Représentation parcimonieuse et procédures de tests multiples : application à la métabolomique

Patrick Tardivel

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

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

Présentée et soutenue le *24/11/2017* **par :**

PATRICK TARDIVEL

Représentation parcimonieuse et procédures de tests multiples : application à la métabolomique

---

JURY

DIDIER CONCORDET	Professeur des Universités	Directeur de thèse
MAGALIE FROMONT	Professeur des Universités	Examinateur
BÉATRICE LAURENT	Professeur des Universités	Président du Jury
HOLGER RAUHUT	Professeur des Universités	Rapporteur
JÉRÔME SARACCO	Professeur des Universités	Examinateur
RÉMI SERVIEN	Chargé de Recherche	Co-directeur de thèse
SARA VAN DE GEER	Professeur des Universités	Rapporteur

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

Unité de Recherche :
Toxalim (UMR 1331)

Directeur(s) de Thèse :
Didier CONCORDET et *Rémi SERVIEN*

Rapporteurs :
Holger RAUHUT et *Sara VAN DE GEER*

Table des matières

Remerciements

Je termine l'écriture de ce manuscrit par la partie la plus importante : les remerciements. Pour une grande partie des collègues, amis, et membres de ma famille qui feuilletteront ce manuscrit c'est sans doute la partie que vous lirez en priorité ; ces remerciements vous sont dédiés. L'achèvement d'une thèse est une véritable aventure ; cette aventure a commencé il y a cinq années lorsque je suis retourné sur les bancs de la fac après avoir enseigné quelques années en lycée et collège. À l'époque, je comptais passer l'agrégation puis retourner enseigner au collège, au lycée ou peut-être, avec un peu de chance, en classe préparatoire. Après l'obtention de l'agrégation j'ai décidé de pousser le bouchon un peu plus loin et de continuer en master 2 recherche en espérant pouvoir faire une thèse. Ces deux années ont été difficiles financièrement mais ont été payantes puisque j'ai eu la chance de pouvoir continuer en thèse en mathématiques À l'INRA.
J'ai été très bien accueilli au sein de l'équipe 3 et l'équipe 7 de Toxalim. J'ai pu travailler dans des conditions exceptionnelles dans mon bureau à l'école vétérinaire que presque tous les thésards, post-doctorants et maîtres de conférence pourraient m'envier. C'est dans ces excellentes conditions que j'ai pu sereinement travailler sur ma thèse. Aujourd'hui j'ai la chance de soutenir très bientôt mon doctorat en compagnie d'amis, collègues, famille, et membres du jury.

Je tiens en premier lieu à remercier chaleureusement les membres du jury sans lesquels cette soutenance n'aurait pas lieu

Au professeur Jérôme Saracco, je vous suis très reconnaissant de prendre le temps de venir à ma soutenance.

À Magalie Fromont, qui me fait le plaisir de venir de Rennes pour ma soutenance ; je suis ravi de te compter parmi les membres de mon jury.

À Béatrice Laurent, naturellement nous avons pensé à toi lorsque nous avons choisi le jury ; cela me fait très plaisir de pouvoir compter sur toi pour ma soutenance.

Au professeur Sara van de Geer, je suis honoré de vous avoir comme rapporteur, je vous suis infiniment reconnaissant d'avoir rapporté mon manuscrit et d'avoir montré autant d'intérêt pour mon travail.

To professor Holger Rauhut, thank you so much for accepting to review my PhD thesis, and for your interest in my work ; it is a great honor to count you among my thesis committee. I am extremely grateful for your careful reading and for your advices to improve this manuscript.

Ce travail n'aurait pas pu voir le jour sans l'indéfectible travail de mes directeurs de thèse.

À Rémi Servien, je tiens à te remercier pour tout l'investissement dont tu as fait preuve, pour m'avoir accompagné durant tous les congrès et pour m'avoir énormément aidé pour l'article en métabolomique.

À Didier Concordet, je tiens à vous remercier chaleureusement pour vos relectures minutieuses, vos remarques très pertinentes sur la rédaction de ce manuscrit et pour m'avoir si bien formé aux présentations orales.

Je tiens à remercier tous les collègues.

À Aude, Catherine, Véronique Gayrard, Véronique Dupouy, Nicole, Étienne, Alain, Diane, Flore, Béatrice, Emmanuelle, Simone, Valérie, Faouzi, Élodie Pascal, Hervé, Monsieur Toutain, Cédric, Marlène, David, Jean-Pierre, Noémie, Nathalie, Peggy, Marie-Françoise, Christiane, Malika, Glenn et Davy, j'ai partagé d'excellents moments avec vous tous, je sais que vous serez nombreux à venir me soutenir. Parmi mes collègues, je tiens à remercier particulièrement ceux avec qui j'ai collaboré : Cécile, Marie, Laurent, Gaëlle et Nathalie.

Je suis particulièrement touché par mes amis qui viendront de loin pour ma soutenance, je pense en particulier à Julien, Émilie, Mélanie, Anaïs, Charlie et Claire. Je sais que vous ferez un long trajet pour venir me voir, ça me fait très plaisir de vous accueillir ; je vous suis infiniment reconnaissant. Je tiens à remercier tous les amis qui seront présents ce jour là ; je pourrai sans aucun doute compter sur Yannick, Jean-Yves, Willy, Yeny, Romuald, Clément. J'ai une pensée pour Valentine qui à 18h (heure de Tokyo) sera en direct pour écouter ma soutenance. Je te suis particulièrement reconnaissant pour cette délicate attention. Certains comme Nathanaël m'ont énormément soutenu mais ne pourront pas faire le déplacement, je ne doute qu'ils prendront des nouvelles dès que ma soutenance sera terminée. Je tiens à remercier toutes les personnes avec qui j'ai eu l'occasion de cohabiter en particulier Margot, Vivien, Paul, Nicolas, François, Maxime, Cécile. Ces trois dernières années, j'ai souvent usé les parquets des pistes de danse. Je tiens à remercier toutes les danseuses qui m'ont fait la joie de partager une danse avec moi, je ne cesse de louer leurs grâce et leurs sourire. Parmi ces superbes danseuses, je tiens à remercier Mélisande, Aurélie, Caroline et Corinne.

Enfin, je ne peux conclure ces remerciements sans montrer toute ma gratitude envers ma famille qui m'a toujours soutenu durant ces cinq années. Je tiens à remercier mon oncle Jean-Charles ainsi que mes parents et mes soeurs. Cela me fait extrêmement plaisir de vous avoir dans mon auditoire le jour de ma soutenance.

J'espère que la fin de cette thèse se ponctuera par une très bonne soutenance et s'achèvera par une très belle fête.

Introduction

This thesis work has an application in metabolomics which is the identification and quantification of metabolites (kind of molecules) using Nuclear Magnetic Resonance (NMR) spectra. One of the objectives of this work is to provide to metabolomics experts an automatic method able to identify and quantify metabolites in a complex mixture with an unknown composition. The method ASICS (Automatic Statistical Identification in Complex Spectra) has been developed to fulfill this objective (Tardivel et al., 2017a). ASICS is implemented on the Galaxy online infrastructure (Guitton et al., 2017). Before programming ASICS, we had to propose a mathematical modelling for this method. This modelling, explained in detail in the chapter 3, is based on a multiple testing procedure in a linear Gaussian model. A way to identify a metabolite is to test the null hypothesis "the concentration of the metabolite is zero". The rejection of the null hypothesis shows that the concentration of the metabolite is different from zero. There are as many tests to perform as there are metabolites to be analysed (i.e 176 for our application).

The spectra of pure metabolites are modelled by linearly independent fixed vectors $X_1, \ldots, X_p$ of $\mathbb{R}^n$ (with $n = 6000$ and $p = 176$ in the application). Let us denotes X the matrix whose columns are $X_1, \ldots, X_p$. The spectrum of the complex mixture is modelled by Y the response of the Gaussian linear model

$$Y = X\beta^* + \varepsilon,$$

where ε is distributed according to $\mathcal{N}(0, \Gamma)$ and $\beta^* := (\beta_1^*, \ldots, \beta_p^*)$ corresponds to the metabolite's concentrations. The active set of this model is defined as $\mathcal{A} := \{i \in [\![1, p]\!] \mid \beta_i^* \neq 0\}$; this set represents the metabolites to identify. Note that because $X_1, \ldots, X_p$ are linearly independent, the parameter β^* is identifiable meaning that $X\beta = X\beta^*$ implies $\beta = \beta^*$. A natural way to recover $\mathcal{A}$, is to test null hypotheses $\mathcal{H}_i : \beta_i^* = 0$ with $i \in [\![1, p]\!]$. Let $\hat{\beta}_i$ be any statistics to test the null hypothesis $\mathcal{H}_i$ so that $\mathcal{H}_i$ is rejected as soon as $\hat{\beta}_i \in R_i$, where $R_i \subset \mathbb{R}$ is a Borel set called rejection region. To determine $\mathcal{A}$, we aim to reject each element of $\mathcal{A}$ and to reject no element of $\mathcal{A}^c$ (the complementary in $[\![1, p]\!]$ of the set $\mathcal{A}$). Consequently, there are two type of errors :

— The false positives that are the elements of $\mathcal{A}^c$, wrongly rejected, namely $\{i \notin \mathcal{A} \mid \hat{\beta}_i \in R_i\}$.

— The false negatives that are the elements of $\mathcal{A}$, wrongly not rejected, namely $\{i \in \mathcal{A} \mid \hat{\beta}_i \notin R_i\}$.

In the context of the metabolomics, false positives are metabolites with a null concentration that are wrongly identified and false negatives are the unidentified metabolites whereas they have a non-zero concentration. In priority, the metabolomics experts wish a method that does not provide false positives. Moreover, this method should be able to identify the metabolites having a low concentration. The first request led us to control the FamilyWise Error Rate (FWER) defined by

$$\text{FWER} := \mathbb{P}(\exists i \notin \mathcal{A} \text{ such that } T_i \in R_i)$$

which is the probability for a multiple testing procedure to give at least one false positive. The second request motivated the use of the average proportion of good rejection, called the average power, defined by

$$\frac{1}{\text{card}(\mathcal{A})} \sum_{i \in \mathcal{A}} \mathbb{E}\left(\mathbb{1}_{T_i \in R_i}\right) = \frac{1}{\text{card}(\mathcal{A})} \sum_{i \in \mathcal{A}} \mathbb{P}\left(T_i \in R_i\right).$$

Intuitively, when the FWER is low, the average power is low too, consequently it is a challenging issue to control the FWER and to keep an average power as large as possible. A multiple testing procedure that controls the level of the FWER and for which the average power is larger than those of the state-of-the-art procedures (Holm, 1979; Janson and Su, 2016; Romano and Wolf, 2005; Westfall and Young, 1993) is given in the submitted article of Tardivel et al. (2017b) reported in Chapter 2.

In the previous modelling, we are in the classical setting of the linear model in which $p \leq n$ and $X_1, \ldots, X_p$ are linearly independent. In the high-dimensional linear model, the number of explicative variables p is larger than the number of observations n. Contrarily to the classical setting, when $p > n$ the parameter β^* is no longer identifiable, $X^T\Gamma^{-1}X$ is no longer invertible and the maximum likelihood estimator $\left((X^T\Gamma^{-1}X)^{-1}X^T\Gamma^{-1}Y\right)$ is thus not available. Because in the high-dimensional linear model the maximum likelihood estimator is not available and the parameter β^* is not identifiable, the study of this model is a very challenging issue.

The lasso $\hat{\beta}(\lambda)$ is an alternative to the maximum likelihood estimator (Tibshirani, 1996). This estimator minimizes the following expression

$$\|Y - X\beta\|^2 + \lambda\|\beta\|_1.$$

The general position condition is a sufficient condition on X for the uniqueness of the lasso's estimator. When X is a random matrix with a continuous distribution onto the set of the $n \times p$ matrices, the general position holds almost surely (Tibshirani, 2013). Consequently, in practice, this condition always holds even if $p \gg n$. Whatever $\lambda > 0$, $\hat{\beta}(\lambda)$ has a lot of null components (at least $p - n$ null components) because $\text{card}\{i \in [\![1, p]\!] \mid \hat{\beta}_i(\lambda) \neq 0\} \leq n$ (Tibshirani, 2013). Consequently, the active estimator $\hat{\mathcal{A}}(\lambda) := \{i \in [\![1, p]\!] \mid \hat{\beta}_i(\lambda) \neq 0\}$ is tailored to recover an active set $\mathcal{A}$ when $\text{card}(\mathcal{A})$ is small. The irrepresentable condition on β^* (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Zou, 2006) is a necessary and almost

sufficient condition to obtain a consistent estimator $\hat{\mathcal{A}}(\lambda)$ of $\mathcal{A}$. This condition is often assumed in applied and theoretical works (Bach, 2008; Perrot-Dockès et al., 2017; Ollier and Viallon, 2017), whereas a weaker condition than the irrepresentable condition is assumed in Bickel et al. (2009) and Lounici (2008). Because β^* is unknown and not identifiable, these assumptions cannot be checked.

There is no uncheckable assumption in the recent work of Meinshausen (2015). In his work, Y is an observed Gaussian vector distributed according to $\mathcal{N}(m, \sigma^2 Id_n)$ with m and σ unknown and X is a fixed $n \times p$ matrix with $p > n$. Instead to assume $m = X\beta^*$ and try to estimate the parameter β^*, Meinshausen proposes to estimate a l^1 sparse representation of m defined by

$$\underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{p} |\beta_i| \text{ under the constraint that } X\beta = m. \tag{1}$$

Without any other conditions, the set of solutions of (1) can be empty or can have several elements. However, as implicitly assumed by Meinshausen, as soon as the columns of X span $\mathbb{R}^n$, this equation admits at least a solution and a unique solution when the general position condition holds for X.

