
t	Time vector
Φ	Total phase error
φ_0	Frequency-independent phase error
$\varphi_1(\nu)$	Frequency-dependent phase error
f	Spectrum
τ	Signal shift
w	Smoothed solvent signal
Q	Criterion to optimize
λ	Smoothness parameter
$\Delta^2 x$	Squared second order differences on a vector x
Δ_ν	Difference in frequency between 2 data points
δ	Chemical shift
r	Subset of the spectral descriptors index, $r \in \{1, \dots, m\}$
z	Baseline estimator
ψ_j	Weight for the frequency $j = 1, \dots, m$ for baseline estimation
$w(\nu)$	Warping function
K	Polynomial order in warping
β	Polynomial coefficients vector in warping
L	Number of B-splines in warping
α_l	the coefficients for the l^{th} B-spline
$B_l(\nu)$	B-spline curve
g	Reference spectrum in warping
κ	Regularisation parameter
c_i	Constant term to multiply a spectrum $i = 1, \dots, n$ during normalisation

Table B.2 – The main mathematical notations for Chapter 3: *Feature Selection in metabolomics with PLS-derived methods.*, in the order of appearance.

$\mathbf{X}$	Data matrix
n	Number of samples
m	Number of variables
$\mathbf{y}$	Response vector (dependent variable)
$\mathbf{B}$	Feature Scores
$\mathcal{A}$	Subset of size $(1 \times m^*)$ indicating the selected features indices from the m original variables
m^*	Number of retained variables in $\mathcal{A}$ as biomarker candidates
δ	Vector of two-sample Welch's t-statistic
$\bar{x}$	Sample mean
s^2	Sample variance
ρ	Vector of p-values for the t-test
$d_{(i)}$	The index i of an vector $\mathbf{d}$ sorted in decreasing order
a	Number of components (latent or orthogonal variables)
$\mathbf{T}$	x -scores matrix with $\mathbf{T} = (t_1 \dots t_a)$ scores vectors
$\mathbf{P}$	x -loadings matrix with $\mathbf{P} = (p_1 \dots p_a)$ loadings vectors
$\mathbf{E}_X$	Matrix of x -residuals
$\mathbf{W}$	x -weights matrix with $\mathbf{W} = (w_1 \dots w_a)$ weights vectors (also called direction vectors)
$\mathbf{U}$	Matrix of y -scores $\mathbf{U} = (u_1 \dots u_a)$ scores vectors
$\mathbf{q}$	Vector of y -loadings
$\mathbf{e}$	Vector of residuals
$\mathbf{c}$	Vector of y -weights
β	Regression coefficients
$\mathbf{b}$	Estimators of the regression coefficients
K	Krylov subspace
$\mathbf{I}_a$	Identity matrix of size $(a \times a)$
$\mathbf{s}$	Cross-product vector of $\mathbf{X}$ and $\mathbf{y}$
SSY_j	Sum of Squares of y -variance explained by component j
$\mathbf{v}$	Vector of VIP values
SSY_{tot}	Total Sum of Squares of the dependent variable
$\mathbf{w}^*$	Surrogate direction vector
$\mathbf{S}$	Sample covariances matrix
λ_1	L_1 sparsity constraint parameter
λ_2	L_2 sparsity constraint parameter
$\tilde{\mathbf{s}}$	Normed cross-product $\mathbf{X}$ and $\mathbf{y}$
η	Degree of sparsity
R	Number of re-sampled sub-datasets
θ	Mean change in spectra
$\Phi(\mathbf{D})$	Stability function with a binary matrix $\mathbf{D}$ as input
$\bar{m}^*$	Mean number of selected features across $\mathcal{A}$

Table B.3 – The main mathematical notations for Chapter 4: *ASCA-GLM and other GLM-multivariate techniques for advanced designs.*, in the order of appearance.

n	Number of observations
m	Number of original variables
f	Index for fixed effects
F	Number of fixed effects
p	Number of fixed parameters, intercept excluded
M_f	Model effect matrix for an effect f
Y	Multivariate response matrix
E	Multivariate model residuals matrix
X	Model matrix
β	Fixed parameters vector for univariate response, $\beta = (\beta_0, \beta_1, \dots, \beta_F)$
Θ	Fixed parameters matrix for multivariate responses, $\Theta = (\Theta_0, \Theta_1, \dots, \Theta_F)$
P_f	PCA loadings on M_f
T_f	PCA scores on M_f
D	Number of modes in PARAFAC
C	Number of components extracted
Y_G	Three way array with all the ANOVA/GLM parameters (except the constant term)
U, V and W	Loadings/scores matrices of the PARAFAC decomposition
E_G	Residual array of the PARAFAC model
M_f^a	Augmented model effect matrix for an effect f
Ψ_f	Sample variance-covariance matrix for an effect f in AComDim
λ_f^c	AComDim salience for effect f and component c
q_c	Orthogonal dimensions common to all the effect matrices in AComDim
Q	Matrix containing the common dimensions q_c : $Q = (q_1, q_2, \dots, q_C)$
R	Response matrix in AMOPLS
K	Kernel matrix
Ψ_f	“Association” matrix for effect f in AMOPLS
Ψ_G	KOPLS predictor matrix in AMOPLS

