From non-parametric estimation to biostatistics

Rémi Servien

▶ **To cite this version:**

Rémi Servien. From non-parametric estimation to biostatistics. Statistics [math.ST]. Université Toulouse 3 Paul Sabatier, 2018. tel-01928342

HAL Id: tel-01928342
https://theses.hal.science/tel-01928342

Submitted on 20 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

MANUSCRIT

En vue de l'obtention de

l'Habilitation à Diriger les Recherches

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *15 Novembre 2018* **par :**

Rémi Servien

From non-parametric estimation to biostatistics

JURY

Avner Bar-Hen	CNAM	Rapporteur
Anne-Laure Boulesteix	LMU München	Rapportrice
Didier Concordet	ENVT	Examinateur
Fabrice Gamboa	Université Toulouse III	Président
Jérôme Saracco	INP Bordeaux	Rapporteur
Nathalie Vialaneix	INRA	Examinatrice

École doctorale et spécialité :
MITT : Domaine Mathématiques : Mathématiques appliquées

Unité de Recherche :
Intheres (UMR 1436)

Rapporteurs :
Avner Bar-Hen, *Anne-Laure Boulesteix* et *Jérôme Saracco*

Remerciements

Je souhaite commencer par remercier Anne-Laure Boulesteix, Avner Bar-Hen et Jérôme Saracco d'avoir accepté de remplir le rôle chronophage et ingrat de rapporteur. Je suis honoré de voir de tels scientifiques faire partie de mon jury et je les remercie chaleureusement pour leur participation. Je suis également très heureux que Fabrice Gamboa ait accepté spontanément de faire partie de ce jury et de son intérêt pour mes travaux. Je voudrais également adresser un immense merci à Didier Concordet et à Nathalie Vialaneix pour leurs précieux conseils pour la rédaction de ce manuscrit ou la gestion administrative (pas toujours évidente) de ce dossier.

La recherche est bien évidemment un travail collectif et je n'aurais pu mener à bien mes différents travaux sans l'aide des nombreuses personnes avec lesquelles j'ai collaborées. Il est toujours délicat de citer certains collaborateurs et de risquer d'en oublier d'autres au risque de minimiser leur importance. Néanmoins je vais essayer de remplir cette tâche sans trop d'oublis. Je vais tout d'abord commencer par le commencement et remercier mes encadrants de stage de M2, Nicolas et Christophe, en espérant que les collaborations qui se sont poursuivies depuis continuent encore. Bien évidemment je ne peux pas oublier ici Thomas qui soutiendra également d'ici peu. Merci pour tout mon ami (et sache que ces 15 jours d'avance me donnent la possibilité d'être dans ton jury, si jamais ...). La transition thésard/chercheur est à mon avis un moment clef dans la vie d'un chercheur. Grâce à Didier, tout s'est passé pour le mieux et mon insertion dans ma nouvelle unité a pu se faire dans les meilleures conditions possibles, je souhaite donc l'en remercier. Nathalie m'a permis de rencontrer beaucoup de chercheurs (notamment toulousains) et de monter une collaboration fructueuse, qu'elle en soit remerciée. Enfin, un grand merci à Alain, Patrick, Laure, Pierre, Eric, Virginie, Cécile, Marie, Victor (SFCB ftw), Malika, Elena, Gaëlle et tous les autres pour les différentes discussions que nous avons pu avoir et que nous aurons encore ainsi qu'à l'ensemble des membres des UMR InTheRes et Toxalim pour leur chaleureux accueil et la convivialité dans laquelle nous travaillons au quotidien. Dans une case un peu à part je souhaite également remercier Thierry Klein bien que je ne le connaisse que très peu. Lire l'introduction de son manuscrit d'HdR a été tout d'abord un soulagement pour moi puis une source d'inspiration et son expérience des bizarreries de la procédure de soutenance d'HdR à l'UPS a également été précieuse.

D'un point de vue plus personnel, j'ai une pensée particulière pour ceux qui ne liront probablement (au mieux) que les remerciements de ce manuscrit et que je suis donc dans l'obligation de citer ici : tous mes amis du GRC, de l'INSA, de Gruissan ou du rugby. Je n'aurais pas pu arriver là sans l'amour que me portent mes parents et sans la totale liberté qu'ils m'ont laissé dans mes études malgré mes changements de direction. Merci Maman, Papa mais je n'oublie pas non plus mes grands-parents ainsi que Benja, Manon et leur petite famille.

Et enfin, je remercie Maëlys (qui compte déjà jusqu'à 39, une future matheuse à n'en pas

douter) et Clément (qui s'arrête à 14 pour l'instant mais qui progresse vite) pour m'apporter joie de vivre et dynamisme à défaut de m'avoir amené le calme nécessaire durant cette rédaction... Enfin comment ne pas finir en remerciant Christelle. Pour avoir su m'épauler pendant les moments difficiles et m'apporter du bonheur au quotidien, je te remercie du fond du coeur.

Table of contents

List of contributions

This is the list of my contributions. All the references cited as [RSxx] in the following manuscript are referred to this list.

Publications

[RS01] V. Picheny, R. Servien and N. Villa-Vialaneix. Interpretable sparse SIR for functional data. *To appear in Statistics and Computing*, [see paper].

[RS02] C. Abraham, R. Servien and N. Molinari. A clustering Bayesian approach for radiotherapy x-ray beam bouquets. *To appear in Statistical Modelling*, [see paper].

[RS03] P. Tardivel, R. Servien and D. Concordet. Sparsest representations and approximations of an underdetermined linear system, *Inverse Problems* (2018), **34**(5), 055002, [see paper].

[RS04] H. Traore, O. Crouzet, L. Mamy, C. Sireyjol, V. Rossard, R. Servien, E. Latrille, F. Martin-Laurent, D. Patureau and P. Benoit. Clustering pesticides according to their molecular properties, fate and effects by considering additional ecotoxicological parameters in the TyPol method, *Environmental Science and Pollution Research* (2018), **25**(5), 4728-4738, [see paper].

[RS05] Y. Guitton, M. Tremblay-Franco, G. Le Corguillé, J.-F. Martin, M. Pétéra, P. Roger-Mele, A. Delabrière, S. Goulitquer, M. Monsoor, C. Duperier, C. Canlet, R. Servien, P. Tardivel, C. Caron, F. Giacomoni and E.A. Thévenot. Create, run, share, publish, and reference your LC-MS, GC-MS, and NMR data analysis workflows with Workflow4Metabolomics 3.0, the Galaxy online infrastructure for metabolomics, *International Journal of Biochemistry and Cell Biology* (2017), **93**, 89-101, [see paper].

[RS06] G. Gauderat, N. Picard-Hagen, P.-L. Toutain, R. Servien, C. Viguié, S. Puel, M.Z. Lacroix, A. Bousquet-Melou and V. Gayrard. Prediction of human prenatal exposure to bisphenol A and bisphenol A glucuronide from an ovine semi-physiological pharmacokinetic model, *Scientific Reports* (2017), **7**, 15330, [see paper].

[RS07] P. Tardivel, C. Canlet, G. Lefort, M. Tremblay-Franco, L. Debrauwer, D. Concordet and R. Servien. ASICS : an automatic method for identification and quantification of metabolites in complex 1D ^{1}H NMR spectra, *Metabolomics* (2017), **13**(10), 109, [see paper].

[RS08] P. Benoit, L. Mamy, R. Servien, Z. Li, E. Latrille, V. Rossard, F. Bessac, D. Patureau and F. Martin-Laurent. Categorizing chlordecone potential degradation products to explore their environmental fate, *Sciences of the Total Environment* (2017), **574**, 781-795, [see paper].

[RS09] T. Laloë and R. Servien. A note on the asymptotic law of the histogram without continuity assumptions, *Brazilian Journal of Probability and Statistics* (2016), **30**(4), 562-569, [see paper].

[RS10] V. Storck, L. Lucini, L. Mamy, F. Ferrari, E. S. Papadopoulou, S. Nikolaki, P. A. Karas, R. Servien, D. G. Karpouzas, M. Trevisan, P. Benoit and F. Martin-Laurent. Identification and characterization of tebuconazole transformation products in soil by combining suspect screening and molecular typology, *Environmental Pollution* (2016), **208** B, 537-545, [see paper].

[RS11] E. Di Bernardino, T. Laloë and R. Servien. Estimating covariate functions associated to multivariate risks : a level sets approach, *Metrika* (2015), **78**(5), 497-526, [see paper].

[RS12] D. Concordet and R. Servien. Individual prediction regions for multivariate longitudinal data with small samples, *Biometrics* (2014), **70**(3), 629-638, [see paper].

[RS13] R. Servien, L. Mamy, Z. Li, V. Rossard, E. Latrille, F. Bessac, D. Patureau and P. Benoit. TyPol - a New Methodology for Organic Pollutants Clustering based on their Molecular Characteristics and Environmental Behavior, *Chemosphere* (2014), **111**, 613-622, [see paper].

[RS14] T. Laloë and R. Servien. Nonparametric estimation of regression level sets using kernel plug-in estimator, *Journal of the Korean Statistical Society* (2013), **42**(3), 301-311, [see paper].

[RS15] T. Laloë and R. Servien. The X-Alter Algorithm : A Parameter-Free Method of Unsupervised Clustering, *Journal of Modern Applied Statistical Methods* (2013), **12**(1), 90-102, [see paper].

[RS16] C. Abraham, N. Molinari and R. Servien. Unsupervised clustering of multivariate circular data, *Statistics in Medicine* (2013), **32**(8), 1376-1382, [see paper].

[RS17] A. Berlinet and R. Servien. Empirical estimator of the regularity index of a probability measure, *Kybernetika* (2012), **48**(4), 589-599, [see paper].

[RS18] A. Berlinet and R. Servien. Necessary and sufficient condition for the existence of a limit distribution of the nearest neighbour estimator, *Journal of Nonparametric Statistics* (2011), **23**(3), 633-643, [see paper].

[RS19] R. Servien. Estimation de la fonction de répartition : revue bibliographique, *Journal de la Société Française de Statistique* (2009), **150**(2), 84-104, [see paper].

Preprints

[RS20] P. Tardivel, R. Servien and D. Concordet. A powerful multiple testing procedure in linear Gaussian model. *Submitted,* [see paper].

[RS21] G. Chagny, T. Laloë and R. Servien. Multivariate adaptive warped kernel estimation. *Submitted*, [see paper].

PhD thesis

[RS22] R. Servien. Estimation de régularité locale, Thèse de l'Université Montpellier II (2010), [see PhD].

Software

[RS23] R package **SISIR** on CRAN.

[RS24] R package **ASICS** on Bioconductor.

Introduction

In theory, this kind of manuscript is expected to give a smart and unified synthesis between all our past research works. So, I have to found a logical link between :

- the regulariy index of a probability measure ;
- the estimation of level sets ;
- the warped estimation of the regression function ;
- a clustering problem over the multivariate circle ;
- the construction of multivariate prediciton region using a mixed linear model ;
- the definition and the control of some sparse procedures on functional data ;
- and other more applicative works in metabolomics or on pollutants.

Some boxes are easy to build (clustering problems here, non-parametric estimation there) but, at the end, I still have three different chapters with very few statistical links. In fact, there is obviously a link but it is very different. This link is my way to produce research : never alone. My different works always start by a discussion with another human being in front of a glass of water (or anything but coffee). It is first a human choice. So, here I want to deeply thank all my co-workers and to recall that this manuscript would not exist without them.

This manuscript presents in a (tentative) synthetic fashion my scientific production developed during and after my PhD thesis, defended in March 2010 at the University Montpellier II. This dissertation is organized around three distinct but complementary themes :

- nonparametric estimation ;
- clustering ;
- statistical learning for functional data.

Note that the two lasts could be regrouped on a "Biostatistics" part as main of these works are inspired by an application in biology or medicine but I decided to keep three less general parts.

Some ongoing works or leads for future research are mentioned throughout the manuscript in each section or subsection of the different chapters. Nevertheless, to clarify and highlight which of them I'm going to invest in the short and medium term, each chapter is concluded with a research perspective section. In a sake of compactness all the different proofs, most of the simulations and some mathematical details are omitted. They could be found in the corresponding references.

In Chapter 1, my contributions to nonparametric statistics are synthetized. In this field of statistics, one of the main challenge is to define new estimators without much assumptions or restrictions. It was my starting point in the world of research, during my PhD under the

supervision of Alain Berlinet at the University of Montpellier. The main objective of my PhD was to extend some old convergence results using weaker assumptions, mainly the notion of the regularity index [RS09, RS17, RS18, RS19, RS22]. During my studies at Montpellier, I met Thomas Laloë (now Assistant Professor at Nice). During his PhD he worked on non-parametric problems, focusing on the estimation of level sets. During my postdoctoral years we started working together on this very wide subject. Our first work was the estimation of the level sets of the regression function [RS14]. Then we were interested in the estimation of the level sets of the distribution function. After discussions with Elena Di Bernardino (Assistant Professor at CNAM) we made a link between this problem and the risk theory, with an application to an hydrological issue [RS11]. This work raised an issue about estimators without compacity assumptions on their support. During a conference, a talk by Gaëlle Chagny (CNRS researcher at Rouen) caught our attention as she used warped estimator to address this issue for other cases. So, we adapt her estimator to our specific problem of regression function estimation [RS21].

Chapter 2 is dedicated to my publications in the field of clustering. In fact, my first research problem was a clustering one, during my Master internship under the supervision of Christophe Abraham (Professor at SupAgro Montpellier) and Nicolas Molinari (Professor at the University of Montpellier). The problem was to cluster non-ordered circular multivariate data obtained from radiotherapy x-beams bouquets. During my postdoctoral years, we reworked on this problem and made a first publication using a frequentist approach [RS16] and, nowadays, a second one using a Bayesian approach [RS02] has been accepted. During his PhD, Thomas Laloë defined a L^1-based clustering algorithm. Nevertheless, the computation of this algorithm was intractable. To overcome this problem, we define a parameter-free clustering method based on his algorithm [RS15]. During my master internship I met Virginie Rossard who was also in master internship and then was recruited as an assistant engineer at the LBE INRA unit at Narbonne. During my postdoctoral years I visited the LBE unit and we started talking about a project they have with Eric Latrille on the clustering of micropollutants. That's how I started to work with them on this project, that leads us to one publication to explain the dedicated clustering approach [RS13] and three about applications of this approach on micropollutants [RS04, RS08, RS10] in collaboration with Laure Mamy and Pierre Benoit (INRA Versailles).

We could see here an important shift in my research interests : starting from very theoretical works during my PhD (without any datasets or potential applications) and continuing mainly with statistical problems driven by applications (radiotherapy, clustering of pollutants) or, at least, applicable to some datasets (hydrological one for example). I think that it is, for me, the most interesting part in statistical research : starting from an applicative problem and then defining and studying an *ad hoc* statistical procedure addressing this problem. That is why I was very enthusiastic when I was recruited as a permanent researcher at the INRA Toulouse, in a unit with a lot of biologists. Since my recruitment and thanks to this very stimulating workplace, I am involved in very interesting projects that mixed problems in an application domain and statistics. In this spirit, the area of the omics (more precisely metabolomics) or the precision livestock farming are very promising : as they are based on new technologies in constant evolution they constantly raised new problems in their data analysis, mainly in the field of functional data analysis. Indeed, longitudinal follow-up, metabolomic spectrum or daily measurements can be viewed as functional data. These different questions are at the thematic center of my different INRA unities (Toxalim until 2018 then InTheRes) and are mentioned

on Chapter 3. On these problems, I mainly collaborate with Didier Concordet (Professor at Toulouse Vet School (ENVT)). The first question we addressed was the building of multivariate individual prediction regions for functional data based on a mixed effect model. This question is of high interest in the actual field of individualized medicine for humans or animals [RS12] and mixed effect models are also widely used in our unit to build pharmacokinetic models [RS06]. The question of variable selection/multiple testing is also of major interest in this field. It was the main subject of the PhD of Patrick Tardivel (now holding a postdoctoral position at the University of Wroclaw) that I supervised with Didier Concordet. The dedicated statistical procedure [RS20] was driven by an application for the idenfication and quantification of metabolites in metabolomics [RS05, RS07] and leaded us to a theoretical problem about the L^0-norm minimization [RS03]. Another variable selection problem for functional data was also studied with Nathalie Villa-Vialaneix and Victor Picheny (researchers at the MIAT INRA unit at Toulouse). The problem was to select interesting (but no predefined) intervals on functional data to predict a variable of interest and was driven by an application in smart farming (*i.e.* predict the yield of a field given the temperature, the rainfall ...)[RS01].

Nowadays, all my research projects shared a common methodology. First, I investigate the applied question by trying to fully understand the nature and the type of data. Second, I translate this problem in statistical terms trying to be as close as posible to the initial applied problem. Then, I develop and study a statistical procedure to address this statistical problem. Finally, I test my statistical approach on the applied question trying to analyze which part of the problem are solved and which are not. That is why I mainly define myself as a biostatistician now and that is how I enjoy research.

Chapitre 1

Contributions to non-parametric estimation

1.1 Introduction

As briefly explained in the general introduction, Section 1.2 of this chapter is devoted to results obtained during my PhD (or just after) on the regularity index [RS09, RS17, RS18, RS19, RS22]. This notion of regularity index, which is weaker than the notion of continuity, help us to extend some well-known convergence results. For example, we provide a necessary and sufficient condition for having a limit distribution for the nearest neighbor density estimate. The results of the Section 1.3 are related to the estimation of level sets. A kernel estimator of the level sets of the regression function is first defined and studied [RS14]. This estimator is simpler and has weaker assumption than the existing one. Then, we studied the level sets of the distribution function with the additional problem of the non-compacity of these level sets. An associated multivariate risk measure is also studied on these level sets [RS11]. Section 1.4 is devoted to the estimation of the regression function without any compacity assumption on the support. To achieve this goal, we defined and studied a warped estimator [RS21].

1.2 Regularity index

1.2.1 General framework

The problem of estimating the probability density from a sample $(X_i)_{1\leq i\leq n}$ has received considerable attention in the literature : Many methods have been developed such as histograms [Ioannidis, 2003], kernel estimators [Nadaraya, 1964, Watson, 1964], statistically equivalent blocks [Gessaman, 1970], the Barron estimator [Barron, 1988] ... For reviews on this subject we refer the interested reader to Silverman [1986], Scott [1992], Hastie et al. [2009]. My PhD work finds its motivations on the study of estimation problems, when usual regularity assumptions are not verified. Indeed, a lot of convergence results are based on some continuity assumptions that could not be checked in practice and that could be weakened. In this purpose, I studied the regularity index of a probability measure applied to some nonparametric estimation problem where it could be useful.

Let μ be a probability distribution and λ be the Lebesgure measure on $\mathbb{R}^d$ equipped with the Euclidean norm $||.||$. We denote by $B(x,\delta)$ the open ball with center at x and radius δ. To

evaluate the local behaviour of $\mu(B(x,\delta))$ in relation to $\lambda(B(x,\delta))$ one can consider the ratio of these two quantities. If, for fixed x, the following limit

$$f(x) = \lim_{\delta \to 0} \frac{\mu(B(x,\delta))}{\lambda(B(x,\delta))} \tag{1.1}$$

exists and is finite, then x is called a Lebesgue point of the measure μ [Rudin, 1987, Dudley, 1989]. If μ is absolutely continuous with respect to λ, we can select a specific density f that checks (1.1) where this limit exists. In Berlinet and Levallois [2000], examples where the density has a bad local behaviour at Lebesgue points are examined. To evaluate rates of convergence or investigate asymptotic normality of estimators, not only the convergence of the ratio of ball measures is required but also information on its higher order behaviour. In this context, Berlinet and Levallois [2000] define a ρ-regularity point of the measure μ as any Lebesgue point x of μ satisfying

$$\left| \frac{\mu(B(x,\delta))}{\lambda(B(x,\delta))} - f(x) \right| \leq \rho(\delta), \tag{1.2}$$

where ρ is a measurable fonction such that $\lim_{\delta \downarrow 0} \rho(\delta) = 0$. To specify an exact rate of convergence of the ratio of ball measures, Beirlant et al. [2008] assumed that a more precise relation than (1.2) holds at the Lebesgue point x ; namely

$$\frac{\mu(B(x,\delta))}{\lambda(B(x,\delta))} = f(x) + C_x \delta^{\alpha_x} + o(\delta^{\alpha_x}) \text{ when } \delta \downarrow 0, \tag{1.3}$$

where C_x is a non-zero constant and α_x is a positive real number called *regularity index*. These constants are unique (provided they exist). The index α_x controls the degree of smoothness of the symmetric derivative of μ with respect to λ. The larger the value of α_x , the smoother the derivative of μ is at the point x (see examples in Berlinet and Levallois [2000]). Note that (1.3) is clearly equivalent to the small ball probability expansion :

$$P(||X - x|| \leq \delta) = V_d \delta^d (f(x) + C_x \delta^{\alpha_x} + o(\delta^{\alpha_x})),$$

where X has density f and $V_d = \pi^{d/2}/\Gamma(1 + d/2)$ denotes the volume of the unit ball in $\mathbb{R}^d$. In other words, the second-order term in the expansion of the small ball probability of radius δ at x is equal, up to a multiplicative constant, to $\delta^{d+\alpha_x}$.

Nevertheless, the definition (1.3) suffers some flaws. First, some measures with ρ-regularity have no regularity index α_x, for example if in (1.3) we replace δ^{α_x} by $\log(\delta)$. Second, many density estimates require a development for a ratio of set measures which are not centered around the estimation point x and which are not balls. The definition of the regularity index is useless in these cases. These flaws represent a major restriction in practice, since we can not obtain similar results for an estimate such as the histogram, even for measures that could have a regularity index α_x. To circumvent these problems, we propose the following definition. Given $x \in \mathbb{R}$ we set $\mathcal{I}_x$ the set of all the intervals which contain x and we define E_x by

$$E_x = \Big\{ r > 0 \text{ such that } \exists C > 0, \exists \lambda_0 > 0, \text{ such that } \forall I \in \mathcal{I}_x$$
$$\text{verifying } \lambda(I) < \lambda_0 \text{ we have } \left| \frac{\mu(I)}{\lambda(I)} - f(x) \right| \leq C \lambda(I)^r \Big\}.$$

If there exists a real r_x such that

$$r_x = \sup E_x, \tag{1.4}$$

r_x is the r-regularity index of the measure μ at x. If $\sup E_x = +\infty$, we set $r_x = +\infty$.

With this definition, the r-regularity can be viewed as an intermediate stage between the ρ-regularity and the regularity index : it gives us a bound for the rate of convergence of the measures. Furthermore, the r-regularity does not involve a ball centered on x and, consequently, can be used with a larger class of density estimates. Note that, as for the regularity index, the larger the value of r_x, the more regular the derivative of μ with respect to λ.

1.2.2 Limit distribution for density estimators

Here, we shall consider the well-known nearest-neighbour estimator f_{k_n} [Loftsgaarden and Quesenberry, 1965] defined by

$$f_{k_n}(x) = \frac{k_n}{n\lambda(\overline{B}_{k_n}(x))},$$

where $\overline{B}_{k_n}(x)$ is the smallest closed ball with center x containing at least k_n sample point. The estimate $f_{k_n}(x)$ is the ratio of the frequency of sample points falling into $\bar{B}_{k_n}$ to the Lebesgue measure of $\bar{B}_{k_n}$. The integer k_n plays the role of a smoothing parameter : When it is chosen too large, the data are oversmoothed ; they are undersmoothed in the opposite case. The choice of k_n is by consequence critical. Different papers [Loftsgaarden and Quesenberry, 1965, Moore and Yackel, 1977, Mack, 1980, van Es, 1992] states consistency results for f_{k_n} based on a global convergence hypotheses for k_n (*i.e.* the same on the whole definition domain) and the hypothesis of a continuous density function. Then, Berlinet and Levallois [2000] states the asymptotic normality of f_{k_n} in cases where the density has a bad local behaviour using definition (1.2). We take advantage of the definition of the regularity index to extend this result and to obtain the following theorem.

Theorem 1.2.1 *Suppose that x is a Lebesgue point where (1.3) is satisfied with $f(x) > 0$. Then, under the conditions $\lim_{n\to\infty} k_n = \infty$ and $\lim_{n\to\infty} k_n/n = 0$, as n tends to infinity,*

$$T_n(x) = \sqrt{k_n}\frac{f_{k_n}(x) - f(x)}{f(x)}$$

converges in distribution if and only if the sequence

$$\left(\frac{k_n^{1+1/2\alpha_x}}{n}\right)$$

has a finite limit κ. When the last condition is satisfied, the asymptotic law of $T_n(x)$ is

$$\mathcal{N}\left(\frac{C_x\kappa^{\alpha_x}}{2^{\alpha_x}}\left(\frac{1}{f(x)}\right)^{\alpha_x+1}, 1\right).$$

This result provides a necessary and sufficient condition for having a limit distribution and explicitly gives this distribution when it does exist. As expected, what is important is the local behaviour of the associated measure, more precisely, the rate at which the derivative of the underlying measure is approximated by ratios of ball measures and its estimation by the regularity index. This rate as a strong impact on the choice of the number of neighbors k_n. Thus, this choice has to be made locally with great care and, whenever the set of data is large enough, a preliminary estimation of α_x (as developed in the next section) is strongly recommended.

As already explained in Subsection 1.2.1 , the definition of the regularity index has some flaws and could not be used for density estimators such has histograms. The histograms are nevertheless probably the oldest and simplest method to estimate an unknown density. The simplest histogram methods partition the space into congruent intervals or cubes whose size and position depends on the number of available data points, but not on the data itself [Ioannidis, 2003].

A histogram f_h consists of a partition of the space $\mathbb{R}$ of Borel-measurable subsets of $\mathbb{R}$, referred to as cells. We consider here partitions with the same size h_n such that

$$B_{nq} = [(q-1)h_n, qh_n[, q \in \mathbb{Z}$$

with the property that (i) $\cup_{q\in\mathbb{Z}} B_{nq} = \mathbb{R}$ and (ii) $B_{nq} \cap B_{nq'} = \emptyset$ if $q \neq q'$. Using these notations, the histogram estimate is

$$f_h(x) = \frac{\nu_{nq}}{nh_n}$$

with $x \in B_{nq}$ and ν_{nq} the number of X_i in the B_{nq} cell. By consequence, the function f_h is constant in a cell. So, to obtain the consistency of f_h towards f, the cells need to become smaller and smaller with n. Asymptotic results have been derived under conditions on the sequence $(h_n)_{1\leq i\leq n}$ with a continuity assumption on the density to estimate [Stadtmüller, 1983, Devroye and Györfi, 1985, Bosq and Lecoutre, 1987]. In Theorem 1.2.2 below we state the asymptotic normality of the histogram estimate of the density function, removing this continuity assumption by using the r-regularity index defined in (1.4).

Theorem 1.2.2 *Suppose that x is a Lebesgue point in $\mathbb{R}$ where (1.4) is satisfied with $f(x) > 0$. Then the condition*

$$\lim_{n\to\infty} nh_n^{2r+1} = 0 \tag{1.5}$$

for some $r \in]0; r_x[$ implies that

$$H_n(x) = \sqrt{nh_n}\frac{f_h(x) - f(x)}{\sqrt{f(x)}}$$

converges in distribution towards a centered gaussian distribution with unit variance.

A major point is that we obtain the asymptotic normality of the histogram without a continuity assumption on the density function f at the point of estimation x. Nevertheless, this result provides a necessary condition for having a limit distribution, but not a sufficient one. This comes from the fact that, unlike the regularity index, the r-regularity does not provide us with an exact rate, but only an upper bound of the rate. This definition could be used with other density estimates but, to my knowledge, no other asymptotic results has been stated using it.

1.2.3 Estimation of the regularity index

A nice estimation of the regularity index is needed to check the previous conditions of Theorem 1.2.1. The only estimate available was the one of Beirlant et al. [2008] based on k_n nearest neighbor density estimate. They define their estimate $\bar{\alpha}_{n,x}$, whatever $\tau > 1$, as

$$\bar{\alpha}_{n,x} = \frac{d}{\log\tau}\log\frac{f_{\lfloor\tau^2 k_n\rfloor}(x) - f_{\lfloor\tau k_n\rfloor}(x)}{f_{\lfloor\tau k_n\rfloor}(x) - f_{\lfloor k_n\rfloor}(x)},$$

if $[f_{\lfloor\tau^2 k_n\rfloor}(x) - f_{\lfloor\tau k_n\rfloor}(x)]/[f_{\lfloor\tau k_n\rfloor}(x) - f_{\lfloor k_n\rfloor}(x)] > 0$ and $\bar{\alpha}_{n,x} = 0$ otherwise, with $\lfloor . \rfloor$ the floor function. This estimate is proven to be consistent in probability and its asymptotic normality is exhibited. Their result are stated under the assumption of absolute continuity of the measure μ with respect to the Lebesgue measure. Inspired by this previous estimate, we defined a new estimate for the regularity index based on an empirical one. The empirical measure μ_n associated with $X_1, \ldots, X_n$ is defined by

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(\mathbf{X_i}\in\mathbf{A})}, \qquad \mathbf{A} \subseteq \mathbb{R}^{\mathbf{d}}$$

where

$$\mathbf{1}_{(\mathbf{X_i}\in\mathbf{A})} = \begin{cases} 1 \text{ if } X_i \in A \\ 0 \text{ otherwise,} \end{cases}$$

and the associated empirical estimator of

$$\varphi_{n,\delta} = \frac{\mu_n(B(x,\delta))}{\lambda(B(x,\delta))}.$$

This estimate is very simple as it does not need the calibration of any parameter. Using this estimate, we define, whatever $\tau > 1$,

$$\widehat{\alpha}_{n,x} = \frac{1}{\ln \tau} \ln \frac{\varphi_{n,\tau^2\delta_n}(x) - \varphi_{n,\tau\delta_n}(x)}{\varphi_{n,\tau\delta_n}(x) - \varphi_{n,\delta_n}(x)}$$

and state the following results.

Theorem 1.2.3 *Suppose that $x \in \mathbb{R}^d$ is a Lebesgue point of μ with regularity index α_x.*

- *Then, under the conditions*

$$\lim_{n\to\infty} \delta_n = 0 \quad \text{and} \quad \lim_{n\to\infty} n\delta_n^{d+2\alpha_x} = +\infty$$

the empirical estimator $\widehat{\alpha}_{n,x}$ converges to α_x in probability.

- *Then, under the conditions*

$$\lim_{n\to\infty} \delta_n = 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{n\delta_n^{2(d+\alpha_x)}}{\log n} = \infty$$

the empirical estimator $\widehat{\alpha}_{n,x}$ converges to α_x almost surely.

Note that these results does not need the absolute continuity of the measure μ with respect to the Lebesgue measure. Simulations show the good performances of this estimate and it needs for large datasets. Note that using the known estimates of the regularity index, a bound could be trivially obtained for the r-regularity index.

According to the specific expression of $\widehat{\alpha}_{n,x}$, one can guess that a convergent estimate of the distribution function can lead us to a new convergent estimate of the regularity index, under appropriate conditions. By consequence, a large bibliography on the estimator of the distribution function was made, with for example, the spline estimate [Berlinet, 1981, Restle, 2001], the support vector machines [Mohamed and Farag, 2004, Mohamed et al., 2004], the level-crossing [Huang and Brill, 2004], the iterated function systems [Iacus and La Torre, 2005] ... Nevertheless, to my knowledge, the two presented estimators of the regularity index are still the only ones that have been studied in details. In a future work, comparing the different estimators obtained using the review would be of interest.

1.3 Estimation of level sets

The estimation of level sets of an interest function has been widely studied in the literature. In particular, for the estimation of density level sets, one can cite for example the work of Polonik [1995], Tsybakov [1997], Cuevas and Fraiman [1997], Baíllo [2003], Biau et al. [2007], Cadre [2006], Rigollet and Vert [2009] ... This large number of works on this subject is motivated by the high number of possible applications. Estimating these level sets can be useful in mode estimation [Polonik, 1995], or in clustering [Biau et al., 2007] to estimater the number of clusters for example.

The same applications are possible with the regression function. Moreover, it is for instance possible to use an estimator of the level sets of the regression function to determine the path of water flow from a digital representation of an area. In the same vein, in medical imaging, a lot of applications exist. For example, people want to estimate the areas where some function of the image exceeds a fixed threshold. For instance, the severity of the cancer is characterized by a variable Y which directly impacts the choice of standard or aggressive chemotherapy. For osteosarcoma [Man et al., 2005], Y is the percent necrosis in the tumor after a first round of treatment. If $Y > 0.9$ (this threshold has been fixed by experts and is now the convention), the aggressive chemotherapy will be chosen. The problem is that Y is measured using an invasive biopsy. If we can collect from the patient a feature vector X (which acquisition is easier), such as gene expression levels or a magnetic resonance image, knowledge of the regression level sets would allow the choice of an efficient treatment planning without a biopsy.

1.3.1 Level sets of the regression function

We first consider the problem of estimating the level sets of a regression function. More precisely, we consider a random pair (X, Y) taking values in $\mathbb{R}^d \times J$, where $J \subset \mathbb{R}$ is supposed to be bounded. The goal of our work was then to build a simple estimator of the level sets of the regression function r of Y on X, defined for all $x \in \mathbb{R}^d$ by

$$r(x) = \mathbb{E}\left[Y|X = x\right].$$

For $t > 0$, a level set for r is defined by

$$\mathcal{L}_r(t) = \{x \in \mathbb{R}^d : r(x) > t\}.$$

Assume that we have an independent and identically distributed sample (i.i.d.) $((X_1, Y_1), \ldots, (X_n, Y_n))$ with the same distribution as (X, Y). We then consider a plug-in estimator of $\mathcal{L}_r(t)$. More precisely, we use a consistent estimator $\hat{r}_n$ of r, in order to estimate $\mathcal{L}_r(t)$ by

$$\mathcal{L}_{\hat{r}_n}(t) = \{x \in \mathbb{R}^d : \hat{r}_n(x) > t\}.$$

Despite the many potential applications, the estimation of the level sets of the regression function has not been widely studied. Müller and Sawitzki [1991] mentioned it briefly in his survey. Nowak and Willett [2007] obtained minimax rates (for different smoothness classes) for estimators based on recursive dyadic partitions. Scott and Davenport [2007] use a cost sensitive approach and a different measure of risk. Cavalier [1997], Polonik and Wang [2005]

used estimators based on the maximization of the excess mass which was introduced by Hartigan [1987]. Cavalier demonstrated asymptotic minimax rate of convergence for piecewise polynomial estimators using smoothness assumptions on the boundary of the level sets. We used a different approach and construct a plug-in estimator using the kernel estimator of the regression. The main advantage of our estimator is the simplicity of his calculation, inherited from the plug-in approach. Moreover, our estimator does not require strong assumptions on the shape of level sets. All our consistency results are in the sense of the volume (in the Lebesgue measure sense) of the symmetrical difference, defined by

$$\lambda\left(\mathcal{L}_{\hat{r}_n}(t)\Delta\mathcal{L}_r(t)\right) = \lambda\left[\left(\mathcal{L}_{\hat{r}_n}(t)\cap\mathcal{L}_r^C(t)\right)\cup\left(\mathcal{L}_{\hat{r}_n}^C(t)\cap\mathcal{L}_r(t)\right)\right]$$

where λ stands for the Lebesgue measure on $\mathbb{R}^d$ and Δ for the symmetrical difference.

Our goal is to establish some consistency results under reasonable assumptions on r and $\hat{r}_n$. Using a kernel estimator for r, we get a rate of convergence equivalent to the one obtained for the density function [Cadre, 2006].

Construction of the estimator

As announced, we use a plug-in approach. That is, given an estimator r_n of r we estimate $\{x \in \Lambda : r(x) > t\}$ by $\{x \in \Lambda : r_n(x) > t\}$. To estimate r, we choose to consider a kernel estimator.

Assume that we can write

$$r(x) = \frac{\varphi(x)}{f(x)},$$

where f is the density function of X, and φ is defined by $\varphi(x) = r(x)f(x)$.

Let K be a kernel on $\mathbb{R}^d$, that is a probability density on $\mathbb{R}^d$. We denote $h = h_n$ and $K_h(x) = K(x/h)$. From an i.i.d. sample $\Big((X_1, Y_1), \ldots, (X_n, Y_n)\Big)$, we define, for all $x \in \mathbb{R}^d$,

$$\varphi_n(x) = \frac{1}{nh^d}\sum_{i=1}^{n} Y_i K_h(x - X_i) \text{ and } f_n(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K_h(x - X_i).$$

For all $x \in \mathbb{R}^d$, the kernel estimator of r is then defined by

$$r_n(x) = \begin{cases} \varphi_n(x)/f_n(x) & \text{if } f_n(x) \neq 0 \\ 0 \text{ otherwise.} \end{cases}$$

The properties of this estimator are already well studied in the literature [Gasser and Müller, 1984, Bosq and Lecoutre, 1987].

Under the assumption

A0 There exists $t^- < t$ such that $\mathcal{L}_r(t^-)$ is compact. Besides, $\lambda(\{r = t\}) = 0$ (where λ stands for the Lebesgue measure),

a first consistency result can be obtained.

Proposition 1.1 *Under Assumption* **A0**, *if* K *is bounded, integrable, with compact support and Lipschitz, and if* $h \to 0$ *and* $nh^d / \log n \to \infty$, *then*

$$\mathbb{E}\,\lambda\Big(\mathcal{L}_{\hat{r}_n}(t)\Delta\mathcal{L}_r(t)\Big) \underset{n\to\infty}{\to} 0.$$

Note that the last part of assumption **A0** means that the regression function can not have a null derivative at the estimated level set.

Rate of convergence

From now on, $\Theta \subset (0, \sup_{\mathbb{R}^d} r)$ is an open interval. Let us introduce the following assumptions :

A1 The functions r and f are twice continuously differentiable, and, $\forall t \in \Theta\,, \exists 0 < t^- < t$: $\inf_{\mathcal{L}(t^-)} f > 0$;

A2 For all $t \in \Theta$,

$$\inf_{r^{-1}(\{t\})} \|\nabla r\| > 0,$$

where, $\nabla\psi(x)$ stands for the gradient at $x \in \mathbb{R}^d$ of the differentiable function $\psi : \mathbb{R}^d \to \mathbb{R}$.

The assumptions **A1** on the regularity are inherited from the classical assumptions in kernel estimation [Bosq and Lecoutre, 1987]. Note that stronger assumptions on the regularity of r and f will not improve the obtained rate of consistency. Moreover, let us mention that under Assumptions **A1** and **A2**, we have (Proposition $A.2$ in Cadre [2006])

$$\forall t \in \Theta : \quad \lambda(r^{-1}[t-\varepsilon, t+\varepsilon]) \to 0 \quad \text{as } \varepsilon \to 0.$$

We are now in a position to establish a rate of convergence for $\mathbb{E}\,\lambda(\mathcal{L}_{\hat{r}_n}(t)\Delta\mathcal{L}_r(t))$.