The convex problem (1) is usually used to recover the sparsest representation of m in X that is the solution of the following intractable problem (2)

$$\underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \operatorname{card}\{i \in [\![1, p]\!] \mid \beta_i \neq 0\} \text{ under the constraint that } X\beta = m. \tag{2}$$

As explained by Meinshausen, under some conditions (null space property(Donoho and Elad, 2003; Gribonval and Nielsen, 2003), restricted isometry property (Candes, 2008)...) these problems (1) and (2) have the same solution. In the noiseless case when $\sigma = 0$ implying thus $Y = m$, the problem (1) is convex consequently, its solutions could be efficiently obtained. Contrarily to the problem (1), the problem (2) is not convex. Even in the noiseless case, solving (2) is a challenging issue.

Developing efficient methods to solve the problem (2) when σ is null is a preliminary work to estimate the solution of the problem (2) when σ is not null. In a more general context, this problem led us to determine the sparsest representation of a vector $y \in \mathbb{R}^n$ in a family $d_1, \ldots, d_p$ spanning $\mathbb{R}^n$ (thus $p \geq n$). This theoretical study has led to the article reported in Chapter 6 (Tardivel et al., 2017c).

This manuscript has three parts, each part is related to a submitted/accepted article. The organization of this manuscript is the following.

Organization of the manuscript

The first part deals with the multiple testing procedures that control the FWER. This part has two chapters :

— There are two different way to recover an active set : provide an active set estimator or provide a multiple testing procedure for the null hypotheses $i \notin \mathcal{A}$ with $i \in [\![1, p]\!]$. In the chapter 1 we give the bibliography of the multiple testing procedures that control the FWER and of the bibliography of the active set estimator building from the lasso.

— The chapter 2 is the submitted article "A powerful multiple testing procedure in linear Gaussian model" (Tardivel et al., 2017b).

In the second part the multiple testing procedure developed in the first part is applied to metabolomics. This part has two chapters :

— In the chapter 3 we tackle the identification and quantification of metabolites using NMR spectra. In the modelling, we see the necessity to have a powerful procedure that controls the FWER.

— The chapter 4 is the article accepted in *Metabolomics* "ASICS : an automatic method for identification and quantification of metabolites in complex 1D ^{1}H NMR spectra" (Tardivel et al., 2017a).

The third part deals with the high-dimensional linear Gaussian model. This part has two chapters :

— In the chapter 5 we explain why there is a challenging issue for the high-dimensional linear model to recover the sparsest representation of a vector in a generating family. The end of this chapter is the bibliography of conditions for which a representation is the l^1 sparse representation or the sparsest representation.

— The chapter 6 is the submitted article "Sparsest representations and approximations of a high-dimensional linear system" (Tardivel et al., 2017c).

Bibliographie

Bach, F. R. (2008). Bolasso : model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, pages 33–40. ACM.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732.

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

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

Gribonval, R. and Nielsen, M. (2003). Sparse representations in unions of bases. *IEEE Transactions on Information Theory*, 49(12) :3320–3325.

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

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2) :65–70.

Janson, L. and Su, W. (2016). Familywise error rate control via knockoffs. *Electronic Journal of Statistics*, 10(1) :960–975.

Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of statistics*, 2 :90–102.

Meinshausen, N. (2015). 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.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3) :1436–1462.

Ollier, E. and Viallon, V. (2017). Regression modelling on stratified data with the lasso. *Biometrika*, 104(1) :83–96.

Perrot-Dockès, M., Lévy-Leduc, C., Sansonnet, L., and Chiquet, J. (2017). Variable selection in multivariate linear models with high-dimensional covariance matrix estimation. *arXiv preprint arXiv :1707.04145*.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469) :94–108.

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

Tardivel, P., Servien, R., and Concordet, D. (2017b). A powerful multiple testing procedure in linear Gaussian model. *Submitted*.

Tardivel, P., Servien, R., and Concordet, D. (2017c). Sparsest representations and approximations of a high-dimensional linear system. *Submitted*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1) :267–288.

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7 :1456–1490.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing : Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7 :2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476) :1418–1429.

Première partie

Une nouvelle procédure de tests multiples contrôlant le FWER

Chapitre 1

Identification des éléments de l'active set d'un modèle linéaire gaussien

On considère le modèle linéaire gaussien

$$Y = X\beta^* + \varepsilon, \tag{1.1}$$

avec X une matrice de dimension $n \times p$ et $\varepsilon \sim \mathcal{N}(0, \sigma^2 Id_n)$. L'active set défini par $\mathcal{A} := \{i \in [\![1, p]\!] \mid \beta_i^* \neq 0\}$; c'est l'ensemble que l'on souhaite identifier. Une façon naturelle de déterminer $\mathcal{A}$ est de tester pour chaque entier $i \in [\![1, p]\!]$ l'hypothèse nulle $\beta_i^* = 0$. Dans ce chapitre, nous nous intéresserons aux procédures de tests multiples qui contrôlent le FamilyWise Error Rate (FWER).

Une seconde façon de déterminer $\mathcal{A}$ est d'estimer l'active set. Lorsque $\hat{\beta}$ est un estimateur parcimonieux (c'est-à-dire un estimateur pour lequel chaque composante peut être nulle avec une probabilité non nulle), on peut estimer l'active set en posant $\mathcal{A}(\hat{\beta}) := \{i \in [\![1, p]\!] \mid \hat{\beta}_i \neq 0\}$. L'estimateur lasso (Tibshirani, 1996) est probablement l'estimateur parcimonieux le plus connu. Dans ce chapitre, nous nous intéresserons aussi à l'estimateur de l'active set construit à partir du lasso.

1.1 Procédures de tests multiples contrôlant le FWER

Dans cette partie nous supposerons que $\ker(X) = 0$. Sous cette hypothèse, le paramètre β^* du modèle est identifiable au sens où $X\beta = X\beta^*$ implique que $\beta = \beta^*$. Par ailleurs cette hypothèse garantit que la matrice de Gram $X^T X$ est inversible et donc que l'estimateur du maximum de vraisemblance $\hat{\beta}^{\text{mle}} := (X^T X)^{-1} X^T Y$ est bien défini. Pour $i \in [\![1, p]\!]$, on teste l'hypothèse nulle $\mathcal{H}_i : \beta_i^* = 0$ avec la statistique de test $T_i := \hat{\beta}_i^{\text{mle}} / \text{se}(\hat{\beta}_i^{\text{mle}})$. L'hypothèse nulle $\mathcal{H}_i : \beta_i^* = 0$ est rejetée dès que $|T_i| \geq s_\alpha$. Le FWER de cette procédure, défini ci-dessous,

dépend de s_α.

$$\text{FWER} := \mathbb{P}\left(\exists i \notin \mathcal{A} \text{ tel que } |T_i| \geq s_\alpha\right) = \mathbb{P}\left(\bigcup_{i \notin \mathcal{A}} |T_i| \geq s_\alpha\right).$$

Soit $\alpha \in [0,1]$, le seuil s_α doit être choisi pour que FWER $\leq \alpha$. La puissance moyenne, définie ci-dessous, est la proportion moyenne de bons rejets.

$$\frac{1}{\text{card}(\mathcal{A})} \sum_{i \in \mathcal{A}} \mathbb{E}\left(\mathbb{1}_{|T_i| \geq s_\alpha}\right) = \frac{1}{\text{card}(\mathcal{A})} \sum_{i \in \mathcal{A}} \mathbb{P}\left(|T_i| \geq s_\alpha\right).$$

On remarque que plus le seuil s_α est petit plus la puissance moyenne est grande. Ainsi, il est naturel de chercher le plus petit seuil s_α permettant de contrôler le FWER au niveau α.

1.1.1 Procédures de tests multiples "single step"

Les procédures "single step" que nous allons décrire prescrivent un seuil s_α pour contrôler le FWER au niveau α. Dans le cadre du modèle linéaire gaussien, pour chaque $i \notin \mathcal{A}$ la statistique T_i suit une loi $\mathcal{N}(0,1)$. On note ϕ la fonction de répartition d'une loi normale $\mathcal{N}(0,1)$. La procédure de Bonferroni (Bonferroni, 1936; Dunn, 1961) repose sur l'inégalité du même nom suivante qui permet de majorer la probabilité $\mathbb{P}\left(\bigcup_{i \notin \mathcal{A}} |T_i| \geq s_\alpha\right)$.

$$\mathbb{P}\left(\bigcup_{i \notin \mathcal{A}} |T_i| \geq s_\alpha\right) \leq \sum_{i \notin \mathcal{A}} \mathbb{P}\left(|T_i| \geq s_\alpha\right).$$

En posant $s_\alpha = \phi^{-1}(1 - \alpha/2p)$ on a $\mathbb{P}(|T_i| \geq s_\alpha) = \alpha/p$, ce qui implique que

$$\text{FWER} = \mathbb{P}\left(\bigcup_{i \notin \mathcal{A}} |T_i| \geq s_\alpha\right) \leq \alpha.$$

Dans le modèle nul *i.e.* lorsque $\mathcal{A} = \emptyset$ et lorsque les statistiques $T_1, \ldots, T_p$ sont indépendantes on a

$$\mathbb{P}\left(\forall i \in [\![1,p]\!], |T_i| \leq 1 - \alpha/2p\right) = (1 - \alpha/p)^p = 1 - \alpha + \circ(\alpha).$$

Ainsi, lorsque α est petit on a FWER $\approx \alpha$. La procédure de Dunn-Šidák (Šidàk, 1967) raffine le contrôle du FWER. Cette procédure préconise de prendre $s_\alpha := \phi^{-1}(\sqrt[p]{1 - \alpha/2})$. Lorsque $\mathcal{A} = \emptyset$, cette procédure donne un contrôle exact du FWER.

Lorsque la matrice de Gram $X^T X$ n'est pas diagonale, les statistiques $T_1, \ldots, T_p$ ne sont pas indépendantes. Dans ce cas on choisit s_α comme le $1 - \alpha$ quantile de $\max\{|Z_1/\text{se}(Z_1)|, \ldots, |Z_p/\text{se}(Z_p)|\}$, avec $(Z_1, \ldots, Z_p) \sim \mathcal{N}\left(0, \sigma^2 (X^T X)^{-1}\right)$ (Lehmann et Romano, 2005). Dans le modèle nul complet, ce seuil nous donne un contrôle exact du FWER au niveau α. Cette procédure généralise celle de Dunn-Šidák lorsque les statistiques $T_1, \ldots, T_p$

ne sont pas indépendantes. La puissance moyenne de la dernière procédure est supérieure aux puissances moyennes des autres procédures "single step" qui contrôlent le FWER.

Pour les trois procédures décrites, le choix du seuil est fait en se plaçant dans le modèle nul complet (*i.e.* $\mathcal{A} = \emptyset$) qui est le pire des cas. En effet, pour chacune des trois procédures, lorsque $\mathcal{A} \neq \emptyset$ le seuil s_α prescrit est trop grand car le FWER est inférieur à α. Supposons qu'une partie $S \subset \mathcal{A}$ telle que $\text{card}(S) = k$ soit connue alors, le seuil $s_\alpha(S)$ fourni pour chacune des trois procédures décrites précédemment serait respectivement égal à :

— $s_\alpha(S) := \phi^{-1}\left(1 - \frac{\alpha}{2(p-k)}\right)$,

— $s_\alpha(S) := \phi^{-1}\left(\sqrt[p-k]{1-\alpha/2}\right)$,

— $s_\alpha(S)$ le $1-\alpha$ quantile de $\max\{|Z_i/\text{se}(Z_i)|\}_{i\notin S}$.

Lorsque $S = \emptyset$, on retrouve les seuils $s_\alpha(\emptyset)$ donnés au début du paragraphe. Lorsque $S \neq \emptyset$ pour chacun des cas, le seuil $s_\alpha(S)$ est plus petit que le seuil $s_\alpha(\emptyset)$. Ainsi, la connaissance d'une partie S de l'active set permet de choisir un seuil $s_\alpha(S)$ plus petit ce qui implique un gain de puissance moyenne. Le raffinement stepdown est une méthode générique qui permet d'améliorer la puissance moyenne d'une procédure single step (Romano et Wolf, 2005 ; Westfall et Young, 1993 ; Lehmann et Romano 2005 page 352). Ce gain de puissance moyenne est obtenu en estimant l'active set ce qui permet de choisir un seuil plus petit que le seuil proposé par la méthode single step.

1.1.2 Le raffinement stepdown

On note $s_\alpha(\emptyset)$ un seuil fournit par une procédure single step qui permet un contrôle du FWER au niveau $\alpha \in [0,1]$, c'est-à-dire

$$\text{FWER} = \mathbb{P}\left(\bigcup_{i\notin\mathcal{A}} |T_i| \geq s_\alpha(\emptyset)\right) \leq \alpha.$$

Lorsque S est une partie de l'active set $\mathcal{A}$, on souhaite ajuster le seuil $s_\alpha(\emptyset)$ par $s_\alpha(S)$. La fonction s_α, définie sur $\mathcal{P}([\![1,p]\!])$, permet d'effectuer cet ajustement. L'algorithme décrit dans la figure 1.1 donne les instructions de la méthode raffinement stepdown.

Le théorème 1.1 donne des conditions sur la fonction $s_\alpha : \mathcal{P}([\![1,p]\!]) \to \mathbb{R}$ pour que le seuil s fourni par l'algorithme permette de contrôler le FWER au niveau α (Romano and Wolf, 2005).

Théorème 1.1 *Supposons que la fonction s_α définie sur $\mathcal{P}([\![1,p]\!])$ soit telle que*

1. $\mathbb{P}(\exists i \notin \mathcal{A}$ *tel que* $|T_i| \geq s_\alpha(\mathcal{A})) \leq \alpha$.