Table B.4 – The main mathematical notations for Chapter 5: *LiMM-PCA : combining ASCA⁺ and linear mixed models to analyse high dimensional designed data.*, in the order of appearance.

n	Number of observations
m	Number of original variables
m^*	Number of PCA principal components retained to form the new response matrix
f	Index for fixed effects
$\tilde{f}$	A set of fixed effects
r	Index for random effects
$\tilde{r}$	A set of random effects
p	Number of fixed parameters, intercept excluded
q	Number of random parameters
F	Number of fixed effects
R	Number of random effects
g	One fixed or random effect(s), $= \{f, r\}$
$\tilde{g}$	One or several fixed or random effect(s), $= \{\tilde{f}, \tilde{r}\}$ or $g \subseteq \{1, \dots, F\} \vee \{1, \dots, R\}$
c_i	Number of principal components retained for PCA on the effect matrix $i \in \{f, r\}$
y	Univariate response vector
X	Model matrix for the fixed intercept and effects, $X = (X_0 \ X_1 \ \dots \ X_F)$
β	Fixed parameters vector for univariate response, $\beta = (\beta_0, \beta_1, \dots, \beta_F)$
ϵ	Univariate residuals vector
Y	Multivariate response matrix
Θ	Fixed parameters matrix for multivariate responses, $\Theta = (\Theta_0, \Theta_1, \dots, \Theta_F)$
E	Multivariate model residuals matrix
M_i	Model effect matrix for an effect i
$\sum_i$	PCA loadings on M_i
$\mathbf{t}_i$	PCA scores on M_i
σ_ϵ^2	Residuals variance for the variance components mixed model
σ_r^2	Variance of random effect r for $r = 1, \dots, R$ for the variance components mixed model
Z	Model matrix for the random effects, $Z = (Z_1 \ Z_2 \ \dots \ Z_R)$
α	Random parameters vector for univariate response, $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_R)$
Σ	General residuals covariance matrix for LMM
G	General random effects covariance matrix for LMM
γ	Vector of random effects variance components for the variance components mixed model
$\{d \ x_i\}_{i=1}^t$	Diagonal matrix with t diagonal elements x_i
V	Marginal variance-covariance matrix of y in LMM
L	Model likelihood
Y^*	Scores of the Y PCA decomposition
Q^*	Loadings of the Y PCA decomposition
H^*	Residuals of the Y PCA decomposition
Γ	Random parameters matrix for multivariate responses, $\Gamma = (\Gamma_1, \dots, \Gamma_R)$
R^2	Coefficient of determination
$\Lambda_{\tilde{i}}$	(Restricted) Log-Likelihood Ratio statistic for a set of effects $\tilde{i} \in \{\tilde{f}, \tilde{r}\}$
B	Number of bootstrap simulations

APPENDIX C

Datasets

Description of the Human Serum and Urine databases

From Féraud et al. (2015) and Martin et al. (2018)

This article presents the Human Serum and Human Urine databases originally published in Féraud *et al.* (2015) and Martin *et al.* (2018). These databases were designed with spectroscopists from the University of Liège and the University of Nantes (France).

Motivations for creating this database

Samples from two well-known human biofluid (serum and urine) are explicitly designed to evaluate the pre-processing methodology performances in Chapter 2. These datasets were both designed to collect multiple measures from the same experimental unit, the donor, and capture other technical sources of variation, allowing the comparison of inter- and intra-units variability.

Design, sample preparation and data acquisition

Human Serum

Design

In the Human Serum dataset, a blood sample was collected from 4 different donors. For each sample, 8 sub-samples were measured across 8 days with one sub-sample from each donor per day and permutations according to a latin hypercube sampling method (Iman, 2014) in order to avoid confusion between donors and times of analysis. Data have been acquired at different moments of the day, thus creating different delays before spectral measurement. The total number of available 1D FID signals is then $4 \times 8 = 32$.

Data preparation

This experimental design is based on human serum. Peripheral blood was collected in serum separating tubes (Greiner). Sera were distributed into 0.5 ml 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 200 μl of deuterated phosphate buffer (0.1 M, $\text{pH} = 7.4$) containing 1 % of sodium azide and 30 μl of TMSP (10 mg/ml). The solutions were then transferred into 5 mm NMR tubes before NMR measurement.

Data acquisition

1D spectra were recorded at 298 K on a Bruker Avance spectrometer operating at 500.13 MHz 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, shims were tuned using auto-tune (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 minimise 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).

The ^1H NMR spectra were acquired using 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 CPMG experiment (pulse sequence cpmgpr1d supplied by Bruker) used a $\text{RD-}90^{\circ}\text{-(t-}180^{\circ}\text{-t}_n\text{)}$ -sequence with a relaxation delay (RD) of 2s, a spin echo delay (t) of 400 μs 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 3.982555 s and a quantity of four dummy scans was chosen. FIDs were collected in 64 K time data points.