Theorem 1.3.1 *Under Assumptions* **A0** $-$ **A2** *and some assumptions on* K, *if* $nh^d/(\log n) \to \infty$ *and* $nh^{d+4}\log n \to 0$, *then for almost all* $t \in \Theta$

$$\mathbb{E}\,\lambda(\mathcal{L}_{\hat{r}_n}(t)\Delta\mathcal{L}_r(t)) = O(1/\sqrt{nh^d}).$$

Remarks :

- Roughly speaking, the assumptions about the bandwidth impose to take h between $(\frac{\log n}{n})^{\frac{1}{d}}$ and $(n\log n)^{\frac{-1}{d+4}}$. Moreover, if we take $h = O((n\log n)^{\frac{-1}{d+4}})$, we get

$$\sqrt{nh^d} \;=\; O\left(\frac{n^{1/3}}{(\log n)^{1/6}}\right) \text{ with } d = 2,$$

that is a rate of the same order as Cadre [2006] in the density case.

- A remaining and crucial problem is the research of an optimal bandwidth h for our estimator. Indeed, if they are already results in the literature about an optimal bandwidth for the estimation of r, this bandwidth is not necessarily optimal for estimating $\mathcal{L}_r(t)$. A data-driven adaptive procedure using a Goldenshluger-Lepski approach [Goldenshluger and Lepski, 2011] (such as in Subsection 1.4) would be of great interest. However, in a fist time, we used a cross-validation procedure to choose the bandwidth in the simulations.

According to the symmetrical difference, the estimator of Cavalier [1997] is proven to be optimal. Nevertheless, this estimator has some major drawbacks : it is always star-shaped and it is rather difficult (and more often impossible) to calculate without any *a priori* knowledge on the dataset.

Note that Mason and Polonik [2009] obtained the asymptotic normality of plug-in level set estimates in the density case, it would be interesting to see if we could extend their result to this regression framework. Another interesting future work will be to replace the level t by an estimated level t_n and to study how the convergence rate is affected by this new plug-in estimate.

1.3.2 Level sets of the multivariate cumulative distribution function

All previous works on the consistency of the level sets are based on a compactness assumption. But, in the case of the cumulative distribution function, this assumption seems no more reasonable and we have to deal with this non-compact setting. Considering a consistent estimator F_n of the distribution function F, we propose a plug-in approach to estimate

$$\mathcal{L}_F(t) = \{x \in \mathbb{R}_+^d : F(x) \geq t\},$$

by

$$\mathcal{L}_{F_n}(t) = \{x \in \mathbb{R}_+^d : F_n(x) \geq t\}$$

for $t \in (0,1)$. As remarked above, to deal with this non-compact setting, we define, given $T > 0$,

$$\mathcal{L}_F(t)^T = \{x \in [0,T]^d : F(x) \geq t\}, \;\; \mathcal{L}_{F_n}(t)^T = \{x \in [0,T]^d : F_n(x) \geq t\}.$$

Using these notations, we establish our consistency result with a convergence rate. The following theorem can be interpreted as a generalization of the results of Cuevas et al. [2006] in the case of non-compact level sets.

Theorem 1.3.2 *Let $t \in (0,1)$. Let $F \in \mathcal{F}$ be a twice differentiable distribution function on $\mathbb{R}_+^{d*}$ satisfying some further regularity conditions on its gradient vector and its Hessian matrix. Assume that for each n, F_n is measurable. Assume that there exists a positive increasing sequence $(w_n)_{n\in\mathbb{N}^*}$ such that $w_n \|F - F_n\|_\infty \xrightarrow[n\to\infty]{\mathbb{P}} 0$. Then, it holds that*

$$p_n \, \lambda(\mathcal{L}_F(t)^{T_n}, \mathcal{L}_{F_n}(t)^{T_n}) \xrightarrow[n\to\infty]{\mathbb{P}} 0,$$

where the convergence rate p_n depends on w_n, T_n and d.

This theorem provides a convergence rate, which obviously suffers from the well-known curse of dimensionality and is closely related to the choice of the truncation sequence T_n. The review of the estimators of the distribution function already made for the first section of this chapter could then provide a wide range of estimators of these level sets. Obviously, a better result would have been

$$u_n \, \lambda(\mathcal{L}_F(t), \mathcal{L}_{F_n}(t)^{T_n}) \underset{n\to\infty}{\overset{\mathbb{P}}{\to}} 0,$$

but it is not possible to derive such results without strong assumptions on the tail behaviour of F. As we were focused on obtaining results without this kind of assumption, it has been kept for future work.

1.3.3 Estimation procedures for multivariate risk measures

In the last decade, much research has been devoted to the construction of risk measures that account both for marginal effects and dependence between risks and many extensions to multidimensional settings have been suggested [Jouini et al., 2004, Embrechts and Puccetti, 2006, Nappo and Spizzichino, 2009, Ekeland et al., 2009]. Traditionally, risk measures were thought of as mappings from a set of real-valued random variables to the real numbers. However, it is often insufficient to consider a single real measure to quantify risks, especially when the risk-problem is affected by other external risk factors. Note that the evaluation of an individual risk may strongly be affected by the degree of dependence amongst all risks. Modeling the dependency structure of multivariate data is helpful to obtain meaningful and accurate inference and prediction results in risk analysis.

An important univariate risk measure, based on the quantile notion, is the *Conditional-Tail-Expectation* (CTE) defined by

$$\mathrm{CTE}_t(X) = \mathbb{E}[\, X \,|\, X > Q_X(t) \,], \qquad \text{for } t \in (0,1).$$

This definition has recently been adapted to the multivariate case by Di Bernardino et al. [2013] and Cousin and Di Bernardino [2013]. It is constructed as the conditional expectation of a multivariate random vector given that the latter is located in the c-upper level set of the associated multivariate distribution function. In this sense this measure is essentially based on a "multivariate distributional approach". More precisely they define, for $i = 1, \ldots, d$ and for $t \in (0,1)$,

$$\mathrm{CTE}_t^i(\mathbf{X}) = \mathbb{E}[\, X_i \,|\, \mathbf{X} \in \mathcal{L}_F(t) \,], \tag{1.6}$$

where $\mathbf{X} = (X_1, \ldots, X_d)$ is a non-negative multivariate risk portfolio with distribution function F. In particular, Cousin and Di Bernardino [2013] proved that properties of the multivariate Conditional-Tail-Expectation in (1.6) turn to be consistent with existing properties on univariate risk measures (positive homogeneity, translation invariance, increasing in risk-level $t, \ldots$). We try to go further to study the behavior of a covariate Y on the level sets of a d-dimensional vector of risk-factors $\mathbf{X}$. More precisely, adapting the multivariate risk measure in (1.6), we deal with the multivariate Covariate-Conditional-Tail-Expectation (CCTE) defined by

Definition 1 *Consider a random vector* $\mathbf{X}$ *with distribution function* F *and a random variable* Y*. For* $t \in (0,1)$*, we define the theoretical multivariate* t*-Covariate-Conditional-Tail-Expectation as*

$$\mathrm{CCTE}_t(\mathbf{X}, Y) = \mathbb{E}\left[Y \,|\, \mathbf{X} \in \mathcal{L}_F(t) \right],$$

and its associated truncated estimate as

$$\widehat{\mathrm{CCTE}}_{t,n}^{T_n}(\mathbf{X}, Y) = \mathbb{E}_n \left[Y | \mathbf{X} \in \mathcal{L}_{F_n}(t)^{T_n} \right],$$

where $\mathbb{E}_n$ denotes the empirical version of the expected value.

Using these definitions, one can show the following result.

Theorem 1.3.3 *Let $t \in (0,1)$. Let $F \in \mathcal{F}$ be a twice differentiable distribution function on $\mathbb{R}_+^{d*}$ satisfying some further regularity conditions on its gradient vector and its Hessian matrix with an associated density f such that $||f||_{1+\varepsilon,\lambda} < \infty$ with $\varepsilon > 0$. Assume that for each n, F_n is measurable. Let $(v_n)_{n\in\mathbb{N}^*}$ and $(T_n)_{n\in\mathbb{N}^*}$ positive increasing sequences such that $v_n \int_{[0,T_n]^d} |F(x) - F_n(x)|^p \lambda(dx) \xrightarrow{\mathbb{P}} 0$, for some $1 \leq p < \infty$. It holds that*

$$\beta_n \left| \widehat{\mathrm{CCTE}}_{t,n}^{T_n}(\mathbf{X}, Y) - \mathrm{CCTE}_t^{T_n}(\mathbf{X}, Y) \right| \xrightarrow[n\to\infty]{\mathbb{P}} 0,$$

where the convergence rate β_n depends on v_n, T_n, d and on conditions for f.

Using this result, a tractable convergence rate in the case of the empirical distribution function F_n can be easily derived. This result is then applied to an engineering problem in the design of a sea defence [Hawkes et al., 2002]. The regression function $r(x) := \mathbb{E}[Y \mid \mathbf{X} = x]$ represents the relationship between the sea conditions X (*i.e.* significant wave height, still water level and the wave period) and the overtopping Y at a given time. We analyzed a dataset recorded on the Dutch coast during storm events [Draisma et al., 2004, Tau and Dam, 2011] and studied the mean overtopping discharge conditionnally to the sea variable conditions.

This application, as well as the theoretical results, highlights the importance of the parameter T_n (which helped solving the problem of the non-compactness of the level sets) as well as the curse of dimensionality. An interesting future work could be a deep investigation about these points, with a focus on the optimal choice for this parameter. Furthermore, the proposed methods are based on an *i.i.d.* samples framework. We remark that in real applications such as seasonal pattern in the temperature and water level rise series, data can have different types of serial correlations like nonlinear or non-stationary correlations [Fan and Yao, 2003].

1.4 Adaptive warped kernel estimation for multivariate regression

We have seen that a commonly shared assumption for regression analysis is that the support of $\mathbf{X}$ is a compact subset of $\mathbb{R}^d$ [Györfi et al., 2002, Guyader and Hengartner, 2013, Furer and Kohler, 2015]. To weaken this assumption, we could proceed as in the previous subsection or use the results of Kohler et al. [2009] that assume some smoothness properties on the regression function. In another hand, "warped" estimators have been developed [Yang, 1981, Kerkyacharian and Picard, 2004] and require very few assumptions on the support of $\mathbf{X}$. If we assume, in a sake of clarity, that $d = 1$, the warped method is based on the introduction of the auxiliary function $g = r \circ F_{\mathbf{X}}^{-1}$, where $F_{\mathbf{X}} : x \in \mathbb{R} \mapsto \mathbb{P}(\mathbf{X} \leq x)$ is the c.d.f. of the design $\mathbf{X}$. First, an estimator $\hat{g}$ is proposed for g, and then, the regression r is estimated using $\hat{g} \circ \hat{F}$, where $\hat{F}$ is the empirical c.d.f. of $\mathbf{X}$. This strategy has already been applied in the regression setting using projection

methods [Kerkyacharian and Picard, 2004, Pham Ngoc, 2009, Chagny, 2013] but also for other estimation problems (conditional density estimation, hazard rate estimation based on randomly right-censored data and c.d.f. estimation from current-status data, see *e.g.* Chesneau and Willer 2015, Chagny 2015). If the warping device permits to weaken the assumptions on the design support, the warped estimates also depend on a unique bandwidth, for $d = 1$, whereas the ratio form of the well-known Nadaraya-Watson kernel estimate, is defined by

$$\widehat{r}^{NW}(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\mathbf{x} - \mathbf{X}_i)}{\sum_{i=1}^{n} K_{\mathbf{h}}(\mathbf{x} - \mathbf{X}_i)}, \tag{1.7}$$

where $\mathbf{h} = {}^t(h_1, \dots, h_d)$ is the bandwidth of the kernel K, $K_{\mathbf{h}}(\mathbf{x}) = K_{1,h_1}(x_1)K_{2,h_2}(x_2)\dots K_{d,h_d}(x_d)$, with $K_{l,h_l}(x) = K_l(x/h_l)/h_l$ for $h_l > 0$, and $K_l : \mathbb{R} \to \mathbb{R}$ such that $\int_{\mathbb{R}} K_l(x)dx = 1$, $l = 1, \dots, d$.

So, this requires the selection of two smoothing parameters (one for the numerator, one for the denominator). In return, the c.d.f. $F_{\mathbf{X}}$ of $\mathbf{X}$ has to be estimated, but this can simply be done using its empirical counterpart. This does not deteriorate the optimal convergence rate, since this estimate converges at a parametric rate. A data-driven selection of the unique bandwidth involved in the resulting warped kernel estimator, in the spirit of Goldenshluger and Lepski [2011] leads to non-asymptotic risk bounds when $d = 1$ [Chagny, 2015]. To our knowledge, this adaptive estimation has never been carried out for a ratio regression estimator, the only reference on this subject is Ngoc Bien [2014] who assumes that the design $\mathbf{X}$ has a known uniform distribution.

Multivariate warping strategy

If $d = 1$, the warping device is based on the transformation $F_{\mathbf{X}}(X_i)$ of the data X_i, $i = 1, \dots, n$. For $d > 1$, a natural extension is to use $F_l(X_{l,i})$, for $l = 1, \dots, d$ and $i = 1, \dots, n$, where F_l is the marginal c.d.f. of X_l. Let us introduce $\widetilde{F}_{\mathbf{X}} : \mathbf{x} = (x_l)_{l=1,\dots,d} \in \mathbb{R}^d \mapsto (F_1(x_1), \dots, F_d(x_d))$. Assume that $\widetilde{F}_{\mathbf{X}}^{-1} : \mathbf{u} \in [0,1]^d \mapsto (F_1^{-1}(u_1), \dots, F_d^{-1}(u_d))$ exists, and let

$$g = r \circ \widetilde{F}_{\mathbf{X}}^{-1},$$

in such a way that $r = g \circ \widetilde{F}_{\mathbf{X}}$. If we consider that the marginal variables X_l of $\mathbf{X}$ are independent, the estimator of Yang [1981] can immediately be adapted to the multivariate setting. We set

$$\mathbf{u} \mapsto \frac{1}{n} \sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\mathbf{u} - \widetilde{F}_{\mathbf{X}}(\mathbf{X}_i)) \tag{1.8}$$

to estimate g, and it remains to compound by the empirical counterpart of $\widetilde{F}_{\mathbf{X}}$ to estimate r. However, a dependence between the coordinates $X_{l,i}$ of $\mathbf{X}_i$ generally appears. The usual model for this dependence using a copula C and the c.d.f $F_{\mathbf{X}}$ of $\mathbf{X}$ can be written

$$F_{\mathbf{X}}(\mathbf{x}) = C(F_1(x_1), \dots, F_d(x_d)) = C(\widetilde{F}_{\mathbf{X}}(\mathbf{x})). \tag{1.9}$$

Denoting the copula density by c, we have

$$c(\mathbf{u}) = \frac{\partial^d C}{\partial u_1 \dots \partial u_d}(\mathbf{u}), \quad \mathbf{u} \in [0;1]^d,$$

and the density $f_{\mathbf{X}}$ of $\mathbf{X}$ can be expressed as

$$f_{\mathbf{X}}(\mathbf{x}) = c(\widetilde{F}_{\mathbf{X}}(\mathbf{x}))\prod_{l=1}^{d} f_l(x_l), \quad \mathbf{x} = (x_l)_{l=1,\dots,d} \in \mathbb{R}^d,$$

where $(f_l)_{l=1,\dots,d}$ are the marginal densities of $\mathbf{X} = (X_1, \dots, X_d)$. It can then be proved that the previous estimator given by (1.8) estimates cg and not g. As a consequence, we propose to set, as an estimator for g,

$$\widehat{g}_{\mathbf{h}}(\mathbf{u}) = \frac{1}{n\widehat{c}(\mathbf{u})}\sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\mathbf{u} - \widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{X}_i)), \quad \mathbf{u} \in [0,1]^d,$$

where $\widehat{c}$ is an estimator of the copula density. We denote by $\widehat{\widetilde{F}}_{\mathbf{X}} : \mathbb{R}^d \to [0;1]^d$ the empirical multivariate marginal c.d.f. :

$$\widehat{\widetilde{F}}_{\mathbf{X}} = (\widehat{\widetilde{F}}_{\mathbf{X},1}, \dots, \widehat{\widetilde{F}}_{\mathbf{X},d}), \ \widehat{\widetilde{F}}_{\mathbf{X},l}(x_l) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_{l,i} \le x_l}, \ x_l \in \mathbb{R}, l \in \{1, \dots, d\}, \tag{1.10}$$

and finally set

$$\widehat{r}_{\mathbf{h}}(\mathbf{x}) = \widehat{g}_{\mathbf{h}} \circ \widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{x}) = \frac{1}{n\widehat{c}(\widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{x}))}\sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{x}) - \widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{X}_i)) \tag{1.11}$$

to rebuild our target function r from the data. In the sequel, we denote by $\|\cdot\|$ the (unweighted) L^2-norm on $L^2(\mathbb{R}^d)$ and, more generally, by $\|\cdot\|_{L^p(\Theta)}$ the classical L^p-norm on a set Θ.

For the sake of clarity, we first consider the regression estimation problem with a known design distribution. In a first time, the copula density c and the marginal c.d.f. $\widetilde{F}_{\mathbf{X}}$ are consequently considered to be known. We first proved a first classical convergence result for $\widehat{r}_{\mathbf{h}(\beta)}$ that could achieved the usual convergence rate in multivariate nonparametric estimation provided that its bandwidth is carefully chosen. But the challenge of adaptive estimation is to propose a data-driven choice for the bandwidth that leads to an estimator with the same optimal convergence rate. So, using a Goldenshluger-Lepki approach, we then proved an oracle-type inequality that leads us to the following result.

Let $\mathcal{H}_n \subset (\mathbb{R}_+^*)^d$ be a finite bandwidth collection such that

$$\exists \alpha_0 > 0, \kappa_1 > 0, \sum_{\mathbf{h}\in\mathcal{H}_n} \frac{1}{h_1 \cdots h_d} \le \kappa_1 n^{\alpha_0}$$
$$\text{and } \forall \kappa_1 > 0, \exists C_0 > 0, \sum_{\mathbf{h}\in\mathcal{H}_n} \exp\left(-\frac{\kappa_1}{h_1 \cdots h_d}\right) \le C_0.$$

For example, $\mathcal{H}_n = \{k_1^{-1} \cdots k_d^{-1}, k_l \in \{1, \dots, \lfloor n^{1/r} \rfloor\}, \ l = 1, \dots, d\}$ satisfies them with $\alpha_0 = 2d/r$ and let $\widetilde{\mathbf{h}} \in \mathcal{H}_n$.

Corollary 1.4.1 *Under some technical assumptions we have*

$$\mathbf{E}[\|\widehat{r}_{\widetilde{\mathbf{h}}} - r\|_{f_X}^2] = O\left(n^{-\frac{2\bar{\beta}}{2\bar{\beta}+d}}\right),$$

where $\bar{\beta}$ is the harmonic mean of $\beta_1, \dots, \beta_d$: $d\bar{\beta}^{-1} = \beta_1^{-1} + \cdots + \beta_d^{-1}$.

Here the smoothness index β is not required : our estimator automatically adapts to unknown smoothness of the function cg and performs as the best estimator of the collection $(\widehat{r}_{\mathbf{h}})_{\mathbf{h}\in\mathcal{H}_n}$.

Technical assumptions are not reminded here in a sake of simplicity. Nevertheless, these assumptions are very common to derive such estimators [Autin et al., 2010, Comte and Lacour, 2013, Chagny, 2015] and additional assumptions on the copula are verified for copula such as the Frank one.

As explained previously, the estimator defined by (1.11) involves an estimator $\widehat{c}$ of the copula density c that was assumed to be known in the previous result. So, the question is now of copula density estimation. To be consistent with the previous kernel regression estimator already chosen, we propose to use the kernel estimator defined by Fermanian [2005]. Consider $\mathbf{b} = {}^t(b_1,\dots,b_d) \in (\mathbb{R}_+^*)^d$ a multivariate bandwidth, a kernel $W_{\mathbf{b}}(\mathbf{u}) = W_{1,b_1}(u_1)W_{2,b_2}(u_2)\dots W_{d,b_d}(u_d)$, with $W_{l,b_l}(u) = W_l(u/b_l)/b_l$ for $b_l > 0$, and $W_l : \mathbb{R} \to \mathbb{R}$ such that $\int_0^1 W_l(u)du = 1, l \in \{1,\dots,d\}$. Let us introduce

$$\widehat{c}_{\mathbf{b}}(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^{n} W_{\mathbf{b}}(\mathbf{u} - \widehat{\widetilde{F}}_{\mathbf{X}}(\mathbf{X}_i)), \quad \mathbf{u} \in [0,1]. \tag{1.12}$$

Using this definition and similar calculations than previously, we proved the same kind of oracle-type inequality with a data-driven bandwidth that could, under some technical conditions, automatically achieves the minimax optimal convergence rate.

Now we are in position to consider the general case of unknown copula density c to estimate the regression function r. The idea is to plug the kernel estimator $\widehat{c}_{\mathbf{b}}$ (defined by (1.12)) of c. We first consider the simpler case of fixed bandwidth, both for the regression and the copula estimators. Let us plug in $\widehat{r}_{\mathbf{h}}$ the estimate $\widehat{c}_{\mathbf{b}}$: for any $\mathbf{b},\mathbf{h} > 0$,

$$\widehat{r}_{\mathbf{h},\mathbf{b}}(\mathbf{x}) = \frac{1}{n\widehat{c}_{\mathbf{b}}(\widetilde{F}_{\mathbf{X}}(\mathbf{x}))}\sum_{i=1}^{n} Y_i K_{\mathbf{h}}(\widetilde{F}_{\mathbf{X}}(\mathbf{x}) - \widetilde{F}_{\mathbf{X}}(\mathbf{X}_i))\mathbf{1}_{\widehat{c}_{\mathbf{b}}(\widetilde{F}_{\mathbf{X}}(\mathbf{x}))\geq m_c/2}, \quad \mathbf{x} \in A. \tag{1.13}$$

This means that $\widehat{r}_{\mathbf{h},\mathbf{b}}(\mathbf{x}) = ((c\times\widehat{g}_{\mathbf{h}})/\widehat{c}_{\mathbf{b}}) \circ \widetilde{F}_{\mathbf{X}}(\mathbf{x})\mathbf{1}_{\widehat{c}_{\mathbf{b}}(\widetilde{F}_{\mathbf{X}}(\mathbf{x}))\geq m_c/2}$. We obtain the following upper-bound for our ratio estimator.

Proposition 1.2 *Under the same assumptions that Corollary 1.4.1, we have*

$$\mathbf{E}[\|\widehat{r}_{\mathbf{h},\mathbf{b}} - r\|_{f_{\mathbf{X}}}^2] \leq \frac{4M_c}{m_c^2} \quad \Big\{ 2M_c\mathbf{E}[\|\widehat{r}_{\mathbf{h}} - r\|_{f_{\mathbf{X}}}^2] + (2\|g\|_{L^\infty(\widetilde{F}_{\mathbf{X}}(A))}^2 + \|g\|_{L^2(\widetilde{F}_{\mathbf{X}}(A))}^2)\mathbb{E}\left[\|\widehat{c}_{\mathbf{b}} - c\|_{L^2([0,1]^d)}^2\right]\Big\}.$$

This risk has the order of magnitude of the worst risk between the risk of $\widehat{r}_{\mathbf{h}}$ and $\widehat{c}_{\mathbf{b}}$ which is not surprising, and we cannot expect to obtain a sharper bound for the plug-in estimator. We thus have to add smoothness assumptions both on the regression function and on the copula density to derive the convergence rate of the plug-in estimator.

The final step of this work would have logically been the proposition of a data-driven selection method for the bandwidth of the regression estimator computed with an adaptive copula estimate. This reflexion is under way and would probably implies a penalization due to the

plug-in, but is not straightforward at all.

And, as explained at the beginning of this section, this final warped estimator of the regression function could also be used to compute the CCTE in a non-compact setting.

1.5 Ongoing projects and prospects

Some prospects on each subject have been proposed in the dedicated sections or subsections. But, unfortunately, each days has only twenty-four hours and I will not be able to follow any of these ideas despite their interest. Here, I'm going to develop one of them, that is ongoing on the master internship of Hai Dang Dau (Polytechnic School) under the co-supervision of Thomas Laloë. As briefly mentioned in Subsection 1.3.1, this first result is just a step towards the asymptotic normality of the level sets of the regression function. Then, we hope to obtain a result that states that, under suitable (but the weakest possible) regularity assumptions, we have that

$$\kappa_n \lambda\Big(\mathcal{L}_{\hat{r}_n}(t) \Delta \mathcal{L}_r(t)\Big) \to \mathcal{Z},$$

where $\mathcal{Z}$ denotes a centered normal random variable with a standard deviation σ_Z, where κ_n depends on n, h_n and d and σ_Z is expressed using known quantities (the dimension, the regularity of the regression function or of the kernel ...). Based on Mason and Polonik [2009], who obtained this result for the special case of density functions, we pave the way to prove such results. To achieve this final goal, we first need to obtain, as Cadre [2006] for the density case, an exact limit for our convergence result stated previously in Proposition 1.1. It will also be interesting to study if this kind of asymptotic result could be extended to plug-in estimators of general level-sets (density, regression, distribution function of Subsection 1.3.2 ...) in the spirit of the approach of Cuevas et al. [2006].

Bibliography

F. Autin, E. Le Pennec, and K. Tribouley. Thresholding methods to estimate copula density. *Journal of Multivariate Analysis*, 101(1) :200–222, 2010.

A. Baíllo. Total error in a plug-in estimator of level sets. *Statistics & Probability Letters*, 65 (4) :411–417, 2003.

A. Barron. The convergence in information of probability density estimators. In *Proceedings of the International Symposium of IEEE on Information Theory*, Kobe, Japan, June 1988.

J. Beirlant, A. Berlinet, and G. Biau. Higher order estimation at lebesgue points. *Annals of the Institute of Statistical Mathematics*, 60 :651–677, 2008.

A. Berlinet. Convergence des estimateurs splines de la densité. *Publications de l'Institut de Statistique de l'Université de Paris*, 26 :1–16, 1981.

A. Berlinet and S. Levallois. Higher order analysis at Lebesgue points. In M. Puri, editor, *G. G. Roussas Festschrift - Asymptotics in Statistics and Probability*, pages 1–16. 2000.

G. Biau, B. Cadre, and B. Pelletier. A graph-based estimator of the number of clusters. *ESAIM : Probability and Statistics*, 11 :272–280, 2007.

D. Bosq and J.-P. Lecoutre. *Théorie de l'Estimation Fonctionnelle*. Economica, 1987.

B. Cadre. Kernel estimation of density level sets. *Journal of Multivariate Analysis*, 97(4) : 999–1023, 2006.

L. Cavalier. Nonparametric estimation of regression level sets. *Statistics*, 29(2) :131–160, 1997.

G. Chagny. Penalization versus Goldenshluger-Lepski strategies in warped bases regression. *ESAIM. Probability and Statistics*, 17 :328–358, 2013.

G. Chagny. Adaptive warped kernel estimators. *Scandinavian Journal of Statistics*, 42(2) : 336–360, 2015.

C. Chesneau and T. Willer. Estimation of a cumulative distribution function under interval censoring "case 1" via warped wavelets. *Communications in Statistics. Theory and Methods*, 44(17) :3680–3702, 2015.

F. Comte and C. Lacour. Anisotropic adaptive kernel deconvolution. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 49(2) :569–609, 2013.

A. Cousin and E. Di Bernardino. On multivariate extensions of Value-at-Risk. *Journal of Mulitivariate Analysis*, 119 :32–46, 2013.

A. Cuevas and R. Fraiman. A plug-in approach to support estimation. *The Annals of Statistics*, 25(6) :2300–2312, 1997.

A. Cuevas, W. González-Manteiga, and A. Rodríguez-Casal. Plug-in estimation of general level sets. *Australian & New Zealand Journal of Statistics*, 48(1) :7–19, 2006.

L. Devroye and L. Györfi. *Nonparametric density estimation : the L^1 view*. Wiley series in probability and mathematical statistics, New-York, 1985.

E. Di Bernardino, T. Laloë, V. Maume-Deschamps, and C. Prieur. Plug-in estimation of level sets in a non-compact setting with applications in multivariable risk theory. *ESAIM : Probability and Statistics*, 17 :236–256, 2013.

G. Draisma, H. Drees, A. Ferreira, and L. de Haan. Bivariate tail estimation : dependence in asymptotic independence. *Bernoulli*, 10(2) :251–280, 2004.

R. Dudley. *Real Analysis and Probability*. Chapman and Hall, New-York, 1989.

I. Ekeland, A. Galichon, and M. Henry. Comonotonic measures of multivariate risks. *Mathematical Finance*, 22(1), 2009.

P. Embrechts and G. Puccetti. Bounds for functions of multivariate risks. *Journal of Multivariate Analysis*, 97(2) :526–547, 2006.

J. Fan and Q. Yao. *Nonlinear time series*. Springer series in statistics. Springer, New York, NY, 2003.

J.-D. Fermanian. Goodness-of-fit tests for copulas. *Journal of Multivariate Analysis*, 95(1) : 119–152, 2005.

D. Furer and M. Kohler. Smoothing spline regression estimation based on real and artificial data. *Metrika*, 78(6) :711–746, 2015.

T. Gasser and H. Müller. Estimating regression function and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, 3 :171–185, 1984.

M. Gessaman. A consistent nonparametric multivariate density estimator based on statistically equivalent block. *The Annals of Mathematical Statistics*, 41 :1344–1346, 1970.

A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation : Oracle inequalities and adaptive minimax optimality. *The Annals of Statistics*, 39(3) :1608–1632, 2011.

A. Guyader and N. Hengartner. On the mutual nearest neighbors estimate in regression. *Journal of Machine Learning Research*, 14 :2361–2376, 2013.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer New York, 2002.

J. A. Hartigan. Estimation of a convex density contour in two dimensions. *Journal of the American Statistical Association*, 82(397) :267–270, 1987.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer, 2009.

J. Hawkes, B. P. Gouldby, J. A. Tawn, and M. W. Owen. The joint probability ofwaves andwater levels in coastal engineering design. *Journal of Hydraulic Research*, 40(3) :241–251, 2002.

M. Huang and P. Brill. A distribution estimation method based on level crossings. *Journal of Statistical Planning and Inference*, 124 :45–62, 2004.

S. Iacus and D. La Torre. A comparative simulation study on the IFS distribution function estimator. *Nonlinear Analysis : Real World Applications*, 6 :858–873, 2005.

Y. E. Ioannidis. The history of histograms (abridged). In *VLDB*, pages 19–30, 2003.

E. Jouini, M. Meddeb, and N. Touzi. Vector-valued coherent risk measures. *Finance and Stochastics*, 8(4) :531–552, 2004.

G. Kerkyacharian and D. Picard. Regression in random design and warped wavelets. *Bernoulli*, 10(6) :1053–1105, 2004.

M. Kohler, A. Krzyzak, and H. Walk. Optimal global rates of convergence for nonparametric regression with unbounded data. *Journal of Statistical Planning and Inference*, 139(4) : 1286–1296, 2009.

D. Loftsgaarden and C. Quesenberry. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, pages 1049–1051, 1965.

Y. Mack. Asymptotic normality of multivariate k-nn density estimates. *Sankhya*, 42 :53–63, 1980.

T. Man, M. Chintagumpala, J. Visvanathan, J. Shen, L. Perlaky, J. Johnson, N. Davino, J. Murray, L. Helman, W. Meyer, T. Triche, K. Wong, and C. Laus. Expression profiles of osteosarcoma that can predict response to chemotherapy. *Cancer Research*, 65 :8142–8150, 2005.

M. Mason and W. Polonik. Asymptotic normality of plug-in level set estimates. *Annals of Applied Probability*, pages 1108–1142, 2009.

R. Mohamed and A. Farag. Mean field theory for density estimation using support vector machines. *Seventh International Conference on Information Fusion, Stockholm*, pages 495–501, 2004.

R. Mohamed, A. El-Baz, and A. Farag. Probability density estimation using advanced support vector machines and the EM algorithm. *International Journal of Signal Processing*, 1 :260–264, 2004.

D. Moore and J. Yackel. Large sample properties of nearest neighbour density function estimates. In S. Gupta and D. Moore, editors, *Statistical Decision Theory and Related Topics II*. Academic Press, New-York, 1977.

D. Müller and G. Sawitzki. Excess mass estimates and tests for multimodality. *Journal of the American Statistical Society*, 86 :738–746, 1991.

E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9 : 141–142, 1964.

G. Nappo and F. Spizzichino. Kendall distributions and level sets in bivariate exchangeable survival models. *Information Sciences*, 179 :2878–2890, 2009.

N. Ngoc Bien. *Adaptation via des inéqualités d'oracle dans le modèle de régression avec design aléatoire*. PhD thesis, Université d'Aix-Marseille, 2014.

R. D. Nowak and R. M. Willett. Minimax optimal level-set estimation. *IEEE Transactions on Image Processing*, 16(12) :2965–2979, 2007.

T. M. Pham Ngoc. Regression in random design and Bayesian warped wavelets estimators. *Electronic Journal of Statistics*, 3 :1084–1112, 2009.

W. Polonik. Measuring mass concentrations and estimating density contour clusters–an excess mass approach. *The Annals of Statistics*, 23(3) :855–881, 1995.

W. Polonik and Z. Wang. Estimation of regression contour clusters : an application of the excess mass approach to regression. *Journal of Multivariate Analysis*, 94 :227–249, 2005.

E. Restle. *Estimating cumulative distributions by spline smoothing*. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2001.

P. Rigollet and R. Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, 15(4) :1154–1178, 2009.

W. Rudin. *Real and Complex Analysis*. McGraw-Hill, New York, 1987.

C. Scott and M. Davenport. Regression level set estimation via costsensitive classification. *IEEE Transaction on Signal Processing*, 55 :2752–2757, 2007.

D. Scott. *Multivariate Density Estimation. Theory, Practice and Visualization.* Wiley, New-York, 1992.

B. Silverman. *Density Estimation.* Chapman and Hall, London, 1986.

U. Stadtmüller. Asymptotic distributions of smoothed histograms. *Metrika*, 30 :145–158, 1983.

V. Tau and G. C. Dam. Preliminary design study. *Project group Flood Defence, Institute for Water and Environment Research*, pages 1–144, 2011.

A. B. Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3) :948–969, 1997.

C. van Es. Asymptotics for least squares cross-validation bandwidths in nonsmooth cases. *The Annals of Statistics*, pages 1647–1657, 1992.

G. S. Watson. Smooth regression analysis. *Sankhyā Series*, 26 :359–372, 1964.

S.-S. Yang. Linear combination of concomitants of order statistics with application to testing and estimation. *Annals of the Institute of Statistical Mathematics*, 33(1) :463–470, 1981.

Chapitre 2

Clustering of complex datasets

2.1 Introduction

Clustering consists in partitioning a data set into subsets (or clusters), so that the data in each subset share some common trait. Proximity is determined according to some distance measure. For a thorough introduction to the subject, we refer to Kaufman and Rousseeuw [1990], Xu and Wunsch [2005]. The origin of clustering goes back to decades, when some biologists and sociologists began to search for automatics methods to build different groups with their data. Today, clustering is used in many fields. For example, in medical imaging, it can be used to differentiate between types of tissue and blood in a three dimensional image. Market researchers use it to partition the general population of consumers into market segments and to better understand the relationships between different groups of consumers/potential customers. There are also many different applications in artificial intelligence, sociology, medical research, or political sciences.

In this chapter we present our contributions in this field. In Section 2.2, we define a new parameter-free clustering algorithm called X-Alter [RS15], based on the convergent Alter algorithm Laloë [2010]. In Section 2.3, we study a clustering problem for multivariate non-ordered circular data based on real data that came from radiotherapy. We provide two different solutions to this problem : one based on an appropriated distance on the circle combined with a simulated annealing algorithm [RS16] and the other one using a Bayesian strategy [RS02]. Section 2.4 is devoted to the definition of a clustering algorithm on micropollutants, called TyPol, that has been implemented [RS13] and then used in several biological publications [RS04, RS08, RS10].

2.2 Robust parameter-free clustering algorithm

The K-means clustering is the most popular clustering method [Hartigan and Wong, 1979, MacQueen, 1967]. Its attractiveness lies in its symplicity and its fast execution. It has however two main drawbacks. On the one hand, the number of clusters K has to be supplied by the user. Thus, different ways to determine K have been studied in the litterature [Li et al., 2008, Pham et al., 2005]. On the other hand, the algorithm strongly depends on the initialisation and can easily converges to a local minimum. Pelleg and Moore [2000] offer a solution for the first problem with a building-block algorithm called X-means which quickly estimates K. After each run of 2-means, local decisions are done whether subsets of the current centroid should be

splitted or not. The splitting decision is done by computing the Bayesian Information Criterion (BIC) [Schwarz, 1978]. In a different approach, Laloë [2010] proposes a consistent algorithm, called Alter, which also needs the specification of K.

The purpose of our work was to combine the X-means and the Alter algorithm in order to overcome the drawbacks of both algorithms. The complexity of the Alter algorithm decreases and an automatic selection of the number of clusters simultaneously performed. Moreover, the convergence properties of the Alter algorithm will overcome the local optimality problem of the X-means algorithm, inherited from the K-means one.

2.2.1 The Alter algorithm

Let us detail the Alter algorithm Laloë [2010]. The method is based on quantization, a commonly used technique in signal compression [Graf and Luschgy, 2000, Linder, 2002]. Consider $(\mathcal{H}, \|.\|)$ a normed space. We let X be a $\mathcal{H}$-valued random variable with distribution μ. Given a set $\mathcal{C}$ of points in $\mathcal{H}^k$, any Borel function $q : \mathcal{H} \to \mathcal{C}$ is called a quantizer. The set $\mathcal{C}$ is called a codebook, and the error made by replacing X by $q(X)$ is measured by the associated distortion. From Laloë [2010] we know that we can consider only nearest neighbor quantizers. Thus, a quantizer can be defined by its codebook only and the aim is to minimize the distortion among all possible nearest neighbor quantizers.

However, in practice, the distribution μ of the observations is unknown, and we only have at hand n independent observations $X_1, \ldots, X_n$ with the same distribution than X. The goal is then to minimize the empirical distortion :

$$\frac{1}{n}\sum_{i=1}^{n} d(X_i, q(X_i)).$$

The chosen distortion was the L^1-based distortion to obtain more robust estimators (see Kemperman [1987] for a discussion on this fact). Then, clustering is done by regrouping the observations that have the same image by q. More precisely, we define a cluster $\mathcal{C}$ by $\mathcal{C} = \{X_i : q(X_i) = \hat{x}_{\mathcal{C}}\}$, $\hat{x}_{\mathcal{C}}$ being the representant of cluster $\mathcal{C}$. Theoretical results of consistency and rate of convergence have been proved in Laloë [2010]. In particular, it is stated that the rate of convergence is closely related to the metric entropy. However, the minimization of the empirical distortion is not possible in practice and an alternative has been proposed with the Alter algorithm. The idea is to select an optimal codebook among the data set. More precisely the outline of the algoritm is :

1. List all possible codebooks , *i.e.*, all possible K-tuples of data ;
2. Compute the empirical distortion associated to the first codebook. Each observation X_i is associated with its closed center ;
3. For each successive codebook, compute the associated empirical distortion. Each time a codebook has an associated empirical distortion smaller than the previous smallest one, store the codebook ;
4. Return the codebook that has the smallest distortion.