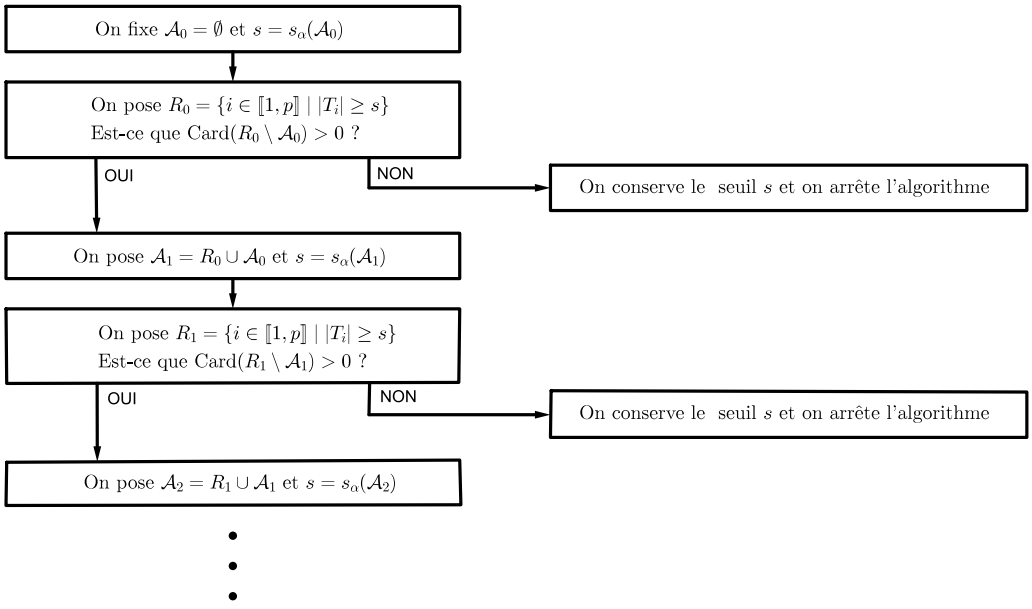


FIGURE 1.1 – Cette figure donne les instructions de l'algorithme de raffinement stepdown. Comme les ensembles $\mathcal{A}_0, \mathcal{A}_1, \ldots$ de $[\![1,p]\!]$ forment une suite strictement croissante pour l'inclusion, cet algorithme s'arrête avant la p^{eme} itération.

2. Si $B \subset C$ *alors* $s_\alpha(B) \geq s_\alpha(C)$

alors, le seuil s donné par l'algorithme décrit par la figure (1.1) contrôle le FWER à un niveau α, c'est à dire

$$\mathbb{P}(\exists i \notin \mathcal{A} \text{ tel que } |T_i| \geq s) \leq \alpha.$$

La première condition signifie que lorsque l'active set $\mathcal{A}$ est connu, en prenant $s_\alpha(\mathcal{A})$ comme seuil, la procédure de tests multiples contrôle le FWER au niveau α. La seconde condition signifie que plus on connaît d'éléments dans l'active set plus le seuil s contrôlant la FWER est petit. Le seuil s fourni par le raffinement stepdown est plus petit que le seuil $s_\alpha(\emptyset)$. Ainsi, le raffinement stepdown permet un gain pour la puissance moyenne.

La méthode du raffinement stepdown généralise la procédure de rejet séquentiel proposée par Holm (1979). On retrouve la procédure de rejet séquentiel en appliquant la méthode du raffinement stepdown à la procédure de Bonferroni.

1.1.3 Procédures utilisant un estimateur lasso

Soit $\lambda > 0$, l'estimateur lasso $\hat{\beta}(\lambda)$ est solution de

$$\hat{\beta}(\lambda) := \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|Y - X\beta\|^2 + \lambda \|\beta\|_1. \tag{1.2}$$

En toute rigueur, pour que $\hat{\beta}(\lambda)$ soit bien défini, il faut que le minimum de l'expression $\|Y - X\beta\|^2 + \lambda\|\beta\|_1$ soit atteint en un unique point. Avoir une matrice X en position générale est

une condition suffisante pour que le minimiseur de $\|Y - X\beta\|^2 + \lambda\|\beta\|_1$ soit unique. C'est une condition très faible, en effet si X est une matrice aléatoire de loi continue sur les matrices $n \times p$ alors, presque sûrement, X est en position générale. Par conséquent, en pratique, l'estimateur lasso est toujours bien défini. Les procédures knockoffs (Barber et Candès, 2015 ; Janson et Su, 2016) et les procédures construites à partir des noeuds du lasso (Lockhart et al., 2014; G'Sell et al., 2015) sont des procédures de tests multiples récentes qui sont construites à partir de l'estimateur lasso. Nous allons développer les procédures construites à partir des noeuds de l'estimateur lasso ; la procédure knockoffs (Barber et Candès, 2015 ; Janson et Su, 2016) contrôlant le FWER est expliquée en détail dans l'article (Tardivel et al., 2017b).

La fonction aléatoire $\lambda > 0 \mapsto \hat{\beta}(\lambda)$ a la propriété d'être affine par morceaux. Cette propriété permet de définir les noeuds du lasso $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots$ de telle sorte que la restriction de la fonction $\lambda > 0 \mapsto \hat{\beta}(\lambda)$ à l'intervalle $[\hat{\lambda}_{i+1}, \hat{\lambda}_i]$ soit affine. Lorsque la matrice de planification X est orthogonale cette fonction a pour expression

$$\lambda > 0 \mapsto \left(\text{sign}(\hat{\beta}_1^{\text{mle}})(|\hat{\beta}_i^{\text{mle}}| - \lambda)_+, \dots, \text{sign}(\hat{\beta}_p^{\text{mle}})(|\hat{\beta}_p^{\text{mle}}| - \lambda)_+\right).$$

On considère une permutation $\hat{\rho}$ de $[\![1, p]\!]$ (aléatoire) telle que $|\hat{\beta}^{\text{mle}}_{\hat{\rho}(1)}| \geq \dots \geq |\hat{\beta}^{\text{mle}}_{\hat{\rho}(p)}|$, le noeud $\hat{\lambda}_i$ est égal à $|\hat{\beta}^{\text{mle}}_{\hat{\rho}(i)}|$. Le théorème suivant est un résultat élégant proposé par Lockhart et al. (2014). Pour la suite de cette partie, on note $\text{Exp}(i)$ la loi exponentielle de paramètre i.

Théorème 1.2 *Soit $Y = X\beta^* + \varepsilon$, avec X une matrice orthogonale de dimension $n \times p$ et ε de loi $\mathcal{N}(0, \sigma^2 Id_n)$ où σ connu. Soient k un entier fixé, $p > k$ et $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_p$ les noeuds du lasso. Pour tout $j \in [\![1, k]\!]$, on pose*

$$T_j := \frac{\hat{\lambda}_i \left(\hat{\lambda}_i - \hat{\lambda}_{i+1}\right)}{\sigma^2}.$$

Lorsque pour tout $i \in [\![1, p]\!]$, $\beta_i^* = 0$ et lorsque p tend vers $+\infty$ on a la convergence suivante

$$(T_1, T_2, \dots, T_k) \xrightarrow[p \to +\infty]{\mathcal{L}} (Z_1, Z_2, \dots, Z_k),$$

avec $Z_1, Z_2, \dots, Z_k$ des variables aléatoires indépendantes telles que Z_i suit une loi $\text{Exp}(i)$.

D'après ce théorème, les variables aléatoires $T_1, \dots, T_k$ peuvent être utilisées pour tester asymptotiquement la nullité du paramètre β^*. Toujours dans le cadre où X est orthogonale, des procédures de tests multiples utilisant le théorème précédent ont été développées par G'Sell et al. (2015).

Lorsque $|\beta_i^*|$ est grand, la statistique $|\hat{\beta}_i^{\text{mle}}|$ devrait être grande. Ainsi, intuitivement les noeuds du lasso les plus grands devraient être associés à des éléments de l'active set $\mathcal{A}$. Rappelons que $\hat{\rho}$ est une permutation (aléatoire) de $[\![1, p]\!]$ telle que $|\hat{\beta}^{\text{mle}}_{\hat{\rho}(1)}| \geq \dots \geq |\hat{\beta}^{\text{mle}}_{\hat{\rho}(p)}|$ ainsi, $\hat{\rho}(1)$ correspond au noeud du lasso le plus grand, $\hat{\rho}(2)$ au deuxième plus grand noeud du lasso...

Une procédure utilisant les statistiques $T_1, \ldots, T_k$ pour tester les hypothèses nulles emboîtées suivantes $\mathcal{H}^i : \mathcal{A} \subset \{\hat{\rho}(1), \hat{\rho}(2), \ldots, \hat{\rho}(i)\}$ et $\mathcal{H}^0 : \mathcal{A} = \emptyset$ a été proposée par G'Sell et al. (2015). Cette procédure repose sur le résultat admis mais non démontré suivant :

Conditionnellement à l'événement $\mathcal{A} \subset \{\hat{\rho}(1), \hat{\rho}(2), \ldots, \hat{\rho}(s)\}$ et pour tout $k < s$,

$\mathcal{A} \not\subset \{\hat{\rho}(1), \hat{\rho}(2), \ldots, \hat{\rho}(k)\}$ (*i.e.* $\mathcal{H}^s$ est vraie et pour tout $k < s$, $\mathcal{H}^k$ est faux) alors lorsque p converge vers $+\infty$, les lois des statistiques de test $T_{s+1}, \ldots, T_{s+l}$ convergent respectivement vers des lois Exp(1), ..., Exp(l) et deviennent indépendantes.

L'exemple suivant semble pourtant invalider ce résultat. On considère un paramètre $\beta^* \in \mathbb{R}^p$ tel que pour tout $i \in \mathcal{A}$ on ait $\beta_i^* = \epsilon$. Si ϵ est très petit la loi du vecteur $(T_1, \ldots, T_s, T_{s+1}, \ldots, T_{s+l})$ est presque égale à la loi que l'on aurait si $\beta^* = 0$. Intuitivement, le théorème 1.2 suggère que lorsque p est grand et ϵ petit, le vecteur aléatoire $(T_{s+1}, \ldots, T_{s+l})$ a approximativement des composantes indépendantes de lois respectives $\text{Exp}(s+1), \ldots, \text{Exp}(s+l)$. Les simulations de la figure 1.2 ont été faites en prenant p grand et β_i^* petit quand $i \in \mathcal{A}$. Ces simulations montrent que conditionnellement à l'événement $\mathcal{A} \subset \{\hat{\rho}(1), \hat{\rho}(2), \ldots, \hat{\rho}(s)\}$ et pour tout $k < s$, $\mathcal{A} \not\subset \{\hat{\rho}(1), \hat{\rho}(2), \ldots, \hat{\rho}(k)\}$, les lois marginales des variables aléatoires $T_{s+1}, \ldots, T_{s+l}$ sont approximativement égales à $\text{Exp}(s+1), \ldots, \text{Exp}(s+l)$. Ainsi lorsque p est grand et ϵ petit, même en conditionnant, les lois marginales des variables aléatoires $T_{s+1}, \ldots, T_{s+l}$ sont approximativement les lois que l'on aurait obtenues sans conditionner et avec $\beta^* = 0$.

Les simulations de la figure 1.2 contredisent le résultat admis de G'Sell et al. (2015). En effet, si ce résultat était vrai, conditionnellement à l'événement $\{\mathcal{A} \subset \hat{\rho}(1)\}$ et lorsque p tend vers $+\infty$, les lois marginales des statistiques T_2 et T_3 devraient être respectivement égales à Exp(1) et Exp(2). Cependant, sur nos simulations, les lois marginales conditionnelles de T_2 et T_3 sont respectivement proches des lois Exp(2) et Exp(3).

Nous avons d'autres remarques sur la procédure de tests multiples fournie par G'Sell et al. (2015).

— L'hypothèse $\mathcal{H}^k : \mathcal{A} \subset \{\hat{\rho}(1), \ldots \hat{\rho}(p)\}$ dépend de la permutation aléatoire $\hat{\rho}$ qui est construite à partir du maximum de vraisemblance $\hat{\beta}^{\text{mle}}$. Par ailleurs, les statistiques de tests $T_1, \ldots, T_k$ sont construites à partir des noeuds du lasso et donc de $\hat{\beta}^{\text{mle}}$. Comme l'indique Bühlmann et al. (2014), tester ces hypothèses ne peut être fait que conditionnellement à la variable aléatoire $\hat{\beta}^{\text{mle}}$. Il semble donc surprenant que l'estimateur $\hat{\beta}^{\text{mle}}$ soit utilisé simultanément pour formuler les hypothèses $\mathcal{H}^0, \ldots, \mathcal{H}^p$ et pour tester ces hypothèses.

— Le rejet de l'hypothèse $\mathcal{H}^k$ indique que $\mathcal{A} \not\subset \{\hat{\rho}(1), \ldots, \hat{\rho}(s)\}$, ce qui ne fournit pas un ensemble contenant l'active set $\mathcal{A}$.

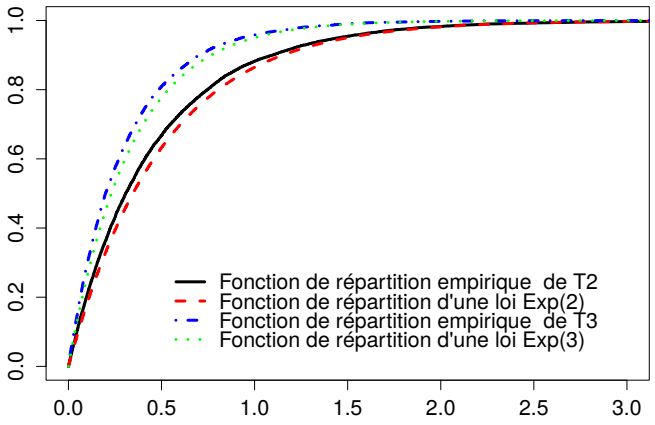


FIGURE 1.2 – Pour ce graphique on a choisi $\beta_1^* = 0.01$ et $\beta_2^* = \cdots = \beta_{100}^* = 0$. Les quatre courbes représentent les fonctions de répartition empirique des variables aléatoire T_2 et T_3 conditionnellement à l'événement $\{\hat{\rho}(1) = 1\}$ (*i.e.* $\mathcal{H}^1$ est vraie) ainsi que les fonctions de répartition des lois Exp(2) et Exp(3). Les fonctions de répartition empirique ont été obtenues avec 1000000 simulations. Ces graphiques semblent illustrer que, conditionnellement à l'événement $\{\hat{\rho}(1) = 1\}$, les lois marginales des variables aléatoires T_2 et T_3 sont respectivement Exp(2) et Exp(3). Ces simulations sont en contradiction avec le résultat admis dans l'article de G'Sell et al. (2015). En effet, conditionnellement à l'événement $\{\hat{\rho}(1) = 1\}$, les statistiques T_2 et T_3 devraient avoir pour lois respectives Exp(1) et Exp(2).