Human Urine

This dataset 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 ULiège.

Design

In this second dataset, morning urine from 3 different donors was collected. Each donor sample was divided into 2 sub-samples: one was kept pure and the other 25 % diluted. The 4 aliquotes of each dilution were analysed on 4 different days. For each day, the order of measurement was held constant across the 6 sub-samples. The total number of collected FID signals is then: $3 \times 2 \times 4 = 24$.

Data preparation

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 μ l and four aliquots of 320 μ l 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 μ l were supplemented with 300 μ l of deuterated phosphate buffer (DPB, pH 7.4), while 320 μ l urine aliquots were supplemented with 380 μ l of the same buffer. 10 μ l 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 analysed. For each day, the order of measurement was held constant across the six sub-samples. A total of 24 1D collected signals are finally available.

Data acquisition

1D spectra were recorded at Ceisam, Nantes on a Avance III NMR spectrometer operating at 700.28 MHz for the proton signal acquisition, and with the NOESYPR1D (1D Nuclear Overhauser effect spectroscopy with water pre-saturation) sequence (number of scans: 64).

The final databases

The final databases are composed of ^1H NMR raw Bruker FID signals (in the *fids* folder) and a design .RData file.

The Human Serum spectra are labelled as: *JxDy* where *x* is the day of measurement and *y* is the donor label. The design file contains the *HS_design* data.frame with variables *ID*, *donor* and *day*.

The Human Urine, spectra are labelled as: *SxDy* where *x* is the donor label and *y* is the dilution (0: no dilution; 1: 25 % diluted). The design file contains the *HU_design* data.frame with variables *ID*, *donor* and *dilution*.

Manual pre-processing

Both datasets were manually pre-processed with a combination of Topspin 3.1 and AMIX 3.9.14. at the resolution of 500 buckets. The manual pre-processing included the following sequence: exponential apodization, Fourier transform, first and zero order phase corrections, internal calibration, polynomial baseline correction, bucketing and spectral region removal.

The manually pre-processed spectra are available under the form of a .RData file for Human Serum *HumanSerumManualPretreat* and as a .csv file *Nantes.Manual.B500.Spectra.data* for Human Urine.

References

- Féraud, B., Govaerts, B., Verleysen, M., & de Tullio, P. 2015. Statistical treatment of 2D NMR COSY spectra in metabolomics: data preparation, clustering-based evaluation of the Metabolomic Informative Content and comparison with 1H-NMR. *Metabolomics*, **11**(6), 1756–1768.
- Iman, R. L. 2014. Latin Hypercube Sampling. *Wiley StatsRef: Statistics Reference Online*.
- Martin, M., Legat, B., Leenders, J., Vanwinsberghe, J., Rousseau, R., Boulanger, B., Eilers, P. H.C., Tullio, P. De, & Govaerts, B. 2018. PepsNMR for ¹H NMR metabolomic data pre-processing. *Analytica Chimica Acta*, **1019**, 1 – 13.

Description of the Semi-artificial Urine database

From Rousseau (2011) and Martin & Govaerts (2019)

This article presents the Semi-artificial Urine database originally used in Rousseau (2011) and Martin & Govaerts (2019). This database was designed with spectroscopists from Eli Lilly and the University of Liege (ULg).

Incentive

The *semi-artificial urine database* was built to assess the quality of potential biomarkers' identification with known in advance biomarkers' location. This exact localisation enables to precisely quantify the statistical methods performances to retrieve the true spectral descriptors that were altered. To build this database, random artificial alterations were added to "control" rats urine spectra.

Sample acquisition and pre-processing

The primary dataset represents the status of physiological stability in rat urine. It is composed of FID signals issued from the COMET database (Lindon *et al.*, 2003) where all samples come from "control" rats that did not receive any treatment. These data were obtained with a flow injection process and a 600 MHz ^1H NMR spectroscope at the Imperial College in London.

To obtain interpretable spectra, pre-treatments (suppression of the residual water signal, apodization, baseline correction, and warping, window selection and bucketing to $m = 500$) of the raw data were handled in an automatic fashion by BUBBLE (Vanwinsberghe, 2005), the PepsNMR's precursor. A typical pre-processed spectrum can be seen in Figure 1. Note that non-informative areas with water and urine peaks (4.5 - 6 ppm) that are displayed in red in this graph were also removed from the spectra. Finally, some outliers were removed based on PCA and Euclidian/Mahalanobis distances.