Again, theoretical results of consistency and rate of convergence have been proved for the Alter algorithm. In particular it is stated that the convergence rate is of the same order than the

theoretical method described above. Moreover, this algorithm does not depend on initial conditions (unlike the K-means algorithm) and it converges to the optimal distortion. Unfortunately its complexity is $O(n^{K+1})$ and it is impossible to use it for high values of n or K.

2.2.2 The X-Means algorithm

In a different approach, Pelleg and Moore [2000] define the X-means algorithm which is adapted from K-means one. It goes into action after each run of K-means, making local decisions about which subset of the current centers should split themselves in order to better fit the data. The splitting decision is done by computing the BIC criterion. This new approach proposes an efficient solution to one major drawbacks of K-means : the search of the number of clusters K. Moreover, X-means has a low computational cost. But results suffer from the non-convergence property of the K-means algorithm. The outline of this algorithm is :

1. Perform 2-means. This gives us clustering C ;
2. Evaluate the relevance of the classification C with a BIC Criterion ;
3. Iterate step one and two in each cell of C. Keep going until there is no more relevant discrimination.

2.2.3 The X-Alter Algorithm

Following the idea of X-means, a recursive use of Alter with $K = 2$ can simultaneously allow us to combine both advantages of these two methods : estimation of K/low computational cost for X-means and convergence/parameter-free character for Alter. Using this idea, we define a new clustering algorithm called X-Alter. Obviously, the convergence properties of Alter are valid on each iteration separetely but we can not know if the whole X-Alter is convergent. We also add an aggregation step at the end of our algorithm to prevent the creation of too many clusters. Note that no parameter is needed by the algorithm. Though, the user can specify a range in which the true K reasonably lies if he wishes to (this is $[2, +\infty[$ if one had no information).

More precisely, the outline of the algorithm is the following :

1. Perform Alter with $K = 2$. This gives us clustering C ;
2. Evaluate the relevance of the classification C (Subfigure (a) of Figure 2.2.3) with a BIC Criterion ;
3. Iterate step one and two in each cell of C ((Subfigure (b) of Figure 2.2.3)). Keep going until there is no more relevant discrimination (Subfigure (c) of Figure 2.2.3) ;
4. Final step of aggregation : aggregation can be considered if $BIC(K = 1) > BIC(K = 2)$. The aggregations are successively made according to the decreasing values of $BIC(K = 1) - BIC(K = 2)$ (Subfigure (d) of Figure 2.2.3).

The algorithm starts by performing Alter with $K = 2$ centers. At this point, a model selection criterion (BIC) is performed on all the data set. Using this criterion, we check the suitability of the discrimination by comparing $BIC(K = 1)$ and $BIC(K = 2)$. In another way, the criterion tests if the model with the two clusters is better than the one with only one. If

the answer is yes, the iterative procedure occurs in the two subsets.

The structure improvement operation begins by splitting each cluster into two subsets. The procedure is local on that the children are fighting each other for the points in the parent's region, no others. When the discrimination is not validated by BIC criterion, the algorithm ends in this region. Up to there, the only difference with X-means is that we use Alter instead of 2-means because the consistent property of Alter must improve results. Finally, when all regions are asleep and no more clusters are needed, the aggregative step starts to prevent the creation of too many clusters or the presence of splitted clusters (as in Figure 2.2.3). The complexity of this algorithm in the worst case scenario (that is when it creates n clusters with one data set) is $O(n^4)$, which is less than the inital Alter algorithm. However, the computational cost is still higher than for X-means. For several thousand points, this complexity is not an important practical concern. But, if the database exceeds several tens of thousand points, it could still be too high.

This empirical algorithm was first tested on different simulated datasets that assessed its robustness compared to classimal k-means. Then, we used it on the well-known wine or iris datasets from UCI Machine Learning Repository [Frank and Asuncion, 2010]. More precisely, we compare the methods on the Iris data set. Pelleg and Moore show that X-means performs better and faster than repeateadly using accelerated K-means for different values of K. So, we compare our X-Alter algorithm to X-means and to X-means with the aggregation step, called X-means-R. We have 150 instances and 4 variables of 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other two ; the latter are not linearly separable from each other which makes it more difficult to classify. The results are gathered in Table 2.1.

TABLE 2.1 – Results for Iris data set.

Algorithm	Number of clusters	A.R.I.	Dunn
X-means	13.7 (var=6.2)	0.46 (var=0.07)	0.0405 (var=6.10^{-5})
X-means-R	8 (var=1.56)	0.57 (var=0.03)	0.0398 (var=0)
3-means	-	0.46 (var=0.0036)	0.04 (var=0)
X-Alter	6	1	0.402

It appears that our method do not find the real number of clusters but gets closer to it than others. Furthermore the high value of the Adjusted Rand Index [Hubert and Arabie, 1985] (A.R.I.) informs us that the great majority of iris plant are well-classified, the 3 additional clusters are in fact very small and do not affect the A.R.I and the global quality of the obtained clustering. In Dy and Brodley [2004], the estimation of the number of clusters is slighty better but, as discussed above, the quality of our clustering seems (as we don't use the same criterions) to be better. Moreover, we observe the interest of the aggregation step in X-means-R and it seems to appear that the spherical gaussian assumption required for the BIC is acceptable and that X-Alter can be tested with every complex data set.

Nevertheless, this X-Alter method was shown to be computationally expensive. A way to overcome this problem could be the adaptation of the Alter-Fast algorithm [Laloë, 2010] instead

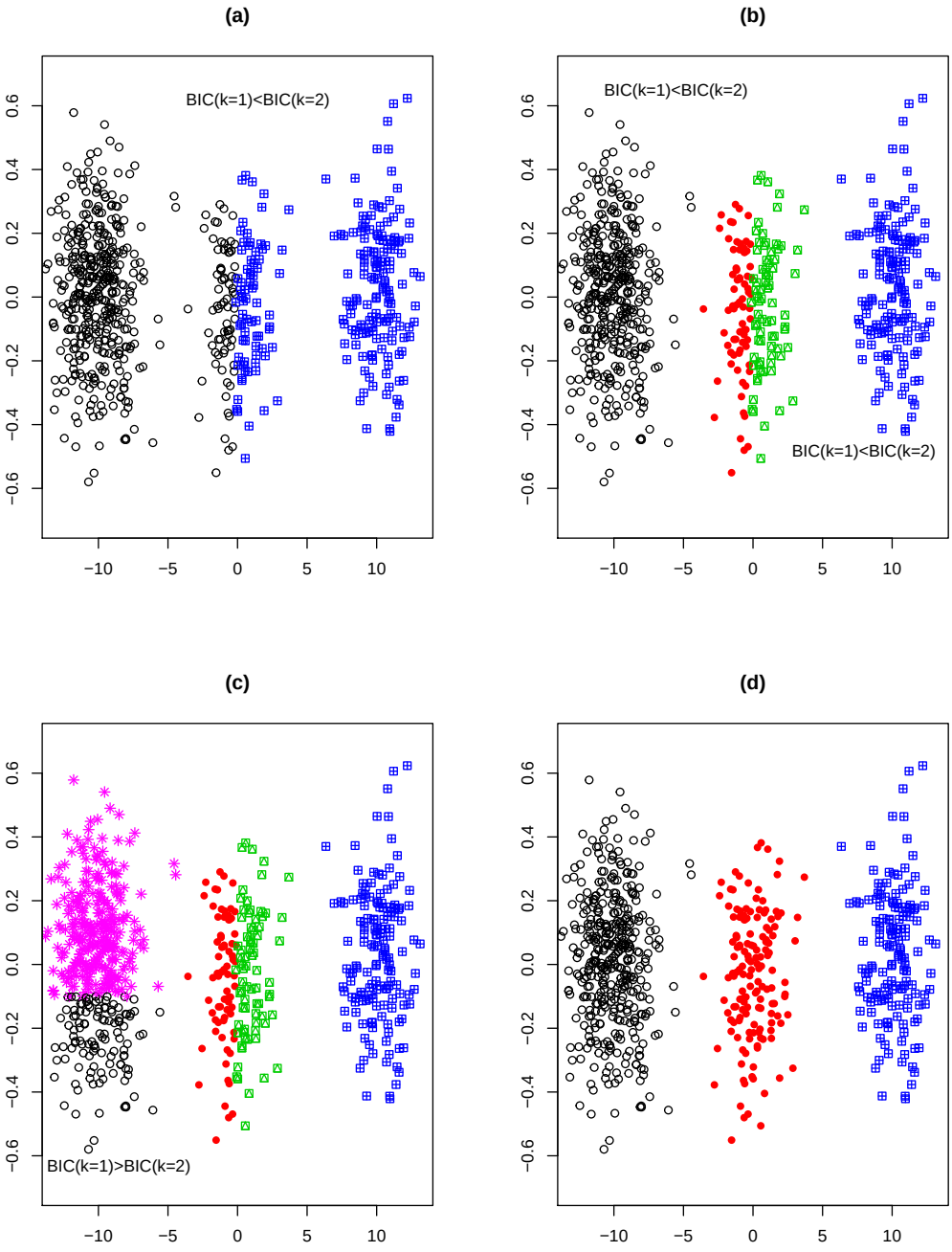

FIGURE 2.1 – (a) First iteration of X-Alter. The discrimination in 2 clusters (Step 1) is validated by BIC criterion (Step 2). In each cluster, observations are represented by a different symbol. (b) Second iteration of X-Alter : the sub-classification is done in the two relevant clusters (Step 1). Sub-classifications are validated by BIC (Step 2) so we obtain four clusters. (c) No relevant sub-classification in the left cluster according to BIC. In the three other clusters, we obtain the same rejection of sub-classification (Step 3). (d) Final discrimination. The two middle clusters have been aggregated in Step 4.

of Alter. It runs several times Alter in several randomly chosen partitions of the dataset resulting in a gain of time but in a lost of efficiency. Another approach could be the use of recent fast procedures to perform a greedy search such as the mixed integer programming [Bourguignon et al., 2016, Liu et al., 2017].

2.3 Clustering for multivariate non-ordered circular data

2.3.1 Motivation

Circular and directional data arise in a number of different fields such as oceanography (wave direction), meteorology (wind direction), biology (animal movement direction). The present works are motivated by circular data in medicine. Nowadays, intensity-modulated radiation therapy (IMRT) has demonstrated its effectiveness for cancer treatment. The latest generation of radiotherapy machines projects multiple rays. Multiplying beams allows to concentrate radiation on the tumor while avoiding the massive irradiation of healthy areas. However, the selection of the incident angles of the treatment beams may be a crucial component of IMRT planning. Due to variations in tumor locations, size and patient anatomy, repositioning for the multiple beams takes long time and is based on the planner's experience to find an optimal set of beams. So, establishing a small set of standardized beam bouquets for planning could be of valuable help. The set of beam bouquets could be determined by learning the beam configuration features from previous IMRT datasets. The multiple beams are fixed on a circle in the transverse plane around the patient. Consequently, an observation is composed of the k beams of a patient, that is k circular measurements. A real data set from post-operative treatment of liver cancer at the Institute of Sainte Catherine in Avignon, France, is represented in Figure 2.2.

One actual observation consists of a (non-ordered) set of k angles rather than of a vector (ordered) of length k but to cope with the technical difficulty of dealing with sets, it is convenient to store the angles of each patient in a vector in increasing order (or in any other given order). Of course, the derived vectors may be very different even for similar sets of angles. This is easily seen by considering a simple case of two patients with angles $\{1°, 60°, 100°, 150°, 180°\}$ and $\{60°, 100°, 150°, 180°, 359°\}$: the two patients should share the same cluster as the sets of angles are very similar (modulo 360) although the derived vectors are very different and, if any classical clustering method was applied, are not likely to share the same cluster.

Several algorithms have been developed to make an exhaustive search and determine the best beams compositions [Wang et al., 2004, Liu et al., 2006, Lee et al., 2006, Lei and Li, 2009] which are different for each patient. But the pratical implementation of these methods is hindered by the excessive computing time associated with the calculation. There is no other tools to assist the selection of beam orientations other than the therapist's experience and intuition whereas it could be very helpful [Pugachev et al., 2001] and accelerate previous algorithms. For example, these algorithms could be sped up by using appropriate initial presets.

Circular data have first been studied using classical non-Bayesian approaches. Three main models for circular data can be found in the litterature : the von-Mises distributions, the wrapped distributions and the projected normal distributions. The von-Mises distributions, first introduced by Von Mises [1918] and extended by Singh et al. [2002] and Mardia et al. [2008], are the natural analogues of the normal distribution on the sphere. The wrapped distributions [Mardia and Jupp, 2009] are based on a simple fact that a probability distribution on a circle

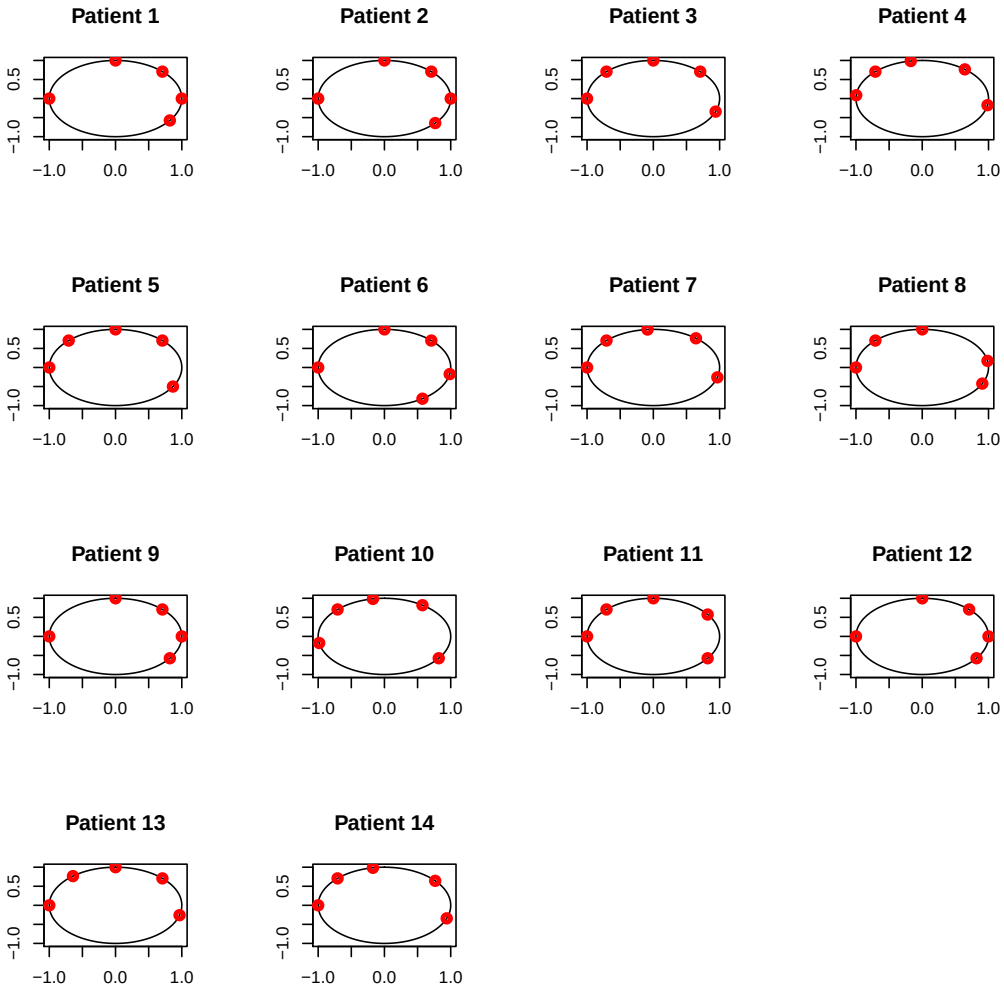

FIGURE 2.2 – Real data set of 14 patients with $k = 5$ angles. A point on the circle represents the location of a treatment beam.

can be obtained by wrapping a probability distribution defined on the real line. Projected normal distributions are obtained by projecting multivariate normal random variables radially onto the sphere [Presnell et al., 1998]. These latter distributions allow for asymmetric and possible bimodal models. We refer the reader to Mardia and Jupp [2009] for a complete review on probability distributions of circular data.

Even if our problem has similarities with some previously treated, the specificity of our data requires a specific method. Data are defined by the ballistic of the five angles $\{x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}\}$. To define sets of recurrent angles used by radiotherapy technicians, and so predefine settings, we used an unsupervised clustering method to obtain patient groups with homogeneous ballistics.

2.3.2 Clustering based on simulated annealing

To achieve this goal, we must consider two problems : the importance of the modulo 2π in the distance between two points on the circle and the permutations between two subsets, which is a novel feature, and is detailed below.

Data can be viewed as subset of $k = 5$ points on the circle. Note it can be easily extended to a different number of beams k. First, we define a distance δ between two points on the circle as follows :

$$\delta(a, b) = \min_{m \in \mathbb{Z}} |a - b + m2\pi| \text{ for all } a, b \in \mathbb{R}$$

where a and b denote the angle in radians with respect to an arbitrary origin. Note that δ can be viewed as a L^1-distance on the circle. Also note that the fact that points are angles is immaterial in the rest of this study and only affects metric δ. So, the following method could be used for any configurations of points lying in any space that has a distance defined between points.
Then, we define a distance between two subsets of five points on the circle. The chosen distance has to test all the permutations between the two subsets. For example, the distance between $x_1 = \{x_{11}, x_{12}, x_{13}, x_{14}, x_{15}\}$ and $x_2 = \{x_{12}, x_{13}, x_{14}, x_{15}, x_{11}\}$ must be zero. Taking into account these specificities, we propose the following function between two items x_1 and x_2 :

$$d(x_1, x_2) = \inf_{\sigma \in \mathcal{F}} \sum_{l=1}^{5} \delta \left(x_{1\sigma(l)}, x_{2l} \right),$$

where $\mathcal{F}$ is the set of permutations. The function d is shown to be a distance. This definition allows us to test all permutations between two angle sets and retain that which corresponds to the smallest distance.

If $x_1, x_2, \ldots, x_n$ denote the n observations to be classified in J clusters, the problem consists in determining the set of cluster centers $\Omega = \{c_1, c_2, \ldots, c_J\}$ which minimizes the distortion D defined by :

$$D(\Omega) = \sum_{i=1}^{n} \min_{c \in \Omega} d(x_i, c).$$

If we set

$$C_j = \{x_i : d\left(x_i, c_j\right) = \min_{c \in \Omega} d\left(x_i, c\right)\},$$

note that

$$D(\Omega) = \sum_{j=1}^{J} \sum_{x_i \in C_j} d\left(x_i, c_j\right)$$

and that $C_1, \ldots, C_J$ define a partitition of $\{x_1, x_2, \ldots, x_n\}$. However, there is no explicit solution for optimizing this criterion, again because the distances involved are non-Euclidean. The k-medoids clustering methods, like PAM [Kaufman and Rousseeuw, 1990] or CLARANS [Ng and Han, 1994], can solve this problem using the most central data of the cluster as centroids. But, because our real data set is small, we fear that few of the observations will be next to their centroids. This can produce bad clustering. For these reasons, and also because these methods only identify local optima, we chose not to used k-medoids clustering methods. Instead we use a simulated annealing type algorithm described below, which can find a better approximation of the cluster centers. So, given the chosen distance and its characteristics mentioned above,

we use, with a fixed number of clusters J, a clustering algorithm based on simulated annealing [Bartoli and Del Moral, 2001]. The $\nu - 1^{th}$ iteration of the algorithm ends giving us a set of J centers Ω^a. We describe below the ν^{th} iteration :

1. All data are assigned to their nearest center according to distance d. This provides us with a distortion D_ν^a defined by

$$D_\nu^a(\Omega^a) = \sum_{i=1}^{n} \min_{c \in \Omega^a} d(x_i, c).$$

2. A cluster j with center $c_j = \{c_{j1}, c_{j2}, c_{j3}, c_{j4}, c_{j5}\}$ is randomly chosen according to a discrete Uniform distribution. Then, a new center c'_j is proposed for this cluster, with coordinates $c'_{js} \sim \mathcal{N}^w(c_{js}, \sigma_a^2)$ for $1 \leq s \leq 5$.
3. The new distortion

$$D_\nu^b(\Omega^b) = \sum_{i=1}^{n} \min_{c \in \Omega^b} d(x_i, c)$$

is computed with $\Omega^b = \{c_1, \ldots, c_{j-1}, c'_j, c_{j+1}, \ldots, c_J\}$.

   (a) The new center is accepted with probability $1 \wedge \exp\left(-(D_\nu^b - D_\nu^a)/(t_\nu)\right)\Big)$, where t_ν is the so-called temperature, and we return to step 1.

   (b) If rejected, we return to step 2 and another center is taken.

The distribution $\mathcal{N}^w(c_{js}, \sigma_a^2)$ is the wrapped normal distribution on the circle [Mardia and Jupp, 2009]. It is obtained by wrapping a common normal distribution $\mathcal{N}(c_{js}, \sigma_a^2)$ onto the circle. Its probability density function is

$$f(x; c_{js}, \sigma_a^2) = \frac{1}{\sqrt{2\pi}\sigma_a} \sum_{l=-\infty}^{\infty} \exp\left\{-\frac{(x - c_{js} + 2l\pi)^2}{2\sigma_a^2}\right\}.$$

This distibution is unimodal and symmetric about its mode c_{js}.
The set of centers $\{c_1, \ldots, c_J\}$ which provides the lowest distortion D over all the chain is retained. This algorithm requires that the user sets in advance the number of clusters J, the shape of the temperature t_ν and the variance of normal distributions σ_a^2.
We provide a study of the convergence of the algorithm from a theoretical point of view. Let K be the transition kernel associated with the described algorithm. And let us define $\operatorname{osc}_K(D)$ as follows

$$\operatorname*{osc}_K(D) = \sup\{|D(x) - D(y)|, x \in E, y \in \operatorname{supp} K(x, .)\}$$

where $\operatorname{supp} K(x, .)$ denotes the support of $K(x, .)$. We state the following Proposition 2.1.

Proposition 2.1 *Taking* $t_\nu = \frac{C_0}{\log(\nu + e)}$ *with* $C_0 > J \operatorname{osc}_K(D)$, *then, for all* $\varepsilon > 0$, $Pr(x_\nu \in D^\varepsilon) \to 1$ *as* $\nu \to \infty$ *where*

$$D^\varepsilon = \{x \in E, D(x) \leq \operatorname*{essinf}_\lambda(D) + \varepsilon\} \text{ and } \operatorname*{essinf}_\lambda(\mathrm{D}) = \sup\{\mathrm{a} \geq 0, \lambda(\mathrm{a} \leq \mathrm{D}) = 1\}.$$

The choice of C_0 is a known problem for the convergence of the algorithm. If C_0 is chosen too large, the algorithm will take a long time to converge because the denominator is $\log(\nu + e)$. On the other hand, if C_0 is chosen too small, the algorithm converges too quickly and does not sufficiently explore the space of possible values to find the optimal clustering. In our problem, it is clear that we have $\operatorname{osc}_K(D) \leq 5n\pi$, which leads us to the sufficient condition $C_0 > 25n\pi$. This

is a rather crude bound, but we cannot obtain a better one without making strong assumptions about the data distribution. In order to reasonably estimate C_0, we run a chain of ν_0 sets of centers Ω and we calculate the variation of the distortion D at each iteration which leads to the following estimate of $\text{osc}_K(D)$:

$$\hat{\text{osc}}_K(D) = \sup_{1 \leq h \leq \nu_0} |D(\Omega^h) - D(\Omega^{h+1})|$$

where $\Omega^{h+1} \sim K(\Omega^h, .)$. This enables us to estimate C_0.
Note that in our algorithm only one randomly chosen center is updated. This provides us with an acceptable trade-off between exploration and convergence. Other strategies could be considered like updating all the centers at each iteration. In any case, the variance σ_a^2 of the proposal distribution must be carefully chosen in order to balance between exploration and convergence.

The performances of our clustering procedure were assessed on simulated datasets. Then, it was applied to our real dataset. The number of clusters was chosen according to BIC criterion. Running our algorithm with $J = 2$ we find the following two groups : one containing data 1,2,6,9 and 12, the second containing data 3,4,5,7,8,10,11,13 and 14. These results are relatively independent of the input parameters, such as initial centers or variance of wrapped normal distributions σ_a. We obtained two presets corresponding to the centers of these two groups :

$$c_1 = \{\pi/4, \pi/2, \pi, 1.81\pi, 1.99\pi\} \text{ and } c_2 = \{\pi/4, 0.51\pi, 3/4\pi, \pi, 1.88\pi\}.$$

We remark that the two centers have three common angles : $\pi/4, \pi/2$ and π and one slightly different from 1.85π. The principal difference resides in only one angle whose presets are $\pi/4$ or 0. Thus, using these preset positions should be fairly easy for praticians, with four fixed values and two choices for the last one. They should only have to make a few minor adjustments around these presets to correctly position beams. Each new patient should be affected to the first cluster with a probability 5/14 and to the second with a probability 9/14. In the first tests, the practitioners will realize quickly a possible wrong assignment of a patient and have just a few quick changes to be done to correct this.

2.3.3 Bayesian clustering

As already mentioned, this first approach have some drawbacks. First, the number of clusters has to be supplied by the user. An additional procedure of model selection (AIC, BIC, RIC, silhouette index, ...) can be used to select the number of clusters but an appropriate methodology that automatically finds this number would be very useful. Second, the final result is only one unique clustering whereas there are probably other clusterings that could be acceptable. A final result with all possible clusterings and a probability of appearance for each could be of great help for the practitioner. These problems can naturally be solved with a Bayesian clustering method based on Dirichlet Process as it does not require a preselected number of clusters and provides different clusterings (possibly with different numbers of clusters) with their posterior probabilities. Also note that the Bayesian framework is well adapted to our application as the sample size is low and can be compensated to some extent by prior information. To our knowledge, such a clustering Bayesian model has never been applied for multivariate circular data in the literature. So, we study a Bayesian clustering extension of this problem.

Bayesian litterature on circular data is more recent. Von Mises distributions are used in the univariate case in Damien and Walker [1999] and are applied to a change-point problem

in SenGupta and Laha [2008]. Wrapped distributions appear in Ravidran and Ghosh [2011], with a data augmentation algorithm to overcome some computational difficulties, and in Jona-Lasinio et al. [2012], to handle structured dependences between spatial measurements. Nuñez-Antonio and Gutiérrez-Peña [2005], Wang and Gelfand [2013] adapted the projected normal distributions in a Bayesian framework. A more sophisticated model was considered in Wang and Gelfand [2014] to capture structured spatial dependence for modeling directional data at different spatial locations. This model was then upgraded to capture joint structured spatial and temporal dependence [Wang et al., 2015].

Note that, for all the models cited above, each observation is simply a point on a circle or on a sphere while in our case, a single observation is made up of k ($k \geq 2$ and $k = 5$ in our dataset) non-ordered points on the circle. For this reason these models cannot straightforwardly be adapted to our dataset.

A simple way of generating distributions on the p-dimensional unit sphere $\mathcal{S}^p$ is to radially project probability distributions originally defined on the p-dimensional space $\mathbb{R}^p$ [Presnell et al., 1998]. Let x be a random p-dimensional vector, then $x/||x||$ is a random point on $\mathcal{S}^p$. If x has a p-variate Normal distribution $\mathcal{N}_p(\mu, \Sigma)$ then $x/||x||$ is said to have a projected normal distribution, denoted by $P\mathcal{N}_p(\mu, \Sigma)$. The literature has been first confined to the special case where $p = 2$ and $\Sigma = \mathbf{I}$ [Presnell et al., 1998, Nuñez-Antonio and Gutiérrez-Peña, 2005, Nuñez-Antonio et al., 2011]. Then, Wang and Gelfand [2013] studied the projected normal family with a general covariance matrix Σ and refer to this richer class $P\mathcal{N}_p(\mu, \Sigma)$ as the general projected normal distribution. This general version allows asymmetry and bimodality [see Figure 2. in Wang and Gelfand, 2014]. The general projected normal distribution is not identifiable because $x/||x||$ is invariant to scale transformation. To overcome this problem, Wang and Gelfand [2013] fixed some variance parameters in Σ to provide identifiability.

In a first step of simplification, we assume that the ith of the n observations is given by a vector of k angles $\theta_i = (\theta_{i1}, \ldots, \theta_{ik})' \in [0, 2\pi[^k$ instead of a non-ordered set $\{\theta_{i1}, \ldots, \theta_{ik}\}$. Using a projected normal distribution, we denote by $x_i = (x_{i1}, \ldots, x_{ik})' \in (\mathbb{R}^2)^k$ a random vector with distribution $\mathcal{N}_{2k}(\mu_i, I_{2k})$ where θ_{ij} is defined as the radial projection of x_{ij} on the unit circle of $\mathbb{R}^2$. In other words, we have $x_{ij} = (x_{ij1}, x_{ij2})' = (r_{ij} \cos\theta_{ij}, r_{ij} \sin\theta_{ij})'$ for all $i \in \{1, \ldots, n\}$ and all $j \in \{1, \ldots, k\}$ where r_{ij} denotes the Euclidean norm of x_{ij}. Note that θ_i is observed while $r_i = (r_{i1}, \ldots, r_{ik})'$ is not and is treated as an unknown parameter. We denote by $P\mathcal{N}_{2k}(\mu_i, I_{2k})$ the joint distribution of (θ_i, r_i). Clustering analysis will be based on a Dirichlet process mixture (DPM) model described as follows :

$$\begin{aligned} \theta_i, r_i|\mu &\sim P\mathcal{N}_{2k}(\mu_i, I_{2k}) \\ \mu_i|P &\sim P \\ P &\sim DP(n_0 P_0), \end{aligned} \tag{2.1}$$

where $\mu = (\mu_1, \ldots, \mu_n)$ and where $DP(n_0 P_0)$ denotes the Dirichlet process (DP) introduced by Ferguson [1973] with center $P_0 = \mathcal{N}_{2k}(0, \Sigma_0)$ and precision parameter n_0. The clustering properties of the DP are well known and date back to Blackwell and MacQueen [1973]. It is shown that the parameter $\mu = (\mu_1, \ldots, \mu_n)$ follows the Pólya urn scheme :

$$\begin{aligned} \mu_1 &\sim P_0 \\ \mu_{i+1}|\mu_1, \ldots, \mu_i &\sim \tfrac{1}{n_0+i} \textstyle\sum_{j=1}^{i} \delta_{\mu_i} + \tfrac{n_0}{n_0+i} P_0, \text{ for } i \geq 2. \end{aligned} \tag{2.2}$$

with δ_{μ_i} indicating the point measure on μ_i. So, μ_{i+1} may be equal to one of the previous μ_i's or may be drawn from P_0. This results in a positive probability of sharing the parameter value

with previous observations ; hence the clusters. In the sequel, we will denote by Pólya($n_0 P_0$) the distribution of μ given by (2.2). Although the DPM is very popular for Bayesian clustering, other model-based cluster methods exist. For a review of these methods, we refer the reader to Quintana [2006], Lau and Green [2007], Fritsch and Ickstadt [2009] and references therein. Note that the DPM model does not require choosing the number of clusters. On the other hand, it is well known that the number of clusters can be controlled by n_0. Learning about n_0 from the data may be addressed by assuming a Gamma prior distribution $n_0 \sim G(a_{n_0}, b_{n_0})$ [Escobar and West, 1995].

Now recall that the actual ith observation consists of a (non ordered) set of the form $\{\theta_{i1}, \ldots, \theta_{ik}\}$ rather than of a vector (ordered) $\theta_i = (\theta_{i1}, \ldots, \theta_{ik})'$. The impact of this simplification is quite easy to understand. Using model (2.1), two observations i_1 and i_2 with the same angles but in different orders would have a very low posterior probability of sharing the same cluster, that is $\mu_{i1} = \mu_{i2}$. We treat the observations as vectors for convenience but we have to introduce a permutation parameter τ_i to compensate this simplification. More precisely, for all $\mu_i = (\mu'_{i1}, \ldots, \mu'_{ik})'$ and all permutation τ_i of $\{1, \ldots, k\}$, we set $\mu_i^{\tau_i} = (\mu'_{i\tau_i(1)}, \ldots, \mu'_{i\tau_i(k)})'$; $\mu_i^{\tau_i}$ can be viewed as a random permutation of the coordinates of μ_i. Therefore, the clustering model becomes :

$$\begin{array}{rcl} \theta_i, r_i | \mu, \tau & \sim & P\mathcal{N}_{2k}(\mu_i^{\tau_i}, I_{2k}) \\ \mu_i | P & \sim & P \\ P & \sim & DP(n_0 P_0), \end{array} \tag{2.3}$$

where $\tau = (\tau_1, \ldots, \tau_n)$ and $\mu = (\mu_1, \ldots, \mu_n)$. The permutations τ_i are assumed to be *a priori* independent with a uniform distribution $\mathcal{U}_{\mathcal{P}}$ on the set $\mathcal{P}$ of permutations of $\{1, \ldots, k\}$. The posterior probability that two observations i_1 and i_2 with the same angles but in different orders would share the same cluster is increased with model (2.3) as there exist some values of τ_{i1} and τ_{i2} such that $\mu_{i1}^{\tau_{i1}} = \mu_{i2}^{\tau_{i2}}$.

Prior information It is natural to assume that the k angles $\theta_{i1}, \ldots, \theta_{ik}$ are *a priori* roughly equally spaced on the unit circle. This prior information can be incorporated into the covariance matrix Σ_0 of P_0 as follows. From (2.3), it is well known that the marginal distribution of μ_i is $P_0 = \mathcal{N}_{2k}(0, \Sigma_0)$. Denote by R the 2×2-matrix of the rotation in $\mathbb{R}^2$ with angle $2\pi/k$ and center 0. Set $\mu_{i1} \sim \mathcal{N}_2(0, \rho I_2)$ where ρ is a positive number and $\mu_{ij} | \mu_{i,j-1} \sim \mathcal{N}_2(R\mu_{i,j-1}, I_2)$ for $j \in \{2, \ldots, k\}$. Then, roughly, $\mu_{i1}, \ldots, \mu_{ik}$ are approximately equally spaced on the circle with center 0 and radius $\sqrt{\rho}$. Note that the variance parameter ρ has an important impact on the prior : a large value of ρ enables to generate $\mu_{i1}, \ldots, \mu_{ik}$ approximately situated on a circle with a large radius. For such a large radius, the angles θ_{ij} of the projections on the unit circle have small variances. Hence, ρ can also be viewed as a precision parameter for θ_i. We have shown that the derived matrix Σ_0, also denoted by $\Sigma_0(\rho)$ in the sequel to highlight the dependence on ρ, can be expressed as a closed-form expression as well as the inverse Σ_0^{-1} and the determinant $|\Sigma_0|$. Inference on ρ can then be performed using an inverse gamma prior $\rho \sim IG(a_\rho, b_\rho)$ for which the full posterior conditional distribution will be calculated in the following section.

Finally, the complete Bayesian model can be expressed as follows :

$$\begin{array}{rcl} \theta_i, r_i | \mu, \tau & \sim & P\mathcal{N}_{2k}(\mu_i^{\tau_i}, I_{2k}) \\ \mu | n_0, \rho & \sim & \text{Pólya}(n_0 P_0(\rho)) \\ \tau_i & \sim & \mathcal{U}_{\mathcal{P}} \\ \rho & \sim & IG(a_\rho, b_\rho) \\ n_0 & \sim & G(a_{n_0}, b_{n_0}). \end{array} \tag{2.4}$$

where $P_0(\rho) = \mathcal{N}_{2k}(0, \Sigma_0(\rho))$. By convention, it is assumed that the random variables at a stage of the hierarchy are independent.

Inference

We set $\theta = (\theta_1, \ldots, \theta_n)$, $r = (r_1, \ldots, r_n)$, $\mu = (\mu_1, \ldots, \mu_n)$, $\tau = (\tau_1, \ldots, \tau_n)$ and $\xi = (r, \mu, \tau, \rho, n_0)$. Thus, the parameter is ξ and the observation is θ. We sample from the posterior distribution of ξ with a Metropolis-Hastings-Within-Gibbs algorithm. In what follows, p stands for a generic notation for a density distribution.

Simulations of μ We can restrict our attention to model (2.3) instead of the full model (2.4) for the simulations of μ as every component of ξ except μ remains fixed. An alternative parameter setting of μ, θ and ρ will prove useful. Denote $x = (x_1, \ldots, x_n)$ where $x_i = (x'_{i1}, \ldots, x'_{ik})'$. Firstly, note that the full conditional distribution of μ reduces to the conditional distribution of μ given (x, n_0, ρ, τ) as there is a natural bijection between x_{ij} and (θ_{ij}, r_{ij}). Secondly, if we denote by $\mathcal{N}_{2k}(x_i; \mu_i, I_{2k})$ the value of the density of $\mathcal{N}_{2k}(\mu_i, I_{2k})$ at x_i, it is easy to check that :

$$\mathcal{N}_{2k}(x_i; \mu_i^{\tau_i}, I_{2k}) = \mathcal{N}_{2k}(x_i^{\tau_i^{-1}}; \mu_i, I_{2k}). \tag{2.5}$$

Consequently, if we set $y_i = x_i^{\tau_i^{-1}}$, sampling from the posterior distribution of μ in the DPM model (2.3) reduces to sampling from the posterior distribution of μ in the following conjugate DPM model :

$$\begin{array}{rcl} y_i|\mu & \sim & \mathcal{N}_{2k}(\mu_i, I_{2k}) \\ \mu_i|P & \sim & P \\ P & \sim & DP(n_0 P_0). \end{array} \tag{2.6}$$

There are several samplers for conjugate DPM models ; for a review, we refer the reader to Griffin and Holmes [2010]. Following the notations of Dahl [2003], we use a parameter setting of μ in terms of :

- a set partition $\eta = \{S_1, \ldots, S_q\}$ for $\{1, \ldots, n\}$ where each S_j represents a cluster, *i.e.*, $\mu_i = \mu_j$ if there exists $j_1 \in \{1, \ldots, q\}$ such that $i, j \in S_{j_1}$ and $\mu_i \neq \mu_j$ if there exist $j_1, i_1 \in \{1, \ldots, q\}$, $i_1 \neq j_1$ such that $i \in S_{i_1}$, $j \in S_{j_1}$,
- a vector $\phi = (\phi_1, \ldots, \phi_q)$ composed of the distinct values of μ, *i.e.*, $\phi_j = \mu_i$ for all $i \in S_j$.