1.2 Estimateur de l'active set

Dans leurs travaux, Zhao et Yu (2006) et Zou (2006) s'intéressent aux propriétés asymptotiques de l'estimateur $\mathcal{A}(\hat{\beta}(\lambda))$ lorsque p est fixé et que n tend vers $+\infty$. Afin de marquer la dépendance en n, on note X_n la matrice de planification du modèle 1.1. Les théorèmes 1.3 et 1.4 sont obtenus sous la condition que $\lim_{n\to+\infty} X_n^T X_n = C$ avec C une matrice symétrique définie positive de dimension $p \times p$. Le théorème 1.3, démontré dans les articles (Zhoa et Yu, 2006 ; Zou, 2006) donne une condition nécessaire pour pouvoir construire un estimateur convergent de l'active set. On note respectivement C_{11} et C_{21} les matrices $(C_{ij})_{i\in\mathcal{A},j\in\mathcal{A}}$ et $(C_{ij})_{i\notin\mathcal{A},j\in\mathcal{A}}$; on note $\text{sign}(\beta_{\mathcal{A}}^*)$ le vecteur $\text{sign}(\beta_{\mathcal{A}}^*) := (\text{sign}(\beta_i^*))_{i\in\mathcal{A}}$ avec $\text{sign}(x) = 1$ si $x > 0$, $\text{sign}(x) = -1$ si $x < 0$ et $\text{sign}(x) = 0$ si $x = 0$.

Théorème 1.3 (Condition nécessaire) *Supposons qu'il existe une suite* $(\lambda_n)_{n\in\mathbb{N}^*}$ *telle que* $\lim_{n\to+\infty} \mathbb{P}\left(\mathcal{A}(\hat{\beta}(\lambda_n)) = \mathcal{A}\right) = 1$ *alors l'inégalité suivante a lieu*

$$\|C_{21}C_{11}^{-1}\text{sign}(\beta_{\mathcal{A}}^*)\|_\infty \leq 1.$$

L'inégalité $\|C_{21}C_{11}^{-1}\text{sign}(\beta_{\mathcal{A}}^*)\|_\infty \leq 1$ est appelée condition d'irreprésentabilité. Le théorème 1.4, démontré dans l'article (Zhao et Yu, 2006), donne une condition suffisante pour avoir un

estimateur convergent de l'active set.

Théorème 1.4 (Condition suffisante) *Soit $(\lambda_n)_{n\in\mathbb{N}^*}$ une suite telle que $\lambda_n = an^\gamma$ avec $a > 0$ et $\gamma \in (1/2, 1)$.*

$$\text{Si } \|C_{21}C_{11}^{-1}\text{sign}(\beta^*_{\mathcal{A}})\|_\infty < 1 \text{ alors } \lim_{n\to+\infty} \mathbb{P}\left(\mathcal{A}(\hat{\beta}^n(\lambda_n) = \mathcal{A}\right) = 1.$$

Ces deux théorèmes montrent que la condition d'irreprésentabilité est nécessaire et "presque" suffisante pour construire un estimateur convergent de l'active set à partir de l'estimateur lasso. Bien qu'intéressants, ces résultats ont quelques défauts.

— La condition d'irreprésentabilité est invérifiable puisqu'elle dépend du paramètre β^* inconnu.

— Parce que le choix des constantes a et γ n'est pas discuté, il est difficile de préconiser un paramètre de régularisation. Voir l'introduction de l'article (Tardivel et al., 2017b) pour plus de détails.

— Pour l'étude asymptotique, le nombre p de variables explicatives est fixé et le nombre n des observations tend vers $+\infty$. Ainsi ces résultats sont valides lorsque $n \gg p$ mais n'ont aucun intérêt pour le modèle linéaire en grande dimension lorsque $p > n$.

L'exemple suivant illustre que lorsque $p \leq n$, il est plus judicieux d'estimer l'active set à partir du maximum de vraisemblance qu'à partir du lasso.

Soient $\beta^* = (1, 0)$ et X_n une matrice de dimension $n \times 2$ telle que

$$\lim_{n\to+\infty} \frac{1}{n} X_n^T X_n = C, \text{avec } C = \begin{pmatrix} 1 & 2 \\ 2 & 5 \end{pmatrix}, (\text{C est définie positive}).$$

La condition d'irreprésentabilité n'est pas satisfaite en effet, $C_{21}C_{11}^{-1}\text{sign}(\beta^*_{\mathcal{A}}) = 2 \times 1 \times 1 = 2$. D'après le théorème 1.3, il n'existe pas de suite $(\lambda_n)_{n\in\mathbb{N}^*}$ telle que $\lim_{n\to+\infty} \mathbb{P}(\mathcal{A}(\hat{\beta}(\lambda_n)) = \mathcal{A}) = 1$. Il n'est donc pas possible de construire un estimateur de $\mathcal{A}$ convergent en utilisant l'estimateur lasso. Pourtant, en posant $\hat{\mathcal{A}}_n = \{i \in \{1, 2\} \mid |\hat{\beta}_i^{\text{mle}}| > t_n/\sqrt{n}\}$, on obtient aisément un estimateur convergent de l'active set à partir du maximum de vraisemblance. En effet,

$$\text{si } \lim_{n\to+\infty} t_n = +\infty \text{ et } \lim_{n\to+\infty} \frac{t_n}{\sqrt{n}} = 0 \text{ alors } \lim_{n\to+\infty} \mathbb{P}(\hat{\mathcal{A}}_n = \mathcal{A}) = 1.$$

Traditionnellement, les résultats asymptotiques sont obtenus en faisant tendre n vers $+\infty$. Cette façon d'obtenir des résultats asymptotiques doit être repensée pour le modèle linéaire en grande dimension. Une autre façon d'avoir des résultats asymptotiques est de faire tendre σ vers 0. Dans ce cadre, le cas non bruité (*i.e.* avec $\sigma = 0$) peut être vu comme un cas limite. Lorsque $\sigma = 0$, Bühlmann et van de Geer (2011) à la page 192 donnent le théorème suivant

Théorème 1.5 *Soient $\beta^* \in \mathbb{R}^p$, X une matrice $n \times p$. Soit $\beta(\lambda)$ (non aléatoire) défini comme suit*

$$\beta(\lambda) := \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \|X\beta^* - X\beta\|^2 + \lambda\|\beta\|_1.$$

On pose $C = X^T X$, on note respectivement C_{11} et C_{21} les matrices $(C_{ij})_{i\in\mathcal{A}, j\in\mathcal{A}}$ et $(C_{ij})_{i\notin\mathcal{A}, j\in\mathcal{A}}$. Si la matrice C_{11} est inversible et si la condition $\|C_{21}C_{11}^{-1}\operatorname{sign}(\beta^*_{\mathcal{A}})\|_\infty < 1$ est vérifiée alors $\lim_{\lambda\to 0} \mathcal{A}(\beta(\lambda)) = \mathcal{A}$.

Remarquons que $\operatorname{card}(\mathcal{A}) \leq n$ est une condition nécessaire pour que C_{11} soit inversible. Dans le théorème 1.5 la matrice X peut avoir un nombre de colonnes p plus grand que le nombre de lignes n. Ainsi, contrairement aux théorèmes 1.3 et 1.4, le théorème 1.5 est un résultat de grande dimension.

Zhao et Yu (2006) s'intéressent à l'estimation de l'active set lorsque n et p tendent vers $+\infty$ avec p qui dépend de n. Parce que $\mathcal{A} \subset [\![1, p]\!]$, l'active set varie avec n ; cette dépendance en n de l'active set rend les résultats asymptotiques plus difficiles à obtenir.

L'estimateur lasso est utilisé dans la construction de la procédure de tests multiples de Tardivel et al. (2017b). Cependant, cette procédure pourrait être construite à partir du maximum de vraisemblance sans faire aucune référence à l'estimateur lasso. L'avantage d'introduire le lasso est de mettre en avant le lemme 2.1 qui montre qu'à une transformation près, lorsque $\ker(X) = 0$, l'estimateur lasso est un seuillage doux du maximum de vraisemblance. Ce lemme ainsi que certains résultats de ce chapitre concernant le lasso nous amène à la conclusion suivante :
Lorsque $p \leq n$ et que $\ker(X) = 0$, l'estimateur lasso n'a aucun intérêt ; le maximum de vraisemblance est plus performant pour construire un estimateur de l'active set ou pour construire procédure de tests multiples.
Nous montrerons que la procédure de tests multiples développée dans l'article Tardivel et al. (2017b) a une puissance moyenne supérieure aux puissances moyennes des procédures de l'état de l'art (Holm, 1979 ; Romano et Wolf, 2005 ; Janson et Su, 2016).

Chapitre 2

A powerful multiple testing procedure in linear Gaussian model

Patrick J.C. Tardivel [1], Rémi Servien and Didier Concordet
TOXALIM, Université de Toulouse, INRA, ENVT, Toulouse, France.

Summary : We study the control of the FamilyWise Error Rate (FWER) in the linear Gaussian model when the $n \times p$ design matrix is of rank p. A procedure based on a lasso-type estimator is optimized with respect to the volume of the multidimensional acceptance region. An important result of this article states that, even if the design is not orthogonal, even if residuals are not i.i.d, this optimization leads to a soft thresholded maximum likelihood estimator. Consequently, when the design matrix is of rank p, we build directly a powerful multiple testing procedure based on the maximum likelihood estimator instead to optimizing a lasso-type procedure. However, the lasso procedure optimization allows us to understand how to build a powerful multiple testing procedure based on the maximum likelihood estimator. Numerical experiments highlight the performance of our approach compared to the state-of-the-art procedures. An application to the detection of metabolites in metabolomics is provided.

Keywords : Familywise error rate, Multiple testing, Lasso, Maximum likelihood estimator, Metabolomics.

2.1 Introduction

Let us consider the linear Gaussian model

$$Y = X\beta^* + \varepsilon, \tag{2.1}$$

1. corresponding author : patrick.tardivel@inra.fr

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. In this article, we focus on the Familywise Error Rate (FWER) defined as the probability to reject wrongly at least one hypothesis $\mathcal{H}_i$.

The lasso estimator (Tibshirani, 1996), defined by

$$\hat{\beta}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2} \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \right\} \tag{2.2}$$

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 $\operatorname{Vect}\{X_i,\ i \in \mathcal{A}\}$. When the design matrix is close to an orthogonal matrix (which implies the irrepresentable condition), an explicit λ has been provided in the SLOPE multiple testing procedure (Bogdan et al., 2015; Su and Candes, 2016) or to estimate the active set (Lounici, 2008). However, such a results are not available for a general matrix X of rank p.

The lasso knots were first introduced by Lockhart et al. (2014) for the covariance test. The knots $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots$ correspond to values of $\hat{\lambda}$ at which the estimated active set $\mathcal{A}(\hat{\beta}(\hat{\lambda}))$ changes. In the same setting as ours ($\operatorname{rank}(X) = p$), recent multiple testing procedures developed by Barber and Candès (2015); Janson and Su (2016) use lasso knots. Both procedures compare knots of the original lasso ($\hat{\lambda}_i$) to the knockoff lasso knots ($\tilde{\lambda}_i$). One can view knots of the knockoff lasso ($\tilde{\lambda}_i$) as knots of the lasso when $\forall i \in [\![1,p]\!], \beta_i^* = 0$.

As discussed above, recent multiple testing procedures such as the SLOPE, the knockoffs or the procedure derived from the covariance test (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, 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. In section 3, 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

$$\forall i \in [\![1,p]\!], \hat{\beta}_i^{U^*}(\lambda) = \ \mathrm{sign}(\hat{\beta}_i^{\mathrm{mle}}) \left(|\hat{\beta}_i^{\mathrm{mle}}| - \lambda/\delta_i^*\right)_+, \text{where } \hat{\beta}^{\mathrm{mle}} := (X^T\Gamma^{-1}X)^{-1}X^T\Gamma^{-1}Y. \quad (2.3)$$

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 $|\hat{\beta}_i^{\mathrm{mle}}| > \lambda/\delta_i^*$, 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 $|\hat{\beta}_i^{\mathrm{mle}}|/\mathrm{se}(\hat{\beta}_{\mathrm{i}}^{\mathrm{mle}}) > \mu$, where $\mathrm{se}(\hat{\beta}_{\mathrm{i}}^{\mathrm{mle}})$ is the standard error of $\hat{\beta}_i^{\mathrm{mle}}$. One should notice that in these decisions rules, the critical value μ is the same for all i.

In contrast, the value δ^* in (2.3) giving a multidimensional acceptance region with a minimal volume leads to decision rules where μ varies with the tested hypothesis $\mathcal{H}_i$.