Half of the spectra are normal and half are altered. For the altered profiles, 46 descriptors are artificially positively altered with random heights from Gamma distributions. They are scattered across 10 different spectral regions, half localised in a noise-free area (6.48 - 7.26 ppm) and half in the noisy part (2.56 - 3.37 ppm) of the spectrum. Furthermore, some pairs of altered regions are correlated by using the same Gamma distribution to generate peaks, leaving only 6 independent biomarkers in the dataset (only six independent Gamma distributions have been used instead of ten). Note also that four peaks have a width of seven descriptors and the six other peaks have a smaller width of three descriptors, for a total of 46 descriptors or biomarkers. Finally, 2 out of the 6 independent biomarkers are randomly chosen to alter each spectrum in order to mimic the natural variety of body response to a same stimulus. Figure 2 from Rousseau (2011) presents the location and the mean amplitude of all possibly added alterations. The index and ppm values of those biomarkers are given in Table 1.

Table 1: Column indexes for artificial alterations peaks in the ^1H NMR spectra.

[illegible]

Note that after the alterations were added, a constant sum normalisation was applied as well.

The final semi-artificial urine database

The total size of the final database is (800×500) , where half of the samples (460) are artificially altered spectra. Finally, random draws of sizes 200 (2×100) and 60 (2×30) are extracted to create the sub-samples used in Chapter 3. The design matrix is represented by a binary variable coded $[1, 0]$ that indicates the presence/absence of the alterations. This dataset is available in the *DataSimul.RData* file that contains the x matrix with the NMR spectra and the binary vector y indicating if the spectra are altered ($y = 1$) or not ($y = 0$).

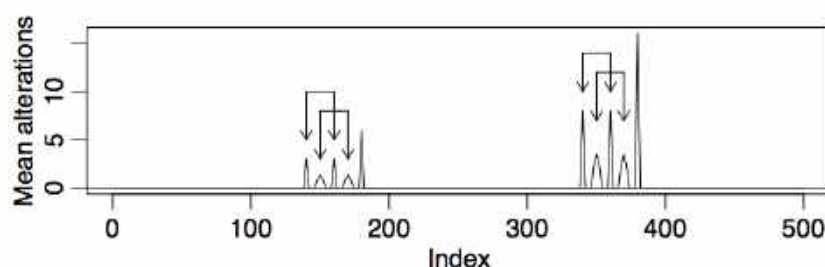


Figure 2: Representation of the artificial alterations in the spectrum. Figure from Rousseau (2011).

References

- Lindon, J. C., Nicholson, J. K., Holmes, E., Antti, H., Bollard, M. E., Keun, H., Beckonert, O., Ebbels, T. M., Reily, M. D., Robertson, D., *et al.* . 2003. Contemporary issues in toxicology the role of metabolomics in toxicology and its evaluation by the COMET project. *Toxicology and Applied Pharmacology*, **187**(3), 137–146.
- Martin, M., & Govaerts, B. 2019. *Feature Selection in metabolomics with PLS-derived methods*. Discussion Paper DP 2019/20. Institute of Statistics, Biostatistics and Actuarial Sciences, UCLouvain, Belgium.
- Rousseau, R. 2011. *Statistical contribution to the analysis of metabolomics data in ^1H NMR spectroscopy*. Ph.D. thesis, Institute of Statistics, Biostatistics and Actuarial Sciences, Université catholique de Louvain, Belgium.
- Vanwinsberghe, J. 2005. *Bubble: development of a matlab tool for automated ^1H -NMR data processing in metabolomics*. Traineeship report (unpublished results), Strasbourg University.

Description of the citrate-hippurate database

From Rousseau (2011) and Guisset et al. (2019)

This article presents the Semi-artificial Urine database originally used in Rousseau (2011) and Guisset *et al.* (2019). This database was designed with spectroscopists from Eli Lilly and the University of Liege (ULg).

Introduction

The database has been experimentally created in order to control the spectral locations of the biomarkers to find. This property allows us to evaluate the performances of the data analysis of various statistical methods.

This urine experimental database is also designed in order to explore the influence on spectra of diuretic fluctuations, intra-sample ^1H NMR replications, and inter-day ^1H NMR measurements.

Motivations for creating this database

This database was collected according to an experimental design with three objectives:

- ① The first one was to study the ability of statistical methods to find as biomarkers the descriptors of the spectra for which a variability was experimentally controlled. In this experiment homogenised medium urine samples were spiked with two products at different levels of concentration and analysed by spectroscopy. The “concentrations” factor of each added product has to mimic the variability focused in a biomarker search study.

② The second objective was to study the quality of normalisation techniques. Two dilutions in water of a same pool of urine were considered in these experiments.

③ The third objective was to study the natural variability of spectra related to several factors: the day of analysis, and the repetition of a spectral acquisition (replication).

Statistical experimental design

Four factors of variability were considered in the experimental design:

1. The medium: two media are considered: the urine from a pool (“B04-331, fisher 344 rats, female”) and the urine coming from the same pool but diluted with a dilution rate of 50 %.
2. The concentration of each of the two products into the samples: hippuric acid and citric acid were added to the samples in different concentrations described in Figure 1. The maximal concentrations are 8mM for the citric acid (Qc) and 4mM for the hippuric acid (Qh). Four levels of each acid were considered, determining fourteen experimental conditions.