Then, the conjugate DPM model (2.6) may be expressed as :

$$\begin{array}{rcl} y_i|\eta, \phi & \sim & \mathcal{N}_{2k}(\sum_{j=1}^q \phi_j \mathbf{1}_{\{i \in S_j\}}, I_{2k}) \\ \phi_j|\eta & \sim & P_0 \\ \eta & \sim & p(\eta) \propto \prod_{i=1}^q n_0 \Gamma(|S_j|), \end{array} \tag{2.7}$$

where $|S_j|$ is the cardinal of S_j, $\mathbf{1}_A$ is the indicator function for the event A, Γ denotes the gamma function and p stands for the generic notation for any density. We can integrate over the cluster location parameter ϕ analytically in (2.7) as P_0 is conjugate to the normal distribution of y_i given η and ϕ. Then, we run the SAMS sampler of Dahl [2003] for simulating η. Once a simulation of η is obtained, it is easy to simulate the cluster location parameter ϕ from its full conditional which reduces to sample independently each ϕ_j from a $\mathcal{N}_{2k}(\Sigma_j \sum_{i \in S_j} y_i/|S_j|, \Sigma_j)$ distribution with $\Sigma_j^{-1} = |S_j|^{-1} I_{2k} + \Sigma_0^{-1}(\rho)$. As recommended, we combine three runs of the Metropolis-Hastings update of the SAMS sampler with a full scan of Gibbs sampling for μ [MacEachern, 1994].

Simulations of r We show that the r_{ij} are independent given $(\theta, \tau, \mu, \rho, n_0)$ with density :

$$p(r_{ij}|\theta, \tau, \mu, \rho, n_0) \propto r_{ij} e^{-\frac{1}{2}\left(r_{ij} - u'_{ij}\mu_{i\tau_i(j)}\right)^2}, \tag{2.8}$$

with $u'_{ij} = (\cos\theta_{ij}, \sin\theta_{ij})$. If we denote by $\mathcal{N}_1^+(m, v)$ the univariate normal distribution truncated to $[0, \infty)$, we remark that (2.8) is close to the value of the density of $\mathcal{N}_1^+(u'_{ij}\mu_{i\tau_i(j)}, 1)$ at r_{ij}. It is then natural to simulate from (2.8) by a Metropolis-Hastings step with a $\mathcal{N}_1^+(u'_{ij}\mu_{i\tau_i(j)}, 1)$ as the proposal distribution. Clearly, the probability of acceptance reduces to the ratio $\min\{r_{ij}^{new}/r_{ij}^{old}, 1\}$ where r_{ij}^{old} and r_{ij}^{new} are, respectively, the current and the proposed values of r_{ij} in the algorithm.

Simulations of τ As the prior distribution of τ is uniform, we have :

$$p(\tau|\theta, r, \mu, \rho, n_0) \quad \propto \quad \prod_{i=1}^{n} \mathcal{N}_{2k}(x_i; \mu_i^{\tau_i}, I_{2k}).$$

Thus, given $(\theta, r, \mu, \rho, n_0)$, the τ_i are independent with density (with respect to the counting measure on the set T of permutations of $\{1, \ldots, k\}$) :

$$p(\tau_i|x, \mu) = \frac{\mathcal{N}_{2k}(x_i; \mu_i^{\tau_i}, I_{2k})}{\sum_{t\in T} \mathcal{N}_{2k}(x_i; \mu_i^{t}, I_{2k})}. \tag{2.9}$$

Simulations of ρ From (2.4), it is clear that the full conditional distribution of ρ reduces to the conditional distribution of ρ given μ. Then, using the parametrization of μ in terms of (η, ϕ), (2.7), and a few calculations, we show that the full conditional of ρ is

$$IG\left(a_\rho + q, b_\rho + \frac{1}{2}\sum_{i=1}^{q} \phi'_{i1}\phi_{i1}\right). \tag{2.10}$$

Simulations of n_0 Using the arguments of Escobar and West [1995], under the $G(a_{n_0}, b_{n_0})$ prior, n_0 is updated at each Gibbs iteration by sampling first an additional variable ζ from a Beta distribution and then a new value of n_0 from a mixture of Gamma distributions as follows :

$$\begin{array}{rcl} \zeta|n_0 & \sim & B\left(n_0 + 1, n\right) \\ n_0|\zeta, q & \sim & \pi_n G(a_{n_0} + q, b_{n_0} - \log\zeta) + (1 - \pi_n)G(a_{n_0} + q - 1, b_{n_0} - \log\zeta), \end{array} \tag{2.11}$$

with weights π_n defined by $\pi_n/(1 - \pi_n) = (a_{n_0} + q - 1)/[n(b_{n_0} - \log\zeta)]$.

Theoretical study of the symmetrized model

To investigate the impact of the symmetrization induced by the variables τ_i, we consider a simple model of the following form :

$$\begin{array}{rcl} x_i|\eta, \phi & \sim & \mathcal{N}_{2k}(\sum_{j=1}^{q} \phi_j \mathbf{1}_{\{i\in S_j\}}, I_{2k}) \\ \phi_j|\eta & \sim & P_0 \\ \eta & \sim & G \end{array} \tag{I}$$

and its symmetrized version :

$$\begin{array}{rcl} x_i|\eta,\phi & \sim & \mathcal{N}_{2k}(\sum_{j=1}^q \phi_j^{\tau_i}\mathbf{1}_{\{i\in S_j\}}, I_{2k}) \\ \phi_j|\eta & \sim & P_0 \\ \eta & \sim & G \\ \tau_i & \sim & \mathcal{U}_{\mathcal{P}}, \end{array} \tag{II}$$

where $\phi_j^{\tau_i} = (\phi'_{j\tau_i(1)}, \ldots, \phi'_{j\tau_i(k)})'$ is obtained by random permutation of the coordinates of $\phi_j = (\phi'_{j1}, \ldots, \phi'_{jk})' \in (\mathbb{R}^2)^k$. In both models, $P_0 = \mathcal{N}_{2k}(0, \Sigma_0)$ and G is any distribution of the partition $\eta = \{S_1, \ldots, S_q\}$ of $\{1, \ldots, n\}$. Such distributions include the distribution derived from the Dirichlet process given by (2.7). Model (II) can be viewed as a simplified and reparametrized version of (2.4). Now consider an idealized sample $x_1, \ldots, x_n$ for which every observation x_i is simply a random permutation of one unique observation $x_0 = (x'_{01}, \ldots, x'_{0k})' \in (\mathbb{R}^2)^k$; in other words, for every i, there exists a permutation α_i such that $x_i = (x'_{0\alpha_i(1)}, \ldots, x'_{0\alpha_i(k)})'$. As the coordinates x_{ij} of all the x_i are the same but in a different order, it is expected that all the observations are put together in one unique cluster. The aim of this section is to study whether model (II) is more appropriate than model (I) for this purpose.

Let p_0 and $p_{\mathrm{I}}(x|\eta)$ denote respectively the density of P_0 and the conditional density of $x = (x_1, \ldots, x_n)$ given η for model (I). We have :

$$p_{\mathrm{I}}(x|\eta) \;=\; \int \prod_{j=1}^q \prod_{i\in S_j} \mathcal{N}_{2k}(x_i; \phi_j, I_{2k}) p_0(\phi_j) d\phi_j = \prod_{j=1}^q m(x_{S_j}),$$

where $x_{S_j} = (x_i, i \in S_j)$ and

$$m(x_{S_j}) \;=\; \int \prod_{i\in S_j} \mathcal{N}_{2k}(x_i; \phi_j, I_{2k}) p_0(\phi_j) d\phi_j.$$

Denote by $p_{\mathrm{II}}(x|\eta)$ the conditional density of x given η for model (II). By (2.5) and noting that $\{\tau_i^{-1}, \tau_i \in \mathcal{P}\} = \mathcal{P}$, we have :

$$p_{\mathrm{II}}(x|\eta) \;=\; \frac{1}{(k!)^n} \sum_{\tau} \prod_{j=1}^q m(x^{\tau}_{S_j}),$$

where the sum above is taken for all the values of $\tau = (\tau_1, \ldots, \tau_n)$ in $\mathcal{P}^n$, $x^{\tau}_{S_j} = (x_i^{\tau_i}, i \in S_j)$ and $x_i^{\tau_i} = (x'_{i\tau_i(1)}, \ldots, x'_{i\tau_i(k)})'$. Therefore, models (I) and (II) reduce to

$$\begin{array}{rcl} x|\eta & \sim & \prod_{j=1}^q m(x_{S_j}) \\ \eta & \sim & G, \end{array} \tag{I'}$$

and

$$\begin{array}{rcl} x|\eta & \sim & \frac{1}{(k!)^n} \sum_{\tau} \prod_{j=1}^q m(x^{\tau}_{S_j}). \\ \eta & \sim & G. \end{array} \tag{II'}$$

For all partition $\eta = \{S_1, \ldots, S_q\}$ and all observation x, we set

$$f(x, \eta) = \frac{1}{(k!)^n} \sum_{\tau\in\mathcal{P}^n} \exp \frac{1}{2} \sum_{j=1}^q \left(\| \sum_{i\in S_j} x_i^{\tau_i} \|^2_{S_j} - \| \sum_{i\in S_j} x_i \|^2_{S_j} \right) \tag{2.12}$$

where $\Sigma_S = \left(\Sigma_0^{-1} + |S| I_{2k}\right)^{-1}$ for all subset $S \subset \{1, \ldots, n\}$ and $\|t\|^2_S = t'\Sigma_S t$ for all $t \in (\mathbb{R}^2)^k$.

Proposition 2.2 *a) For all partition* $\eta = \{S_1, \ldots, S_q\}$ *and all observation* $x = (x_1, \ldots, x_n)$, *we have :*

$$\frac{p_{\mathrm{II}}(x|\eta)}{p_{\mathrm{I}}(x|\eta)} = f(x, \eta).$$

b) For all distribution G, *there exists a positive number* B_G *such that :*

$$\frac{p_{\mathrm{II}}(\eta|x)}{p_{\mathrm{I}}(\eta|x)} = B_G \; f(x, \eta),$$

for all partition η *and all observation* x.

c) For all distribution G, *all partition* η *and all observation* x, *we have :*

$$\frac{p_{\mathrm{II}}(\eta|x)}{p_{\mathrm{I}}(\eta|x)} \geq f(x, \eta) \; \frac{1}{\max_\eta f(x, \eta)}$$

where the maximum is taken over all partitions of $\{1, \ldots, n\}$.

From *a*) of Proposition 2.2, we see that $f(x, \eta)$ is the likelihood ratio of models (II') and (I'). From *b*), we know that the posterior odds ratio is large when $f(x, \eta)$ is large. It would be of interest to know whether this ratio is greater than one. Unfortunately, this is not an easy task except for a few particular cases given below. Indeed, although the factor B_G is actually known (see the proof of Proposition 2.2 for more details), it is rather intractable. From *c*), we deduce that the posterior odds is actually greater or equal to one at least for the partition η_x that maximizes $f(x, \eta)$. This partition does exist for any observation x and is independent of G. In other words, for any x, there exists a partition η_x such that $p_{\mathrm{II}}(\eta_x|x) \geq p_{\mathrm{I}}(\eta_x|x)$ for all prior G.

Consider the partition $\bar{\eta}$ with a single cluster : $q = 1$ and $S_1 = \{1, \ldots, n\}$. From (2.12), the posterior odds ratio when $\eta = \bar{\eta}$ is likely to be large when $\sum_{i=1}^n x_i \approx 0$ and small when all the $x_i \approx x_0$ for all $i \in \{1, \ldots, n\}$. Assume from now that $\sum_{i=1}^n x_i = 0$ and that $\Sigma_0 = I_{2k}$. Remenber that Σ_0 models the prior information about the mutual positions of the angles on the circle. Therefore $\Sigma_0 = I_{2k}$ can be viewed as a non informative prior. In this case, $\|t\|_{S_j}^2 = (1 + |S_j|)^{-1} t't = (1 + |S_j|)^{-1} \|t\|$ for all $t \in (\mathbb{R}^2)^k$ and we have :

$$f(x, \bar{\eta}) = \frac{1}{(k!)^n} \sum_{\tau \in \mathcal{P}^n} \exp \frac{1}{2(n+1)} \left(\| \sum_{i=1}^n x_i^{\tau_i} \|^2 \right). \tag{2.13}$$

Example 1 below provides a typical sample $x = (x_1, \ldots, x_n)$ for which the posterior probabilty of a unique cluster is greater with model (II) than with model (I) independently of the prior distribution G.

Example 1 *Assume* $n = k$ *and that* $x_1 = (x'_{11}, \ldots, x'_{1k})' \in (\mathbb{R}^2)^k$ *is made up of* k *consecutive points on the unit circle separated from an angle of* $2\pi/k$, x_2 *is obtained by a rotation with angle* $2\pi/k$ *of each point of* x_1 *and so on. Therefore, we have* $\sum_{i=1}^n x_i = 0$. *Our conjecture is that* $\max_\eta f(x, \eta) = f(x, \bar{\eta})$ *for all integer* k *which implies, from c) of Proposition 2.2, that the probability of a unique cluster is greater for model (II) than for model (I) for any distribution* G. *For* $n = k = 2$ *the conjecture reduces to* $f(x, \eta) \leq f(x, \bar{\eta})$ *for a single partition* $\eta = \{\{x_1\}, \{x_2\}\}$.

As $\|x_i\|_{S_j} = \|x_i^{\tau_i}\|_{S_j}$ *for all* i *and* τ_i*, it is easily seen from (2.12) that* $f(x,\eta) = 1$*. On the other hand, as* $\|x_1\|^2 = k$ *and* $x_1 = -x_2$*, we see from (2.13) that*

$$\begin{aligned} f(x,\bar{\eta}) &= \frac{1}{4}\left(2\exp\frac{1}{6}\|x_1+x_2\|^2 + 2\exp\frac{1}{6}\|2x_1\|^2\right) \\ &= \frac{1}{2}\left(1+2\exp\frac{4}{3}\right), \end{aligned}$$

hence the proof of the conjecture for $n = k = 2$*. We also proved the conjecture for* $n = k = 3$ *with a rather large amount of calculations (not given here) to take into account all the partitions* η *and all the permutation* $\tau = (\tau_1, \tau_2, \tau_3)$*. We are not in a position to provide general proof of the conjecture for* $n = k \geq 4$*.*

Real dataset results

Some simulations enhanced the performances of the whole clustering methodology and its robustness to hyperparameter values. Then it was applied to the post-operative treatment of liver cancer at the Institute of Sainte Catherine in Avignon, France (see Figure 2.2). Its results are compared to our previous method in which the number of clusters was preselected to $q = 2$. We used non informative priors and investigated the MCMC convergence with the clustering entropy $-\sum_{i=1}^{q} \frac{|S_i|}{n} \log\left(\frac{|S_i|}{n}\right)$.

The majority clustering (mode of the posterior distribution of the clusterings) is the same that is been obtained previously with the simulated annealing with a posterior probability equal to 30.5%. This result was awaited and is coherent with the choice of 2 clusters in the previous method. But the real gain from our Bayesian approach is to look beyond this majority clustering. Here there are 3 more clusterings that are significant and that could give some information on this real dataset. The second majority clustering is nearly the same as the previous one : the clusters are the same but data 6 is alone in a third cluster. Indeed, this data is very atypical because it is the only one that contains an angle near 1.69π. The posterior probability for this clustering is 14.9%. The third majority clustering gives nearly the same information with a posterior probability of 13.5%. There are two clusters : one with data 6 and a second with all the others. Finally, another clustering with a posterior probability of 12.0% is made up of only one cluster. Even with other choices for the hyperparameters a_{n_0} and b_{n_0}, the posterior probability of this clustering remains high. It highlights the fact that all the data share some common traits and the main difference in the two clusters of the majority clustering only concerns one angle. It can be noted that a credible region with a posterior probability of 71% is composed of the 4 previous clusterings. As explained with the previous approach, using these preset positions should be fairly easy for praticians, with four fixed values and only two choices for the last one. Furthermore, the results suggest another preset position that should be added and tested if the two previous one do not fit : the beam angles of data 6.

Note that between our two publications on this subject, Yuan et al. [2015] generalized the first approach using k-medoids to cluster beam configuration features with different numbers of beams. The efficiency of this approach was tested using an appropriate clinical trial and they stated that the dosimetric quality of the plans using the standardized beam bouquets have comparable quality to that of usual clinical plans. These standardized beam configuration bouquets will by consequence help improve plan efficiency and facilitate automated planning. They also recorded a US Patent [Wu et al., 2015] that cites our first work [RS16]. Very recently,

they also improved this approach by considering noncoplanar beams [Yuan et al., 2018]. However some improvements could be considered, such as, incorporating covariates (shape or size of the tumor, stage of the cancer, sex, age, ...) to preselect the beam positions and/or refine the prior probabilities of assignment in each cluster.

2.4 Clustering of micropollutants

2.4.1 Methodology

I was also involved in an applied project called TyPol whose goal was to cluster micropollutants. This project was based on the fact that new legislations such as the REACH (Registration, Evaluation, Authorization and restriction of CHemicals) improved the needed information on chemical substances [Muir and Howard, 2006]. Consequently, a high number of different *in silico* approaches have been developed to estimate the behavior of organic compounds in the environment such as QSAR [Eriksson et al., 2002, Pavan et al., 2008] or other numerical models [Jarvis and Larsbo, 2012]. Therefore, approaches able to classify compounds according to their environmental behavior or eco/toxicological effects will help both regulators and scientists facing the problem of the constant increase in the diversity and in the number of the chemical substances which will be concerned by environmental risk assessment. The objective of this work was thus to develop a new simple approach, TyPol (Typology of Pollutants), to classify organic compounds and their degradation products according to both their behavior in the environment and their structural molecular properties. TyPol is based on a large database containing environmental endpoints (*i.e.* environmental parameters such as sorption coefficient, degradation half-life or bioconcentration factor), and structural molecular descriptors (number of atoms in the molecule, molecular surface, dipole moment, energy of orbitals, etc.). The calculation of molecular descriptors is based on *in silico* approach, and the environmental parameters are extracted from available databases and from literature.

The problematic of TyPol is that it considers two sets of variables (molecular descriptors and environmental parameters), which are different by nature. Partial least squares regression (PLS) [Wold, 1996, Boulesteix and Strimmer, 2007] can be used to find the fundamental relation between two sets of variables using a latent variable approach to model the covariance structures in these two spaces. To be more specific, the general underlying model of multivariate PLS is

$$X = TP^{\mathrm{T}} + E$$
$$Y = UQ^{\mathrm{T}} + F$$

where X is an $n \times m$ matrix of predictors (here the molecular descriptors), Y is an $n \times p$ matrix of responses (here the environmental parameters) ; T and U are $n \times l$ matrices that are, respectively, projections of X and Y ; P and Q are, respectively, $m \times l$ and $p \times l$ orthogonal loading matrices ; and matrices E and F are the error terms, assumed to be independent and identically distributed random normal variables. The decompositions of X and Y are made so as to maximise the covariance between T and U. PLS model tries to find the multidimensional directions in the observable variables (*i.e.* molecular descriptors) space that explain the maximum multidimensional variance direction in the predicted variable (*i.e.* environmental parameters) space. So PLS, as the most-known PCA, constructs uncorrelated variables which summarizes the information, but PLS takes into account the information of both observable and predictive

variables. After the PLS analysis, a hierarchical clustering algorithm [Ward, 1963] is performed on the new constructed variables to cluster the organic compounds. The robustness of this procedure was assessed using the A.R.I. as described in Section 2.2. This clustering procedure is connected to a database containing now more than three hundred molecules. The whole procedure is implemented in a RStudio version and is available online on a dedicated server after an identification. Nevertheless, this application has several drawbacks for the user which mainly concern the difficulties to exchange the results or the configuration tested with another user. Thus, an upgrade of the code and a migration to a more adapted Galaxy platform is now studied through a master internship. Different versions of penalized PLS [Kraemer et al., 2008, Mehmood et al., 2012] are also about to be tested to bring sparsity and ease the interpretation of the results.

2.4.2 Applications

TyPol was widely used since its birth. First, it was used in combination with mass spectrometry to identify and categorize tebucanozole products in soil. TyPol was used to group the detected transformation products according to common molecular descriptors and to indirectly elucidate their environmental properties by analogy to known pesticide compounds having similar molecular descriptors. Our approach was then evaluated via the identification of the tranformation products of the triazole fungicide tebuconazole occurring in a field dissipation study. Overall, 22 empirical and 12 yet unknown transformation products were detected and categorized into three groups with defined environmental properties.

Second, TyPol was applied to chlordecone and its transformation products. Starting from the list of putative chlordecone transformation products and considering available data on degradation routes of other organochlorine compounds, we used TyPol to explore the potential environmental behaviour of putative chlordecone transformation products from the knowledge on their molecular descriptors. Our findings suggest that some transformation products of chlordecone (namely mono and di-hydrochlordecone), often found in contaminated soils, may have similar environmental behaviour in terms of persistence.

Then, TyPol was extended to the ecotoxicological effects of pesticides on non-target organisms, based on data analysis from available literature and databases. It revealed that relevant ecotoxicological endpoints for terrestrial organisms (e.g., soil microorganisms, invertebrates) that support a range of ecosystemic services are lacking compared to aquatic organisms. Consequently, seven parameters were included for acute and chronic ecotoxicological effects for terrestrial and three aquatic organisms. With this new configuration, we used TyPol to classify 50 pesticides into different clusters that gather molecules with similar environmental behaviors and ecotoxicological effects. The classification results evidenced relationships between molecular descriptors, environmental parameters, and the added ecotoxicological endpoints.

2.5 Ongoing projects and prospects

As explained in each subsection, different leads exist to improve the dedicated approaches. For the TyPol algorithm, the migration to a new version of the code and to a Galaxy platform is also made to delete the need of a statistician in the exportation and the first interpretation of the results. On a more theoretical framework, it would be of interest to have a radiotherapy dataset with interesting covariables (and eventually noncoplanar beam angles) and to see how

to integrate this prior information in the defined clustering model and how it can affect it. Nevertheless, to be clear, such a study is not planned for now.

Bibliography

N. Bartoli and P. Del Moral. *Simulations et algorithmes stochastiques : une introduction avec applications*. CEPADUES, Toulouse, 2001.

D. Blackwell and J. B. MacQueen. Ferguson distributions via Polya urn schemes. *The Annals of Statistics*, 1(2) :353–355, 1973.

A.-L. Boulesteix and K. Strimmer. Partial least squares : a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8(1) :32–44, 2007.

S. Bourguignon, J. Ninin, H. Carfantan, and M. Mongeau. Exact Sparse Approximation Problems via Mixed-Integer Programming : Formulations and Computational Performance. *IEEE Transactions on Signal Processing*, 64(6) :1405–1419, 2016.

D. B. Dahl. An improved merge-split sampler for conjugate Dirichlet process mixture models. *Technical Report, University of Wisconsin - Madison*, 1086 :1–32, 2003.

P. Damien and S. Walker. A full bayesian analysis of circular data using the von Mises distribution. *The Canadian Journal of Statistics*, 27(2) :291–298, 1999.

J. Dy and C. Brodley. Feature selection for unsupervised learning. *Journal of Machine Learning Research*, 5 :845–889, 2004.

L. Eriksson, P. Andersson, E. Johansson, and M. Tysklind. Multivariate biological profiling and principal toxicity regions of compounds : the PCB case study. *Journal of Chemometrics*, 16 :497–509, 2002.

M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430) :577–588, 1995.

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2) :209–230, 1973.

A. Frank and A. Asuncion. UCI Machine Learning Repository, 2010.

A. Fritsch and K. Ickstadt. Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis*, 4(2) :367–392, 2009.

S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*, volume 1730 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2000.

J. Griffin and C. Holmes. Computational issues arising in bayesian nonparametric hierarchical models. In N. Hjort, C. Holmes, P. Muller, and S. G. Walker, editors, *Bayesian Nonparametrics*, pages 208–222, Cambridge University Press, 2010.

J. Hartigan and M. Wong. A k-means clustering algorithm. *Journal of the Royal Statistical Society*, 28 :100–108, 1979.

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1) :193–218, 1985.

N. Jarvis and M. Larsbo. Macro (v5.2) : model use, calibration and validation. *Transactions of the ASABE*, 55 :1413–1423, 2012.

G. Jona-Lasinio, A. Gelfand, and M. Jona-Lasinio. Spatial analysis of wave direction data using wrapped gaussian processes. *The Annals of Applied Statistics*, 6(4) :1478–1498, 2012.

L. Kaufman and P. Rousseeuw. *Finding Groups in Data : an Introduction to Cluster Analysis*. John Wiley & Sons, New-York, 1990.

J. H. B. Kemperman. The median of a finite measure on a Banach space. In *Statistical Data Analysis Based on the L_1-norm and Related Methods (Neuchâtel, 1987)*, pages 217–230. North-Holland, Amsterdam, 1987.

N. Kraemer, A.-L. Boulesteix, and G. Tutz. Penalized partial least squares with applications to B-spline transformations and functional data. *Chemometrics and Intelligent Laboratory Systems*, 94(60-69), 2008.

T. Laloë. L_1 quantizationand clustering in banach spaces. *Mathematical Methods of Statistics*, 19(2) :136–150, 2010.

J. W. Lau and P. J. Green. Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics*, 16(3) :526–558, 2007.

E. Lee, T. Fox, and I. Crocker. Simultaneous beam geometry and intensity map optimization in intensity-modulated radiation therapy. *International Journal of Radiation Oncology, Biology, Physics*, 64, 2006.

J. Lei and Y. Li. An approaching genetic algorithm for automatic beam angle selection in IMRT planning. *Computer Methods and Programs in Biomedicine*, 93, 2009.

M. Li, M. Ng, Y.-M. Cheung, and J. Huang. Agglomerative fuzzy k-means clustering algorithm with selection of number of clusters. *IEEE transactions on knowledge and data engineering*, 20 :1519–1534, 2008.

T. Linder. Learning-theoretic methods in vector quantization. In *Principles of Nonparametric Learning (Udine, 2001)*, volume 434 of *CISM Courses and Lectures*, pages 163–210. Springer, Vienna, 2002.

H. Liu, M. Jauregui, X. Zhang, Z. Wang, L. Dong, and R. Mohan. Beam angle optimization and reduction for intensity-modulated radiation therapy of non-small-cell lung cancers. *International Journal of Radiation Oncology, Biology, Physics*, 65, 2006.

Y. Liu, S. Canu, P. Honeine, and S. Ruan. Une véritable approche ℓ_0 pour l'apprentissage de dictionnaire. In *Proceedings of GRETSI'2017*, Juan-Les-Pins, 2017.

S. N. MacEachern. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics : Simulation and Computation*, 23(3) :727–741, 1994.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, Berkeley, University of California Press, 1967.

K. Mardia and P. Jupp. *Directional Statistics*. John Wiley & Sons, New-York, 2009.

K. V. Mardia, G. Hugues, C. C. Taylor, and H. Singh. A multivariate von Mises distribution with applications to bioinformatics. *The Canadian Journal of Statistics*, 36(1) :99–109, 2008.

T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø. A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 118 : 62–69, 2012.

D. Muir and P. Howard. Are there other persistent organic pollutants ? A challenge for environmental chemists. *Environmental Science & Technology*, 40 :7157–7166, 2006.

R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. In *Proceedings of the 20th International Conference on Very Large Data Bases*, San Francisco, 1994.

G. Nuñez-Antonio and E. Gutiérrez-Peña. A Bayesian analysis of directional data using the projected normal distribution. *Journal of Applied Statistics*, 32(10) :995–1001, 2005.

G. Nuñez-Antonio, E. Gutiérrez-Peña, and G. Escarela. A Bayesian regression model for circular data based on the projected normal distribution. *Statistical Modeling*, 11(3) :185–201, 2011.

M. Pavan, T. Netzeva, and A. Worth. Review of literature-based quantitative structure-activity relationship models for bioconcentration. *QSAR & Combinatorial Science*, 27 :21–31, 2008.

D. Pelleg and A. Moore. X-means : Extending k-means with efficient estimation of the number of clusters. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 727–734, San Francisco, 2000. Morgan Kaufmann.

T. Pham, S. Dimov, and C. Nguyen. Selection of K in K-means clustering. *Proceedings of the Institution of Mechanical Engineers, Part C : Journal of Mechanical Engineering Science*, 219 :103–119, 2005.

B. Presnell, S. P. Morrison, and R. C. Littell. Projected multivariate linear models for directional data. *Journal of the American Statistical Association*, 93(443) :1068–1077, 1998.

A. Pugachev, G. Li, A. Boyer, S. Hancock, Q.-T. Le, S. Donaldson, and L. Xing. Role of beam orientation optimization in intensity-modulated radiation therapy. *International Journal of Radiation Oncology, Biology, Physics*, 50, 2001.

F. A. Quintana. A predictive view of bayesian clustering. *Journal of Statistical Planning and Inference*, 136(8) :2407–2429, 2006.

P. Ravidran and S. K. Ghosh. Bayesian analysis of circular data using wrapped distributions. *Journal of Statistical Theory and Practice*, 5(4) :547–560, 2011.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2) :461–464, 03 1978.

A. SenGupta and A. K. Laha. A bayesian analysis of the change-point problem for directional data. *Journal of Applied Statistics*, 35(6) :693–700, 2008.

H. Singh, V. Hnizdo, and E. Demchuk. Probabilistic model for two dependant circular variables. *Biometrika*, 89(3) :719–723, 2002.

R. Von Mises. Über die ganzzahligkeit der atomgewicht und verwandte fragen. *Physikalische Zeitschrift*, 19 :490–500, 1918.

F. Wang and A. E. Gelfand. Directional data analysis under the general projected normal distribution. *Statistical Methodology*, 10(1) :113–127, 2013.

F. Wang and A. E. Gelfand. Modeling space and space-time directional data using projected gaussian processes. *Journal of the American Statistical Association*, 109(508) :1565–1580, 2014.

F. Wang, A. E. Gelfand, and G. Jona-Lasinio. Joint spatio-temporal analysis of a linear and a directional variable : space-time modeling of wave heights and wave directions in the adriatic sea. *Statistica Sinica*, 25(1) :25–29, 2015.

Z. Wang, X. Zhang, L. Dong, H. Liu, Q. Wu, and R. Mohan. Development of methods for beam angle optimization for imrt using an accelerated exhaustive search strategy. *International Journal of Radiation Oncology, Biology, Physics*, 60, 2004.

J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301) :236–244, 1963.

H. Wold. Estimation of principal component and related models by iterative least squares. In *Krishnaiaah, P.R. (Ed.), Multivariate Analysis*, pages 391–420. Academic Press, New York, 1996.

Q. Wu, Y. Ge, Y. Fang-Fang, L. Yuan, Y. Sheng, T. Li, and J. Liu. Systems and methods for automated radiation treatment planning with decision support. *US Patent*, US20180043182A1, 2015.

R. Xu and D. Wunsch, II. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3) :645–678, 2005.

L. Yuan, Q. J. Wu, F. Yin, Y. Li, Y. Sheng, C. R. Kelsey, and Y. Ge. Standardized beam bouquets for lung IMRT planning. *Physics in Medicine & Biology*, 60(5) :1821–1843, 2015.

L. Yuan, W. Zhu, Y. Ge, Y. Jiang, Y. Sheng, F.-F. Yin, and Q. J. Wu. Lung IMRT planning with automatic determination of beam angle configurations. *Physics in Medicine & Biology*, Forthcoming, 2018.

Chapitre 3

Statistical learning for functional data

3.1 Introduction

At my arrival at the INRA, I did not know anything about metabolomics or precision livestock farming. I had no idea that I will be involved in statistical modeling problems caming from these applications. Indeed, practical problems that came from these applications often need the careful building of an *ad hoc* statistical procedure that raises very interesting statistical issues. The building of new technologies to obtain data (omics data, functional follow-up data ...) leads to numerous questions for the statistician and the data really feeds the statistician.

In this permanent concern of studying statistical problems raised by a direct application, an expanding domain is the personalized (or precision) medicine (human or veterinary). My first work within my new appointment consisted in the study of medical follow-up data : several variables are measured in a longitudinal way in a subject (animal or human) and the question is to build a region of prediction allowing to detect a health modification (disease, doping....)[RS12]. This modeling, based on a mixed effect model, also allowed me to familiarize with these models that are the cultural environment of my current team [RS06]. This work can also be put in the more general context of the personalized medicine/precision livestock farming, where the joint and simultaneous analysis of several sensor measures is derived to allow an early and individual detection of some pathologies. A major part of my research perspectives takes place in this framework.

The use of sensors is also generalizing in farming. This allows a more personalized management of each field and a less dependence on the climatic hazards due to a better knowledge of the crop needs. To answer this kind of questions it is important to be able to select, in multivariate temporal data, ranges of time with regard to a specific factor of interest (for example the yield of a field). Following this idea, we define a new method of variables selection, based on Sliced Inverse Regression (SIR) combined with a sparse criterion. Furthermore, this method integrates a data-driven algorithm that automatically defines the relevant intervals in a functional framework [RS01, RS23]. Another variable selection problem is at the center of the PhD of Patrick Tardivel : in metabolomics, a complex mixture spectrum is composed of the weighted sum of the spectra of all the metabolites that are therein. The difficulty is, from the complex mixture spectrum and a database of all the metabolite spectra, to reconstruct the complex mixture composition. During this PhD, we developed and studied a new dedicated multiple testing procedure based on the thresholded maximum likelihood [RS20] and its practical use on

metabolomics data [RS05, RS07, RS24]. This PhD also led us to a more theoretical article on how to minimize the L^0 norm in high dimension, that is a common issue to perform variable selection [RS03].

3.2 Individual Prediction Regions for multivariate longitudinal data

3.2.1 Background and motivations

Individualized or preventive medecine are expanding domains [Hanczar and Bar-Hen, 2016, Pritchard et al., 2017, Ginsburg and Phillips, 2018] that could be achieved using a longitudinal individual follow-up of biological variables. It consists of monitoring the markers of important functions for the early detection of slowly progressive diseases with a subclinical phase. For example, the prostate specific antigen (PSA) is used to detect prostate cancer in men. The same kind of follow-up is systematically done with teenagers using their weight and height to detect the beginning of obesity. In sport, like cycling or athletics, anti-doping control authorities try to generalize the use of a biological passport which consists of a longitudinal follow-up of some endogenous substances of interest in order to detect abnormal variations in an individual [Sottas et al., 2007, Zorzoli and Rossi, 2010].

A standard method of doing these follow-ups is to use the so-called reference intervals [CLSI, 2008]. These intervals contain a fixed percentage (usually 95%) of measurements that can be observed in healthy individuals. However, this method suffers from several flaws. First, it does not use individual information *i.e.* a healthy individual can have extreme values, outside the reference interval, while for some other individuals values inside the reference interval are pathologic. Second, these intervals are built in an univariate framework (*i.e.* variable by variable) without taking into account the possible correlations between them. Finally, it does not account for their evolution over time within a given individual.

The individual reference intervals (or prediction intervals) mitigate this flaw by allowing the construction of a reference individual based on the observed values in a healthy individual and taking into account some covariables (such as sex, age). The literature on this subject is plentiful and the usual methodology is to use linear/nonlinear mixed effects models [Verbeke and Molenberghs, 2000, Davidian and Giltinan, 1995]. In these models the observations are usually assumed to be independent conditional to the individual specific parameters (compound symmetry assumption). To our knowledge, the development of reliable methods to detect abnormal variations of longitudinal variables has remained limited. Sottas et al. [2007] proposed a Bayesian approach to combine population-derived limits and individual-based thresholds. Nevertheless, this method is built in an univariate framework whereas a follow-up is usually performed on several markers. Intuition suggests that building regions using simultaneous information on correlated variables could help to better detect abnormal values. More recently, Wang and Fan [2010] proposed a method to build prediction regions. They used a p order autoregressive process to model the autocorrelation of a variable with time while the correlations between different variables is assumed to be fixed over time.

3.2.2 The model

We propose to build an individual prediction region from previous observations of these variables carried out in the same individual and model parameter estimates. The observations obtained in an individual are assumed to be correlated over time. The correlation between a variable X_1 at time t_1 and a variable X_2 at time t_2 is not assumed to be equal to the correlation between X_1 at time t_2 and X_2 at time t_1. This leads to highly structured autocorrelations that cannot be directly estimated by conventional methods (the NLMIXED procedure in SAS or the R package **nlme**). Therefore, we also proposed a specific estimation method (not detailed here).

Let us denote $\mathbf{X}_i = [\mathbf{X}_{i1} : \cdots : \mathbf{X}_{ir}]$ the measurements performed in the i^{th} individual of a sample of size N. The vector $\mathbf{X}_{ij}$ contains the n_i observations carried out over time for the j^{th} variable. More precisely, X_{ijk} is the value observed for the i^{th} individual for the j^{th} variable at time t_{ik}. Without loss of generality, we can assume that $t_{i1} \leq t_{i2} \leq \ldots \leq t_{in_i}$. Note that all the variables are supposed to be measured at the same time for an individual, but time measures may differ from one individual to another. We assume that, up to a monotonic transformation

$$\mathbf{X}_i = \mathbf{B}_i\boldsymbol{\beta} + \mathbf{T}_i\boldsymbol{\Phi}_i + \boldsymbol{\zeta}_i \tag{3.1}$$

where $\mathbf{B}_i$ and $\mathbf{T}_i$ are known full-rank covariate matrices of dimensions $n_i \times p$ and $n_i \times q$ respectively, $\boldsymbol{\beta} = [\boldsymbol{\beta}_1 : \cdots : \boldsymbol{\beta}_r]$ is a $p \times r$ matrix of parameters used to describe the population mean, $\boldsymbol{\Phi}_i = [\boldsymbol{\Phi}_{i1} : \cdots : \boldsymbol{\Phi}_{ir}]$ and $\boldsymbol{\zeta}_i = [\boldsymbol{\zeta}_{i1} : \cdots : \boldsymbol{\zeta}_{ir}]$ are respectively $q \times r$ and $n_i \times r$ matrices of unobserved Gaussian random effects. The variance of the components of the random matrix $\boldsymbol{\zeta}_i$ is assumed to be highly structured :

$$\text{cov}\left(\zeta_{ijk}, \zeta_{ij'k'}\right) = \Sigma_{jj'}\rho_{jj'}^{t_{ik}-t_{ik'}} \text{ if } k > k' \text{ and } \text{cov}\left(\zeta_{ijk}, \zeta_{ij'k'}\right) = \Sigma_{jj'}\rho_{j'j}^{t_{ik'}-t_{ik}} \text{ if } k < k'$$

where $\rho_{jj'} \in [0,1]$ and $\Sigma_{jj'} = \alpha_{jj'}\sigma_j\sigma_{j'}$. The numbers $\alpha_{jj} = 1, \forall j \in \{1,\ldots,r\}$, $\alpha_{jk} = \alpha_{kj} \in [-1,1] \forall j \neq k \in \{1,\ldots,r\}$ and $\rho_{jk} \in [0,1] \forall j,k \in \{1,\ldots,r\}$, σ_j represents the standard deviation of the j^{th} variable at each measurement time. The correlation between ζ_{ijk} and $\zeta_{ij'k'}$ is assumed to be $\rho_{jj'}^{t_{ik}-t_{ik'}}$ for $k > k'$ and $\rho_{j'j}^{t_{ik'}-t_{ik}}$ for $k < k'$. This means that we do not assume that the correlation between the j^{th} variable in $\boldsymbol{\zeta}_i$ measured at time k and the j'^{th} variable in $\boldsymbol{\zeta}_i$ measured at time k' is the same as the correlation between the j^{th} variable in $\boldsymbol{\zeta}_i$ measured at time k' and the j'^{th} variable in $\boldsymbol{\zeta}_i$ measured at time k. The major difference with the paper of Wang and Fan [2010] is that they assume that the observation times t_{ik} are equally spaced integer numbers, and that for all j and j', $\text{cov}\left(\zeta_{ijk}, \zeta_{ij'k'}\right) = \Sigma_{jj'}\rho_{|t-t'|}$ where $\rho_{|t-t'|}$ is the correlation of an auto-regressive process of order p.