This article is organized as follows. In section 2, we study the particular case in which the design matrix X has orthogonal columns (i.e. X^TX is diagonal). In this setting, we provide a "lasso-type" procedure which controls the FWER. Section 3 addresses the general case where X is a design matrix of rank p. We establish that the lasso-type estimator obtained by minimizing the volume of the multidimensional acceptance region is just a soft thresholded maximum likelihood estimator. Section 4 gives the construction of the new multiple testing procedure based on the maximum likelihood estimator. Section 5 is devoted to simulation experiments : we compare our multiple testing procedure with 1) the stepdown multiple testing procedure of Holm (1979) and the generic stepdown multiple testing procedure of Romano and Wolf (2005) and Lehmann and Romano (2005) (p. 352), 2) the active set estimation provided by Lounici (2008), 3) the multiple testing procedure that uses knockoff knots described in Janson and Su (2016). Section 6 details the analysis of metabolomic data which motivated this work.

2.2 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 (2.1) 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)

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

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 2.1 provides a closed form for the lasso estimator and an explicit tuning parameter λ_0 to control the FWER.

Proposition 2.1 *Let X be a $n \times p$ matrix such that $X^TX = \text{ diag}(d_1, \ldots, d_p)$ then*

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

Let $Z^{\text{ols}} := (Z_1^{\text{ols}}, \ldots, Z_p^{\text{ols}})$ 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,

$$\mathbb{P}(\forall i \notin \mathcal{A}, \hat{\beta}_i(\lambda_0) = 0) \geq 1 - \alpha. \tag{2.4}$$

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.

2.3 General case : when the lasso is a soft thresholded likelihood estimator

In this section, 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 (2.1). This leads to the new linear Gaussian model

$$\tilde{Y} = \tilde{X}\beta^* + \tilde{\varepsilon} \text{ with } \tilde{Y} = UY, \tilde{X} = UX \text{ and } \tilde{\varepsilon} = U\varepsilon. \tag{2.5}$$

Because $\tilde{X}$ has orthogonal columns, it is possible to use the Proposition 2.1 of the previous section. 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 (2.5), 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 $\lambda_0^U/d_1(U), \ldots, \lambda_0^U/d_p(U)$ 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. Of course, 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 $\beta^* = 0$,

$$\begin{aligned}
1-\alpha &= \mathbb{P}(\forall i \in [\![1,p]\!], \hat{\beta}_i^U(\lambda_0^U) = 0), \\
&= \mathbb{P}(\forall i \in [\![1,p]\!], d_i(U) \times |\hat{\beta}_i^{\text{ols}}(U)| \leq \lambda_0^U), \\
&= \mathbb{P}\left(\hat{\beta}^{\text{ols}}(U) \in \left[-\frac{\lambda_0^U}{d_1(U)}, \frac{\lambda_0^U}{d_1(U)} \right] \times \cdots \times \left[-\frac{\lambda_0^U}{d_p(U)}, \frac{\lambda_0^U}{d_p(U)} \right] \right).
\end{aligned}$$

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 2.1 *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 $\hat{\beta}^{\mathrm{mle}}$ is the maximum likelihood estimator of the model (2.1).

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 $|\hat{\beta}_i^{\mathrm{mle}}| \geq \lambda/\delta_i^*$ thus lasso-type estimator is useless. Consequently, to manage this new procedure, it is not useful to construct the transformation U^*; discussions about this matrix and an explicit construction of U^* are given in Appendix 1.

In general, the optimal parameter δ^* of the theorem 2.1 is not collinear to $1/\mathrm{se}(\hat{\beta}_1^{\mathrm{mle}}), \ldots, 1/\mathrm{se}(\hat{\beta}_p^{\mathrm{mle}})$. Consequently the random variables $\delta_1^*\hat{\beta}_1^{\mathrm{mle}}, \ldots, \delta_p^*\hat{\beta}_p^{\mathrm{mle}}$ have different variances. This remark is the main difference with the classical procedures for which statistical tests $\hat{\beta}_1^{\mathrm{mle}}/\mathrm{se}(\hat{\beta}_1^{\mathrm{mle}}), \ldots, \hat{\beta}_p^{\mathrm{mle}}/\mathrm{se}(\hat{\beta}_p^{\mathrm{mle}})$ are re-scaled to have unit variance. To provide a mutiple testing procedure which reject $\mathcal{H}_i : \beta_i^* = 0$ as soon as $|\hat{\beta}_i^{\mathrm{mle}}| \geq \lambda/\delta_i^*$ the parameter λ have to be chosen as the $1-\alpha$ quantile of $\max\{\delta_1^*|Z_1^{\mathrm{mle}}|, \ldots, \delta_p^*|Z_p^{\mathrm{mle}}|\}$. From now on, we denote $\lambda_0(\delta)$ the $1-\alpha$ quantile of $\max\{\delta_1|Z_1^{\mathrm{mle}}|, \ldots, \delta_p|Z_p^{\mathrm{mle}}|\}$ 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 to compute the optimal parameter δ^*. The next section deals with this issue.

2.4 A new procedure based on the maximum likelihood estimator

The theorem 2.1 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 $P\left(Z^{\mathrm{mle}} \in B\right) = 1-\alpha$. 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 2.2 *Let* $b^* = \left(b_1^*, \ldots, b_p^*\right)$ *be a solution of the following optimisation problem*

$$\min \prod_{i=1}^{p} b_i \text{ subject to } \mathbb{P}\left(|Z_1^{\text{mle}}| \leq b_1, \ldots, |Z_p^{\text{mle}}| \leq b_p\right) = 1 - \alpha. \tag{2.6}$$

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 $\Sigma^{-1}\text{var}(T^{b^*})$ *should be equal.*

Notice that if the variance matrix of T^{b^*} (here denoted by $\text{var}(T^{b^*})$) was equal to Σ, all the diagonal coefficients of $\Sigma^{-1}\text{Var}(T^{b^*})$ would be equal, indicating that b^* is a solution of (6.4). Because the diagonal terms of $\text{var}(T^{b^*})$ are always smaller than the diagonal terms of Σ, $\text{var}(T^{b^*})$ cannot be equal to Σ. However, the condition given by Proposition 2.2 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}. In the general case, the optimal B^* cannot be explicitly calculated. Nevertheless, there are some simple cases of interest where its computation can be performed by hand. Let us give the optimal parameter δ^* in the following three examples. For convenience, we denote $M(a, b)$ a matrix whose diagonal coefficients are equal to a and whose non-diagonal coefficients are equal to b.

1) In the independent case : the components $\hat{\beta}_1^{\text{mle}}, \ldots, \hat{\beta}_p^{\text{mle}}$ are independent thus, Σ is the diagonal matrix $\text{diag}(\text{var}(\hat{\beta}_1^{\text{mle}}), \ldots, \text{var}(\hat{\beta}_1^{\text{mle}}))$. From Proposition 2.2, the vector b^* must satisfy

$$\frac{1}{\text{var}(\hat{\beta}_1^{\text{mle}})}\text{var}(T_1^{b^*}) = \cdots = \frac{1}{\text{var}(\hat{\beta}_p^{\text{mle}})}\text{var}(T_p^{b^*}).$$

One deduces that $b_1^* = \text{se}(\hat{\beta}_1^{\text{mle}}), \ldots, b_p^* = \text{se}(\hat{\beta}_p^{\text{mle}})$. Consequently, the vector $\delta^* = (\delta_1^*, \ldots, \delta_p^*)$ is collinear to $(1/\text{se}(\hat{\beta}_1^{\text{mle}}), \ldots, 1/\text{se}(\hat{\beta}_p^{\text{mle}}))$. In this particular case, the variances of $\delta_1^*\hat{\beta}_1^{\text{mle}}, \ldots, \delta_p^*\hat{\beta}_p^{\text{mle}}$ are equals

2) In the equicorrelated case : the components of $\hat{\beta}_1^{\text{mle}}, \ldots, \hat{\beta}_p^{\text{mle}}$ have unit variance and $\forall i \neq j$, we set $\text{cov}(\hat{\beta}_i^{\text{mle}}, \hat{\beta}_j^{\text{mle}}) = \rho$ thus, $\Sigma = M(1, \rho)$. It follows that $\Sigma^{-1} = M(a, b)$ for some a and b. When $\delta^* = (1, \ldots, 1)$, we have $\text{var}(T^{b^*}) = M(c, d)$ for some c and d. In this case, all the diagonal coefficients of $\Sigma^{-1}\text{var}(T^{b^*}) = M(a, b)M(c, d)$ are equal. As in the previous case 1), the variances of $\delta_1^*\hat{\beta}_1^{\text{mle}}, \ldots, \delta_p^*\hat{\beta}_p^{\text{mle}}$ are equals.

3) In the block diagonal equicorrelated case : the covariance matrix Σ of $\hat{\beta}^{\text{mle}}$ is the following block diagonal matrix $\text{diag}(M(1, \rho), M(1, \rho'))$ where $M(1, \rho)$ and $M(1, \rho')$ are respectively a $s \times s$ and a $p - s \times p - s$ matrices. It follows that Σ^{-1} is block diagonal with $\Sigma^{-1} = \text{diag}(M(a, b), M(a', b'))$. If we set $\delta_1^* = \cdots = \delta_s^* = k_1$ and $\delta_{s+1}^* = \cdots = \delta_p^* = k_2$,

one deduces that $\mathrm{var}(T^{b^*})$ is block diagonal with $\mathrm{var}(T^{b^*}) = \mathrm{diag}(M(c,d), M(c',d'))$ for some c, d, c', d'. Consequently, whatever k_1 and k_2, the s first diagonal coefficients of $\Sigma^{-1}\mathrm{var}(T^{b^*})$ are equal and the $p-s$ last diagonal coefficients of $\Sigma^{-1}\mathrm{var}(T^{b^*})$ are equal. It remains to tune k_1 and k_2 such that all the diagonal coefficients of $\Sigma^{-1}\mathrm{var}(T^{b^*})$ become equal. Conversely to the cases 1) and 2), the variances of $\delta_1^*\hat{\beta}_1^{\mathrm{mle}}, \ldots, \delta_p^*\hat{\beta}_p^{\mathrm{mle}}$ are not equals. Because in this case variances are not all equals, comparison with classical procedures for which components of $\hat{\beta}^{\mathrm{mle}}$ are re-scaled to have unit variance is interesting.

When the computation of the optimal B^* cannot be carried out explicitly, 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 (2.1), namely $\Sigma := (X^T\Gamma X)^{-1}$.
2. The parameter $\delta^* \in (0, +\infty)^p$ is obtain by solving the problem (6.4). This optimal parameter must satisfies the relation $\Sigma^{-1}\mathrm{var}(T^{b^*})$ given in the proposition 2.2.
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 of the following section show that the gain of volume for the acceptance region provides a gain in power.

2.5 Comparison with other multiple testing procedures

In this section, we compare the performances of our method to the ones of existing methods. Comparisons with the Lounici's active set estimator (Lounici, 2008) and with the multiple testing procedure *via* knockoffs (Janson and Su, 2016) are carried out using different criteria but also different simulations. This is because 1) contrarily to knockoffs, the generic stepdown and the Holm's procedures that control the FWER, Lounici's work provides an active set estimator and aims at controlling the probability to recover exactly the active set 2) the knockoffs

procedure requires a long computer time that precludes its performances evaluation with large values of p.

2.5.1 Comparison with Holm's and generic stepdown procedure

In the Gaussian linear model, the hypothesis $\mathcal{H}_i : \beta_i^* = 0$ is associated to the p-value $P_i := 2\bar{\phi}\left(|\hat{\beta}_i^{\text{mle}}|/\text{se}(\hat{\beta}_i^{\text{mle}})\right)$, where $\bar{\phi}$ is the complementary cumulative distribution function of a $\mathcal{N}(0,1)$ distribution. The Holm multiple testing procedure (Holm, 1979) is a stepdown procedure for which p-values are sorted from the most significant to the least significant, namely $P_{s(1)} \leq P_{s(2)} \leq \cdots \leq P_{s(p)}$. The rejection of the hypotheses $\mathcal{H}_{s(1)}, \ldots, \mathcal{H}_{s(p)}$ is carried-out sequentially as explain hereafter. The hypothesis $\mathcal{H}_{s(1)}$ is rejected if and only if $P_{s(1)} \leq \alpha/p$. The hypothesis $\mathcal{H}_{s(2)}$ is rejected if and only if $P_{s(1)} \leq \alpha/p$ and $P_{s(2)} \leq \alpha/(p-1)$ and so on. This procedure insures a FWER control at a level α and improves the Bonferroni procedure since the cutoff $\alpha/(p-i+1)$ associated to the hypothesis $\mathcal{H}_{s(i)}$ is smaller than α/p.

The generic stepdown procedure defined by Romano and Wolf (2005), Lehmann and Romano (2005) p. 352 and Dudoit and Van Der Laan (2007) p. 126 takes into account the joint distribution of $\hat{\beta}^{\text{mle}}$. Because the Holm's multiple testing procedure only takes into account the marginal distribution of $\hat{\beta}^{\text{mle}}$, the generic stepdown procedure has a higher power than the Holm's multiple testing procedure. To describe the generic stepdown procedure, let us denote $T_i = \hat{\beta}_i^{\text{mle}}/\text{se}(\hat{\beta}_i^{\text{mle}})$ the statistical test and $Z = (Z_1, \ldots, Z_p)$ a centered Gaussian vector with the same covariance matrix as $T := (T_1, \ldots, T_p)$. The statistical tests are sorted from the most significant to the least significant, namely $|T_{r(1)}| \geq \cdots \geq |T_{r(p)}|$. The rejection of the hypotheses $\mathcal{H}_{r(1)}, \ldots, \mathcal{H}_{r(p)}$ is done sequentially as explain hereafter. The hypothesis $\mathcal{H}_{r(1)}$ is rejected if $|T_{r(1)}| \geq t_{r(1)}$. The hypothesis $\mathcal{H}_{r(2)}$ is rejected if $|T_{r(1)}| \geq t_{r(1)}$ and $|T_{r(2)}| \geq t_{r(2)}$ and so on. In the previous expressions, the threshold $t_{r(s)}$ is the $1-\alpha$ quantile of $\max\{|Z_{r(s)}|, \ldots, |Z_{r(p)}|\}$.