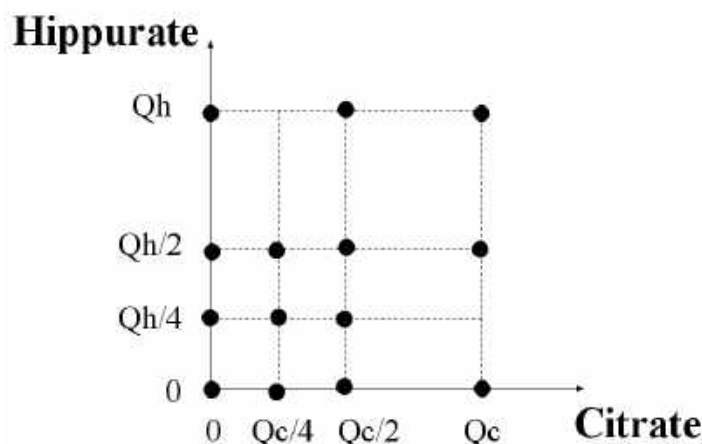


Figure 1: *The urine experimental design.*

As shown in Figure 2, the peaks corresponding to each product are located in distant ppm domains. The hippurate has three peaks with two in the region of high ppms containing a low noise level. On the contrary, citrate peaks are located in the noisy region of low ppms. Note that in the spectral pre-treatments the citrate peaks are aggregated in one peak to avoid alignment problems.

3. The replicates: each mixture or experimental condition described in Figure 1. was repeated at least twice. Consequently, 28 samples were prepared for each medium.

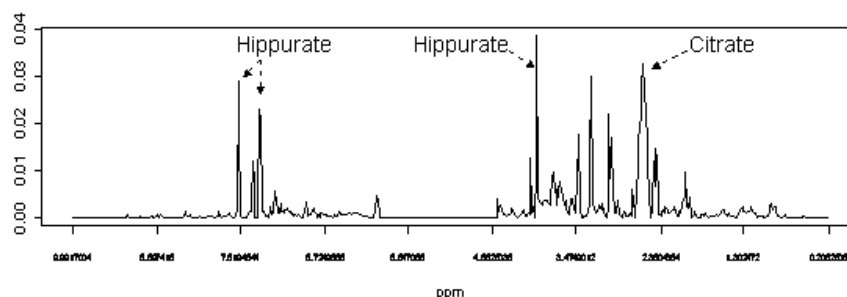


Figure 2: *A typical urine spectrum with spiked citrate and hippurate.*

4. The day of ^1H NMR measurement: for each medium, the preparation of the 28 mixtures (14 experimental conditions $\times$ 2 replicates) were performed in three series. The two media samples of each series correspond to a plate. Each day, a plate was defrosted in order to be analysed by the ^1H NMR spectrometer. Additionally, plate 1 was analysed twice: the plate was analysed on day 1, refrosted, then defrosted and analysed again on day 5. Values for the day of measurement factor are: day 1, day 2, day 3, day 5.

In each plate, according to the previous description, we have 56 samples (2 media $\times$ 14 experimental conditions $\times$ 2 replicates). In addition, other replicates were included for experimental conditions with no spiked compounds. In this way, over the four days, we obtained 269 samples to analyse.

Sample preparation and acquisition of the ^1H NMR data

Each sample had a final volume of 1200 μl . For each prepared sample, a mixture of 600 μl was added to 600 μl of media. The media was either urine coming from a pool of 344 female Fischer rats or the same urine diluted at 50%. The mixture included the two products (citrate and hippurate) in the chosen concentrations and phosphate buffer containing TMSP. The volume of the phosphate buffer was adapted in order to obtain a volume of 600 μl to add to the 600 μl of urine. For example, a final urine media sample contains 600 μl of urine and 600 μl of mixture composed of 150 μl of citrate, 150 μl of hippurate and 300 μl of buffer. The corresponding diluted media sample was obtained by taking half of the citrate volume, half of the hippurate and a larger volume of buffer to keep the mixture volume at 600 μl . In the example, the corresponding final diluted urine media sample contains 600 μl of the diluted urine media and 600 μl of the mixture, here composed by 75 μl of citrate, 75 μl of hippurate and 450 μl of buffer. In each preparation, the mixture was added to the urine, centrifuged, frozen at -80°C and unfrozen at 4°C the day before the ^1H NMR analysis.

Sample tubes were analysed within each day of experiment with a NOESY presaturation sequence.

The pre-treatments

Each spectrum was processed using the pre-treatment procedure advised in Rousseau (2011). The part of the spectrum between 0.2 and 10 ppm has been reduced to 600 descriptors. Attention must be paid to the facts that:

- The ppm values corresponding to the large non-informative urea and to the water peak (4.5 - 6.0 ppm) were set to zero.
- The region around the citrate resonances (2.56 - 2.72 ppm) was integrated into one peak to suppress the high shifts of the citrate peaks.
- the data were normalised by a constant sum normalisation.