The matrix of the covariance of the $\boldsymbol{\zeta}_i$ is a variance/covariance matrix because it is a symmetric and positive-definite matrix as the Kronecker and Shur products of two positives matrices [Bhatia, 2009]. If this model writing is easy to understand, its multidimensional nature does not facilitate the estimation of parameters and the distribution definition of $\boldsymbol{\Phi}_i$ and $\boldsymbol{\zeta}_i$. Thus, we rewrite this model in a vectorial framework to facilitate further estimations. Let us define $\boldsymbol{\psi}_i = \text{vec}(\boldsymbol{\Phi}_i)$ the vector obtained by stacking the columns of $\boldsymbol{\Phi}_i$ columnwise. We assume that $\boldsymbol{\psi}_i \underset{iid}{\sim} N(0, \boldsymbol{\Omega})$. The variance matrix $\boldsymbol{\Omega} = [\boldsymbol{\omega}_{jm}]_{jm}$ is block-partitioned with $q \times q$ variance matrices $\boldsymbol{\omega}_{jm} = \text{cov}(\boldsymbol{\Phi}_{\text{ij}}, \boldsymbol{\Phi}_{\text{im}})$.

Similarly, the within subject error $\boldsymbol{\zeta}_i$ can be stored columnwise into a vector $\boldsymbol{\varepsilon}_i = \text{vec}(\boldsymbol{\zeta}_i) \sim N(0, \boldsymbol{\Lambda}_i(\boldsymbol{\rho}, \boldsymbol{\Sigma}))$. The matrix $\boldsymbol{\Lambda}_i(\boldsymbol{\rho}, \boldsymbol{\Sigma})$ can be written as $\mathbf{D}_i^{-1}\mathbf{R}_i^{-1}\mathbf{D}_i^{-1}$ where $\mathbf{D}_i^{-1} =$

$\text{diag}(\sigma_1, \ldots, \sigma_1, \sigma_2, \ldots, \sigma_2, \sigma_r, \ldots, \sigma_r)$ with each σ_j repeated n_i times. Thus, it is assumed to be constant over time. The matrix $\mathbf{R}_i^{-1}(\boldsymbol{\rho})$ is block-partitioned with $n_i \times n_i$ matrices $\boldsymbol{\omega}_{ijk}$ with

$$\boldsymbol{\omega}_{ijk}(\rho) = \left(\text{corr}\left(\zeta_{ijl}, \zeta_{ikf}\right)\right)_{l,f \in \{1,\ldots,n_i\}} \quad = \quad \alpha_{jk}\rho_{jk}^{|t_{if}-t_{il}|}$$

and $\alpha_{jj} = 1$. The matrix $\boldsymbol{\omega}_{ijk}(\rho)$ contains the correlation between the j^{th} and k^{th} variable at the different sampling times.

The $\boldsymbol{\varepsilon}_i$'s are assumed mutually independent and independent of the $\boldsymbol{\psi}_i$'s. Let us define $\mathbf{Y}_i = \text{vec}(\mathbf{X}_i)$, $\mathbf{A}_i = \mathbb{1}_r \otimes \mathbf{B}_i$ and $\mathbf{Z}_i = \mathbb{1}_r \otimes \mathbf{T}_i$. Using these notations, the model (3.1) can be re-written as

$$\mathbf{Y}_i = \mathbf{A}_i\boldsymbol{\theta} + \mathbf{Z}_i\boldsymbol{\psi}_i + \boldsymbol{\varepsilon}_i \tag{3.2}$$

where $\boldsymbol{\theta} = \text{vec}(\boldsymbol{\beta})$. This model may appear to be a standard linear mixed effect model whose parameter $\boldsymbol{\xi} = (\boldsymbol{\theta}, \boldsymbol{\Omega}, \boldsymbol{\Sigma}, \boldsymbol{\rho}) \in \Xi$ can be easily estimated using standard statistical software. However, the covariance matrices of this model are highly structured and their estimation needs careful development.

Assume that n_w observations of the r variables are available at times $(t_1, \ldots, t_{n_w})$ in a new individual. Let us denote $\mathbf{U} \in \mathbb{R}^{r\times 1}$ the future values that will be observed at time $t_u > t_{n_w}$ for the r variables in this new individual. We assume that

$$\left(\mathbf{W}'\,\mathbf{U}'\right)' = \mathbf{A}\boldsymbol{\theta} + \mathbf{Z}\boldsymbol{\psi} + \boldsymbol{\varepsilon}$$

where $\mathbf{Z} = \left(\mathbf{Z}_w'\,\mathbf{Z}_u'\right)'$, $\mathbf{A} = \left(\mathbf{A}_w'\,\mathbf{A}_u'\right)'$ are known matrices and $\boldsymbol{\varepsilon} = \left(\boldsymbol{\varepsilon}_w'\,\boldsymbol{\varepsilon}_u'\right)'$. The random matrix $\left(\mathbf{W}'\,\mathbf{U}'\right)'$ is assumed to be independent of the $\mathbf{Y}_i$'s. We are looking for a region $\mathcal{R}_\xi^\alpha(\mathbf{W})$ so that $P\left(\mathbf{U} \in \mathcal{R}_\xi^\alpha(\mathbf{W}) \middle| \mathbf{W}\right) = 1 - \alpha$.
To build such a region, we need two things : a random sample of individuals $(\mathbf{Y}_i)_{i\in\{1,\ldots,N\}}$ that enables the population parameters $\boldsymbol{\xi}$ to be estimated and some observations performed in the individual of interest $\mathbf{W}$. We proceed in three steps : first, we build a region $\mathcal{R}_{\boldsymbol{\xi}}^\alpha(\mathbf{W})$ by assuming that $\boldsymbol{\xi}$ is known, secondly, we plug-in the estimate $\widehat{\boldsymbol{\xi}}$ of $\boldsymbol{\xi}$ obtained using the sample $(\mathbf{Y}_i)_{i\in\{1,\ldots,N\}}$ into $\mathcal{R}_{\boldsymbol{\xi}}^\alpha(\mathbf{W})$ to get $\mathcal{R}_{\widehat{\boldsymbol{\xi}}}^\alpha(\mathbf{W})$. Of course, because the estimate $\widehat{\boldsymbol{\xi}}$ is a random variable, this plug-in estimator does not guarantee a coverage of $1 - \alpha$.

We also define an *ad hoc* procedure to estimate the different parameters. It is based on the EM algorithm [Dempster et al., 1977] and on a good choice of starting values to speed the convergence of the algorithm. Finally, the computer time needed for parameter estimation is less than one second using an ordinary laptop.

Building prediction regions

Remind that we assume that observations $\boldsymbol{W}$ for the r variables are available in a new individual. We are going to build a prediction region for the next observation $\boldsymbol{U}$ for this new individual. From the model defined in (3.2), we have

$$\boldsymbol{U} = \boldsymbol{A}_u\boldsymbol{\theta} + \boldsymbol{Z}_u\boldsymbol{\psi}_u + \boldsymbol{\varepsilon}_u. \tag{3.3}$$

We assume here that all the model parameters are known. We denote

$$\boldsymbol{E} = \text{vec}(\boldsymbol{\varepsilon}_w, \boldsymbol{\varepsilon}_u) \quad \sim \quad N\left(0; \begin{pmatrix} \boldsymbol{\Lambda}_w(\boldsymbol{\rho}, \boldsymbol{\Sigma}) & \boldsymbol{M}_{wu}(\boldsymbol{\rho}, \boldsymbol{\Sigma})' \\ \boldsymbol{M}_{wu}(\boldsymbol{\rho}, \boldsymbol{\Sigma}) & \boldsymbol{\Lambda}_u(\boldsymbol{\rho}, \boldsymbol{\Sigma}) \end{pmatrix}\right)$$

where $\boldsymbol{\Lambda}_w(\boldsymbol{\rho},\boldsymbol{\Sigma})$ and $\boldsymbol{\Lambda}_u(\boldsymbol{\rho},\boldsymbol{\Sigma})$ are defined in the first section and $\boldsymbol{M}_{wu}(\boldsymbol{\rho},\boldsymbol{\Sigma})$ is a $r \times (rn_w)$ matrix with

$$\boldsymbol{M}_{wu}(\boldsymbol{\rho},\boldsymbol{\Sigma}) = \operatorname{cov}(\boldsymbol{\varepsilon}_w;\boldsymbol{\varepsilon}_u) = \begin{pmatrix} \operatorname{cov}\left(\varepsilon_u^1;\boldsymbol{\varepsilon}_{k=1,\dots,n_w}^1\right) & \dots & \operatorname{cov}\left(\varepsilon_u^1;\boldsymbol{\varepsilon}_{k=1,\dots,n_w}^r\right) \\ \vdots & \dots & \vdots \\ \operatorname{cov}\left(\varepsilon_u^r;\boldsymbol{\varepsilon}_{k=1,\dots,n_w}^1\right) & \dots & \operatorname{cov}\left(\varepsilon_u^r;\boldsymbol{\varepsilon}_{k=1,\dots,n_w}^r\right) \end{pmatrix}$$

where ε_u^i is the i^{th} term of ε_u and $\boldsymbol{\varepsilon}_{k=1,\dots,n_w}^j$ is a n_w dimensional vector for variable j and individual $\boldsymbol{W}$ and

$$\operatorname{cov}\left(\varepsilon_u^i;\boldsymbol{\varepsilon}_{k=1,\dots,n_w}^j\right) = \left(\Sigma_{ij}\rho_{ij}^{|t_u-t_1|},\dots,\Sigma_{ij}\rho_{ij}^{|t_u-t_{n_w}|}\right).$$

Using these notations and Schur lemma, we obtain the following proposition.

Proposition 3.1 *Let α be any real number in $[0;1]$ and $\chi^2_{r,1-\alpha}$ be the $1-\alpha$ quantile of a chi-square distribution with r degrees of freedom. Let us consider the vector $\boldsymbol{m}(\boldsymbol{\xi},\boldsymbol{W})$ and the matrix $\boldsymbol{V}(\boldsymbol{\xi})$ defined by*

$$\begin{aligned} \boldsymbol{m}(\boldsymbol{\xi},\boldsymbol{W}) &= \boldsymbol{A}_u\boldsymbol{\theta} + \left(\boldsymbol{Z}_u\boldsymbol{\Omega}\boldsymbol{Z}'_w + \boldsymbol{M}_{wu}(\boldsymbol{\rho},\boldsymbol{\Sigma})\right)\left(\boldsymbol{Z}_w\boldsymbol{\Omega}\boldsymbol{Z}'_w + \boldsymbol{\Lambda}_w(\boldsymbol{\rho},\boldsymbol{\Sigma})\right)^{-1}\left(\boldsymbol{W}-\boldsymbol{A}_w\boldsymbol{\theta}\right), \\ \boldsymbol{V}(\boldsymbol{\xi}) &= \left(\boldsymbol{Z}_u\boldsymbol{\Omega}\boldsymbol{Z}'_u + \boldsymbol{\Lambda}_u(\boldsymbol{\rho},\boldsymbol{\Sigma})\right) \\ &- \left(\boldsymbol{Z}_u\boldsymbol{\Omega}\boldsymbol{Z}'_w + \boldsymbol{M}_{wu}(\boldsymbol{\rho},\boldsymbol{\Sigma})\right)\left(\boldsymbol{Z}_w\boldsymbol{\Omega}\boldsymbol{Z}'_w + \boldsymbol{\Lambda}_w(\boldsymbol{\rho},\boldsymbol{\Sigma})\right)^{-1}\left(\boldsymbol{Z}_u\boldsymbol{\Omega}\boldsymbol{Z}'_w + \boldsymbol{M}_{wu}(\boldsymbol{\rho},\boldsymbol{\Sigma})\right)'. \end{aligned}$$

A $(1-\alpha)$ prediction region of $\boldsymbol{U}$, conditionally to $\boldsymbol{W}$, is the set

$$S = \left\{\boldsymbol{u}\in\mathbb{R}^r; \left\|\boldsymbol{V}(\boldsymbol{\xi})^{-1/2}\left(\boldsymbol{u}-\boldsymbol{m}(\boldsymbol{\xi},\boldsymbol{W})\right)\right\|^2 \le \chi^2_{r,1-\alpha}\right\} \tag{3.4}$$

where $\boldsymbol{V}(\boldsymbol{\xi})^{-1/2}$ is the inverse of the Cholesky transformation of $\boldsymbol{V}(\boldsymbol{\xi})$.

When $r>1$, the prediction region for $\boldsymbol{U}$ is thus an ellipsoid centered on $\boldsymbol{m}(\boldsymbol{\xi},\boldsymbol{W})$. This ellipsoid degenerates to the interval $\left[m(\boldsymbol{\xi},\boldsymbol{W}) - \tau_{(1-\alpha/2)}\sqrt{V(\xi)}; m(\boldsymbol{\xi},\boldsymbol{W}) + \tau_{(1-\alpha/2)}\sqrt{V(\xi)}\right]$, where $\tau_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ quantile of the standard gaussian distribution, when one wants to predict the next value U of a single variable (*i.e.* $r=1$).
In this case, if $\rho=0$ and assuming that there is no covariable, the j^{th} observation in the i^{th} individual writes

$$X_{ij} = Y_{ij} = \theta + \psi_i + \varepsilon_{ij}$$

with $\psi_i \sim N(0,\omega^2)$ and $\varepsilon_{ij} \sim N(0,\sigma^2)$. Using Schur complement, the $100(1-\alpha)\%$ prediction interval for the future value when $k-1$ observations are already available in an individual has the following expression :

$$\left[\frac{\theta}{1+\gamma^2(k-1)} + \frac{\gamma^2(k-1)}{1+\gamma^2(k-1)}\overline{W}_{k-1} - \tau_{(1-\alpha/2)}\sqrt{\frac{1+\gamma^2 k}{1+\gamma^2(k-1)}\sigma^2}, \right. \tag{3.5}$$
$$\left. \frac{\theta}{1+\gamma^2(k-1)} + \frac{\gamma^2(k-1)}{1+\gamma^2(k-1)}\overline{W}_{k-1} + \tau_{(1-\alpha/2)}\sqrt{\frac{1+\gamma^2 k}{1+\gamma^2(k-1)}\sigma^2}\right],$$

where $\gamma=\omega/\sigma$ and $\overline{W}_{k-1}$ is the average of the $k-1$ available observations. Note that γ measures the benefit of the individualization compared to the usual reference interval built with a single value per individual [CLSI, 2008]. When γ is high, the prediction interval is close to $\left[\overline{W} \pm \tau_{(1-\alpha/2)}\sigma\right]$ and the individualization is beneficial.

Plug-in corrections

These regions are then built assuming that the model parameters are known while estimates are used to compute it. While this plug-in method is easy to use, its very nature does not guarantee an exact coverage rate for the prediction region because it does not account for the imprecision of the parameter estimates. This can be a real problem when the sample size is small [Barndorff-Nielsen and Cox, 1996]. Therefore, special attention has to be paid to this problem to control the real coverage rate of the built prediction region. By consequence, we proposed three different corrections of the asymptotic confidence region that were compared on the real dataset. These corrections aim at correcting the plug-in estimation of the prediction region. The first two come from Beran [1990], Hall et al. [1999], Ueki and Fueda [2007], Fonseca et al. [2012], and can be read as delta-methods. The third method is an application of a simple parametric bootstrap method. These corrections are then compared on the real data set. It can be noted that the third correction gives the narrower prediction region which was expected because it does not assume any *a priori* distribution. Its only approximation is to substitute the real distribution by its bootstrap counterpart. As expected it achieves a coverage probability very close to the targeted one.

3.2.3 Real dataset application

The data come from a prospective study aimed at evaluating variations over time of several biochemical variables in healthy cats. This study was carried in the clinics of the Veterinary College, which usually received sick animals or healthy animals for sterilization (obviously only once). This is the reason why only $N = 20$ healthy cats could have been included in this study. The main variables for renal follow-up are the urea X_1, the creatinine X_2 and the protein X_3 which are plotted in Figure 3.1 for the 20 healthy cats. There is no reason to think that these variables are not stable over time in healthy cats [Reynolds et al., 2010, Lefebvre, 2011]. Univariate analyses were performed and the effect of time was found not significant. Every cat but three were measured five times : 0, 3, 6, 12 and 24 months after inclusion. The remaining three were sampled only for the first four times. Note that the entire study is performed on the log transformation of the variables as usual.

So, according to previous results, we propose the following model

$$\boldsymbol{X}_i = \boldsymbol{B}_i\boldsymbol{\beta} + \boldsymbol{T}_i\boldsymbol{\Phi}_i + \boldsymbol{\zeta}_i \tag{3.6}$$

where $\boldsymbol{X}_i$ is a $n_i \times 3$ matrix with n_i the number of observation for the i^{th} cat (4 or 5), $\boldsymbol{B}_i$ and $\boldsymbol{T}_i$ are vectors of length n_i such that $\boldsymbol{B}_i = \boldsymbol{T}_i = (1, \ldots, 1)$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ and $\boldsymbol{\phi}_i = (\phi_{i1}, \phi_{i2}, \phi_{i3})$ and $\boldsymbol{\zeta}_i$ is a $n_i \times 3$ matrix. Here we have $n_i = 4$ or 5, $r = 3$, $p = 1$, $q = 1$ and $N = 20$. As we have no available covariable (age, sex), the matrix $\boldsymbol{B}_i$ does not incorporate any information but this kind of information can easily be inserted in our model as in Sottas et al. [2007].

In this example, the estimation of the parameters provide a correlation between two successive measurements carried out in the same individual is rather low for practical use with $t' - t > 1$ month. More surprisingly, it appears that no variable is an earlier marker than the others to detect kidney insufficiency. In other words, there is no major correlation between two different variables at two different times. This result could not be anticipated. With this result, the benefit of the individualization can be roughly measured by the ratio $\gamma = \omega/\sigma$ (see (3.5)) which is equal to 1.8, 2.0 and 1.7 for urea, creatinine and protein respectively. As these ratios

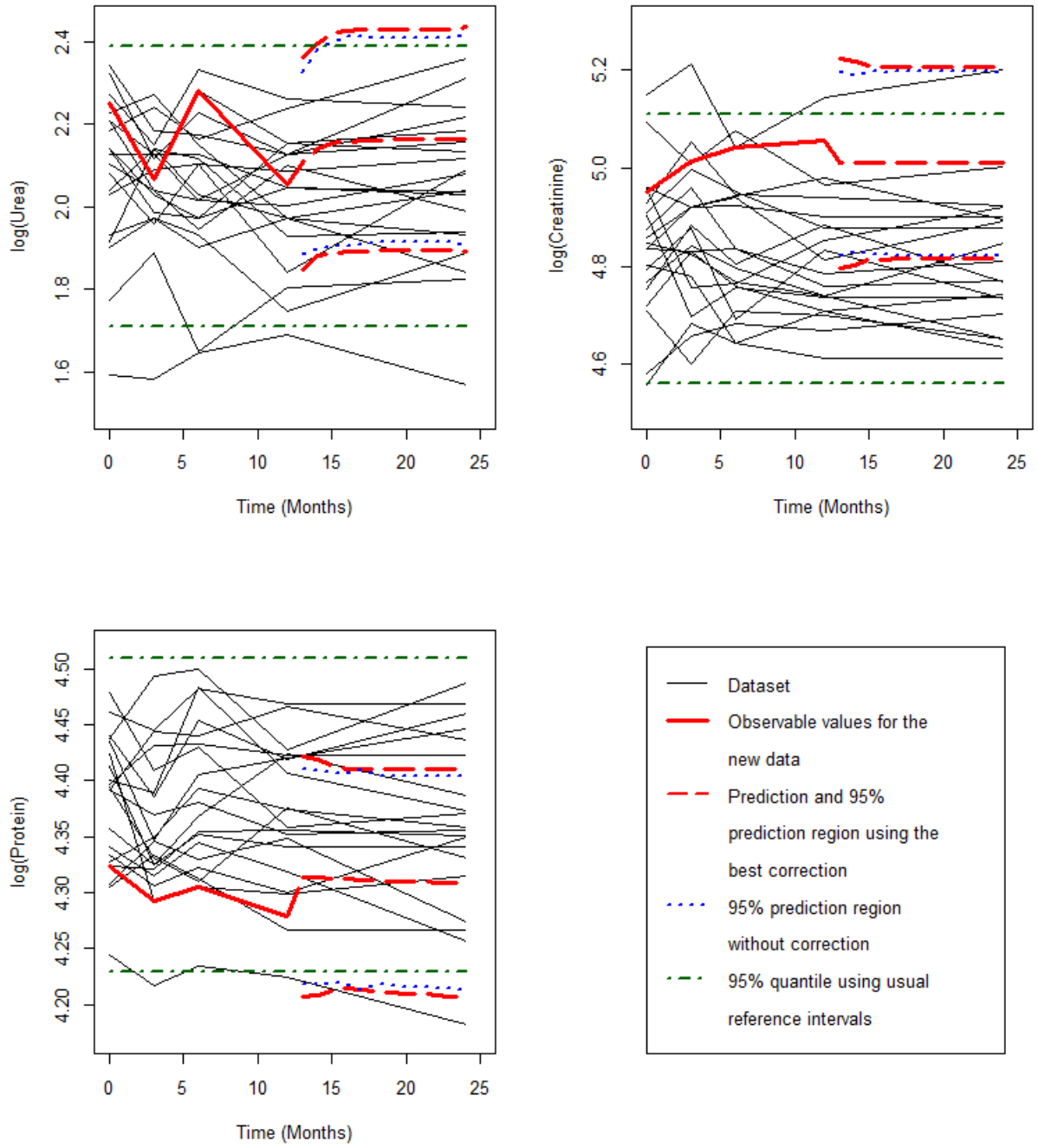

FIGURE 3.1 – The three variables of interest are plotted for the whole dataset and the new cat in bold. The usual reference intervals (in dotted-dashed lines) are wider than the individual ones (in dashed lines).

are greater than one, one can expect the individualized region (3.4) to be narrower than the population counterpart.

Now, there is a new cat for which we possess four measurements (at 0, 3, 6 and 12 months)

for each variable. Using the proposed method, we can build an individual reference region (an ellipsoid) for future values for these variables. If its future measures lie outside this region, this cat has a low probability of being healthy. The results on this new cat are plotted in Figure 3.1. Because clinicians are not accustomed to matrix calculus, it is not easy to check whether or not a new point on the given cat belongs to its prediction region. This is the reason why we proposed to represent the projection of the ellipsoid for each variable. This gives an interval of prediction for each variable and each future time of measurement. Note that these intervals are presented to give a graphical representation. As they were obtained by projection they do not guarantee the right coverage contrarily to the ellipsoids defined by Proposition 4.1. So, they can not be used separately to diagnose a cat as the three variables are strongly related. As soon as a value of a variable is outside the prediction region, the cat can be considered as probably not healthy.

We can remark that our prediction intervals are very different and narrower than the so-called "reference intervals" and therefore lead to different clinical decisions. As an example, a log(Creatinine) of 5.15 at fifteen months would be detected as suspicious for the new cat using the standard reference intervals while the individualization does not trigger such a false alarm. On the other hand, a log(Urea) value of 1.8 would be detected as abnormal by our method but not by the usual reference intervals. The reduction of width for the prediction region decreases the probability for each individual of being detected as a false-negative. Despite the considerable difference between the χ^2 threshold and our estimate, the corresponding prediction regions are very close. In this case, this can be explained by a small variance in a future value conditional on the observations. This cannot easily be anticipated by a simple glance on the parameter estimations because this conditional variance depends on a complicated function of all the variance parameters (see Proposition 3.1).

3.2.4 Discussion

The main novelty of our approach lies in its individualization and multidimensionality. Indeed, every individual gets its own prediction region which takes into account the possible correlations between all the variables at all the different times. These advantages enable us to build narrower prediction regions than the usual "reference intervals" method. Using our methodology, clinicians will be alerted with more precision to a potential unhealthy animal or person. Nevertheless, our model is based on two assumptions which can be false. First, the Gaussian one. An alternative could be the use of a nonparametric framework but it would need more individuals and, by consequence, it can not be applied to our practical problem. This assumption is also a classical one at least up to a Box-Cox transformation [CLSI, 2008]. Second, an assumption was also made on the exponential decrease of the correlation over time that can appear restrictive. To the best of our knowledge, this kind of problem has already been modeled by an $AR(p)$-process [Wang and Fan, 2010] : an assumption difficult to check. In this respect, the model we propose can be seen in continuous time as a first order approximation of such chains.

As mentioned, this method could be of great interest to detect doping. Indeed, the World Anti-Doping Agency biological passport is a follow-up of professional athletes on different hematological or urine markers. By consequence, our multivariate longitudinal approach could produce narrower prediction regions and help in the detection of doping compared to the current methodology, based on Sottas et al. [2007]. So, we contacted P.-E. Sottas (responsible

of the World Anti-Doping Agency biological passport) but he did not want to test our new multivariate approach despite its obvious interest. Note that a very recent paper from Saulière et al. [2018] address the same issue that our work. It is based on maxima of Z-scores and does not rely on the use of an extra population to calibrate some model parameters. Nevertheless, its multivariate extension is based on an independent assumption between variables and, by consequence, does not take into account the possible correlation between variables overtime. It would be interesting to study the building of a new procedure based on both advantages of the two methods and to compare the two different approaches on their database composed of the follow-up of elite soccer players. In a more general way, a lot of problems are raised by applications in sports. Beyond the already mentioned doping issue, one can cite among a large literature the study of the potential number of winners of a tournament using a Bradley-Terry model [Chetrite et al., 2017], the use of spatial statistics to characterize defensive skills in basketball [Franks et al., 2015] or more applied works such as the study of the collective effectiveness in the XV de France [Bar-Hen, 2017] or the risk study of common illnesses for elite swimmers [Hellard et al., 2015] ... Nevertheless, a lot of sport data analyses remain qualitative and, with the development of new technologies and the growth of financial interests in sport, a lot of new data are now measured (such as optical player tracking systems) without any dedicated statistical analysis method. So, this domain seems very promising and attractive and also leads to the very recent creation of the French Sport Statistics Group in the SFdS.

3.2.5 Mixed effect model for pharmacokinetics

Mixed effects model are also widely used by pharmacologists in my unit, and I was involved in a projet to predict the internal exposure to bisphenol A of the human fetus during late pregnancy. Different dose levels are tested on an ovine feto-maternal animal model (on mother and/or fetus) and a longitudinal follow-up of the concentrations of bisphenol A is then carried out on the different compartments. A compartmental human model is then derived based on a non-linear mixed effect model on the ovine dataset and a reparametrization using human pharmacokinetic parameters. The predicted concentrations result in a fetal exposure to BPA during late pregnancy.

3.3 Intervals selection for functional data

3.3.1 Motivation

A challenging agronomic problem is the inference of interpretable climate-yield relationships on complex crop models. Process-based crop model are developed to simulate the annual grain yield Y (in tons per hectare) of sunflower cultivars, as a function of $X = \{$time, environment (soil and climate), management practice and genetic diversity$\}$ [Casadebaig et al., 2011]. This model requires functional inputs in the form of climatic series. These series consist of daily measures of five variables over a year : minimal temperature, maximal temperature, global incident radiation, precipitations and evapotranspiration. Due to the complexity of plant-climate interactions and the strongly irregular nature of climatic data, understanding the relation between yield and climate is a particularly challenging task.

In this practical situations, the relevant information may not correspond to isolated evaluation points of X neither to some of the components of its expansion on a functional basis, but to its value on some continuous intervals, $X([t_a, t_b])$. In that case, variable selection amounts

to identify those intervals. As advocated by James et al. [2009], a desirable feature of variable selection provided by such an approach is to enhance the interpretability of the relation between X and Y. Indeed, it reduces the definition domain of the predictors to a few influential intervals, or it focuses on some particular aspects of the curves in order to obtain expected values for Y. Tackling this issue can be seen as selecting groups of contiguous variables (*i.e.*, intervals) instead of selecting isolated variables. Fraiman et al. [2016], in the linear setting, and Fauvel et al. [2015], Ferraty and Hall [2015], in a nonparametric framework, propose several alternatives to do so. However, no specific contiguity constraint is put on groups of variables.

To solve this problem, we focus here on the functional regression problem, in which a real random variable Y is predicted from a functional predictor $X(t)$ that takes values in a functional space (*e.g.*, $L^2([0,1])$, the space of squared integrable functions over $[0,1]$), based on a set of observed pairs (X,Y), $(\mathbf{x}_i, y_i)_{i=1,\dots,n}$. The main challenge with functional regression lies in its high dimension : the underlying dimension of a functional space is infinite, and even if the digitized version of the curves is considered, the number of evaluation points is typically much larger than the number of observations. A number of classical approaches have been extended to this framework, including linear models [Cardot et al., 1999] or kernel-based methods [Ferraty and Vieu, 2006]. These extensions rely on some kind of dimension reduction by representing the functional predictors on a functional basis, either predefined (splines, wavelets...) or data-driven (using PCA for instance). It is also possible to tailor the basis to the regression problem : this is the idea of the Sliced Inverse Regression [SIR, Li, 1991], which has been extended to the functional framework in Ferré and Villa [2006].

Recently, an increasing number of works have focused on variable selection in this functional regression framework, in particular in the linear setting. The problem is to select parts of the definition domain of X that are relevant to predict Y. Considering digitized versions of the functional predictor X, approaches based on Lasso have been proposed to select a few isolated points of X [Ferraty et al., 2010, Aneiros and Vieu, 2014, Kneip et al., 2016]. Alternatively, other authors proposed to perform variable selection on predefined functional bases. For instance, Matsui and Konishi [2011] used L^1 regularization on Gaussian basis functions and Chen et al. [2015] on wavelets.

In the present work, we propose a semi-parametric model that selects intervals in the definition domain of X with an automatic approach. The method is based on SIR, even though it could easily be extended to linear regression. Our choice for SIR is motivated by the fact that the method is based on a semi-parametric model that is more flexible than linear models. However, at the same time, since it is based on a prior linear dimension reduction, it can be conveniently penalized by L_1-type penalty to select groups of variables corresponding to intervals in the definition domain of the functional predictors. Our second contribution is a fast and automatic procedure based on the full regularization path of the Lasso for building intervals in the definition domain of the predictors without using any prior knowledge.

3.3.2 Sparse Sliced Inverse Regression (SIR)

SIR

In this subsection, we review the standard SIR for multivariate data and its extensions to the high-dimensional setting. Here, (X,Y) denotes a random pair of variables such that X takes values in $\mathbb{R}^p$ and Y is real. We assume given n i.i.d. realizations of (X,Y), $(\mathbf{x}_i, y_i)_{i=1,\dots,n}$.

When p is large, classical modeling approaches suffer from the well-known curse of dimensionality. A standard way to overcome this issue is to rely on dimension reduction techniques.

This kind of approaches is based on the assumption that there exists an Effective Dimension Reduction (EDR) space $\mathcal{S}_{Y|X}$ which is the smallest subspace such that the projection of X on $\mathcal{S}_{Y|X}$ retains all the information on Y contained in the predictor X. More precisely, $\mathcal{S}_{Y|X}$ is assumed of the form $\text{Span}\{\mathbf{a}_1, \ldots, \mathbf{a}_d\}$, with $d \ll p$, such that

$$Y = F(\mathbf{a}_1^\top X, \ldots, \mathbf{a}_d^\top X, \epsilon), \tag{3.7}$$

in which $F : \mathbb{R}^{p+1} \to \mathbb{R}$ is an unknown function and ϵ is an error term independent of X. To estimate this subspace, SIR is one of the most classical approaches when $p < n$: under an appropriate and general enough condition, Li [1991] shows that $\mathbf{a}_1, \ldots, \mathbf{a}_d$ can be estimated as the first d Σ-orthonormal eigenvectors of the generalized eigenvalue problem : $\Gamma\mathbf{a} = \lambda\Sigma\mathbf{a}$, in which Σ is the covariance matrix of X and Γ is the covariance matrix of $\mathbb{E}(X|Y)$.

In practice, Σ is replaced by the empirical covariance, $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \left(\mathbf{x}_i - \overline{X}\right)\left(\mathbf{x}_i - \overline{X}\right)^\top$, and Γ is estimated by "slicing" the observations $(y_i)_i$ as follows. The range of Y is partitioned into H consecutive and non-overlapping slices, denoted hereafter $\mathcal{S}_1$, ..., $\mathcal{S}_H$. An estimate of $\mathbb{E}(X|Y)$ is thus simply obtained by $\left(\overline{X}_1, \ldots, \overline{X}_H\right)$ in which $\overline{X}_h$ is the average of the observations $\mathbf{x}_i$ such that y_i is in $\mathcal{S}_h$ and $\overline{X}_h$ is associated with the empirical frequency $p_h = \frac{n_h}{n}$ with n_h the number of observations in $\mathcal{S}_h$. $\widehat{\Gamma}$ is thus defined as $\sum_{h=1}^H p_h \overline{X}_h \overline{X}_h^\top$.

However, as detailed in Dauxois et al. [2001], Li and Yin [2008], in a high dimensional or functional setting , $\widehat{\Sigma}$ is singular and the SIR problem is thus ill-posed. Solutions to overcome this difficulty include variable selection [Coudret et al., 2014], ridge regularization or sparsity constraints. Indeed, in the high-dimensional setting,if we denote $A \in \mathbb{R}^{p\times d}$ the matrix in which the searched vectors a_j are the columns and $C = (C_1, \ldots, C_H)$, with $C_h \in \mathbb{R}^D$ (for $h = 1, \ldots, H$). Bernard-Michel et al. [2008] shows that the regularization of $\widehat{\Sigma}$ leads to an optimization problem dependint on A and C and that minimizing this optimization problem is also equivalent to finding the first d eigenvectors of $\left(\widehat{\Sigma} + \mu_2 \mathbb{I}_p\right)^{-1}\widehat{\Gamma}$.

Sparse SIR

Sparse estimates of $\mathbf{a}_j$ usually increase the interpretability of the model (here, of the EDR space) by focusing on the most important predictors only. To the best of our knowledge, only two alternatives have been introduced to use such methods.

Li and Yin [2008] derive a sparse ridge estimator from Cook [2004], Ni et al. [2005]. Given $(\hat{A}, \hat{C})$, solution of the ridge SIR, a shrinkage index vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^\top \in \mathbb{R}^p$ is obtained by minimizing a least square error with L_1 penalty :

$$\mathcal{E}_{s,1}(\boldsymbol{\alpha}) \;=\; \sum_{h=1}^H \hat{p}_h \left\| \left(\overline{X}_h - \overline{X}\right) - \widehat{\Sigma}\text{Diag}(\boldsymbol{\alpha})\hat{A}\hat{C}_h \right\|_{\mathbb{I}_p}^2 + \mu_1 \|\boldsymbol{\alpha}\|_{L_1}, \tag{3.8}$$

for a given $\mu_1 \in \mathbb{R}+*$ where $\|\boldsymbol{\alpha}\|_{L_1} = \sum_{i=1}^p |\alpha_p|$. Once the coefficients $\boldsymbol{\alpha}$ have been estimated, the EDR space is the space spanned by the columns of $\text{Diag}(\hat{\boldsymbol{\alpha}})\hat{A}$, where $\hat{\boldsymbol{\alpha}}$ is the solution of the minimization of $\mathcal{E}_{s,1}(\boldsymbol{\alpha})$.

An alternative is described in Li and Nachtsheim [2008] using the correlation formulation of the SIR [Chen and Li, 1998]. After the standard SIR estimates $\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_d$ have been computed, they solve d independent minimization problems with sparsity constraints introduced as an L_1 penalty : $\forall\, j = 1, \ldots, d$,

$$\mathcal{E}_{s,2}(\mathbf{a}_j^s) = \sum_{i=1}^n \left[\mathcal{P}_{\hat{\mathbf{a}}_j}(X|y_i) - (\mathbf{a}_j^s)^\top \mathbf{x}_i\right]^2 + \mu_{1,j}\|\mathbf{a}_j^s\|_{L_1}, \tag{3.9}$$

in which $\mathcal{P}_{\hat{\mathbf{a}}_j}(X|y_i) = \widehat{\mathbb{E}}(X|Y = y_i)^\top \hat{\mathbf{a}}^j$, with $\widehat{\mathbb{E}}(X|Y = y_i) = \overline{X}_h$ for h such that $y_i \in \mathcal{S}_h$ in the case of a sliced estimate of $\widehat{\mathbb{E}}(X|Y)$. Note that both proposals have problems in the high-dimensional setting :

- In their proposal, Li and Yin [2008] avoid the issue of the singularity of $\widehat{\Sigma}$ by working in the original scale of the predictors for both the ridge and the sparse approach (hence the use of the $\|.\|_{\mathbb{I}_p}$-norm in Equation (3.8) instead of the standard $\|.\|_{\widehat{\Sigma}^{-1}}$-norm where $\forall\, u \in \mathbb{R}p$, $\|u\|^2_{\widehat{\Sigma}^{-1}} = u^\top \widehat{\Sigma}^{-1} u$). For the ridge problem, this choice has been proved to produce a degenerate problem [Bernard-Michel et al., 2008].
- Li and Nachtsheim [2008] base their sparse version of the SIR on the standard estimates of the SIR problem that cannot be computed in the high-dimensional setting.

Moreover, the other differences between these two approaches can be summarized in two points :

- using the approach of Li and Yin [2008] based on shrinkage coefficients, the index α_p where $\alpha_p > 0$ are the same on all the d dimensions of the EDR. This makes sense because the vectors $\mathbf{a}_j$ themselves are not relevant : only the space spanned by them is and so there is no interest to select different variables j for the d estimated directions. Moreover, this allows to formulate the optimization in a single problem. However, this problem relies on a least square minimization with dependent variables in a high dimensional space $\mathbb{R}^p$;
- on the contrary, the approach of Li and Nachtsheim [2008] relies on a least square problem based on projections and is thus obtained from d independent optimization problems. The dimension of the dependent variable is reduced but the different vectors which span the EDR space are estimated independently and not simultaneously.