For the numerical experiments, we performed 1000 simulations. The covariance matrix Σ of the maximum likelihood estimator is $\Sigma := \text{diag}(M(1,\rho), Id_{500})$, where $M(1,\rho)$ and Id_{500} are both 500×500 matrices. We set $\beta^* \in \mathbb{R}^{1000}$, $\mathcal{A} = [\![1, 20]\!]$ and $\forall i \in \mathcal{A}, \beta_i^* = c$. We performed simulations for different values of $\rho \in \{0, 0.3, 0.6, 0.9\}$. The optimal parameter δ^* of the lemma 2.2 is $\delta_1^* = \cdots = \delta_{500}^* = k_1$ and $\delta_{501}^* = \cdots = \delta_{1000}^* = k_2$. In the independent case, when $\rho = 0$, k_1 and k_2 can be computed by hand and we obtained $k_1 = k_2 = 1$ while in the other cases, k_1 and k_2 had been computed numerically. When $\rho = 0.3$, $\rho = 0.6$ and $\rho = 0.9$, we obtained respectively $k_1 = 1, k_2 = 0.956$, $k_1 = 1, k_2 = 0.895$ and $k_1 = 1, k_2 = 0.690$. These values of δ^* were used to derive $\lambda_0(\delta^*)$ giving a FWER less that $\alpha = 0.05$. In figure 2.1, the power of each multiple testing procedure is represented as a function of $\beta_i^* = c$, for $i \in \mathcal{A}$ and for different values of ρ . The power is the average proportion of true discoveries that can be written respectively for our procedure, Holm's procedure and generic stepdown procedure as

$$\frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}}\mathbb{E}_c\left(\mathbb{1}_{\{|\hat{\beta}_i^{\mathrm{mle}}|>\lambda_0(\delta^*)/\delta_i^*\}}\right),$$
$$\frac{1}{|\mathcal{A}|}\sum_{s(i)\in\mathcal{A}}\mathbb{E}_c\left(\prod_{j=1}^{i}\mathbb{1}_{\{P_{s(j)}\leq\frac{\alpha}{p+1-j}\}}\right),$$
$$\frac{1}{|\mathcal{A}|}\sum_{s(i)\in\mathcal{A}}\mathbb{E}_c\left(\prod_{j=1}^{i}\mathbb{1}_{\{t_{r(j)}\leq|T_{r(j)}|\}}\right).$$

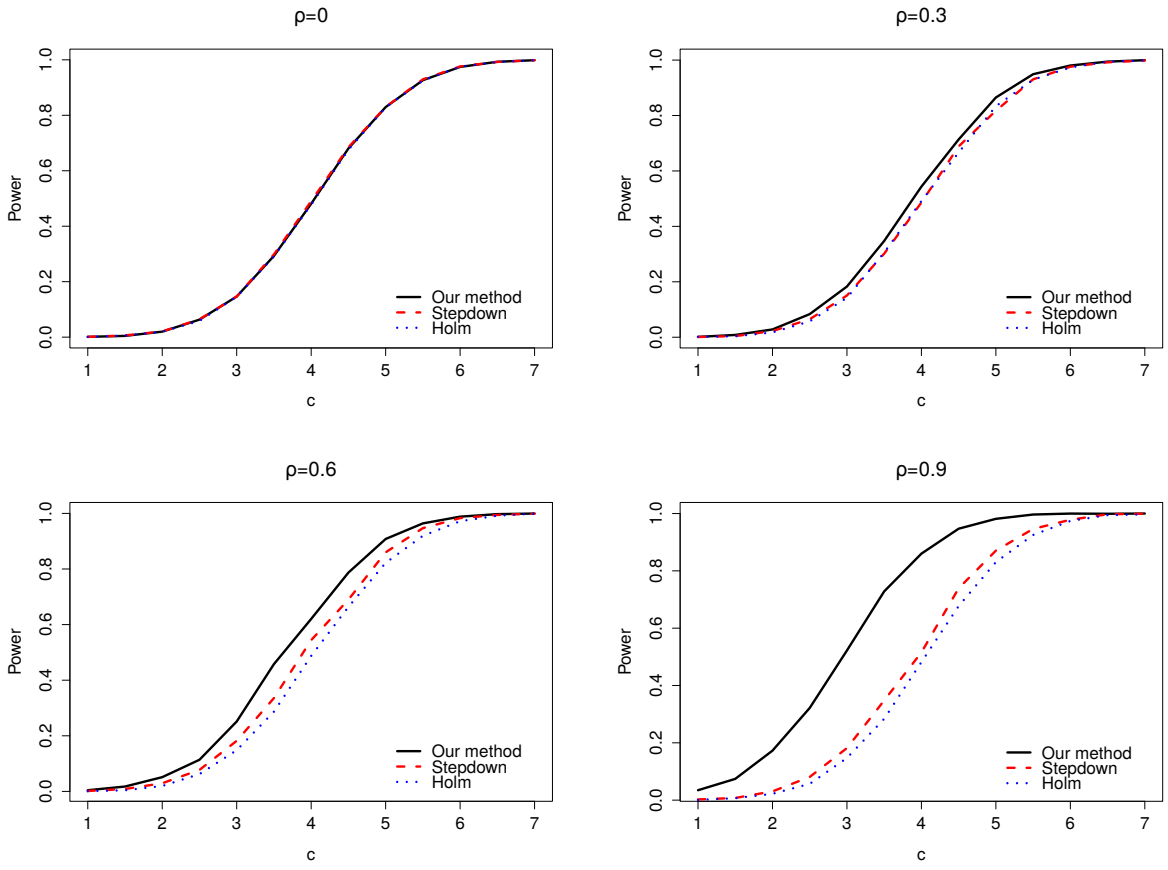


FIGURE 2.1 – This figure shows the power our multiple testing procedure, the power of multiple testing procedures generic stepdown and the power of Holm's procedure. When $\rho = 0$, the three procedures have approximately the same power. When ρ increases, the difference between the power of our procedure and the other one increases.

These numerical experiments illustrates that our procedure is more powerful than the other two procedures, especially when the maximum likelihood estimator owns strong correlated components. Comparison of power of different procedures makes sense only when these procedures share the same FWER. The table 2.1 provides the FWER of the three compared procedures.

	$\rho = 0$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.9$
Holm	0.0496	0.0430	0.034	0.0286
Generic stepdown	0.0491	0.0498	0.0491	0.0505
Our procedure	0.0483	0.0487	0.0502	0.0540

TABLE 2.1 – This table gives the empirical FWER estimated with 1000 simulations. The FWER level of our procedure and the generic stepdown procedure is close to the nominal level of 5%. The FWER level of the Holm procedure decreases when the maximum likelihood estimator has strong correlated components.

2.5.2 Comparison with Lounici's estimator

Lounici (2008) used a thresholded lasso estimator $\hat{\beta}^{\text{th}}$ to build the following estimator of $\mathcal{A}$:

$$\mathcal{A}(\hat{\beta}_i^{\text{th}}(\lambda_L)) := \{i \in [\![1, p]\!] \mid \hat{\beta}_i^{\text{th}}(\lambda_L) \neq 0\}.$$

He proved that the event $\{\mathcal{A}(\hat{\beta}_i^{\text{th}}(\lambda_L)) = \mathcal{A}\}$ has a controlled probability when the design matrix X is close to an orthogonal matrix up to a multiplicative constant, the noise ε is Gaussian standard $\mathcal{N}(0, \sigma^2 Id_p)$, and the smallest non-null parameter $|\beta_i^*|$ is sufficiently large. For the numerical experiments, we took the same setting as the one given in the previous subsection. However, because Lounici's estimator requires a design matrix close to an orthogonal one, we only focused on the particular case where $\rho = 0$. This implies that $\Sigma = Id_{1000}$. In this case, the estimator $\hat{\beta}^{\text{th}}$ has a closed form

$$\forall i \in [\![1, 1000]\!], \hat{\beta}_i^{\text{th}}(\lambda_L) = \begin{cases} \hat{\beta}_i \text{ if } \hat{\beta}_i \geq 3/2\lambda_L \\ 0 \text{ otherwise} \end{cases}, \text{ with } \hat{\beta}_i = \text{ sign}(\hat{\beta}_i^{\text{mle}})(|\hat{\beta}_i^{\text{mle}}| - \lambda_L)_+$$

The tuning parameter λ_L is given by $\lambda_L := A\sigma\sqrt{\log(p)}$ where A has to be determined to fit the desired level. When the smallest non-null parameter $|\beta_i^*|$ is large enough, $\mathbb{P}(\mathcal{A}(\hat{\beta}_i^{\text{th}}(\lambda_L)) = \mathcal{A}) \geq 1 - p^{1-A^2/8}$. From this last expression, we chose A such that $1 - p^{1-A^2/8} = 0.95$. Because Lounici's work proposed to control the probability of $\{\mathcal{A}(\hat{\beta}_i^{\text{th}}(\lambda_L)) = \mathcal{A}\}$, we compared the probability to recover exactly the active set with our method and with the Lounici's one. These probabilities are respectively $\mathbb{P}_c(\{i \in [\![1, p]\!] \mid |\hat{\beta}_i^{\text{mle}}| > \lambda_0(\delta^*)/\delta_i^*\} = \mathcal{A})$ and $\mathbb{P}_c(\mathcal{A}(\hat{\beta}_i^{\text{th}}(\lambda_L)) = \mathcal{A})$ are represented in figure 2.2.

The main explanation of the observed difference between $\mathbb{P}_c(\hat{\mathcal{A}}^{\text{L}}(\lambda_L) = \mathcal{A})$ and $\mathbb{P}_c(\{i \in [\![1, p]\!] \mid |\hat{\beta}_i^{\text{mle}}| > \lambda_0(\delta^*)/\delta_i^*\} = \mathcal{A})$ relies on the choice of the tuning parameter. Indeed, the parameter $\lambda_0(\delta^*)$ is the $1-\alpha$ quantile of $\max\{|Z_1^{\text{mle}}|, \ldots, |Z_p^{\text{mle}}|\}$ ($\delta^* = (1, \ldots, 1)$), whereas Lounici's tuning parameter λ_L bounds above the $1 - \alpha$ quantile of $2\max\{|Z_1^{\text{mle}}|, \ldots, |Z_p^{\text{mle}}|\}$. With our multiple testing procedure, the probability of no false discovery is $\mathbb{P}(\forall i \in [\![21, 1000]\!], \hat{\beta}_i^{\text{mle}} \leq \lambda_0(\delta^*)/\delta_i^*)$ is exactly equal to 0.9510. As one can notice in figure 2.2, when the all the parameters β_i^* in the active set increase, *ie* when c increases, the probability $\mathbb{P}_c(\{i \in [\![1, p]\!] \mid$

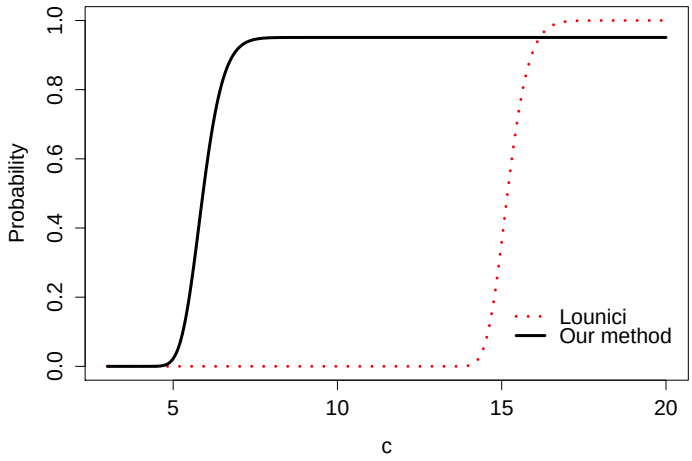


FIGURE 2.2 – This figure represents the probabilities to recover the active set with Lounici's method ($\mathbb{P}_c(\hat{\mathcal{A}}^{\mathrm{L}}(\lambda_L) = \mathcal{A})$) in red dotted line and with our method ($\mathbb{P}_c(\{i \in [\![1,p]\!] \mid |\hat{\beta}_i^{\mathrm{mle}}| > \lambda_0(\delta^*)/\delta_i^*\} = \mathcal{A})$) in black plain line. Our method recovers exactly the active set even when the non null parameters are small (c is small). When c is very large, $\mathbb{P}_c((\hat{\beta}_i^{\mathrm{th}}(\lambda_L)) = \mathcal{A}) \approx 1$ and $\mathbb{P}_c(\{i \in [\![1,p]\!] \mid |\hat{\beta}_i^{\mathrm{mle}}| > \lambda_0(\delta^*)/\delta_i^*\} = \mathcal{A}) \approx 0.95$.

$|\hat{\beta}_i^{\mathrm{mle}}| > \lambda_0(\delta^*)/\delta_i^*\} = \mathcal{A})$ does not go to 1. This is because, when there is at least one false discovery (which occurs with a probability 0.0490), we have $\mathcal{A}(\hat{\beta}(\lambda_0)) \neq \mathcal{A}$, thus, one can not have $\mathbb{P}_c(\mathcal{A}(\hat{\beta}(\lambda_0)) = \mathcal{A}) \approx 1$ even if c is very large.

2.5.3 Comparison with multiple testing procedure via knockoffs

A multiple testing procedure that controls the k-FWER had been proposed by Janson and Su (2016). This procedure compares the solution path $\lambda \in \mathbb{R}_+ \mapsto \hat{\beta}(\lambda)$ of the original lasso with the solution path $\lambda \in \mathbb{R}_+ \mapsto \tilde{\beta}(\lambda)$ of the knockoff lasso. These two estimators are defined as follow

$$(\hat{\beta}(\lambda), \tilde{\beta}(\lambda)) = \underset{\beta \in \mathbb{R}^{2p}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|Y - X_{\mathrm{KO}}\beta\|^2 + \lambda \|\beta\|_1 \right\},$$

where the design matrix $X_{\mathrm{KO}} = [X, \tilde{X}]$ is the concatenation of the original design matrix X with a knockoffs design matrix $\tilde{X}$ whose building is given in Barber and Candès (2015). We can view $\tilde{\beta}(\lambda)$ as the lasso estimator obtained when $\beta^* = 0_{\mathbb{R}^p}$.