The full final urine database

The spectra are labelled as: $Mm - Cxy - Dd - Rr$ where m is the code for the medium (1 or 2 for urine and diluted urine respectively), xy are the product proportions of the maximal concentrations, d is the day, and r is the time index (replicate). Among all the the samples analysed, three presented problems during the spectral acquisition (M1C04D1R2, M2C00D1R7, M2C48D1R1).

The spectral data matrix X of dimensions (269×600) is contained in the data file *HC_PtotP.Rdata*. Indeed, in addition to the 224 samples were 2 replicates are present, blank conditions were replicated 8 times.

The data file *HC_Destot.Rdata* corresponds to the experimental design of dimensions with 14 columns describing the spectra. The variables contained in it are: the name of the spectrum, the medium, the combined level of hippurate and citrate (qualitative variable), the level of hippurate separately (qualitative variable), the level of only citrate (qualitative variable), the volume of added hippurate (quantitative variable), the volume of added citrate (quantitative variable), the index of the replicate and the day.

The UCH sub-dataset used for Chapter 4

In the subset of this database used in Chapter 4, and further referred to as the UCH data set, citrate and hippurate factors have three levels: no added chemical (a concentration of 0), a medium concentration (2 mM for the citrate factor ($Q_c/4$) and 1 mM for the hippurate factor ($Q_h/4$)) and a high concentration (4 mM for the citrate factor ($Q_c/2$) and 2 mM for the hippurate factor ($Q_h/2$)). This leads to an experimental design with 9 possible combinations of concentrations as illustrated in Figure 3(a). Two different days were retained (days 2 and 3) in the subset. The UCH data originally contained 36 spectra *i.e.* 9 combinations

of concentrations $\times$ 2 time points $\times$ 2 replicates. However, two outliers were removed, resulting in 34 observations and a slightly unbalanced experimental design (*cf.* Figure 3(b)). The resulting spectral matrix is of size (34 $\times$ 600).

Finally, Figure 3(d) presents the scores plot of the first two PCs for the spectral matrix. It shows that the information on the citrate and hippurate factor levels from the experimental design can be recovered, but only to some extent.

The data file *UCH.Rdata* contains this subdataset where *design* represents the experimental design with variables *Hippurate*, *Citrate*, *Dilution*, *Day* and *Time*; *outcomes* represents the metabolomic spectral profiles; and *formula* is the model formula that was used in Chapter 4.

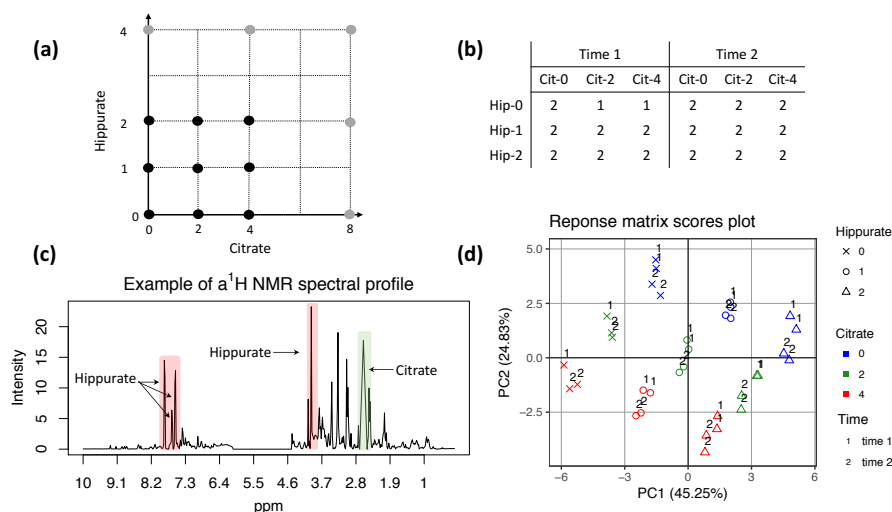


Figure 3: Presentation of the UCH sub-dataset (a) Experimental design (b) Sample sizes for each of the 9 combinations of citrate, hippurate and Time levels (c) Example of a ^1H NMR spectral profile showing the hippurate and citrate peaks (d) Response matrix scores plots for the first two PCs of the UCH data set.

References

Guisset, S., Martin, M., & Govaerts, B. 2019. Comparison of PARAFASCA, AComDim, and AMOPLS approaches in the multivariate GLM modelling of multi-factorial designs. *Chemometrics and Intelligent Laboratory Systems*, **184**, 44 – 63.

Rousseau, R. 2011. *Statistical contribution to the analysis of metabonomics data*

in ^1H NMR spectroscopy. Ph.D. thesis, Institute of Statistics, Biostatistics and Actuarial Sciences, Université catholique de Louvain, Belgium.

Description of the Repeatability/reproducibility database

From Rousseau (2011) and Martin & Govaerts (2019)

This article presents the Repeatability/reproducibility database originally used in Rousseau (2011) and Martin & Govaerts (2019). This database was designed with spectroscopists from Eli Lilly and the University of Liège.