In our proposal, we combine both advantages of these 2 methods using a single optimization problem based on the correlation formulation of SIR. In this problem, the dimension of the dependent variable is reduced (d instead of p) when compared to the approach of Li and Yin [2008] and it is thus computationally more efficient. Identical sparsity constraints are imposed on all d dimensions using a shrinkage approach, but instead of selecting the nonzero variables independently, we adapt the sparsity constraint to the functional setting to avoid selecting isolated measurement points.

Sparse and Interpretable SIR (SISIR)

A functional regression framework is now assumed. X is thus a functional random variable, taking value in a (infinite dimensional) Hilbert space. $(x_i, y_i)_{i=1,\dots,n}$ are n i.i.d. realizations of (X, Y). However, x_i are not perfectly known but observed on a given (deterministic) grid $\tau = \{t_1, \dots, t_p\}$. We denote by $\mathbf{x}_i = (x_i(t_j))_{j=1,\dots,p} \in \mathbb{R}p$ the i-th observation, by $\mathbf{x}^j = (x_i(t_j))_{i=1,\dots,n}$ the observations at t_j and by $\mathbf{X}$ the $n \times p$ matrix $(\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$. Unless said otherwise, the notations are derived from the ones introduced in the multidimensional setting (Section 3.3.2) by using the $\mathbf{x}_i$ as realizations of X.

Contrary to most methods in functional data analysis, we do not assume smoothess on X or on the EDR space. We take advantage of the functional aspects of the data in a different way, using the natural ordering of the definition domain of X to impose sparsity on the EDR space. To do so, we assume that this definition domain is partitioned into D contiguous and non-overlapping intervals, $\tau_1, \dots, \tau_D$. In the present section, these intervals are supposed to be given *a priori* and we will describe later a fully automated procedure to obtain them from the data.

First, using the formulation of Bernard-Michel et al. [2008] we solve the ridge step and obtain $\hat{A}$ and $\hat{C}$.

3.3.3 Interval-sparse estimation

Once $\hat{A}$ and $\hat{C}$ have been computed, the estimated projections of $(\widehat{\mathbb{E}}(X|Y=y_i))_{i=1,\dots,n}$ onto the EDR space are obtained by : $\mathcal{P}_{\hat{A}}(\widehat{\mathbb{E}}(X|Y=y_i)) = (\overline{X}_h - \overline{X})^\top \hat{A}$, for h such that $y_i \in \mathcal{S}_h$. This p dimensional vector will be denoted by $(\mathcal{P}_i^1, \dots, \mathcal{P}_i^p)^\top$. In addition, we will also denote by $\mathbf{P}^j$ (for $j = 1, \dots, d$), $\mathbf{P}^j = (\mathcal{P}_1^j, \dots, \mathcal{P}_n^j)^\top \in \mathbb{R}n$.

D shrinkage coefficients, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_D) \in \mathbb{R}^D$, one for each interval $(\tau_k)_{k=1,\dots,D}$, are finally estimated. This leads to solve the following Lasso problem

$$\arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^D} \|\mathbf{P} - \Delta(\mathbf{X}\hat{A})\,\boldsymbol{\alpha}\|^2 + \mu_1\|\boldsymbol{\alpha}\|_{L_1} \tag{3.10}$$

with $\mathbf{P} = \begin{pmatrix} \mathbf{P}^1 \\ \vdots \\ \mathbf{P}^d \end{pmatrix}$, a vector of size dn and $\Delta(\mathbf{X}\hat{A}) = \begin{pmatrix} \mathbf{X}\Delta(\hat{\mathbf{a}}_1) \\ \vdots \\ \mathbf{X}\Delta(\hat{\mathbf{a}}_p) \end{pmatrix}$, a $(dn) \times D$-matrix with $\Delta(\hat{\mathbf{a}}_j)$ the $(p \times D)$-matrix such that $\Delta_{lk}(\hat{\mathbf{a}}_j)$, is the l-th entry of $\hat{\mathbf{a}}_j$, $\hat{a}_{jl}$, if $t_l \in \tau_k$ and 0 otherwise.

$\hat{\boldsymbol{\alpha}}$ are used to define the $\hat{\mathbf{a}}_j^s$ of the vectors spanning the EDR space by :

$$\forall\, l = 1, \dots, p,\ \hat{a}_{jl}^s = \hat{\alpha}_k\, \hat{a}_{jl} \text{ for } k \text{ such that } t_l \in \tau_k.$$

Once the sparse vectors $(\hat{\mathbf{a}}_j^s)_{j=1,\dots,d}$ have been obtained, an Hilbert-Schmidt orthonormalization approach is used to make them $\widehat{\Sigma}$-orthonormal.

Of note, as a single shrinkage coefficient is defined for all $(\hat{a}_{jl})_{t_l\in\tau_k}$, the method is close to group-Lasso [Simon et al., 2013], in the sense that, for a given $k \in \{1, \dots, D\}$, estimated $(\hat{a}_{jl}^s)_{j=1,\dots,d,\, t_l\in\tau_k}$ are either all zero or either all different from zero. However, the approach differs from group-Lasso because group-sparsity is not controlled by the L_2-norm of the group but by a single shrinkage coefficient associated to that group : the final optimization problem of Equation (3.10) is thus written as a standard Lasso problem (on $\boldsymbol{\alpha}$) with only D coefficients to estimate instead of p for a group-Lasso problem.

An iterative procedure to select the intervals

The previous subsection described our proposal to detect the subset of relevant intervals among a *fixed*, predefined set of intervals of the definition domain of the predictor, $(\tau_k)_{k=1,\dots,D}$. However, choosing *a priori* a proper set of intervals is a challenging task without expert knowledge, and a poor choice (too small, too large, or shifted intervals) may largely hinder interpretability. In the present section, we propose an iterative method to automatically design the intervals, without making any *a priori* choice.

In a closely related framework, Fruth et al. [2015] tackle the problem of designing intervals by combining sensitivity indices, linear regression models and a method called *sequential bifurcation* [Bettonvil, 1995] which allows them to sequentially split in two the most promising intervals (starting from a unique interval covering the entire domain of X). Here, we propose the inverse approach : we start with small intervals and merge them sequentially. Our approach is based on the previosu standard sparse SIR and iteratively performs the most relevant merges in a flexible way (contrary to a splitting approach, we do not need to arbitrary set the splitting positions).

The intervals $(\tau_k)_{k=1,\dots,D}$ are first initialized to a very fine grid, taking for instance $\tau_k = \{t_k\}$ for all $k = 1, \dots, p$ (hence, at the beginning of the procedure, $D = p$). The sparse step defined previously is then performed with the *a priori* intervals $(\tau_k)_{k=1,\dots,D}$: the set of solutions of Equation (3.10), for varying values of the regularization parameter μ_1, is obtained using a regularization path approach[Friedman et al., 2010]. Three elements are derived from the path results :

- $(\hat{\boldsymbol{\alpha}}_k^*)_{k=1,\dots,D}$ are the solutions of the sparse problem for the value μ_1^* of μ_1 that minimizes the GCV error ;
- $(\hat{\boldsymbol{\alpha}}_k^+)_{k=1,\dots,D}$ and $(\hat{\boldsymbol{\alpha}}_k^-)_{k=1,\dots,D}$ are the first solutions, among the path of solutions, such that at most (resp. at least) a proportion P of the coefficients are non zero coefficients (resp. are zero coefficients), for a given small chosen P (0.05 for instance).

Then, the following sets are defined : $\mathcal{D}_1 = \{k : \hat{\boldsymbol{\alpha}}_k^- \neq 0\}$ (called "strong non zeros") and $\mathcal{D}_2 = \{k : \hat{\boldsymbol{\alpha}}_k^+ = 0\}$ (called "strong zeros"). This step is illustrated in Figure 3.2. Intervals are

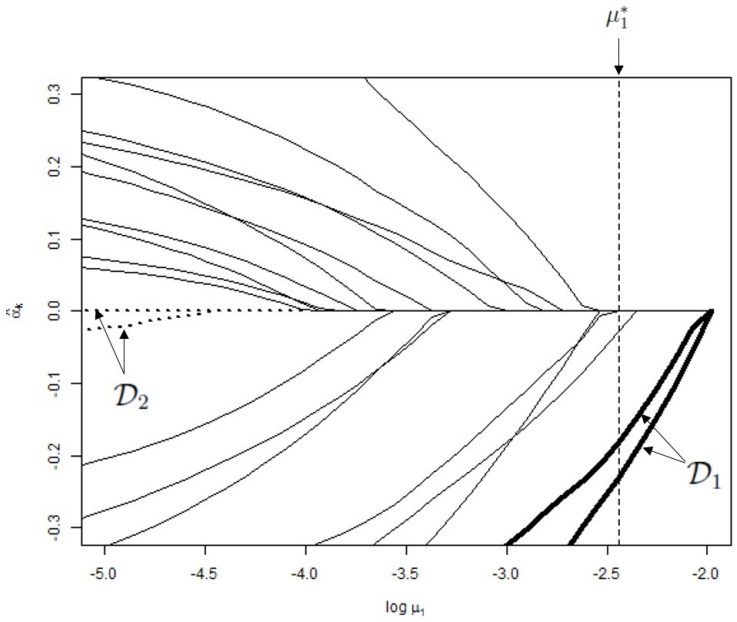

FIGURE 3.2 – Example of regularization path with $D = 20$: $(\hat{\boldsymbol{\alpha}}_k)_{k=1,\dots,D}$ are plotted according to different values of the tuning parameter μ_1. The vertical dotted line represents the optimal value μ_1^* that provides the solutions $(\hat{\boldsymbol{\alpha}}_k^*)_{k=1,\dots,D}$ of the sparse problem. $(\hat{\boldsymbol{\alpha}}_k)_{k\in\mathcal{D}_1}$ and $(\hat{\boldsymbol{\alpha}}_k)_{k\in\mathcal{D}_2}$ are respectively represented in bold and in pointed lines for $P = 0.1$.

merged using the following rules :

- "neighbor rule" : consecutive intervals of the same set are merged (τ_k and τ_{k+1} are merged if both k and $k+1$ belong to $\mathcal{D}_1$ or if they both belong to $\mathcal{D}_2$) (see a) and b) in Figure 3.3) ;
- "squeeze rule" : τ_k, τ_{k+1} and τ_{k+2} are merged if both k and $k+2$ belong to $\mathcal{D}_1$ while $k+1 \notin \mathcal{D}_2$ (or if both k and $k+2$ belong to $\mathcal{D}_2$ while $k+1 \notin \mathcal{D}_1$) and $l_k + l_{k+2} > l_{k+1}$ with $l_k = \max \tau_k - \min \tau_k$ (see c) and d) in Figure 3.3).

If the current value of P does not yield any fusion between intervals, P is updated by $P \leftarrow 2P$. The procedure is iterated until all the original intervals have been merged.

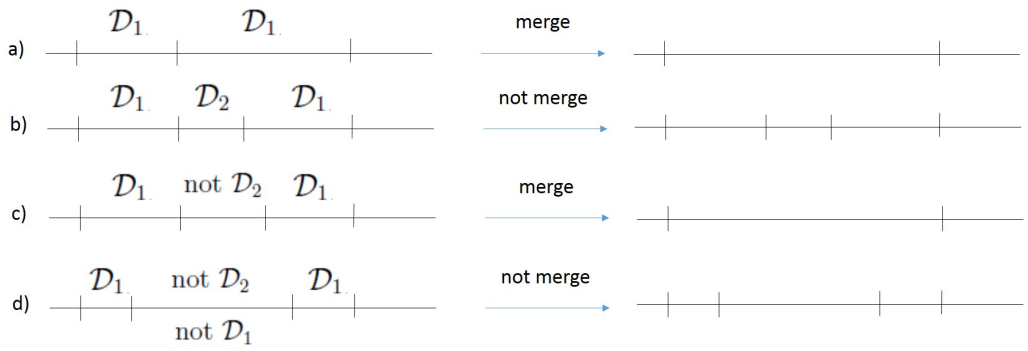

FIGURE 3.3 – Illustration of the merge procedure for the intervals.

The result of the method is a collection of models $(\hat{\boldsymbol{\alpha}}_k^*)_{k=1,\dots,D}$, starting with p intervals and finishing with one. The final selected model is the one that minimizes the CV error. In practice, this often results in a very small number of contiguous intervals which are of the same type (zero or non zero) and are easily interpretable.

Let us remark that the intervals $(\tau_k)_{k=1,\dots,D}$ are not used in the ridge step, which can thus be performed once, independently of the interval search. The whole procedure is summarized in Algorithm 1.

Algorithm 1 Overview of the complete procedure

Ridge estimation
Obtain $\hat{A}$ and $\hat{C}$, ridge estimates of the SIR.
Sparse estimation
Initialize the intervals $(\tau_k)_{k=1,\dots,D}$ to $\tau_k = \{t_k\}$
repeat
    Estimate and store $(\hat{\boldsymbol{\alpha}}_k^*)_{k=1,\dots,D}$ the solutions of the sparse problem that minimizes the GCV error
    Estimate $(\hat{\boldsymbol{\alpha}}_k^+)_{k=1,\dots,D}$ and $(\hat{\boldsymbol{\alpha}}_k^-)_{k=1,\dots,D}$ such that at most (resp. at least) a proportion P of the coefficients are non zero coefficients (resp. are zero coefficients), for a given chosen P
    Update the intervals $(\tau_k)_{k=1,\dots,D}$ according to the "neighbor" and the "squeeze" rules
until $\tau_1 \neq [t_1, t_p]$
Output : A collection of models $(\hat{\boldsymbol{\alpha}}_k^*)_{k=1,\dots,D}$
Select the model $(\hat{\boldsymbol{\alpha}}_k^*)^*_{k=1,\dots,D^*}$ that minimizes the CV error
Active intervals (for interpretation) are consecutive τ_k with non zero coefficients $\hat{\boldsymbol{\alpha}}_k^*$

The method requires to tune four parameters : the number of slices H, the dimension of the EDR space p, the penalization parameter of the ridge regression μ_2 and of the one of the sparse procedure μ_1. Two of these parameters, H and μ_1, are chosen in a standard way [Li, 1991] for further details). This section presents a method to jointly choose μ_2 and d, for which no solution has been proposed that is suited to our high-dimensional framework. Two issues are raised to tune these two parameters : i) they depend from each other and ii) the existing methods to tune them are only valid in a low-dimensional setting ($p < n$). We propose an iterative method inspired from existing approaches [Ferré, 1998, Bernard-Michel et al., 2008, Liquet and Saracco, 2012] only valid for the low dimension framework and combine them to find an optimal joint choice for μ_2 and d.

3.3.4 Experiments and discussion

We evaluate different aspects of the methods on simulated and real datasets. Our procedure shows good performances on simulated datasets and was then tested on the complex crop model. Note that all experiments have been performed using the R package **SISIR**. Datasets and R scripts are provided at `https://github.com/tuxette/appliSISIR`. So, finally, we applied our strategy to the challenging agronomic problem, the inference of interpretable climate-yield relationships on complex crop models.

We consider a process-based crop model called SUNFLO, which was developed to simulate the annual grain yield (in tons per hectare) of sunflower cultivars, as a function of time, environment (soil and climate), management practice and genetic diversity [Casadebaig et al., 2011]. SUNFLO requires functional inputs in the form of climatic series. These series consist of daily measures of five variables over a year : minimal temperature, maximal temperature, global incident radiation, precipitations and evapotranspiration. Globally, the SUNFLO crop model has about 50 equations and 64 parameters (43 plant-related traits and 21 environment-related). The dataset used in the experiment consisted of 111 yield values computed using SUNFLO for different climatic series (recorded between 1975 and 2012 at five French locations). We focused solely on evapotranspiration as a functional predictor because it is essentially a combination of the other four variables [Allen et al., 1998]. The cultural year (*i.e.*, the period on which the simulation is performed) is from weeks 16 to 41 (April to October). We voluntarily kept unnecessary data (11 weeks before simulation and 8 weeks after) for testing purpose (because these periods are known to be irrelevant for the prediction). The resulting curves contained 309 measurement points. Ten series of this dataset are shown in Figure 3.4, with colors corresponding to the yield that we intend to explain : no clear relationship can be identified between the the value of the curves at any measurement point and the yield value.

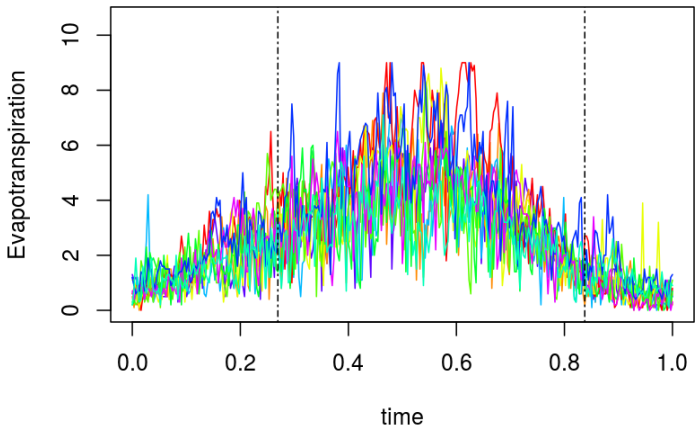

FIGURE 3.4 – Ten series of evaportranspiration daily recordings. The color level indicates the corresponding yield and the dashed lines bound the actual simulation definition domain.

We followed the approach described previously to design the relevant intervals and Figure 3.5 shows the selected intervals obtained after running our algorithm, as well as the points selected using a standard sparse approach. The standard sparse SIR (top of the figure) captures well the simulation interval (with only two points selected outside of it), but fails to identify the important periods within it. In contrast, SISIR (bottom) focuses on the second half of the simulation interval, and in particular its third quarter. This matches well expert knowledge,

that reports little influence of the climate conditions at early stage of the plant growth and almost none once the grains are ripe [Casadebaig et al., 2011].

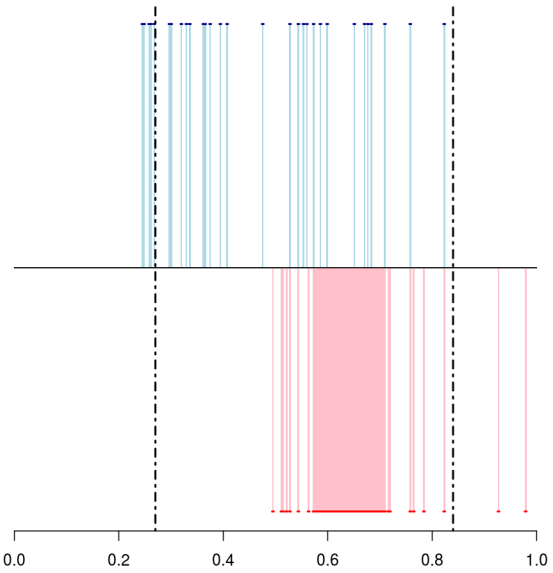

FIGURE 3.5 – *Sunflo*. Top : standard sparse SIR (blue). Bottom : SISIR (pink). The colored areas depict the active intervals. The dashed lines bound the actual simulation definition domain.

Discussion

Perspective of developments would extend the approach to multiple functional predictors, allowing to design common or separated interval selections for the different predictors. The final choice of the best model using a simple CV criterion could also be improved : we own a large collection of model and we only choose one without taking any information from the other ones. A model-averaging procedure could extract the informations included in all models and, maybe, produce a more appropriated selection of relevant intervals. Some technical modifications could also be tested such as the SIR-QZ [Coudret et al., 2014](instead of the ridge penalty) or the variable importance adapted to SIR [Jlassi and Saracco, 2017] (to perform variale selection instead of the LASSO step).

Other approaches could also be developed to achieve intervals selection. For example, we could adapt a clustering procedure on constrained variables [Wagstaff et al., 2001] : for each (t, t'), compute $c_{tt'} = \mathrm{Cor}(X_t, X_{t'})$ that leads to the matrix $\mathbf{C} = (c_{tt'})_{t,t'}$ that could be used as an input for a clustering procedure on constrained variables [Dehman et al., 2015]. This procedure provides a dendrogram of consecutive groupings of variables and each segmentation of this dendrogram leads to intervals of variables.

3.4 Multiple testing to perform variable selection

Using Lasso-type estimate is a first solution to perform variable selection : as it provides sparse estimate, the active set (i.e the set of variables with non-null coefficients) are the selected ones. But it is obviously not the only way to perform variable selection. Multiple testing is an

equivalent way to select some variables. Indeed, testing if each variable coefficients is equal to zero allows to perform a selection on all the variables. In this section we want to construct a multiple testing procedure (in order to control the FWER that is the probability to give at least one false positive) based on a Lasso type estimator. This problem is driven by an application in metabolomics that could be linked to previous section because it could also be seen as intervals (i.e. metabolites pure spectrum) selection in a "functional" framework (as there is a natural order on the predictor range). But, here, these pure spectra (that are stored in the design matrix X) are already known and can overlap.

3.4.1 Motivations

Metabolomics is the science concerned with the detection of metabolites (small molecules) in biological mixtures (e.g. blood and urine). The most common technique for performing such characterization is proton nuclear magnetic resonance (NMR). Each metabolite generates a characteristic resonance signature in the NMR spectra with an intensity proportional to its concentration in the mixture. The number of peaks generated by a metabolite and their locations and ratio of heights are reproducible and uniquely determined : each metabolite has its own signature in the spectra. Each signature spectrum of each metabolite can be stored in a library that could contain hundreds of spectra. One of the major challenges in NMR analysis of metabolic profiles remains to be automatic metabolite assignment from spectra. To identify metabolites, experts use spectra of pure metabolites and manually compare these spectra to the spectrum of the biological mixture under analysis. Such a method is time-consuming and requires domain-specific knowledge. Furthermore, complex biological mixtures can contain hundreds or thousands of metabolites, which can result in highly overlapping peaks.

Recently, automatic methods have been proposed (see Subsection 3.4.5 for details). Nevertheless, most are time-consuming and thus cannot be applied to a large library of metabolites, and/or their statistical properties are not proven. Thus, establishment of a gold-standard methodology with proven statistical properties for identification of metabolites would be very helpful for the metabolomic community as highlighted by Considine et al. [2018].

Because the number of tests is not too much large (one can expect to analysed a mixture with about 200 metabolites), because NMR experts want to recover all metabolites present in the mixture but, did not want to observe a false discovery, we have developed an *ad hoc* multiple testing procedure to identify and quantify metabolites in 1D ^{1}H NMR spectrum.

3.4.2 Statistical background

Let us consider the linear Gaussian model

$$Y = X\beta^* + \varepsilon, \tag{3.11}$$

where $X = (X_1|\dots|X_p)$ is a $n \times p$ design matrix of rank p, ε is a centered Gaussian vector with an invertible variance matrix Γ, and β^* is an unknown parameter. We want to estimate the so-called active set $\mathcal{A} = \{i \in [\![1, p]\!] \mid \beta_i^* \neq 0\}$ of relevant variables. A natural way to recover $\mathcal{A}$ is to test the hypotheses $\mathcal{H}_i : \beta_i^* = 0$, with $1 \leq i \leq p$. Several type I errors can be controlled in such multiple hypotheses tests. As the metabolomic experts did not want to observe a false discovery, we focus on the Familywise Error Rate (FWER) defined as the probability to reject wrongly at least one hypothesis $\mathcal{H}_i$.

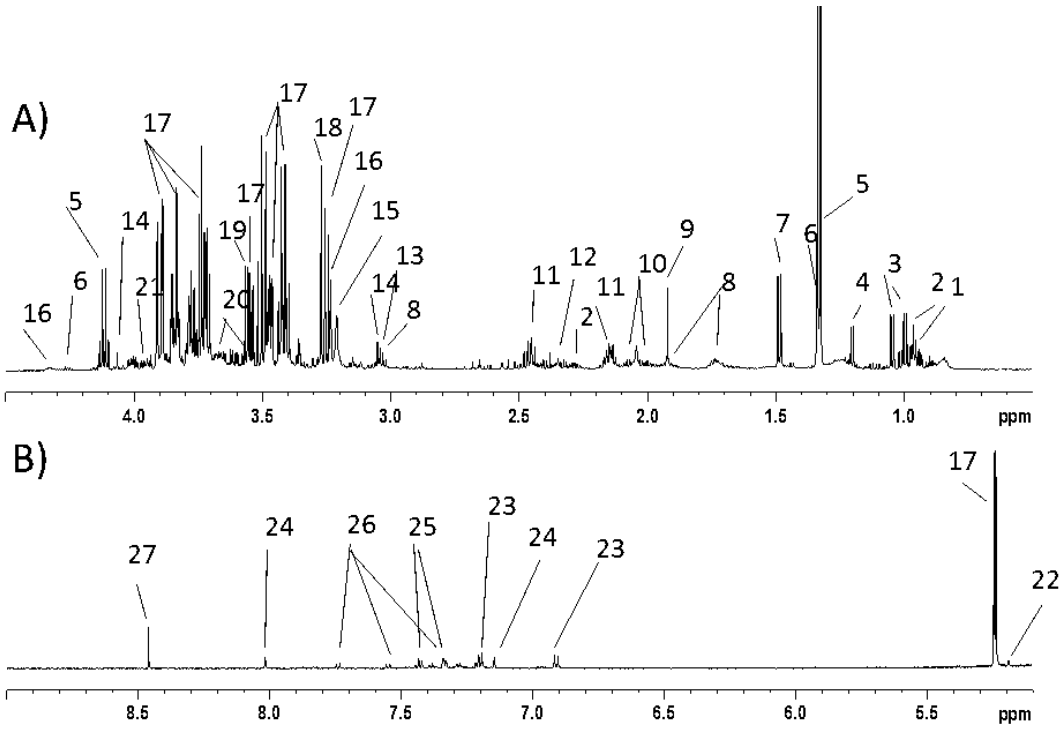

FIGURE 3.6 – Example of mixture spectra. For example, there are overlaps between the peaks of metabolites 5. and 6. and between the peaks of metabolites 25. and 26.

The lasso estimator [Tibshirani, 1996], defined by

$$\hat{\beta}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2} \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \right\} \tag{3.12}$$

has been designed for the high-dimensional setting (*i.e.* $n < p$ that is not our framework). In this case, the lasso is an alternative to the ordinary least squares estimator which is not defined. Some components of $\hat{\beta}(\lambda)$ are exactly null, thus a very simple way to test the hypothesis $\mathcal{H}_i$ is to reject it when $\hat{\beta}_i \neq 0$. This is probably the reason why the lasso has been widely studied both in the high-dimensional and in the small-dimensional setting (*i.e.* $n \geq p$ and $\operatorname{rank}(X) = p$).

Meinshausen and Bühlmann [2006], Zhao and Yu [2006], Zou [2006] showed that the irrepresentable condition is an almost necessary and sufficient condition for $\mathcal{A}(\hat{\beta}(\lambda)) := \{i \in [\![1, p]\!] \mid \hat{\beta}_i(\lambda) \neq 0\}$ to be a consistent estimator of $\mathcal{A}$ when n tends to $+\infty$ and p is fixed (up to a λ correctly chosen). This result could be used when n is very large, thus consistency is not an high-dimensional property. Geometrically, the irrepresentable condition means that each variable X_i with $i \notin \mathcal{A}$ is almost orthogonal to the subspace $\operatorname{Vect}\{X_i,\ i \in \mathcal{A}\}$.

Recent multiple testing procedures such as the SLOPE [Bogdan et al., 2015, Su and Candes, 2016], the knockoffs [Barber and Candes, 2015, Janson and Su, 2016] or the procedure derived from the covariance test [Lockhart et al., 2014, G'Sell et al., 2015] use a lasso-type estimator. These procedures are not restricted to the high-dimensional setting when $p > n$, they are also used when the design matrix X has a rank p. In particular, G'Sell et al. [2015] and Bogdan et al. [2015] studied the case in which X is orthogonal and the knockoffs procedure is only devoted to the case in which $\operatorname{rank}(X)$ is p. In this setting, lasso-type multiple testing procedures are alternative procedures to classical multiple testing procedures based on the maximum likelihood

estimator [Dunn, 1961, Holm, 1979, Romano and Wolf, 2005].

Because lasso-type procedures have been developed recently, one could expect them to be more powerful than classical and older ones. Since our aim is to provide a powerful multiple testing procedure that controls the FWER, we first naively developed a lasso-type procedure. Because the irrepresentable condition means that the design is almost orthogonal and because the lasso has an explicit expression in the orthogonal case, we orthogonalize the design X before using the lasso. So, we prove that, up to a transformation U^* which orthogonalizes the design matrix X and that minimizes the volume of the multidimensional acceptance region, the lasso-type estimator $\hat{\beta}^{U^*}$ has the following expression

$$\forall i \in [\![1,p]\!], \hat{\beta}_i^{U^*}(\lambda) = \operatorname{sign}(\hat{\beta}_i^{\text{mle}}) \left(|\hat{\beta}_i^{\text{mle}}| - \lambda/\delta_i^*\right)_+, \text{where } \hat{\beta}^{\text{mle}} := (X^T\Gamma^{-1}X)^{-1}X^T\Gamma^{-1}Y. \tag{3.13}$$

This expression delivers a simple message, when X is of rank p and when one wants to maximise the "power", the obtained lasso estimator is just the soft thresholded maximum likelihood estimator. This is not so surprising because the maximum likelihood estimator is efficient but it shows that choosing the lasso to optimise the power was definitely a naive idea. Because rejecting $\mathcal{H}_i : \beta_i = 0$ when $\hat{\beta}_i^{U^*}(\lambda) \neq 0$ is equivalent to reject $\mathcal{H}_i$ when $|\hat{\beta}_i^{\text{mle}}| > \lambda/\delta_i^*$, a lasso-type estimator is useless. The construction of this "lasso-type" procedure allowed us to discover a new multiple testing procedure procedure which is only based on the maximum likelihood estimator. General testing procedures (see the book of Lehmann and Romano [2005]) reject $\mathcal{H}_i$ as soon as $|\hat{\beta}_i^{\text{mle}}|/\text{se}(\hat{\beta}_{\text{i}}^{\text{mle}}) > \mu$, where $\text{se}(\hat{\beta}_{\text{i}}^{\text{mle}})$ is the standard error of $\hat{\beta}_i^{\text{mle}}$. One should notice that in these decisions rules, the critical value μ is the same for all i.

In contrast, the value δ^* in (3.13) giving a multidimensional acceptance region with a minimal volume leads to decision rules where μ varies with the tested hypothesis $\mathcal{H}_i$.

3.4.3 Theoretical results

Orthogonal-columns case

By convenience, we write that the X matrix has orthogonal columns when X^TX is diagonal. An orthogonal matrix is thus an orthogonal columns matrix but with $X^TX = Id_p$. When the design matrix X of the Gaussian linear model (3.11) has orthogonal columns, the lasso estimator has a closed form. This closed form allows to choose the tuning parameter in order to control the FWER at a given level. As an example, when X is orthogonal, the lasso estimator has the following expression [Tibshirani, 1996, Hastie et al., 2009, Bühlmann and van de Geer, 2011]

$$\hat{\beta}_i(\lambda) = \operatorname{sign}(\hat{\beta}_i^{\text{ols}}) \left(|\hat{\beta}_i^{\text{ols}}| - \lambda\right)_+$$

where $\hat{\beta}^{\text{ols}}$ is the ordinary least squares estimator of β^*. Let Z^{ols} denotes a centered Gaussian vector with the same covariance matrix as $\hat{\beta}^{\text{ols}}$, the tuning parameter giving a FWER at level α is the $1-\alpha$ quantile of $\max\{|Z_1^{\text{ols}}|, \ldots, |Z_p^{\text{ols}}|\}$. When X has orthogonal columns, the Proposition 3.2 provides a closed form for the lasso estimator and an explicit tuning parameter λ_0 to control the FWER.

Proposition 3.2 *Let X be a $n \times p$ matrix such that $X^TX = \operatorname{diag}(d_1, \ldots, d_p)$ then*

$$\forall i \in [\![1,p]\!], \hat{\beta}_i(\lambda) = \operatorname{sign}(\hat{\beta}_i^{\text{ols}}) \left(|\hat{\beta}_i^{\text{ols}}| - \lambda/d_i\right)_+.$$

Let $Z^{\text{ols}} := (Z_1^{\text{ols}}, \dots, Z_p^{\text{ols}})$ *be a random variable distributed according to a* $\mathcal{N}\left(0, (X^TX)^{-1}X^T\Gamma X(X^TX)^{-1}\right)$ *distribution. Let* $\alpha \in (0,1)$*, if* λ_0 *is the* $1-\alpha$ *quantile of* $\max_{i\in[\![1,p]\!]}\{d_i \times |Z_i^{\text{ols}}|\}$ *then,*

$$\mathbb{P}(\forall i \notin \mathcal{A}, \hat{\beta}_i(\lambda_0) = 0) \geq 1-\alpha. \tag{3.14}$$

When the covariance matrix Γ is given *a priori*, the distribution of Z^{ols} is known and λ_0 can be obtained by simple numerical simulations. In the next section we study the more general case where X has no longer orthogonal columns.

General case : when the lasso vanishes

Now we assume that the design matrix X is a matrix of rank p. Let us consider the set G of applications that orthogonalise X. In other terms, if $U \in G$, the matrix $(UX)^TUX$ is diagonal. For example the matrix $U := (X^TX)^{-1}X^T$ is a transformation of G. Without any other assumption on X, the lasso estimator has no closed form. Consequently, it becomes challenging to choose a tuning parameter λ_0 to control the FWER. To overcome this problem, we propose to apply a linear transformation $U \in G$ to each member of the model (3.11). This leads to the new linear Gaussian model

$$\tilde{Y} = \tilde{X}\beta^* + \tilde{\varepsilon} \text{ with } \tilde{Y} = UY, \tilde{X} = UX \text{ and } \tilde{\varepsilon} = U\varepsilon. \tag{3.15}$$

Because $\tilde{X}$ has orthogonal columns, it is possible to use the previous Proposition 3.2. For all $\lambda \geq 0$, the lasso estimator of β^* is

$$\hat{\beta}^U(\lambda) = \left(\text{sign}(\hat{\beta}_i^{\text{ols}}(U)) \left(|\hat{\beta}_i^{\text{ols}}(U)| - \lambda/d_i(U)\right)_+ \right)_{1\leq i\leq p}.$$

The tuning parameter λ_0^U giving a FWER α is the $1-\alpha$ quantile of $\max_{i\in[\![1,p]\!]}\{d_i(U)\times|Z_i^{\text{ols}}(U)|\}$. In the previous expression, $\hat{\beta}^{\text{ols}}(U)$, $Z^{\text{ols}}(U)$ and $(d_i(U))_{1\leq i\leq p}$ are respectively the ordinary least squares estimator of (3.15), a centered Gaussian vector with the same covariance matrix as $\hat{\beta}^{\text{ols}}(U)$ and the diagonal coefficients of $\tilde{X}^T\tilde{X}$.

Since the hypothesis $\beta_i^* = 0$ is rejected as soon as $\hat{\beta}_i^U(\lambda_0^U) \neq 0$ in other terms when $|\hat{\beta}_i^{\text{ols}}(U)| \geq \lambda_0^U/d_i(U)$, one proposes to look for a linear transformation U such that the thresholds $\lambda_0^U/d_1(U), \dots, \lambda_0^U/d_p(U)$ are as small as possible. Such a choice should increase the "power" of our test procedure : the smaller are the thresholds, the higher is the number of non-null detected components. As a p-uplet can be minimized in several ways, we propose to choose $U \in G$ so that the function $\phi(U) = \prod_{i=1}^p \frac{\lambda_0^U}{d_i(U)}$ is minimal. Intuitively, this choice can be understood by noticing that under the assumption that when $\beta^* = 0$,

$$1-\alpha = \mathbb{P}\left(\hat{\beta}^{\text{ols}}(U) \in \left[-\frac{\lambda_0^U}{d_1(U)}, \frac{\lambda_0^U}{d_1(U)}\right] \times \cdots \times \left[-\frac{\lambda_0^U}{d_p(U)}, \frac{\lambda_0^U}{d_p(U)}\right]\right).$$

The minimization of ϕ thus leads to minimize the volume of the multidimensional acceptance region $\left[-\frac{\lambda_0^U}{d_1(U)}, \frac{\lambda_0^U}{d_1(U)}\right] \times \cdots \times \left[-\frac{\lambda_0^U}{d_p(U)}, \frac{\lambda_0^U}{d_p(U)}\right]$ among those that have a level $1-\alpha$. The following theorem shows that it is possible to pick a transformation U^* for which simultaneously ϕ is minimal and the lasso is a soft thresholded maximum likelihood estimator.

Theorem 3.1 *There exists a linear transformation $U^* \in G$, such that*

$$\forall U \in G, \phi(U^*) \leq \phi(U).$$

Furthermore, for the optimal transformation U^* the lasso estimator has the following expression

$$\exists \delta^* \in (0,+\infty)^p \text{ such that } \forall i \in [\![1,p]\!], \hat{\beta}_i^{U^*}(\lambda) = \text{ sign}(\hat{\beta}_i^{\text{mle}}) \left(|\hat{\beta}_i^{\text{mle}}| - \lambda/\delta_i^*\right)_+,$$

where $\hat{\beta}^{\text{mle}}$ is the maximum likelihood estimator of the model (3.11).

Recovering the maximum likelihood estimator *via* the orthogonalisation U^* is satisfying because the maximum likelihood estimator is efficient. That is why this estimator is usually used for classical multiple testing procedures such as Bonferroni, Holm,.... Rejecting the null hypothesis $\mathcal{H}_i : \beta_i^* = 0$ as soon as $\hat{\beta}_i^{U^*}(\lambda) \neq 0$ is equivalent to reject $\mathcal{H}_i$ when $|\hat{\beta}_i^{\text{mle}}| \geq \lambda/\delta_i^*$ thus lasso-type estimator is useless here. Consequently, to manage this new procedure, it is finally not useful to construct the transformation U^* !

In general, the optimal parameter δ^* of the theorem 3.1 is not collinear to $1/\text{se}(\hat{\beta}_1^{\text{mle}}), \ldots, 1/\text{se}(\hat{\beta}_p^{\text{mle}})$. Consequently the random variables $\delta_1^*\hat{\beta}_1^{\text{mle}}, \ldots, \delta_p^*\hat{\beta}_p^{\text{mle}}$ have different variances. This remark is the main difference with the classical procedures for which statistical tests $\hat{\beta}_1^{\text{mle}}/\text{se}(\hat{\beta}_1^{\text{mle}}), \ldots, \hat{\beta}_p^{\text{mle}}/\text{se}(\hat{\beta}_p^{\text{mle}})$ are re-scaled to have unit variance. To provide a mutiple testing procedure which reject $\mathcal{H}_i : \beta_i^* = 0$ as soon as $|\hat{\beta}_i^{\text{mle}}| \geq \lambda/\delta_i^*$ the parameter λ have to be chosen as the $1-\alpha$ quantile of $\max\{\delta_1^*|Z_1^{\text{mle}}|, \ldots, \delta_p^*|Z_p^{\text{mle}}|\}$. From now on, we denote $\lambda_0(\delta)$ the $1-\alpha$ quantile of $\max\{\delta_1|Z_1^{\text{mle}}|, \ldots, \delta_p|Z_p^{\text{mle}}|\}$ where $\delta = (\delta_1, \ldots, \delta_p) \in (0,+\infty)^p$. To manage the previous multiple testing procedure based on the maximum likelihood estimator, the keystone is now to compute the optimal parameter δ^*.