In this procedure, the number of false discovery is stochastically dominated by a negative binomial distribution $\mathcal{NB}(v, 0.5)$ in which the parameter v is set by the user. This procedure uses the random variables $\hat{\lambda}_j = \sup\{\lambda \mid \hat{\beta}_j(\lambda) \neq 0\}$ and $\tilde{\lambda}_j = \sup\{\lambda \mid \tilde{\beta}_j(\lambda) \neq 0\}$ that are called knots of the lasso solution path. When, $|\beta_i^*| \gg 0$, one would expect that $W_j = \max\{\hat{\lambda}_j, \tilde{\lambda}_j\}$ is large and $\chi_j = \mathbb{1}_{\tilde{\lambda}_j > \hat{\lambda}_j}$ is equal to 0. The random variables $W_1, \ldots, W_p$ are sorted as follow

$W_{s(1)} \geq W_{s(2)} \geq \cdots \geq W_{s(p)}$ and the hypothesis $\mathcal{H}_{s(i)}$ is rejected if and only if $\sum_{j=1}^{i} \chi_{s(i)} < v$.

Because the building of the knockoff matrix needs a normalized matrix X (diagonal coefficients of X^TX must be equal to 1), we can not determine such a matrix and a standard error $\sigma > 0$ such that $\sigma^2(X^TX)^{-1} = \text{diag}(M(1,\rho), Id_{500})$. Indeed, diagonal coefficients of $M^{-1}(1,\rho)$ are not equal to 1 when $\rho \neq 0$. Consequently, whatever $\sigma > 0$, the matrix $X^TX = \sigma^2\text{diag}(M^{-1}(1,\rho), Id_{500})$ can not have diagonal coefficients equal to 1. That is why, we only focus on the equi-correlated case.

In the numerical experiments, we set $n = 250$, $p = 100$ and $\sigma > 0$ is such that $\Sigma = \sigma^2(X^TX)^{-1} = M(1,\rho)$. Different values of ρ have been used $\rho \in \{0, 0.3, 0.6, 0.9\}$. The design matrix X has smaller dimensions than in the previous subsection to avoid a too long computational time. Because we wanted the smallest FWER as possible, we set $v = 1$. In this case, the number of false positive is stochastically dominated by a geometric distribution $\mathcal{NB}(1, 0.5)$ leading to a minimal FWER equals to 0.5. If we had set $v > 1$, the familywise error rate would have been $P(F_v > 0) = 1 - 0.5^v > 0.5$, with F_v distributed according to $\mathcal{NB}(v, 0.5)$. We used the R package knockoff (Barber and Candès, 2015) to build the knockoff matrix and knockoff knots. The optimal parameter δ^* provided by the Lemma 2.2 is $\delta^* = (1, \ldots, 1)$. Then, the parameter $\lambda_0(\delta^*)$ was determined to obtain a FWER equal to 0.5.

The power of each multiple testing procedure is represented in the figure 2.3. The power is the average proportion of true discoveries ; the expression of the power for our procedure and the knockoffs procedure are respectively equal to

$$\frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}} \mathbb{E}_c\left(\mathbb{1}_{\{|\hat{\beta}_i^{\text{mle}}| > \lambda_0(\delta^*)/\delta_i^*\}}\right) \text{ and } \frac{1}{|\mathcal{A}|}\sum_{s(i)\in\mathcal{A}} \mathbb{E}_c\left(\mathbb{1}_{\{\sum_{j=1}^{i}\chi_{\rho(j)} < v\}}\right).$$

These numerical experiments illustrate that our procedure is better, especially when the maximum likelihood estimator has strong correlated components. Comparison of power is meaningful when the FWER is the same for all procedures. An average of 1000 simulations allows to estimate the FWER level of our procedure. This level is equal to $\mathbb{P}_c(\exists i \notin \mathcal{A} \mid \hat{\beta}^{\text{mle}} > \lambda_0(\delta^*)/\delta_i^*) = \mathbb{P}(|Z_i^{\text{mle}}| > \lambda_0(\delta^*)/\delta_i^*)$. This probability does not depend from c, we obtained 0.462, 0.477, 0.482 and 0.495 when the correlation ρ were respectively equal to $\rho = 0, \rho = 0.3, \rho = 0.6$ and $\rho = 0.9$. The figure 2.4 provides the FWER level for the knockoff procedure. Surprisingly, it seems that the knockoff multiple testing procedure does not control the FWER at a level 0.5 for small values of c.

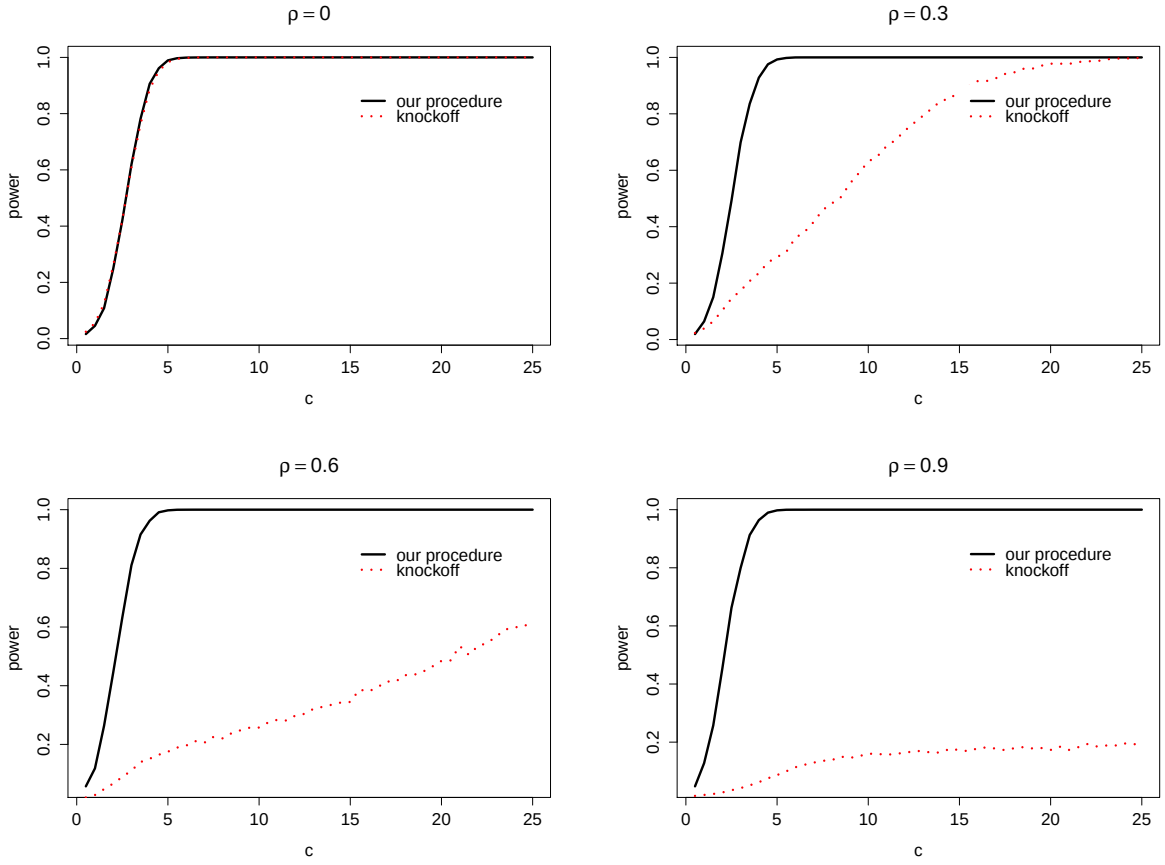


FIGURE 2.3 – In this figure, we compared the power our multiple testing procedure with the power of the knockoff multiple testing procedure. Each point is an average of 1000 simulations. In the case where $\rho = 0$, components of $\hat{\beta}^{\text{mle}}$ are independent and two procedures have approximately the same power. In the case where $\hat{\beta}^{\text{mle}}$ have equi-correlated components, our procedure is more powerful.

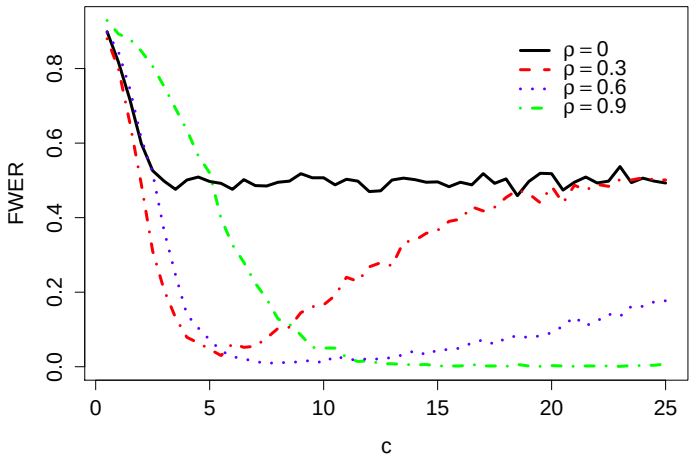


FIGURE 2.4 – In this figure, we have computed the FWER level of the knockoff procedure for all $c > 0$. When non-null parameters are small (i.e c is small), the FWER level is not well controlled. When c is large enough, except in the independent case, the FWER level is largely smaller than its nominal value 0.5. Each point is an average of 1000 simulations.

2.6 Application in metabolomics : detection of metabolites

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. Figure 2.5 gives an example of an annotated spectrum of a mixture.

Recently, automatic methods have been proposed, for example, Metabohunter (Tulpan et al., 2011), BATMAN (Astle et al., 2012; Hao et al., 2012), Bayesil (Ravanbakhsh et al., 2015) or the software Chenomx (Weljie et al., 2006). Most of these methods are based on a modelling using a Lorentzian shape and a Bayesian strategy. Nevertheless, most are time-consuming and thus

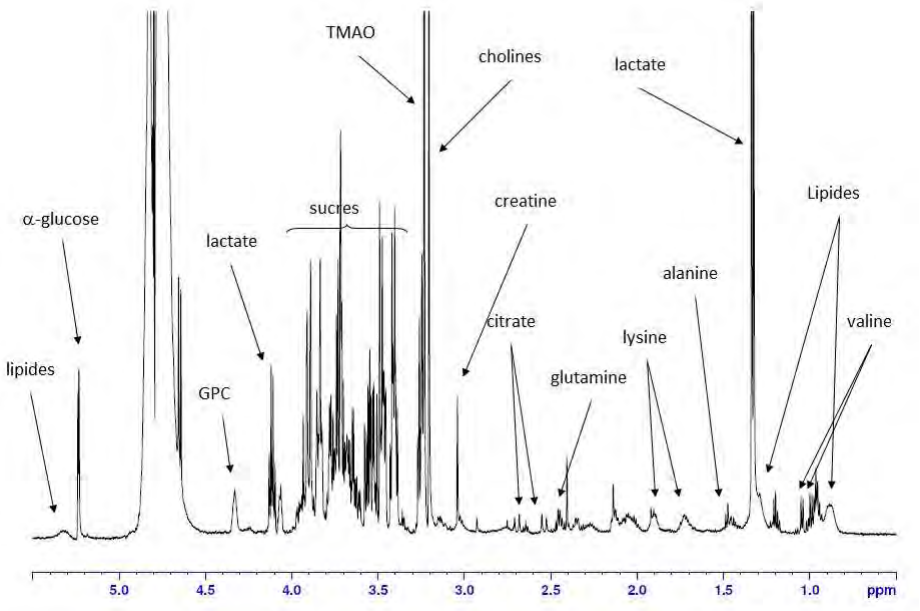


FIGURE 2.5 – Example of an annotated mixture spectrum. There are overlaps between peaks of lipides and valine and between the peaks of glutamine and lysine.

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.

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 a multiple testing procedure that control the FWER.

2.6.1 Modelling

The spectrum of a metabolite (or a mixture) is a nonnegative function defined on a compact interval T. We assume that we have a library of spectra containing all $p = 36$ metabolites $\{f_i\}_{1 \leqslant i \leqslant p}$ (with $\int_T f_i(t)dt = 1$) that can be found in a mixture. This family of p spectra is assumed to be linearly independent. In a first approximation, the observed spectrum of the mixture Y can be modelled as a discretized noisy convex combination of the pure spectra :

$$Y_j = \left(\sum_{i=1}^{p} \beta_i^* f_i(t_j) \right) + \varepsilon_j \text{ with } 1 \leqslant j \leqslant n \text{ and } t_1 < \cdots < t_n \text{ a subdivision of } T.$$

The random vector $(\varepsilon_1, \ldots, \varepsilon_n)$ is a standard Gaussian $\mathcal{N}(0, \sigma^2 Id_n)$. The variance σ^2 is estimated using several observations of a metabolite spectrum.

2.6.2 Real dataset

The method for the detection of metabolites was tested on a known mixture. The NMR experts supplied us with a library of 36 spectra of pure metabolites and a mixture composed of these metabolites. The number of used metabolites and their proportions were unknown to us. The results are presented in Table 2.2.

Metabolites	Actual proportions	Rejection for the nullity of the proportion
Choline chloride	0.545	Yes
Creatinine	0.209	Yes
Benzoic acid	0.086	Yes
L-Proline	0.069	Yes
D-Glucose	0.060	Yes
L-Phenylalanine	0.029	Yes
30 other metabolites	0	No

TABLE 2.2 – This table presents the results for the 36 metabolites of the library. The actual proportions of each metabolite are presented in the first column. For each metabolite, evidence against the nullity of the proportion is given in the second column.