Motivations for creating this database

This serum database is dedicated to the exploration of the impact of multi-source biological factors on the spectral variability (*e.g.* variability between and within healthy patients, etc.). The goal of this database is not to describe all the biological variance components in detail. The aim instead is to get a reasonable idea of the global amplitude of natural biological variability due to inter- and intra-individual sources. Another purpose of this database is to compare the analytical variability and the biological one. For various further use of the discovered metabolomic biomarkers, it is very crucial to have an analytical variability lower than the biological one. If a discovered metabolomic spectral biomarker corresponds to a spectral zone presenting a systematically higher intensity but also a huge variability due to the acquisition or the pre-treatments of data, detection tools built with this biomarker will have a low sensitivity and specificity.

This database was created from the collection of blood samples taken from volunteers. Samplings were organised following a procedure allowing to observe the biological factors of interest. The resulting blood samples were then prepared and measured in order to explore some analytical factor.

Statistical experimental design

Four factors of variability were considered in the experimental design:

1. The volunteer (from 1 to 12): the blood of twelve volunteers was collected. Since the alterations of the spectra due to diseases is outside the scope of the current analysis, all selected volunteers were healthy. Among the volunteers, seven were women and five were men. By controlling the sex balance, sex is not a confounding factor in the study.
2. The sampling for a given volunteer (from 1 to 3): the blood sampling was performed on three non consecutive days under similar conditions. Similar conditions means that the three blood samplings were all carried out the morning around 10 a.m.. The volunteers were encouraged to have fasted. Thanks to these measures, the fluctuation and the variability between the individuals was controlled.
3. The tube for a given volunteer and a given sampling (from 1 to 2): each blood sampling day, two tubes were collected. Since two samples are available for a given volunteer the experimental variance can be estimated. By experimental variance, we mean the variability due to the laboratory operations.
4. The time of measurement: each tube was measured twice the same day with a couple of hours between the two measurements. The replication allows determining whether the sample can be conserved after defrozing. Actually this has been already considered in the previous serum database. The only difference is the use of human serum instead of commercial serum from rats.

The two first factors are biological sources of variability while the two last are analytical sources.

Sample preparation

The preparation of the samples and the spectral acquisition were realised at the Laboratoire de Chimie Pharmaceutique of ULg, Belgium. Once the blood sample had been collected, it was processed into serum. Then phosphate buffer was added to control the pH. The trimethylsilyl propanoate (TMSP) was finally put in as internal reference. The protein signals were eliminated thanks to the pulse sequence CPMG. Serum samples were conserved frozen.

The pre-treatments

Each acquired spectrum was pre-treated according to the procedure proposed in Rousseau (2011). Each spectrum has 750 ppms and is normalised to have a sum equal to 750.

The final database

In total, $12 \times 3 \times 2 \times 2 = 144$ spectra were analysed. Each spectrum is available at a resolution of 750 buckets (*i.e.* variables) and observations are named with a series of 5 digits as: volunteer-sampling-tube-time ; e.g. 01312 corresponds to volunteer 01, sampling 3, tube 1 and time 2.

4 observations had an acquisition problem and were removed beforehand (04221, 08222, 11121,11122), rendering the design unbalanced with a total of 140 observations instead on 144, as represented in Figure 1.

The data are contained in the *HumanSerumData.Rdata* file with the *design* data.frame with the 4 variables from the experimental design and the *outcomes* matrix containing the spectra.

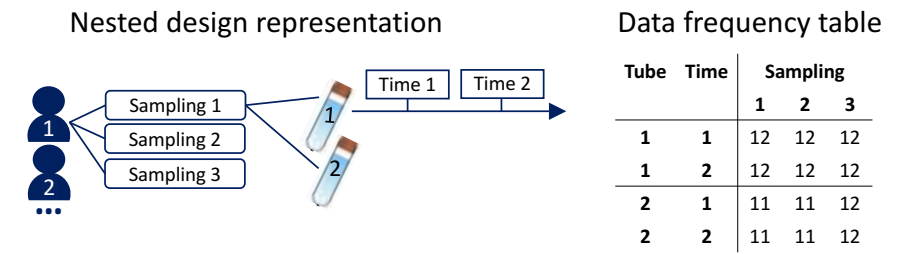


Figure 1: Nested design representation and data frequency table for the Human Serum (Repeatability/reproducibility) dataset.

A stacked representation of the spectra from this dataset is illustrated in Figure 2.

References

Martin, M., & Govaerts, B. 2019. *LiMM-PCA : combining ASCA⁺ and linear mixed models to analyse high dimensional designed data*. Discussion Paper DP 2019/21. Institute of Statistics, Biostatistics and Actuarial Sciences, UCLouvain, Belgium.

Rousseau, R. 2011. *Statistical contribution to the analysis of metabonomics data in ¹H NMR spectroscopy*. Ph.D. thesis, Institute of Statistics, Biostatistics and Actuarial Sciences, Université catholique de Louvain, Belgium.