A new procedure based on the old maximum likelihood estimator

Theorem 3.1 does not explain how to get such an optimal parameter δ^*. We did not manage to obtain a closed form of it. However some simple remarks could help its numerical computation.

First, because whatever $t > 0$ the thresholds $\lambda_0(t\delta^*)/t\delta_1^*, \ldots, \lambda_0(t\delta^*)/t\delta_p^*$ are equal to $\lambda_0(\delta^*)/\delta_1^*, \ldots, \lambda_0(\delta^*)/\delta_p^*$, one only needs to determine an optimal value δ^* for which $\|\delta^*\|_\infty = 1$. Second, this problem can be translated more simply as follows. Let us set $b_1 = \lambda_0(\delta)/\delta_1, \ldots, b_p = \lambda_0(\delta)/\delta_p$ (resp. $b_1^* = \lambda_0(\delta)/\delta_1^*, \ldots, b_p = \lambda_0(\delta)/\delta_p^*$) and consider the acceptance region $B = [-b_1, b_1] \times \ldots \times [-b_p, b_p]$ (resp. $B^* = [-b_1^*, b_1^*] \times \cdots \times [-b_p^*, b_p^*]$). Let Σ be the covariance matrix of the maximum likelihood estimator and let Z^{mle} be distributed according to $\mathcal{N}(0_{\mathbb{R}^p}, \Sigma)$. The rectangular parallelepiped B^* has the smallest volume among rectangular parallelepiped B such that $P\left(Z^{\text{mle}} \in B\right) = 1-\alpha$. This is a constraint optimization problem whose solutions are stationary points of the Lagrangian. The condition given in the following proposition should hold for B^*.

Proposition 3.3 *Let $b^* = \left(b_1^*, \ldots, b_p^*\right)$ be a solution of the following optimisation problem*

$$\min \prod_{i=1}^p b_i \text{ subject to } \mathbb{P}\left(|Z_1^{\text{mle}}| \leq b_1, \ldots, |Z_p^{\text{mle}}| \leq b_p\right) = 1-\alpha. \tag{3.16}$$

Let T^{b^*} denotes the truncated Gaussian vector on B^* having the following density

$$f_{T^{b^*}}(u) = \frac{1}{(1-\alpha)\sqrt{(2\pi)^p \det(\Sigma)}} \exp(-u\Sigma^{-1}u)\mathbb{1}_{u\in B^*}du$$

then all the diagonal coefficients of $\Sigma^{-1}\mathrm{var}(T^{b^*})$ should be equal.

Notice that if the variance matrix of T^{b^*} (here denoted by $\mathrm{var}(T^{b^*})$) was equal to Σ, all the diagonal coefficients of $\Sigma^{-1}\mathrm{Var}(T^{b^*})$ would be equal, indicating that b^* is a solution of (3.16). Because the diagonal terms of $\mathrm{var}(T^{b^*})$ are always smaller than the diagonal terms of Σ, $\mathrm{var}(T^{b^*})$ cannot be equal to Σ. However, the condition given by Proposition 3.3 can be intuitively interpreted. The optimal (with respect to the volume) rectangular parallelepiped should be such that the covariance of the truncated Gaussian variable Z^{mle} restrained to $[-b_1^*, b_1^*] \times \cdots \times [-b_p^*, b_p^*]$ is as close as possible to the non constraint covariance of the random variable Z^{mle}. If we exclude some simple case (independent, equicorrelated and block diagonal equicorrelated), the optimal B^* cannot be explicitly calculated but one can assume that, up to a dilatation of the obtained b^* by the diagonal coefficients of Σ, the diagonal coefficients of Σ are equal to 1. Indeed, one can check that $(b_1^*/\sqrt{\Sigma_{1,1}}, \ldots, b_p^*/\sqrt{\Sigma_{p,p}})$ is the solution of the following problem

$$\min \prod_{i=1}^{p} b_i \text{ subject to } \mathbb{P}\left(\frac{|Z_1^{\mathrm{mle}}|}{\sqrt{\Sigma_{1,1}}} \leq b_1, \ldots, \frac{|Z_p^{\mathrm{mle}}|}{\sqrt{\Sigma_{p,p}}} \leq b_p\right) = 1-\alpha.$$

To summarize, the setting up of our multiple testing procedure is detailed hereafter :

1. One computes the covariance matrix of the maximum likelihood estimator of the model (3.11), namely $\Sigma := (X^T\Gamma X)^{-1}$;
2. The parameter $\delta^* \in (0, +\infty)^p$ is obtain by solving the problem (3.16). This optimal parameter must satisfies the relation $\Sigma^{-1}\mathrm{var}(T^{b^*})$ given in the proposition 3.3 ;
3. One compute $\lambda_0(\delta^*)$ which is the $1-\alpha$ quantile of the random variable $\{\delta_1^*|Z_1^{\mathrm{mle}}|, \ldots, \delta_p^*|Z_p^{\mathrm{mle}}|\}$. The quantile $\lambda_0(\delta^*)$ is computed numerically using a large number of realizations of Z^{mle} distributed according to $\mathcal{N}(0, \Sigma)$;
4. The multiple testing procedure rejects the null hypothesis $\mathcal{H}_i : \beta_i^* = 0$ when $|\hat{\beta}_i^{\mathrm{mle}}| > \lambda_0(\delta^*)/\delta_i^*$. This procedure controls the FWER at a level $1-\alpha$.

As expected, numerical experiments show that the gain of volume for the acceptance region provides a gain in power and that our approach shows better performances than the thresholded Lasso estimate of Lounici [2008] or the knockoff procedures [Janson and Su, 2016].

3.4.4 Discussion

As already mentioned, the keystone of this procedure is to compute the optimal parameter δ^*. However, this computation could be improved. In a future work, we aim to develop a fast and accurate numerical scheme for the computation of δ^*. It is also a challenging issue to provide a useful multiple testing when p is very large. Finally, a stepdown multiple testing procedure based on our procedure could increase the power.

3.4.5 Application in metabolomics : detection of metabolites

As already mentioned, this *ad hoc* procedure has been built to a practical purpose : the identification and quantification of metabolites in NMR spectrum.

Modelling

This method was called ASICS for Automatical Statistical Identification in Complex Spectra. A spectrum can be represented as a function over the range I of chemical shifts. All the spectra were normalized so that their area under the curve over I is 1. To model the spectrum of the complex mixture g, possible slight variations of chemical shifts with the experimental conditions have to be taken into account. The warping function $\phi : I \to I$ allows to model the variation of chemical shift, where ϕ is an increasing function and I is an interval of the chemical shifts associated to a spectrum. If f denotes the spectrum of a metabolite of the library, $fo\phi$ models the warped spectrum of the same metabolite observed in a different experimental condition. The spectrum of a complex mixture g can be written as a combination of the warped spectra of the metabolites belonging to the library where p is the number of metabolites of the library, α_i is a non-negative number depending on the proportion of the i^{th} metabolite in the complex mixture and on its number of hydrogen atoms, f_i is the spectrum of the i^{th} metabolite of the library and ϕ_i represents the corresponding warping function. Although the experimental conditions of the complex mixture spectrum g are controlled, they are slightly different from those used to generate the spectra of the library. Finally, the term ε is a random error term. The structure of the noise ε is very important in the identification and quantification of metabolites in the mixture. Several observations of a spectrum obtained from the same metabolite allowed modeling the noise as

$$\varepsilon = \sqrt{\sum_{i=1}^{p} \alpha_i f_i o \phi_i \varepsilon_1} + \varepsilon_2$$

where ε_1 and ε_2 are standard independent white noises with known standard deviations σ_1 and σ_2. This equation models the signal taking into account both an additive noise ε_2 and a multiplicative one ε_1. The multiplicative noise is proportional to the intensity of the signal. The additive noise is the same whatever the signal and is always present even when the signal is equal to zero. These two noise parameters influence differently the performances of our method. The additive noise has a strong impact on the identification of the metabolites whereas the multiplicative one has a major impact on their quantification. It is very difficult to be more quantitative on the standard deviation of the additive noise on the detection performances because it depends strongly on some experimental conditions (operator, pH, equipment, baseline quality correction ...). The multiplicative noise is commonly used in quantification methods. Usually values between 0.1 and 0.2 (which is quite common in metrology) are considered as acceptable to quantify. An estimation was carried out from our duplicated experiments and led to a value of 0.17.

The first step of the method is to identify the metabolites of the library that cannot belong to the complex spectra. The chemical shift between two spectra of the same metabolites obviously depends on the experimental conditions (pH ...). For a given metabolite, we assume that the maximum variation of the chemical shift is smaller than an upper bound M, which was fixed at 0.02 ppm. It is assumed that a metabolite belonging to a complex mixture must display its related signals in the complex spectra. Thus, a metabolite cannot belong to the complex mixture if at least one peak of its spectrum does not appear in the complex spectra. Consequently, a metabolite displaying a peak at a chemical shift d cannot belong to a complex spectrum which does not present any peak in the interval $[d-M, d+M]$. ASICS quickly detects these metabolites and reduces the number of metabolites of the library that need to be taken into account in the identification and quantification steps.

The i^{th} metabolite is considered as identified in the complex mixture when its coefficient α_i is greater than zero. Using our estimation method defined in the previous subsections, we own a sparse estimate whose some components are exactly zero, leading to simple identification in our complex mixture. However, the warping functions $\phi_1, ...\phi_p$ need to be known to obtain a sparse estimator of $\alpha_1, ...\alpha_p$. To solve this problem, ASICS proceeds in two stages. During the first stage, the warping functions are successively estimated using non sparse estimates of $(\alpha_1, ...\alpha_p)$. At the beginning of the k^{th} step of this first stage, the estimates of the first $k-1$ warping functions $\phi_1^{(1)}, \ldots, \phi_{k-1}^{(k-1)}$ and nonsparse estimates $\alpha_1^{(k-1)}, \ldots, \alpha_p^{(k-1)}$ of $\alpha_1, \ldots, \alpha_p$ are known. The superscript in $\phi_i^{(i)}$ and $\alpha_i^{(k-1)}$ i indicates the step at which the estimate was obtained. The k^{th} warping function is estimated by solving the following optimization problem

$$\arg\min_{\phi_k, \alpha_k} \left| g - \alpha_k f_k o \phi_k - \sum_{i=1}^{k-1} \alpha_i^{(k-1)} f_i o \phi_i^{(i)} - \sum_{i=k+1}^{p} \alpha_i^{(k-1)} f_i \right|^2 .$$

The warping function ϕ_k is estimated so that the maximum variation of the chemical shift is smaller than M. This estimate is then used to update the non-sparse estimates of $\alpha_1, \ldots, \alpha_p$ as shown hereafter

$$\left(\alpha_1^{(k)}, \ldots, \alpha_p^{(k)}\right) = \arg\min_{\alpha_1, \ldots, \alpha_p} \left| g - \sum_{i=1}^{k} \alpha_i f_i o \phi_i^{(i)} - \sum_{i=k+1}^{p} \alpha_i^{(k-1)} f_i \right|^2 .$$

Figure 3.7 provides an illustration of this warping strategy.

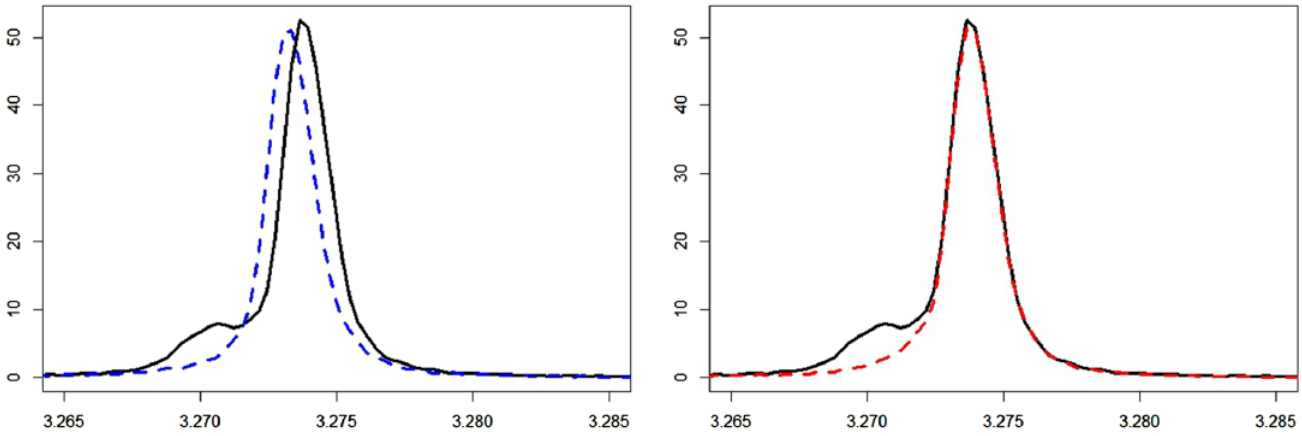

FIGURE 3.7 – On the left on solid line, the main peak of the creatinine in the spectrum of synthetic urine. In dotted line, the same peak observed on the spectrum of the creatinine before the warping stage. On the right on dotted line, the main peak of the creatine spectrum observed after the warping stage.

Note that, using this warping strategy, ASICS is able to take into account a chemical shift variation that is not only a unique translation on the whole spectrum. Local translations, dilations or tightenings would also been adjusted. However, this procedure is not able to create a new peak or to delete an existing one.

These estimations of the warping functions are then used at the second stage to derive sparse estimates of $(\alpha_1, \ldots, \alpha_p)$ using the methodology of previous subsections where $\sum_i \alpha_i f_i$ is replaced by its estimation g in the covariance matrix of the residuals (for more details see Tardivel [2017] page 57).

Thresholded estimators inherited from Lasso ones are known to be biased [Hastie et al., 2009]. For this reason the final quantification of metabolites is performed with a least squares method limited to the metabolites identified (i.e. with estimated proportions greater than zero) at the previous step.

Results

The performances of ASICS were first assessed on duck plasma, where a validated enzymatic method was also available to quantify some metabolites. It shows good correlation that validates the order of magnitude of the quantification carried out using ASICS.

Then, ASICS was compared to other current methods available for the analysis of complex mixtures NMR spectra. Metabohunter [Tulpan et al., 2011] computes a score for each metabolite individually. This score gives the probability of presence of each metabolite in the mixture and is related to the number of signals found in the mixture spectrum for a given metabolite. This simple method is very quick but does not provide quantification. BATMAN [Astle et al., 2012, Hao et al., 2012, 2014] is based on a Bayesian model selection and combines the representation of peaks by Lorentzian curves with a MCMC algorithm. The estimation of proportions of each metabolite using this method provides good results. However, it is time-consuming and requires a careful description of each peak of a metabolite. This step can be very tedious especially with metabolites displaying a large number of peaks. To date, BAYESIL features [Ravanbakhsh et al., 2015] seem to outperform BATMAN ones. BAYESIL handles spectral matching as an inference problem within a probabilistic graphical model that rapidly approximates the most likely metabolic profile. Actually, the most used tool appears to be the Chenomx software [Weljie et al., 2006]. Computations performed by this software are rather fast but it is known to yield many false positive metabolites. Finally, it is a commercial tool that could be quite expensive. The comparisons were carried out using two different biofluids : synthetic urine containing salts with a known concentration of metabolites and a biological human plasma sample (NIST SRM1950 plasma) that is a reference plasma sample already annotated by NMR experts [Simón-Manso et al., 2013]. The results of the different methods on the synthetic urine are gathered in Table 3.1.

TABLE 3.1 – Comparison of the five methods on the synthetic urine

	True positive	False positive	False negative	True negative	Accuracy (%)	Compounds in library	Computing time
ASICS	17	10	4	145	92	176	<3mns
Metabohunter	4	51	17	795	92	867	<1mn
Batman	21	125	0	1	18	147	74 hours
Bayesil	12	17	7	53	73	89	$\sim$ 10mns
Chenomx	15	48	6	269	54	338	<3mns

ASICS was able to identify 17 metabolites out of the 21 actually present, with only 10 false

detections, thus giving an accuracy of 92%. MetaboHunter analysis led to the same accuracy but with very different results : a very poor detection of true positive but a very high exclusion of true negative related to its very large library. BATMAN identified nearly all the metabolites in the mixture as already described in Ravanbakhsh et al. [2015] but yielded a very high number of false positives. Bayesil and Chenomx tools share a good accuracy but also a high number of false positives. In terms of computational time, ASICS lasts four times less than Bayesil for a twice as large library. Spectral processing with BATMAN was very long whereas Chenomx and MetaboHunter were the quickest. The same king of results were obtained for the quantification and ASICS showed the best order of magnitude.

As the composition of the NIST plasma is still an open question, it cannot be used to assess the superiority of any method. Nevertheless, all the main compounds identified by the experts were also identified by ASICS whereas it is not the case for the other methods. In addition to the 21 compounds already known, ASICS allowed identifying L-serine and GPC that were further confirmed by the NMR experts using other analyses.

Note that the ASICS procedure is now implemented in a Bioconductor R package that also provides different statistical tools for the analysis of NMR spectra (more details in the final section of this chapter) and is also avaible on Galaxy on the Workflow4Metabolomics infrastructure.

3.5 Sparse issues in high-dimension

The previous section brought a lot of question for us on the Lasso (i.e a L^1-penalty). As explained, we try to develop a Lasso-type estimate with special properties (powerful and with FWER control) but, when we optimize it, it leads us to a simple thresholded maximum likelihood estimate. In fact, Lasso is nowadays widely used to provide sparse estimates. But, when a sparse estimate is desirable that is the L^0-norm of the solutions that is the real objective. Obviously, minimizing this norm is still an open issue in high dimensions and some other tools (such as the Lasso) have to be used. But does it converge to the optimal L^0 norm solution ? Under which assumptions ? Is it possible to define a more general surrogate function to achieve this objective ? That was the starting point of the following section.

3.5.1 Background and motivation

We consider a vector $y \in \mathbb{R}^n$ and a family of vectors $\mathcal{D} = \{d_1, \dots, d_p\}$ spanning $\mathbb{R}^n$. An $\epsilon-$approximation of y in $\mathcal{D}$ is a vector $x = (x_1, \dots, x_p)$ such that $\|y - (x_1 d_1 + \dots + x_p d_p)\|^2 \leq \epsilon$. The aim of this article is to find at least one of the sparsest $\epsilon-$approximations of y when $p > n$. These sparsest $\epsilon-$approximations are defined as the solutions of

$$S_0^\epsilon := \text{argmin}\|x\|_0 \text{ subject to } \|y - Dx\|^2 \leq \epsilon \qquad (\mathcal{P}_0^\epsilon)$$

where $\|x\|_0 := \text{Card}\{i \in [\![1, p]\!] \mid x_i \neq 0\} = \sum_{i=1}^p \mathbf{1}_{x_i \neq 0}$ is the l^0 "norm" of x and $D := (d_1|\dots|d_p)$ is the $n \times p$ matrix whose columns are the vectors $(d_j)_{1 \leq j \leq p}$.

A first simplified problem is to look for the sparsest representations of y in $\mathcal{D}$ corresponding to the solutions of $\mathcal{P}_0^0$ namely

$$S_0 := \text{argmin}\|x\|_0 \text{ subject to } Dx = y. \qquad (\mathcal{P}_0)$$

Many applications concerning tomography [Burger et al., 2016, Liu and Gao, 2016, Prieto and Dorn, 2016] or radar [Baraniuk and Steeghs, 2007, Herman and Strohmer, 2009] are related to the resolution of the problems $\mathcal{P}_0$ and $\mathcal{P}_0^\epsilon$. Because $n < p$, recovering x from D and y is an ill posed problem. However, when x has a sparse representation in a known basis $\{b_1, \ldots, b_p\}$ of $\mathbb{R}^p$, it is possible to recover x by determining its components $\theta = (\theta_1, \ldots, \theta_p)$ in this basis. These components are obtained by looking for the sparsest representation of $y = DB\theta$, with B the matrix $(b_1|\ldots|b_p)$. When y is corrupted by a noise, a way to recover x is to compute the sparsest $\epsilon-$approximation of y in DB where the number ϵ is calibrated with respect to the noise magnitude [Ender, 2010].

A simple way to solve $\mathcal{P}_0$ is to compute $\tilde{x} = \tilde{D}^{-1}y$ for all $n \times n$ invertible submatrices $\tilde{D}$ of D and to select the $\tilde{x}$ with the lowest l^0 "norm". The number of such $n \times n$ submatrices of D is $\binom{p}{n}$. When $p \gg n$ this number is huge rending the previous approach intractable.

So, other approaches such as the basis pursuit problem, denoted $\mathcal{P}_1$, have been proposed [Gribonval and Nielsen, 2003, Donoho et al., 2006]. Under some conditions, given hereafter, the problem

$$\text{argmin}\|x\|_1 \text{ subject to } Dx = y \tag{$\mathcal{P}_1$}$$

has a unique solution that is also a solution of $\mathcal{P}_0$. The standard approach to know if a solution of $\mathcal{P}_1$ is also a solution of $\mathcal{P}_0$ is to compute s the l^0 "norm" of a solution of $\mathcal{P}_1$ and to check whether or not one of these conditions holds for s. When the solution of $\mathcal{P}_1$ does not meet any of these conditions, we do not know if it belongs to S_0.

The null space property [Donoho and Elad, 2003] is probably the most known condition. However, as pointed out by Tillmann and Pfetsch [2014], this condition is uncheckable. Another condition is the restricted isometry property detailed in Candes [2008], Cai and Zhang [2013]. However, this condition is not easy to use because the computation of the restricted isometry constant is intractable [Tillmann and Pfetsch, 2014]. On the contrary, the mutual coherence condition [Donoho and Elad, 2003, Gribonval and Nielsen, 2003] is easily checkable. Unfortunately, none of these three conditions (null space property, restricted isometry property and mutual coherence) hold for the basis pursuit solution as soon as its l^0 "norm" is greater or equal to $(n+1)/2$. In this case, the solutions of $\mathcal{P}_1$ does not give any information on those of $\mathcal{P}_0$. Moreover, even if the l^0 "norm" of the sparsest representation is strictly smaller than $(n+1)/2$, the numerical comparisons of Candes et al. [2008] illustrate that the solution of the basis pursuit may not be a solution of $\mathcal{P}_0$.

An intuitive alternative approach consists in the approximation of the l^0 "norm" in $\mathcal{P}_0$ by a surrogate function with nice properties. As an example, the function $\sum_{i=1}^p \ln(1+|x_i|/\delta)$ has been studied as an approximation of the l^0 "norm" [Candes et al., 2008, Lobo et al., 2007], leading to the following problem

$$\text{argmin} \sum_{1 \leqslant i \leqslant p} \ln(1+|x_i|/\delta) \text{ subject to } Dx = y. \tag{3.17}$$

With some well chosen δ, simulations show that this heuristic approach gives better results than the basis pursuit. However, nothing guarantees that the solutions of (3.17) are also solutions of $\mathcal{P}_0$ and the choice of δ plays a major role on the performances of the method. A similar surrogate approach is given in Foucart and Lai [2009], Lai [2010], Sun [2012] in which the l^0 "norm" is approximated by a l^α "norm". Numerical experiments show that these performances are very close to the ones of Candes et al. [2008].

When $\epsilon > 0$, the problem $\mathcal{P}_0^\epsilon$ is even more complicated and still intractable. Similarly to the basis pursuit problem $\mathcal{P}_1$, one can substitute in $\mathcal{P}_0^\epsilon$ the l^0 "norm" by a l^1 norm. This leads to

the following problem

$$\text{argmin}\|x\|_1 \text{ subject to } \|y - Dx\|_2^2 \leqslant \epsilon. \tag{$\mathcal{P}_1^\epsilon$}$$

This problem $\mathcal{P}_1^\epsilon$ can be rewritten as a lasso problem [Tibshirani, 1996] :

$$\text{argmin}\|y - Dx\|^2 + \lambda\|x\|_1. \tag{$\mathcal{P}(\lambda)$}$$

Actually, there exists a (not explicit) bijection between λ et ϵ guaranteeing that both problems have the same solution [Bertsekas, 1999].

To our knowledge, there is no theoretical result insuring that $x(\lambda)$, the unique solution of $\mathcal{P}(\lambda)$, is an element of S_0^ϵ. Instead, there exists a lot of conditions that state the convergence of $x(\lambda)$ to a solution $x^* \in S_0$ when λ converges to 0. Among these conditions (for an exhaustive list, see Bühlmann and van de Geer [2011]), the two most known are probably the irrepresentable condition [Meinshausen and Bühlmann, 2006, Zou, 2006] and the compatibility condition [Van de Geer, 2008]. In practice all these conditions are not easily checkable. Furthermore, when these conditions do not hold the solution obtained with the basis pursuit or with the lasso can be very far from the set S_0^ϵ we wish to recover.

The aim of this work is to propose a new tractable problem which allows to catch one of the sparsest representations (element of S_0) or one of the sparsest $\epsilon-$approximations (element of S_0^ϵ).

3.5.2 Theoretical results

Sparsest representations

The substitution in $\mathcal{P}_0$ of the l^0 "norm" by a l^α "norm" with $\alpha < 1$ gives the following problem $\mathcal{P}_\alpha$ which also has sparse solutions

$$S_\alpha := \text{argmin}\|x\|_\alpha \text{ subject to } Dx = y, \tag{$\mathcal{P}_\alpha$}$$

where $\|x\|_\alpha = (\sum_{i=1}^p |x_i|^\alpha)^{1/\alpha}$ is the l^α "norm" of the vector x. The problem $\mathcal{P}_\alpha$ is better than the basis pursuit to recover a solution of $\mathcal{P}_0$. Indeed, when the problem $\mathcal{P}_1$ provides a solution of $\mathcal{P}_0$, the problem $\mathcal{P}_\alpha$ still provides a solution of $\mathcal{P}_0$ [Gribonval and Nielsen, 2007]. The study of this problem has been the subject of an abundant literature, see for example Gribonval and Nielsen [2007], Lai [2010], Sun [2012], Zhang et al. [2015]. The problem $\mathcal{P}_\alpha$ provides a sparsest representation as soon as the null space property condition or the restricted isometry property hold. But, as for the basis pursuit, these conditions are uncheckable.

We can generalize the problem $\mathcal{P}_\alpha$ by substituting the function $|x_i|^\alpha$ by a function $f_\alpha(|x_i|)$. This modification leads to minimize an expression of the form $\sum_{i=1}^p f_\alpha(|x_i|)$. Intuitively, by comparing $\sum_{i=1}^p f_\alpha(|x_i|)$ with the l^α "norm", one sees that the function $\sum_{i=1}^p f_\alpha(|x_i|)$ should simply converge to $\|.\|_0$ and should have level sets that look like spheres for the l^α "norm". So, we focus on the following problem

$$S_{f_\alpha} := \text{argmin} \sum_{1 \leqslant i \leqslant p} f_\alpha(|x_i|) \text{ subject to } y = Dx. \tag{$\mathcal{P}_{f_\alpha}$}$$

Without any condition, we prove that the solutions of $\mathcal{P}_{f_\alpha}$ are also solutions of $\mathcal{P}_0$ as soon as α is small enough.

Theorem 3.2 *Let f_α be a function defined on $\mathbb{R}_+$ strictly increasing and strictly concave such that*

$$\forall x \in \mathbb{R}_+, \lim_{\alpha \to 0} f_\alpha(x) = \mathbf{1}_{x \neq 0}.$$

Then, there exists $\alpha_0 > 0$ such that for all $\alpha \in (0, \alpha_0)$, $S_{f_\alpha} \subset S_0$.

The α_0 threshold depends on D and y and its value is quite hard to infer except in few cases. For example, a lower bound of α_0 is given in Sun [2012]. This minoration requires assumptions on the restricted isometry constant and on the sparsity of S_0. Let us notice that Theorem 3.2 is obtained without assuming anything about the restricted isometry constant or about the sparsity of the sparsest representation. Nevertheless, since the $\mathcal{P}_{f_\alpha}$ allows to capture a part of S_0 for all $\alpha < \alpha_0$, one can choose *a priori* a very small α so that we can expect it is less than α_0. A study of the problem $\mathcal{P}_{f_\alpha}$ where the functions f_α have different properties that those given in the theorem 3.2 is given in Woodworth and Chartrand [2016]. The authors proved that the problem $\mathcal{P}_{f_\alpha}$ catches an element of S_0 under the conditions that the l_0 "norm" of the sparsest representation is smaller than $n/2$ and that the matrix D satisfies the unique representation property. Nevertheless, Theorem 3.2 does not hold once $\mathbb{R}^n$ is substituted by an infinite dimensional space.

Because the numerical resolution of the problem $\mathcal{P}_{f_\alpha}$ requires some regularity, we restrict ourselves to functions f_α which are differentiable on $(0, +\infty)$. Numerically, we solve the problem $\mathcal{P}_{f_\alpha}$ using a MM method [Hunter and Lange, 2004]. This method iteratively alternates two steps. First a function that majorizes the function $\sum_{1 \leqslant i \leqslant p} f_\alpha(|x_i|)$ is defined. Then this majorazing function is minimized.

So, we define a sequence $(x^{(k)})_{k \in \mathbb{N}}$ by "linearising" the function $\sum_{1 \leqslant i \leqslant p} f_\alpha(|x_i|)$ at the point $x^{(k)} \in \mathbb{R}^p$. This "linearisation" (we use quotation because this function is not affine) gives the function $x \in \mathbb{R}^p \mapsto \sum_{1 \leq i \leq p} f_\alpha(|x_i^{(k)}|) + f'_\alpha(|x_i^{(k)}|)(|x_i| - |x_i^{(k)}|)$. Because f is concave on $\mathbb{R}_+$, we have

$$\forall x \in \mathbb{R}^p, \sum_{1 \leqslant i \leqslant p} f_\alpha(|x_i|) \leq \sum_{1 \leq i \leq p} f_\alpha(|x_i^{(k)}|) + f'_\alpha(|x_i^{(k)}|)(|x_i| - |x_i^{(k)}|).$$

Then, this majorizing function is minimized with respect to x leading to $x^{(k+1)}$. More precisely, we choose $x^{(0)} \in \mathbb{R}^p$ and we set $x^{(k+1)}$ as the solution of the following weighted basis pursuit problem

$$\begin{aligned} x^{(k+1)} &:= \operatorname{argmin} \sum_{1 \leq i \leq p} f_\alpha(|x_i^{(k)}|) + f'_\alpha(|x_i^{(k)}|)(|x_i| - |x_i^{(k)}|) \text{ subject to } Dx = y, \\ &= \operatorname{argmin} \sum_{i=1}^{p} f'_\alpha(|x_i^{(k)}|)|x_i| \text{ subject to } Dx = y. \end{aligned}$$

If at iteration k, there are several minimizers, it suffices to choose among them, one minimizer for which the family $(d_i)_{i \in \operatorname{supp}(x^{(k)})}$ is linearly independent. We have shown that such a minimizer always exists. The first iteration of the previous MM method gives a vector $x^{(1)}$ solution of the weighted basis pursuit problem. This vector has a large number of null components. When f is right differentiable at 0, as for small α the quantity $f'_\alpha(0)$ is very large (because $\lim_{\alpha \to 0} f'_\alpha(0) = +\infty$), the null components of $x^{(1)}$ will be strongly weighted implying that the algorithm will get stuck at this point. To avoid this problem, we propose to iteratively solve the following approximate problem that gives less weight on null components

$$x^{(k+1)} := \operatorname{argmin} \sum_{1 \leqslant i \leqslant p} f'_\alpha(|x_i^{(k)}| + \Delta)|x_i| \text{ subject to } Dx = y. \tag{3.18}$$

We have shown that this sequence is stationnary and we obtain the following theorem that states that the limit of this sequence is a local minimum of the problem $\mathcal{P}_0$.

Theorem 3.3 *Let $(x^{(k)})_{k\in\mathbb{N}}$ be the sequence defined in (3.18) and l its limit then, there exists a radius $r > 0$ such that $\forall x \in B_\infty(l, r)$ with $Dx = y$ and $x \neq l$, we have $\|x\|_0 > \|l\|_0$.*

Obviously, this local convergence can be seen as disappointed. This is the price to pay to have a procedure without assuming any of the previously cited assumptions. Nevertheless, we could see in the following subsections that a nice choice for the starting point $x^{(0)}$ seems to drive the sequence onto the global minimum.

Sparsest $\epsilon-$approximations

Similarly to the resolution of $\mathcal{P}_0$, to solve the intractable problem $\mathcal{P}_0^\epsilon$, one substitutes the constraint $Dx = y$ that appears in the problem $\mathcal{P}_{f_\alpha}$ by the constraint $\|y - Dx\|_2^2 \leq \epsilon$. This modification leads to consider

$$S_{f_\alpha}^\epsilon := \operatorname{argmin} \sum_{1\leqslant i\leqslant p} f_\alpha(|x_i|) \text{ subject to } \|y - Dx\|^2 \leq \epsilon. \qquad (\mathcal{P}_{f_\alpha}^\epsilon)$$

The following theorem 3.4 shows that, when α is small enough, the set $S_{f_\alpha}^\epsilon$ is arbitrary close to the set S_0^ϵ of solutions of $\mathcal{P}_0^\epsilon$. For this theorem, we introduce the $\eta-$magnification of the set S_0^ϵ. It is defined as the open set $G_\eta := \bigcup_{x\in S_0^\epsilon} B(x, \eta)$, where $B(x, \eta)$ is an l^2 open ball of radius $\eta > 0$ centered in x.

Theorem 3.4 *Let $(f_\alpha)_{\alpha>0}$ be a family of strictly increasing, strictly concave and continuous functions defined on $\mathbb{R}_+$ such that*

$$0 < \alpha \leq \alpha' \Rightarrow f_\alpha \geq f_{\alpha'} \text{ and } \forall x \in \mathbb{R}_+ \lim_{\alpha\to 0} f_\alpha(x) = \mathbf{1}_{x\neq 0}.$$

Then, for all $\eta > 0$, there exists $\alpha_0 > 0$ such that the following inclusion holds

$$\forall \alpha \leq \alpha_0, S_{f_\alpha}^\epsilon \subset G_\eta.$$

Such families of functions may appear difficult to build, but this is not the case. As an example, the assumptions of Theorem 3.4 hold for the families of functions $f_\alpha : x \in \mathbb{R}_+ \mapsto x/(\alpha + x)$ and $f_\alpha : x \in \mathbb{R}_+ \mapsto \arctan(x/\alpha)$.

To solve numerically the problem $\mathcal{P}_{f_\alpha}^\epsilon$, one uses the same MM method as previously leading to the iterative sequence given hereafter. Let $x^{(0)} \in \mathbb{R}^p$ and define the sequence $(x^{(k)})_{k\in\mathbb{N}}$ as follows

$$x^{(k+1)} := \operatorname{argmin} \sum_{1\leqslant i\leqslant p} f'_\alpha(|x_i^{(k)}| + \Delta)|x_i| \text{ subject to } \|y - Dx\|^2 \leq \epsilon. \qquad (3.19)$$

We have shown, that, whatever Δ, when $x^{(0)}$ is well chosen, one can expect that for k large enough, $x^{(k)}$ is arbitrary close to the set $S_{f_\alpha}^\epsilon$.

3.5.3 Numerical experiments

Choice of the initial point

Whereas by taking $x^{(0)} = x^{\text{bp}}$ the performances of the modified MM method to solve $\mathcal{P}_0$ are better than the performances of the basis pursuit, x^{bp} is not the better initial point. Because the MM algorithm converges to a local minimum of $\mathcal{P}_0$, the choice of its initial point is critical. Candes et al. [2008] took the solution of problem $\mathcal{P}_1$ as the initial point for the iterative sequence (3.18). Another way to choose this initial point is based on the following two remarks.

1. Intuitively, the largest components of $\tilde{x}$ are more easily recovered than the smallest one.
2. When $\mathcal{A}$ is a known set that owns the largest components of $\tilde{x}$, the expression $\sum_{i\notin\mathcal{A}}|\tilde{x}_i|$ becomes small. As a consequence, substituting in $\mathcal{P}_1$ the function $\sum_{i=1}^{p}|x_i|$ by $\sum_{i\notin\mathcal{A}}|\tilde{x}_i|$ should provide a solution closer to $\tilde{x}$ than x^{bp}. So, to insure the uniqueness of the solution, instead of $\sum_{i\notin\mathcal{A}}|x_i|$ we could minimize the expression $\omega\sum_{i\in\mathcal{A}}|x_i|+\sum_{i\notin\mathcal{A}}|x_i|$, with ω very small. This leads to the problem

$$\text{argmin}\omega\sum_{i\in\mathcal{A}}|x_i|+\sum_{i\notin\mathcal{A}}|x_i| \text{ subject to } Dx=y. \tag{$\mathcal{P}_{\mathcal{A}}$}$$

   provides a closer solution of $\tilde{x}$ than the problem $\mathcal{P}_1$.

Using these remarks, we could build a simple procedure to provide an initial point $x^{(0)}$. The input of this procedure is x^{bp}. Ideally, when $\mathcal{A}_1\subset\mathcal{A}_2\subset\cdots\subset\text{supp}(\tilde{x})$, the solutions $x^{\text{init},(1)}, x^{\text{init},(2)}\ldots$ of the problems $\mathcal{P}_{\mathcal{A}_1},\mathcal{P}_{\mathcal{A}_2},\ldots$ should be increasingly close to $\tilde{x}$. When at the j^{th} iteration $\text{Card}(\text{supp}(x^{\text{init},(j)})\setminus\mathcal{A}_j)=0$, it is not possible to find an element i_j to construct the set $\mathcal{A}_{j+1}$ and the algorithm stops. As already mentioned, the sparsest representation of y in D has a l^0 "norm" smaller than n. Consequently, the previous inclusion can not hold after the n^{th} iteration. So we stop the algorithm no later than the n^{th} iteration.

Comparisons

Currently, the basis pursuit $\mathcal{P}_1$ and the reweighted l^1 minimization [Candes et al., 2008] are the reference methods to recover a solution of $\mathcal{P}_0$. So, we compare our method with both the basis pursuit and the reweighted l^1 minimization. For this numerical study, we use the same simulation framework as Candes et al. [2008]. The family $\mathcal{D}=\{d_1,\ldots,d_p\}$ owns $p=256$ vectors of $\mathbb{R}^n$ with $n=100$. Whatever $i\in[\![1,256]\!]$, the vector d_i is random vector $d_i:=X_i/\|X_i\|$ with X_i i.i.d $\mathcal{N}(0,Id_{100})$. Consequently, the vectors $d_1,\ldots,d_p$ are independent and uniformly distributed on the $\mathbb{R}^n$ sphere. The vector $y\in\mathbb{R}^{100}$ that appears in the constraint $y=Dx$ is such that $y=D\tilde{x}$. For a given $s\in[\![1,n-1]\!]$, we choose $\tilde{x}$ as a random vector constructed as follows. Let $Z_1,\ldots,Z_s$ be i.i.d random variables $\mathcal{N}(0,1)$ distributed, we set $\forall i\notin[\![1,s]\!],\tilde{x}_i=0$ and $\forall i\in[\![1,s]\!],\tilde{x}_i:=Z_{(i)}$, where $Z_{(1)},\ldots,Z_{(s)}$ are ordered variables such that $|Z_{(1)}|\geq\cdots\geq|Z_{(s)}|$. Because, by construction, almost surely the unique representation property holds for D (*i.e.* with a probability 1, $\text{spark}(D)=n+1$), when $s<(n+1)/2$ $\tilde{x}$ is almost surely the unique sparsest representation of y in D [Woodworth and Chartrand, 2016]. When $s\in[\![(n+1)/2,n-1]\!]$, one can show that $\tilde{x}$ is still the unique sparsest representation of y in D. The proposed MM method aims to find the sparsest representation of y in D which correspond to $\tilde{x}$.