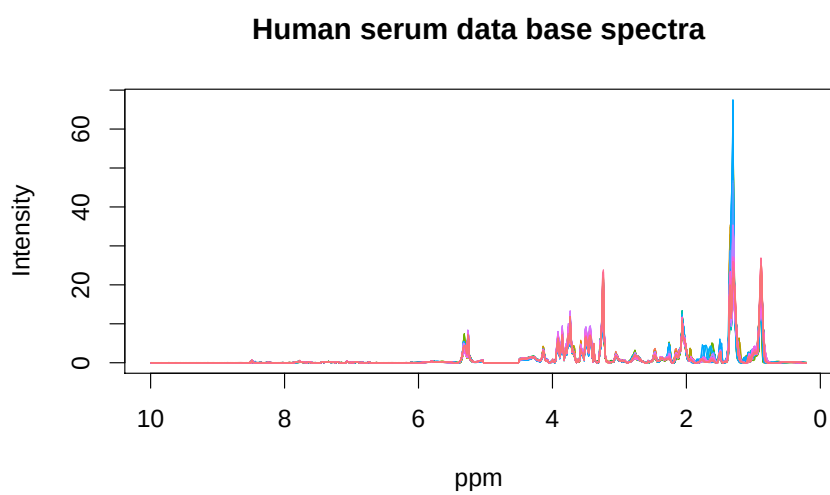


Figure 2: *Plot of the Repeatability/reproducibility database*

Description of the Candies database

From Luciano & Næs (2009) and Martin & Govaerts (2019)

This article presents the Candies database originally published in Luciano & Næs (2009) and Liland *et al.* (2018). It has been used in this thesis in Chapter 5 and in Martin & Govaerts (2019).

Statistical experimental design

Three factors of variability were considered in the experimental design:

1. The Assessors: a panel of 11 judges are rating the candies.
2. The Candies: 5 different candies are rated with respect to a series of 9 attributes: transparency, acidity, sweet taste, raspberry flavour, sugar coated texture tested with a spoon, biting strength in the mouth, hardness, elasticity in the mouth and stick to teeth in the mouth.
3. The replicates: Each candy was evaluated in triplicate by each assessor.

The final Candies database

This dataset is available in the *Candies.RData* data file where the *design* data.frame corresponds to the experimental design matrix with variables *Judges* and *Candies*, and the *outcomes* matrix corresponds to the sensory attributes evaluated by the judges.

The resulting balanced design is presented here in an unfolded data table with $165 = 11 \times 5 \times 3$ observations and 9 variables (attributes). The observations are identified by 4 digits as: judge-candy-replicate, *e.g.* 0712 corresponds to judge 7, candy 1 and replicate 2.

In Figure 1, the raw values of the 9 different attributes are displayed and colour-coded according to the candies. Clearly, distinct patterns are appearing: candies {5, 1} and {2, 3, 4} already seem to share similar attribute rating profiles.

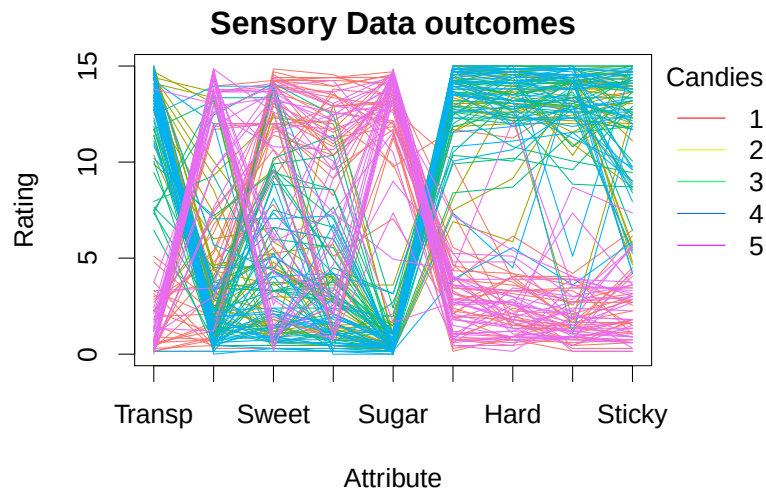


Figure 1: Plot of the sensory attributes matrix.

References

- Liland, K. H., Smilde, A., Marini, F., & Næs, T. 2018. Confidence ellipsoids for ASCA models based on multivariate regression theory. *Journal of Chemometrics*, **32**(5), e2990.
- Luciano, G., & Næs, T. 2009. Interpreting sensory data by combining principal component analysis and analysis of variance. *Food Quality and Preference*, **20**(3), 167 – 175.
- Martin, M., & Govaerts, B. 2019. *LiMM-PCA : combining ASCA⁺ and linear mixed models to analyse high dimensional designed data*. Discussion Paper DP 2019/21. Institute of Statistics, Biostatistics and Actuarial Sciences, UCLouvain, Belgium.