In this section, we propose to slightly modify as follows the MM method given in (3.18).

$$\text{Let } a:=\text{argmin}\sum_{1\leqslant i\leqslant p}f'_\alpha(|x_i^{(k)}|+\Delta)|x_i| \text{ subject to } Dx=y \text{ and set } \begin{cases}x^{(k+1)}=a \text{ if } \|a\|_0\leq\|x^{(k)}\|_0\\ x^{(k+1)}=x^{(k)} \text{ otherwise}\end{cases}. \tag{3.20}$$

The general position condition holds almost surely for D. This condition insure the uniqueness of the weighted basis pursuit solution [Rosset et al., 2004] thus at the iteration k the solution $x^{(k)}$ is unique. The computation of the sequence $(x^{(k)})_{k\geq 0}$ has been performed with the R package **lpSolve**. As for the sequence given in (3.18), when k is large enough, the sequence (3.20) is stationary onto a point l. As defined in (3.20) the sequence $(\|x^k\|_0)_{k\in\mathbb{N}}$ is decreasing, consequently, $\|l\|_0 \leq \|x^{(0)}\|_0$. In particular when the initial point is the solution of $\mathcal{P}_1$, denoted hereafter x^{bp}, the modified MM method allows to catch a representation l better than x^{bp} in the sense that $\|l\|_0 \leq \|x^{\text{bp}}\|_0$.

The simulations were performed for each $s \in \{24, 26, \ldots, 72\}$ using 500 random vectors $\tilde{x}$ such that $\text{supp}(\tilde{x}) = [\![1, s]\!]$, and 500 families $\mathcal{D} = \{d_1, \ldots, d_{256}\}$. These random vectors were ordered so that $|\tilde{x}_1| \geq \cdots \geq |\tilde{x}_s|$. For each family and each $\tilde{x}$, we compute the basis pursuit solution (x^{bp}) of $\mathcal{P}_1$, the reweighted l^1 minimization solution and the solution given by our method as defined by (3.20). The reweighted l^1 solution is the limit of the sequence $(x^{\text{l1,(k)}})_{k\in\mathbb{N}}$ defined by $x^{\text{l1,(0)}} = x^{\text{bp}}$ and

$$x^{\text{l1,(k+1)}} := \text{argmin} \sum_{i=1}^{p} \frac{1}{|x_i^{\text{l1,(k)}}| + \delta} |x_i| \text{ subject to } Dx = y, \text{ with } y = D\tilde{x}.$$

As in Candes et al. [2008] we set $\delta = 0.1$. The number of iterations was set to $k_0 = 8$ for both the reweighted l^1 minimization method and our method. We choose $f_\alpha(x) = x^\alpha$ with $\alpha = 0.01$ and the initial point of (3.20) was computed using the algorithm described previously.

The figure 3.8 shows the performances of the basis pursuit, the reweighted l^1 minimization and our method.

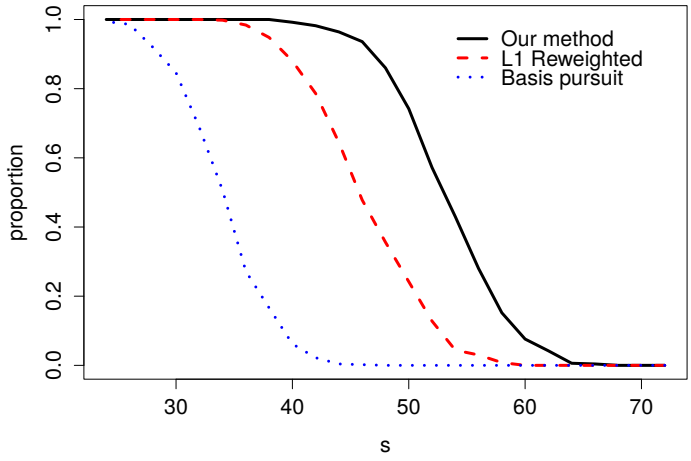

FIGURE 3.8 – The performances of the three competing methods are represented by the proportions of realisations of the events $x^{\text{bp}} = \tilde{x}$, $x^{\text{l1,(8)}} = \tilde{x}$ and $x^{(8)} = \tilde{x}$ as a function of the number of non null components of $\tilde{x}$ denoted s. One notices that the graph of the reweighted l1 minimization method is almost the same as those given in Candes et al. [2008].

Numerical experiments given in the figure 3.8 show that when $\|\tilde{x}\|_0 \leq 22$, $\tilde{x}$ is always recovered by all these three methods. No method recovered $\tilde{x}$ when $\|\tilde{x}\|_0 \geq 68$. When $22 \leq \|\tilde{x}\|_0 \leq 68$, the proportion of times for which our method recovers $\tilde{x}$ is greater than the proportion given by the two other methods. These numerical experiments illustrate that the performances of our method are better than those of the basis pursuit and the reweighted l^1 minimization.

3.5.4 Discussion

In this study, the vector y is not corrupted by any noise. When y is a random vector, Meinshausen [2015] provides an estimation of the representation of its expectation which has the smallest l^1 norm. In a future work, this work could be extended to estimate the sparsest representation (*i.e.* the smallest l^0 norm) of the expectation of y.

3.6 Ongoing projects and prospects

As already explained, this section represents the major part of my research activities right now. So, the two projects I'm going to present in the following subsections will take my major research time in a near future. Their extensions (mainly for the second one, briefly mentioned in the corresponding subsection) will also be the keystone for me for middle and long-term research perspectives.

3.6.1 Statistical methods for RMN spectra analysis

A wide part of the statistical research community is focusing on problems concerning transcriptomics or genomics data. Nevertheless, in metabolomics, some important (and very interesting) statistical problems still remain (for example Blaise et al. [2016] recently developed a first method to objectify the estimation of the statistical power and the sample size for metabolomics study). Considine et al. [2018] also highlighted the lack of a standard procedure to analyse metabolomics data that could hamper the basic understanding of the results or the reuses of protocols or datasets.

This project is the natural extension of the metabolomics project developped in Section 3.4. Indeed, in this section we define a procedure to identify and quantify metabolites in 1D ^{1}H NMR spectrum. In fact, rending this identification tractable *a priori* would lead to a major modification in the whole process of spectrum analysis. Indeed, it would make metabolomics asserting a general approach to test *a priori* formulated hypotheses on the basis of exhaustive metabolome characterization rather than an exploratory tool dealing with unknown metabolic features. To be more precise : usually each generated spectrum is first divided into intervals called buckets [Alves et al., 2009]. Then, the areas under the curve are computed for each bucket. These steps are repeated for each spectrum and multiple comparisons provide a list of buckets that are significantly different between the studied groups. Finally, NMR experts identify the metabolites involved in the significant buckets. By this approach, the identification of metabolites is restricted to significant ones. Another way to proceed would be to identify and quantify all the metabolites in each spectrum and to perform statistical analyses on these data. Due to numerous problems (peak overlapping, warping spectrum ...), these automatic identification was not possible. Using the identification procedure defined in Section 3.4 it is now possible. So, using this procedure, we start to develop a new R package **ASICS** (now available at Bioconductor) that combines all the steps of the analysis of 1D ^{1}H NMR spectra (library of spectra management, preprocessing, identification, quantification, post-quantification statistical analyses). This will allow the understanding of the steps employed during an analysis and/or the reuse of the protocol by an interested researcher. All the package functionalities are summarized in Figure 3.9.

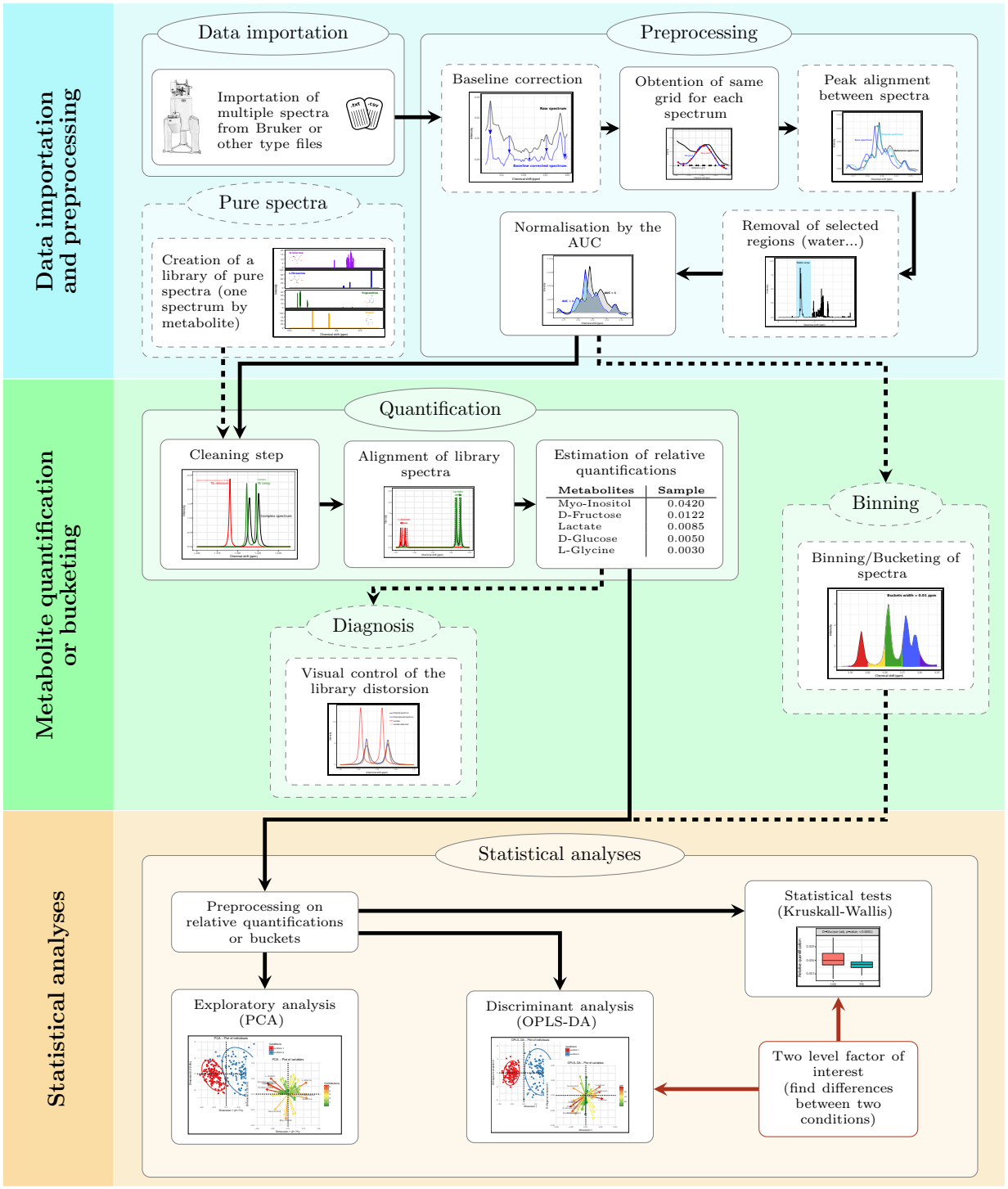

FIGURE 3.9 – Complete workflow of analysis for a 1D ^{1}H NMR spectrum in the **ASICS** package

Nevertheless, problems still remain and some will be addressed during the future PhD of Gaëlle Lefort. Some of these problems are directly linked to the metabolomics application such as the improvement of the warping step or some parameter choices on the preprocessing steps but some others are statistical research problems. First, as explained in Subsection 3.4.5, our quantification of the metabolites is a two step procedure (first selection then quantification)

and the statistical properties of the final quantification estimations are not well established. Studying these estimations using post selection inference theory [Berk et al., 2013] would be of great interest, especially if we can control the FWER with a dedicated approach [Blanchard et al., 2017]. Second, incorportating *a priori* biological information on the model would also help to adress the identifiability issue for example using a Bayesian approach [Grollemund et al., 2018] or a constrained regression problem [Hofner et al., 2016]. All the developed methodologies will be applied to datasets to explain early death in piglets. I can also bet that, studying one of these problem would lead to another one, such as Section 3.5 was derived from Section 3.4 previously.

3.6.2 Statistical methods for precision livestock farming

I'm now part of a new unit called InTheRes (for "Innovations thérapeutiques et résistances"). One of the main goal of this unit is to propose new breeding management tools to decrease the amount of antibiotics used. This would be part of the precision livestock farming (PLF) framework and could be seen as a (maybe far) extension of Section 3.2. Indeed, the modernisation of food production systems is characterized by the development of PLF. PLF systems aim to offer a real-time monitoring and managing system for the farmer, providing a real-time warning of a problem so that immediate action can be taken [Berckmans, 2014, Ellies-Oury et al., 2016]. This requires real-time algorithms that are able to detect or predict problems while the rearing process in ongoing. The basic methods used in PLF involve continuously measuring responses directly produced by the animal. These real-time responses, known as bio-signals, can be temperature measurement, GPS position, accelerometer data, real-time image analysis, sound analysis, or water/food consumption activity. In this spirit, we built a project called PigletDetect with the French pork institute IFIP (that could perform tests and produce datasets on the breeding of piglets) and the manufacturer ASSERVA that produces the connected-feeding system. This project is based on the fact that the individual behaviours of pigs are linked to their health status. So, analysis of individual drinking behaviour could allow these problems to be detected upon occurrence of a pathology and even before the first symptoms are visible by an operator [Madsen et al., 2005] and an early individual detection of the disease would decrease the amount of antibiotics used. Using HF RFID technology, we are now in position to continuously monitor the weights, the food and the water consumption at the individual level in pigs. In the project, we associate this real-time measurements (that could be viewed as functional data) with a clinical evaluation of the health status. Then, by mathematical modelling of the individual time-series produced during the project, we aim at identifying early the individuals or set of individuals becoming diseased, and thus allowing the farmer/veterinary to choose rationally a therapeutic strategy.
Nevertheless, it brings some modelling difficulties. We have to derive the health status from each individual signals. This implies to model how this hidden state (the health status) changes with time and switches between the reference curve of a healthy animal to the reference curve of a diseased one. In the spirit of Aparna et al. [2014], Bartolucci and Farcomeni [2015], hidden Markov processes should be a good simplified modelling to start by but need to be adapted to our problems to provide a dedicated procedure. At the end, this approach should produce alarms for all diseased animals. As for all detection system, false alarms and undetected diseased animals will occur. To be implemented in a breed, these two errors of detection should be minimized. Finally, we anticipate that the breed management has a strong impact on the shape of the reference curves for both healthy and diseased animals. The breed management

should then have an impact on the performances of the detection system that will be built from experiments coming from the IFIP station. This is the reason why, a statistical learning method will be proposed. This learning method will learn with time how to minimize these two errors of detection for the specific conditions of the breed. More practically, the learning method will adjust in real-time the parameters values of the detection system to the breed management. On this ongoing project, Malika Chassan, a post-doctorate student, is now working on these questions.

With the development of precision livestock farming, this kind of projects will take the major part of my future research work. The species could obviously be different (cattle, broilers, lamb ...) as well as the recorded real-time measurement (GPS tracking, video ...) leading to other kind of interesting statistical problems.

Bibliography

R. G. Allen, L. S. Pereira, D. Raes, and M. Smith. Crop evapotranspiration-guidelines for computing crop water requirements-fao irrigation and drainage. *FAO, Rome*, 300(9) :D05109, 1998.

A. Alves, M. Rantalainen, E. Holmes, J. K. Nicholson, and T. M. D. Ebbels. Analytic properties of statistical total correlation spectroscopy based information recovery in 1H NMR metabolic data sets. *Analytical Chemistry*, 81 :2075–2084, 2009.

G. Aneiros and P. Vieu. Variable in infinite-dimensional problems. *Statistics and Probability Letters*, 94 :12–20, 2014.

L. Aparna, J. Pedersen, and E. Jorgensen. Hidden phase-type Markov model for the prediction of onset of farrowing for loose-housed sows. *Computers and Electronics in Agriculture*, 108 : 135–147, 2014.

W. Astle, M. de Iorio, S. Richardson, D. Stephens, and T. Ebbels. A Bayesian model of NMR spectra for deconvolution and quantification of metabolites in complex biological mixtures. *Journal of the American Statistical Association*, 107(500) :1259–1271, 2012.

A. Bar-Hen. Collective effectiveness in the XV de France : selections and time matter. *European Journal of Sport Science*, 17(6) :656–664, 2017.

R. Baraniuk and P. Steeghs. Compressive radar imaging. In *Radar Conference, 2007 IEEE*, pages 128–133. IEEE, 2007.

R. F. Barber and E. J. Candes. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5) :2055–2085, 2015.

O. E. Barndorff-Nielsen and D. R. Cox. Prediction and asymptotics. *Bernoulli*, 2(4) :319–340, 1996.

F. Bartolucci and A. Farcomeni. Information matrix for hidden Markov models with covariates. *Statistics and Computing*, 25(3) :515–526, 2015.

R. Beran. Calibrating prediction regions. *Journal of the American Statistical Association*, 85 (411) :715–723, 1990.

D. Berckmans. Precision livestock farming technologies for welfare management in intensive livestock systems. *Revue scientifique et technique (IOE)*, 33(1) :189–196, 2014.

R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2) :802–837, 04 2013.

C. Bernard-Michel, L. Gardes, and S. Girard. A note on sliced inverse regression with regularizations. *Biometrics*, 64(3) :982–986, 2008.

D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

B. Bettonvil. Factor screening by sequential bifurcation. *Communications in Statistics, Simulation and Computation*, 24(1) :165–185, 1995.

R. Bhatia. *Positive Definite Matrices*. Princeton Series in Applied Mathematics. Princeton University Press, 2009.

B. J. Blaise, G. Correia, A. Tin, J. H. Young, A.-C. Vergnaud, M. Lewis, J. T. M. Pearce, P. Elliott, J. K. Nicholson, E. Holmes, and T. M. D. Ebbels. Power analysis and sample size determination in metabolic phenotyping. *Analytical Chemistry*, 88(10) :5179–5188, 2016.

G. Blanchard, P. Neuvial, and E. Roquain. Post hoc inference via joint family-wise error rate control, 2017. Submitted.

M. Bogdan, E. van den Berg, C. Sabatti, W. Su, and E. J. Candes. Slope - adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3) :1103–1140, 2015.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data : Methods, Theory and Applications*. Springer, 2011.

M. Burger, C. Rossmanith, and X. Zhang. Simultaneous reconstruction and segmentation for dynamic spect imaging. *Inverse Problems*, 32(10) :104002, 2016.

T. T. Cai and A. Zhang. Sharp RIP bound for sparse signal and low-rank matrix recovery. *Applied and Computational Harmonic Analysis*, 35(1) :74–93, 2013.

E. J. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9) :589–592, 2008.

E. J. Candes, M. B. Wakin, and S. P. Boyd. Enhancing sparsity by reweighted l_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5-6) :877–905, 2008.

H. Cardot, F. Ferraty, and P. Sarda. Functional linear model. *Statistics and Probability Letters*, 45(1) :11–22, 1999.

P. Casadebaig, L. Guilioni, J. Lecoeur, A. Christophe, L. Champolivier, and P. Debaeke. Sunflo, a model to simulate genotype-specific performance of the sunflower crop in contrasting environments. *Agricultural and forest meteorology*, 151(2) :163–178, 2011.

C. Chen and K. Li. Can SIR be as popular as multiple linear regression ? *Statistica Sinica*, 8 : 289–316, 1998.

S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis puirsuit. *SIAM Journal on Scientific Computing*, 20(1) :33–61, 2015.

R. Chetrite, R. Diel, and M. Lerasle. The number of potential winners in Bradley-Terry model in random environment. *The Annals of Applied Probability*, 27(3) :1372–1394, 06 2017.

CLSI. *Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory : Approved Guideline*. CLSI document C28-A3. Wayne, PA : CLSI, 2008.

E. C. Considine, G. Thomas, A. L. Boulesteix, A. S. Khashan, and L. C. Kenny. Critical review of reporting of the data analysis step in metabolomics. *Metabolomics*, 14(1) :7, 2018.

R. Cook. Testing predictor contributions in sufficient dimension reduction. *Annals of Statistics*, 32(3) :1061–1092, 2004.

R. Coudret, B. Liquet, and J. Saracco. Comparison of sliced inverse regression aproaches for undetermined cases. *Journal de la Société Française de Statistique*, 155(2) :72–96, 2014.

J. Dauxois, L. Ferré, and A. Yao. Un modèle semi-paramétrique pour variable aléatoire hilbertienne. *Comptes Rendus de l'Académie des Sciences*, 327(I) :947–952, 2001.

M. Davidian and D. M. Giltinan. *Nonlinear models for repeated measurement data*. Monographs on statistics and applied probability. Chapman & Hall, London, New York, 1995.

A. Dehman, C. Ambroise, and P. Neuvial. Performance of a blockwise approach in variable selection using linkage disequilibrium information. *BMC Bioinformatics*, 16(148), 2015.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39(1) :1–38, 1977.

D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5) : 2197–2202, 2003.

D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1) :6–18, 2006.

O. J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293) :52–64, 1961.

M.-P. Ellies-Oury, G. Cantalapiedra-Hijar, D. Durand, D. Gruffat, A. Listrat, D. Micol, I. Ortigues-Marty, J.-F. Hocquette, M. Chavent, J. Saracco, and B. Picard. An innovative approach combining animal performances, nutritional value and sensory quality of meat. *Meat Science*, 122 :163 – 172, 2016.

J. H. Ender. On compressive sensing applied to radar. *Signal Processing*, 90(5) :1402–1414, 2010.

M. Fauvel, C. Deschene, A. Zullo, and F. Ferraty. Fast forward feature selection of hyperspectral images for classification with Gaussian mixture models. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(6) :2824–2831, 2015.

F. Ferraty and P. Hall. An algorithm for nonlinear, nonparametric model choice and prediction. *Journal of Computational and Graphical Statistics*, 24(3) :695–714, 2015.

F. Ferraty and P. Vieu. *NonParametric Functional Data Analysis*. Springer, 2006.

F. Ferraty, P. Hall, and P. Vieu. Most-predictive design points for functiona data predictors. *Biometrika*, 97(4) :807–824, 2010.

L. Ferré. Determining the dimension in sliced inverse regression and related methods. *Journal of the American Statistical Association*, 93(441) :132–140, 1998.

L. Ferré and N. Villa. Multi-layer perceptron with functional inputs : an inverse regression approach. *Scandinavian Journal of Statistics*, 33(4) :807–823, 2006.

G. Fonseca, F. Giummolè, and P. Vidoni. A note about calibrated prediction regions and distributions. *Journal of Statistical Planning and Inference*, 142(9) :2726–2734, 2012.

S. Foucart and M.-J. Lai. Sparsest solutions of underdetermined linear systems via l_q-minimization for $0 < q \leq 1$. *Applied and Computational Harmonic Analysis*, 26(3) :395–407, 2009.

R. Fraiman, Y. Gimenez, and M. Svarc. Feature selection for functional data. *Journal of Multivariate Analysis*, 146 :191–208, 2016.

A. Franks, A. Miller, L. Bornn, and K. Goldsberry. Characterizing the spatial structure of defensive skill in professional basketball. *The Annals of Applied Statistics*, 9(1) :94–121, 03 2015.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1) :1–22, 2010.

J. Fruth, O. Roustant, and S. Kuhnt. Sequential designs for sensitivity analysis of functional inputs in computer experiments. *Reliability Engineering & System Safety*, 134 :260–267, 2015.

G. S. Ginsburg and K. A. Phillips. Precision medicine : from science to value. *Health Affairs*, 37(5) :694–701, 2018.

R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Transactions on Information Theory*, 49(12) :3320–3325, 2003.

R. Gribonval and M. Nielsen. Highly sparse representations from dictionaries are unique and independent of the sparseness measure. *Applied and Computational Harmonic Analysis*, 22 (3) :335–355, 2007.

P. Grollemund, C. Abraham, M. Baragatti, and P. Pudlo. Bayesian functional linear regression with sparse step functions. *Bayesian Analysis*, Forthcoming, 2018.

M. G'Sell, S. Wager, A. Chouldechova, and R. Tibshirani. Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Sociey : Series B*, 78(2) :423–444, 2015.

P. Hall, L. Peng, and N. Tajvidi. On prediction intervals based on predictive likelihood or bootstrap methods. *Biometrika*, 86(4) :871–880, 1999.

B. Hanczar and A. Bar-Hen. Controlling the cost of prediction in using a cascade of reject classifiers for personalized medicine. *7th International Conference on Bioinformatics Models, Methods and Algorithms*, pages 42–50, 2016.

J. Hao, W. Astle, M. De Iorio, and T. M. Ebbels. BATMAN - an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a bayesian model. *Bioinformatics*, 28(15) :2088–2090, 2012.

J. Hao, M. Liebeke, W. Astle, M. De Iorio, J. Bundy, and T. Ebbels. Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nature Protocols*, 9(6) :1416–1427, 2014.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning : Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer, 2009.

P. Hellard, M. Avalos, F. Guimaraes, J. F. Toussaint, and P. David. Training-Related Risk of Common Illnesses in Elite Swimmers over a Four-Year Period. *Medicine and Science in Sports and Exercise*, 47(4) :698–707, 2015.

M. A. Herman and T. Strohmer. High-resolution radar via compressed sensing. *IEEE transactions on signal processing*, 57(6) :2275–2284, 2009.

B. Hofner, T. Kneib, and T. Hothorn. A unified framework of constrained regression. *Statistics and Computing*, 26(1) :1–14, 2016.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2) :65–70, 1979.

D. R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 58(1) : 30–37, 2004.

G. James, J. Wang, and J. Zhu. Functional linear regression that's interpretable. *Annals of Statistics*, 37(5A) :2083–2108, 2009.

L. Janson and W. Su. Familywise error rate control via knockoffs. *Electronic Journal of Statistics*, 10(1) :960–975, 2016.

I. Jlassi and J. Saracco. Variable importance assessment in sliced inverse regression for variable selection. *Communications in Statistics - Simulation and Computation*, pages 1–31, 2017.

A. Kneip, D. Poß, and P. Sarda. Functional linear regression with points of impact. *The Annals of Statistics*, 44(1) :1–30, 2016.

M.-J. Lai. On sparse solutions of underdetermined linear systems. *Journal of Concrete and Applicable Mathematics*, 8(2) :296–327, 2010.

H. P. Lefebvre. Renal function testing. In B. J. and P. D. J., editors, *Nephrology and urology of small animals*, pages 91–96. Blackwell Publishing, Ames, IA, USA, 2011.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, New York, 2005.

K. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414) :316–342, June 1991.

L. Li and C. Nachtsheim. Sparse sliced inverse regression. *Technometrics*, 48(4) :503–510, 2008.

L. Li and X. Yin. Sliced inverse regression with regularizations. *Biometrics*, 64(1) :124–131, 2008.

B. Liquet and J. Saracco. A graphical tool for selecting the number of slices and the dimension of the model in SIR and SAVE approches. *Computational Statistics*, 27(1) :103–125, 2012.

J. Liu and H. Gao. Material reconstruction for spectral computed tomography with detector response function. *Inverse Problems*, 32(11) :114001, 2016.

M. S. Lobo, M. Fazel, and S. Boyd. Portfolio optimization with linear and fixed transaction costs. *Annals of Operations Research*, 152(1) :341–365, 2007.

R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2) :413–468, 2014.

K. Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of Statistics*, 2 :90–102, 2008.

T. Madsen, S. Andersen, and A. Kristensen. Modelling the drinking patterns of young pigs using a state space model. *Computers and Electronics in Agriculture*, 48 :39–62, 2005.

H. Matsui and S. Konishi. Variable selection for functional regression models via the l_1 regularization. *Computational Statistics and Data Analysis*, 55(12) :3304–3310, 2011.

N. Meinshausen. Group bound : confidence intervals for groups of variables in sparse high dimensional regression without assumptions on the design. *Journal of the Royal Statistical Society : Series B*, 77(5) :923–945, 2015.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3) :1436–1462, 2006.

L. Ni, D. Cook, and C. Tsai. A note on shrinkage sliced inverse regression. *Biometrika*, 92(1) : 242–247, 2005.

K. Prieto and O. Dorn. Sparsity and level set regularization for diffuse optical tomography using a transport model in 2D. *Inverse Problems*, 33(1) :014001, 2016.

D. E. Pritchard, F. Moeckel, M. S. Villa, L. T. Housman, C. A. McCarty, and H. L. McLeod. Strategies for integrating personalized medicine into healthcare practice. *Personalized Medicine*, 14(2) :141–152, 2017.

S. Ravanbakhsh, P. Liu, T. C. Bjordahl, R. Mandal, J. R. Grant, M. Wilson, R. Eisner, I. Sinelnikov, X. Hu, C. Luchinat, R. Greiner, and D. S. Wishart. Accurate, fully-automated NMR spectral profiling for metabolomics. *PLoS ONE*, 10(5) :e0124219, 2015.

B. Reynolds, D. Concordet, C. Germain, T. Daste, K. Boudet, and H. Lefebvre. Breed dependency of reference intervals for plasma biochemical values in cats. *Journal of Veterinary Internal Medicine*, 24(4) :809–818, 2010.

J. P. Romano and M. Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469) :94–108, 2005.

S. Rosset, J. Zhu, and T. Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5 :941–973, 2004.

G. Saulière, J. Dedecker, L.-A. Marquet, P. Rochcongar, J.-F. Toussaint, and G. Berthelot. Z-scores-based methods and their application to biological monitoring : an example in professional soccer players. *To appear in Biostatistics*, 2018.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22 :231–245, 2013.

Y. Simón-Manso, M. Lowenthal, L. Kilpatrick, M. Sampson, K. Telu, P. Rudnick, W. Mallard, D. Bearden, T. Schock, D. Tchekhovskoi, N. Blonder, X. Yan, Y. Liang, Y. Zheng, W. Wallace, P. Neta, K. Phinney, A. Remaley, and S. Stein. Metabolite profiling of a NIST standard reference material for human plasma (SRM 1950) : GC-MS, LC-MS, NMR, and clinical laboratory analyses, libraries, and web-based resources. *Analytical Chemistry*, 85(24) : 11725–11731, 2013.

P.-E. Sottas, N. Baume, C. Saudan, C. Schweizer, M. Kamber, and M. Saugy. Bayesian detection of abnormal values in longitudinal biomarkers with an application to t/e ratio. *Biostatistics*, 8(2) :285–296, 2007.

W. Su and E. Candes. Slope is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3) :1038–1068, 2016.

Q. Sun. Recovery of sparsest signals via l_q-minimization. *Applied and Computational Harmonic Analysis*, 32(3) :329–341, 2012.

P. Tardivel. *Représentation parcimonieuse et procédures de tests multiples : application à la métabolomique*. PhD thesis, Université de Toulouse, 2017.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1) :267–288, 1996.

A. M. Tillmann and M. E. Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory*, 60(2) :1248–1259, 2014.

D. Tulpan, S. Léger, L. Belliveau, A. Culf, and M. Čuperlović-Culf. Metabohunter : an automatic approach for identification of metabolites from ^{1}H-NMR spectra of complex mixtures. *BMC Bioinformatics*, 12(1) :400, 2011.

M. Ueki and K. Fueda. Adjusting estimative prediction limits. *Biometrika*, 94(2) :509–511, 2007.

S. A. Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2) :614–645, 2008.

G. Verbeke and G. Molenberghs. *Linear mixed models for longitudinal data*. Springer Series in Statistics. Springer-Verlag, New-York, 2000.

K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained k-means clustering with background knowledge. In *Proceedings of ICML '01*, 2001.

W. L. Wang and T. H. Fan. ECM-based maximum likelihood inference for multivariate linear mixed models with autoregressive errors. *Computational Statistics and Data Analysis*, 54(5) : 1328–1341, 2010.

A. M. Weljie, J. Newton, P. Mercier, E. Carlson, and C. M. Slupsky. Targeted profiling : Quantitative analysis of ^{1}H-NMR metabolomics data. *Analytical Chemistry*, 78(13) :4430–4442, 2006.

J. Woodworth and R. Chartrand. Compressed sensing recovery via nonconvex shrinkage penalties. *Inverse Problems*, 32(7) :075004, 2016.

Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang. A survey of sparse representation : algorithms and applications. *IEEE Access*, 3 :490–530, 2015.

P. Zhao and B. Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7 :2541–2563, 2006.

M. Zorzoli and F. Rossi. Implementation of the biological passport : The experience of the international cycling union. *Drug Testing and Analysis*, 2(11-12) :542–547, 2010.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476) :1418–1429, 2006.

Resume

35 years old.
Mail : remi.servien@inra.fr
Webpage : http ://www.biostat.envt.fr/remi-servien/

UMR 1436 InTheRes, INRA - ENVT, Laboratoire de Physiologie,
23 chemin des Capelles, BP 87614, 31076 Toulouse Cedex 03, France.
Phone number : +33 (0)5.61.19.32.82

Professional experience

Since 2018 : **Permanent researcher INRA**, UMR 1436 InTheRes, Toulouse. Associate member of the Mathematical Institute of Toulouse (UMR CNRS 5219).

2012 - 2017 : Permanent researcher INRA, UMR 1331 Toxalim, Toulouse. Associate member of the Mathematical Institute of Toulouse (UMR CNRS 5219).

2011 - 2012 : Lecturer, University Montpellier I, Medicine faculty.

2010 - 2011 : Lecturer, University Perpignan Via Domitia, STID faculty.

2009 - 2010 : Lecturer, INP Grenoble, ENSIMAG.

2006 - 2009 : PhD in Statistics, University Montpellier II.
Estimation de régularité locale.
Commitee : C. Abraham, A. Berlinet (supervisor), G. Biau, A. Mas, B. Pelletier (reporter), P. Sarda (reporter).
Defended the 12th March 2010.

Contributions

See Page 1 of this manuscript.

Fundings

2018 - 2019 : Member of the project "SuBPiG", **Métaprogramme INRA GISA**, 60k€.

2017 - 2019 : Principal Investigator of the project "PigletDetect", **Institut Carnot France Futur Elevage (F2E)**, 299k€.

2016 : Principal Investigator of the project "MathExpo2", **PEPS FaiDoRA, CNRS**, 13k€.

2015 - 2017 : Principal Investigator of the project "Automatic Statistical Identification of Metabolites in Complex Spectra" (ASICS), **IDEX Toulouse**, 39 k€.

2015 : Principal Investigator of the project "MathExpo", **PEPS FaiDoRA, CNRS**, 15k€.

2014 - 2017 : Work-Package Leader of the research consortium **GMO^{90+}**, funded by **MEDDE** ministry. Total budget : 2.5 M€, budget of my WP : 220 k€.

Duties

2017 - ... : Member of the organizing commitee of the **useR !2019** conference.

2017 - ... : Elected member at the **Scientific commitee** of the Toulouse Vet School (ENVT).

2017 - ... : Elected member at the **management committee** of the Animal Health department of INRA.

2015 - ... : Expert at the **Biotechnology** Working Group of the French Agency for Food, Environmental and Occupational Health & Safety **(ANSES)**.

2013 - ... : Member of the ethic committee of the Toxalim unit.

2015 - 2017 : Member of the scientific committee of the 6th and 7th French Young Statisticians Conference.

2013 - 2017 : Co-webmaster of the French Statistic Society **(SFdS)**.

2015 - 2016 : Member of the organizing committee of the 4th R french conference, Toulouse.

2011 - 2016 : Elected member at the desk of the Youg Research Group of the **SFdS**, webmaster of the group.

2012 - 2013 : Member of the organizing committee of the **French Statistical Conference (JdS)** 2013, Toulouse.

Teaching

2017 - 2018 : **ISAE-SUPAERO** and **University of Toulouse**, Faculty of pharmacy, **15 hours**.

2016 - 2017 : **ISAE-SUPAERO**, **12 hours**.

2015 - 2016 : **ENVT**, **ISAE-SUPAERO** and **University of Toulouse**, Faculty of pharmacy, **34 hours**.

2014 - 2015 : **ENVT, 30 hours.**

2013 - 2014 : **ENVT, 30 hours.**

2011 - 2012 : **University Montpellier I**, Faculty of Medicine, **96 hours.**

2010 - 2011 : **University Perpignan Via Domitia**, IUT STID, **96 hours**.

2009 - 2010 : **INP Grenoble**, ENSIMAG, **96 hours.**

2008 - 2009 : **University Montpellier III**, Faculty of Applied Mathematics, **55 hours.**

2006 - 2007 : **University Montpellier III**, Faculty of Applied Mathematics, **76 hours**.

Supervisions

Summer-School :

2015 : Co-supervisor of the summer-school "Analysis of metabolomics data on Workflow4Metabolomics", Roscoff.

Post-doctorate :

2017 - 2019 : Co-supervisor (with D. Concordet) of the post-doctorate of M. Chassan entitled "Real-time monitoring of pigs to perform early disease detection".

PhD students :

2018 - ... : Co-supervisor (with N. Villa-Vialaneix) of the PhD of Gaëlle Lefort entitled "Statistical development for the analysis of metabolomic data with application to the analysis of early deaths in piglets".

2014 - 2017 : Co-supervisor (with D. Concordet) of the PhD of Patrick Tardivel entitled "Sparse representation and multiple testing prodecures with application to metabolomics". Defended the 24th November 2017.

Master internships :

2018 : Co-supervisor (with T. Laloë) of the Master 2 internship de Hai Dang Dau (Polytechnic School).

2018 : Co-supervisor (with N. Villa-Vialaneix et V. Picheny) of the Master 2 internship of Yousra Khattali (University of Toulouse).

2018 : Co-supervisor (with E. Latrille et V. Rossard) of the Master 2 internship of Sandrine Akouete (University of Grenoble).

2016 : Co-supervisor (with N. Villa-Vialaneix, S. Saby et S. Chastant) of the Master 2 internship of Jessica Wenting (Toulouse School of Economics).

2014 : Co-supervisor (with E. Latrille et V. Rossard) of the Master 2 internship of Enguerrand De Villiers (ISPED, Bordeaux).

2012 : Co-supervisor (with E. Latrille, V. Rossard, L. Mamy et P. Benoit) of the Master 2 internship of Ziang Li (AgroParisTech).