The 6 metabolites that are present in the complex mixture are detected, including those with small proportions. There is no false discovery because any hypothesis associated to the 30 other metabolites was rejected. Because the whole procedure is quite fast, lasting only a few seconds, it could be easily applied to a library containing several hundred metabolites. We refer the interested reader on this application to metabolomics to Tardivel et al. (2017a) where our procedure is compared to the existing ones on more complex datasets.

2.7 Conclusions

When the rank of the $n \times p$ design matrix X is p, we prove that even if X is not orthogonal, even if residuals of the Gaussian model (2.1) are not i.i.d, up to an orthogonalisation, the lasso estimator is just a soft thresholded maximum likelihood estimator. Thus, in this setting, lasso estimator is not useful, maximum likelihood is more appropriate to build a powerful multiple testing procedure. In our new procedure based on the maximum likelihood estimator, one rejects the null hypothesis $\mathcal{H}_i : \beta_i^* = 0$ when $|\hat{\beta}_i^{\text{mle}}| > \lambda_0(\delta^*)/\delta_i^*$. The parameter δ^* is the optimal one given in proposition 2.2 and $\lambda_0(\delta^*)$ is the $1-\alpha$ quantile of $\max\{\delta_1^*|Z_1^{\text{mle}}|, \ldots, \delta_p^*|Z_p^{\text{mle}}|\}$. The keystone of this procedure is to compute the optimal parameter δ^*, an exact computation of δ^* is documented in three particular cases. Numerical comparisons illustrate the benefit of our procedure comparing to the state-of-the-art procedures that control the FWER. Concerning the application in metabolomic a numerical approximation of the parameter δ^* is implemented.

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 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.

2.8 Appendix 1 : construction of the matrix U^*

The theorem 2.1 gives the existence of U^* but does not give a construction of it. The building of an optimal U^* can be performed in two steps. First, because we want a small tuning parameter λ_0^U, we select a set of applications of G that minimize the variance of $\hat{\beta}^{\text{ols}}(U)$. Actually, we will see that there exists a set of transformations that allow $\hat{\beta}^{\text{ols}}(U)$ to become an efficient estimator having thus the same distribution as the maximum likelihood estimator of the model (2.1). Second, we look for an application U^* minimizing $\phi(U)$ among the applications selected at the first step. These two steps are described in the following two lemmas.

Lemma 2.1 *Let P be an invertible $n \times n$ matrix such that $(PX)^T = \begin{pmatrix} Id_p & 0 \end{pmatrix}$ and set A the $n \times n$ invertible matrix*

$$A := (P\Gamma P^T)^{-1} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \text{ with } A_{11} \text{ a } p \times p \text{ matrix. Remind that } \Gamma = \text{var}(\varepsilon).$$

Let $\delta = (\delta_1, \ldots, \delta_p) \in (0, +\infty)^p$ and consider the $p \times n$ matrix V_δ defined by

$$V_\delta = \begin{pmatrix} \Delta & \Delta A_{11}^{-1} A_{12} \end{pmatrix} P, \text{ with } \Delta = diag(\sqrt{\delta_1}, \ldots, \sqrt{\delta_p}).$$

Then, for all $\delta \in (0, \infty)^p$, the matrix V_δ belongs to G, and $\hat{\beta}^{\text{ols}}(V_\delta) = \hat{\beta}^{\text{mle}}$, where β^{mle} is the maximum likelihood estimator of the model (2.1).

The matrix P given in the lemma 2.1 is not unique. To obtain such a matrix P, one completes the linearly independent family $X_1, \ldots, X_p$ with the vectors $v_{p+1}, \ldots, v_n$ of $\mathbb{R}^n$ to obtain a basis and set $P := (X_1 | \ldots | X_p | v_{p+1} | \ldots | v_n)^{-1}$. Lemma 2.1 evidences V_δ transformations that both orthogonalise the design and allow to gain efficiency instead of keeping an ordinary least squares estimator. A traditional transformation to get an efficient estimator in model (2.5) is to apply the linear transformation $\Gamma^{-1/2}$. Because $(\Gamma^{-1/2}X)^T(\Gamma^{-1/2}X) = X^T\Gamma^{-1}X = \text{var}(\hat{\beta}^{\text{mle}})^{-1}$, contrarily to the V_δ transformations, the obtained design matrix $\tilde{X} = \Gamma^{-1/2}X$ in general does not have orthogonal columns. The Puffer transformation $F = UD^{-1}U$, where U and D are given by the singular value decomposition of X, is a transformation given in Jia and Rohe (2012) which relax the irrepresentable condition. When the rank of X is p, FX is orthogonal thus $F \in G$. However contrarily to the V_δ transformations, the estimator $\hat{\beta}^{\text{ols}}(F)$ is not efficient.

As an example for Lemma 2.1, let us set $\Gamma = \text{diag}(1, 2, 3, 4)$ and X the following matrix

$$X := \begin{pmatrix} 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{pmatrix}^T.$$

A (not unique) couple of matrices P and $V_{(1,1)}$ satisfying Lemma 2.1 is

$$P := \begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.5 & -0.5 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix} \text{ and } V_{(1,1)} := \frac{1}{26} \begin{pmatrix} 13 & 6 & -4 & 3 \\ 13 & -6 & 4 & -3 \end{pmatrix}.$$

Let us set $\tilde{X} = V_{(1,1)}X$. The following equality guarantees that $V_{(1,1)} \in G$ and $\hat{\beta}^{\text{ols}}(V_{(1,1)})$ is the maximum likelihood estimator

$$\tilde{X} = Id_2 \text{ and } \hat{\beta}^{\text{ols}}(V_{(1,1)}) = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y} = V_{(1,1)}Y = (X^T\Gamma^{-1}X)^{-1}X^T\Gamma^{-1}Y = \hat{\beta}^{\text{mle}}.$$

The following lemma shows that there exists at least a linear transformation U^* among the linear transformations $(V_\delta)_{\delta \in]0,+\infty[^p}$ that optimizes ϕ.

Lemma 2.2 *Set*

$$U^* = V_{\delta^*} \text{ with } \delta^* = \underset{\delta \in]0,+\infty[^p}{\text{arginf}} \phi(V_\delta), \tag{2.7}$$

then, for all $U \in G$, we have

$$\phi(U^*) \leqslant \phi(U).$$

As shown in the proof (given in the following appendix), there always exists at least a vector $\delta^* \in]0, +\infty[^p$ such that the infimum is reached. Consequently, Theorem 2.1 holds for $U^* = V_{\delta^*}$.

2.9 Appendix 2 : Proofs

Proof (Proposition 2.1) The lasso estimator $\hat{\beta}(\lambda)$ is the point for which the function $\psi(\beta) = \frac{1}{2}\|Y - X\beta\|^2 + \lambda\|\beta\|_1$ reaches its global minimum. Because the penalty term is a L^1 norm, the function ψ is not differentiable everywhere. However, as ψ is a convex function, it has a subdifferential. To find where the global minimum of ψ is reached, we are going to determine $\beta \in \mathbb{R}^p$ for which the subdifferential $\partial\psi(\beta)$ contains $0_{\mathbb{R}^p}$ (Hiriart-Urruty and Lemaréchal, 2013). We have $\partial\psi(\beta) = -X^TY + D\beta + \lambda\partial_{\|.\|_1}(\beta)$ with

$$\partial_{\|.\|_1}(\beta) = C_1 \times \cdots \times C_p, \text{ with } C_i = [-1, 1] \text{ if } \beta_i = 0 \text{ and } C_i = \text{ sign}(\beta_i) \text{ otherwise.}$$

Indeed, the differential of $\beta \mapsto \frac{1}{2}\|Y - X\beta\|^2$ is $-X^TY + X^TX\beta = -X^TY + D\beta$ and $\partial_{\|.\|_1}(\beta)$ is the

subdifferential of $\beta \mapsto \|\beta\|_1$. The function ψ reaches its global minimum at $\hat{\beta}(\lambda)$ consequently $0_{\mathbb{R}^p} \in \partial\psi(\hat{\beta}(\lambda))$; this holds if and only if

$$0_{\mathbb{R}^p} \in \hat{\beta}^{\text{ols}} + \hat{\beta}(\lambda) + \lambda D^{-1}\partial_{\|.\|_1}(\hat{\beta}(\lambda)) \Leftrightarrow \hat{\beta}(\lambda) = \text{ sign}(\hat{\beta}_i^{\text{ols}}) \left(|\hat{\beta}_i^{\text{ols}}| - \frac{\lambda}{d_i}\right)_+ .$$

The multiple testing procedure does not have any false discovery if $\forall i \notin \mathcal{A}, \hat{\beta}_i(\lambda) = 0$. We are going to see that $\{\forall i \notin \mathcal{A}, \hat{\beta}_i(\lambda) = 0\}$ has a probability larger than $1-\alpha$ when the tuning parameter is λ_0. When $i \notin \mathcal{A}$, the Gaussian vector $(\hat{\beta}_i^{\text{ols}})_{i\notin\mathcal{A}}$ has the same distribution as $(Z_i^{\text{ols}})_{i\notin\mathcal{A}}$ because $\beta_i^* = 0$. Therefore, the following inequalities hold

$$\begin{aligned}
\mathbb{P}\left(\forall i \notin \mathcal{A}, \hat{\beta}_i(\lambda_0) = 0\right) &= \mathbb{P}\left(\forall i \notin \mathcal{A}, |\hat{\beta}_i^{\text{ols}}| - \frac{\lambda_0}{d_i} \leqslant 0\right), \\
&= \mathbb{P}\left(\forall i \notin \mathcal{A}, |Z_i^{\text{ols}}| \times d_i \leqslant \lambda_0\right), \\
&\geqslant \mathbb{P}\left(\forall i \in [\![1,p]\!], |Z_i^{\text{ols}}| \times d_i \leqslant \lambda_0\right) = 1-\alpha.
\end{aligned}$$

□

Proof (Lemma 2.1) The matrix V_δ orthogonalises X. Indeed, $\tilde{X} = V_\delta X$ is the following diagonal matrix

$$\tilde{X} = \begin{pmatrix}\Delta & \Delta A_{11}^{-1}A_{12}\end{pmatrix} PX = \begin{pmatrix}\Delta & \Delta A_{11}^{-1}A_{12}\end{pmatrix} \begin{pmatrix} Id_p \\ 0 \end{pmatrix} = \Delta.$$

The estimator $\hat{\beta}^{\text{ols}}(V_\delta)$ is equal to

$$\begin{aligned}
\hat{\beta}^{\text{ols}}(V_\delta) &= (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y}, \\
&= \Delta^{-1}V_\delta Y = \begin{pmatrix} Id_p & A_{11}^{-1}A_{12}\end{pmatrix} PY.
\end{aligned}$$

It remains to show that $\hat{\beta}^{\text{mle}} = \begin{pmatrix} Id_p & A_{11}^{-1}A_{12}\end{pmatrix} PY$.

$$\begin{aligned}
\hat{\beta}^{\text{mle}} &= (X^T\Gamma^{-1}X)^{-1}X^T\Gamma^{-1}Y, \\
&= (X^TP^T(P^T)^{-1}\Gamma^{-1}P^{-1}PX)^{-1}X^TP^T(P^T)^{-1}\Gamma^{-1}P^{-1}PY, \\
&= \left((PX)^TAPX\right)^{-1}(PX)^TAPY, \\
&= \left(\begin{pmatrix} Id_p & 0\end{pmatrix}\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22}\end{pmatrix}\begin{pmatrix} Id_p \\ 0\end{pmatrix}\right)^{-1}\begin{pmatrix} Id_p & 0\end{pmatrix}\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22}\end{pmatrix} PY, \\
&= \begin{pmatrix} Id_p & A_{11}^{-1}A_{12}\end{pmatrix} PY = \hat{\beta}^{\text{ols}}(V_\delta).
\end{aligned}$$

□

The proof of lemma 2.2 relies on two main steps. In the first step, using lemmas A and B given below, we obtain that the function

$$\delta \in (0,+\infty)^p \mapsto \phi(V_\delta)$$

is minimized for at least one element δ^*. In the second step, we prove that the linear transformation V_{δ^*} is such that $\phi(V_{\delta^*})$ is minimal.

Because $(V_\delta X)^T(V_\delta X)$ is the diagonal matrix $\mathrm{diag}(\delta_1,\ldots,\delta_p)$, the quantity $\lambda_0^{V_\delta}$ is equal to $\lambda_0(\delta)$. Let us remind that $\lambda_0(\delta)$ is the $1-\alpha$ quantile of $\max\{\delta_1|Z_1^{\mathrm{mle}}|,\ldots,\delta_1|Z_1^{\mathrm{mle}}|\}$. It is straightforward to show that the function λ_0 verifies the following two properties.

1. The function $\delta \in (0,+\infty)^p \mapsto \lambda_0(\delta)$ is homogeneous :

$$\forall k>0, \forall \delta \in (0,+\infty)^p, \lambda_0(k\delta) = k\lambda_0(\delta).$$

2. The function $\delta \in (0,+\infty)^p \mapsto \lambda_0(\delta)$ is componentwise-increasing :

$$\text{let } \delta, d \in (0,+\infty)^p, \text{ if } \delta \text{ is componentwise-smaller than } d, \text{ then } \lambda_0(\delta) \leqslant \lambda_0(d).$$

The following lemma provides the continuity of the function $\delta \in (0,+\infty)^p \mapsto \lambda_0(\delta)$.

Lemma A *Let g be a function that satisfies the two previous properties ; then, the function g is continuous.*

Proof Let $x=(x_1,\ldots,x_p)\in(0,+\infty)^p$, for an arbitrary $\epsilon>0$, we are going to construct $\eta>0$ such that $\|y-x\|_\infty \leq \eta$ implies $|g(y)-g(x)|\leq\epsilon$ which gives the continuity of g at x. We set $u=(u_1,\ldots,u_p)$ the unit vector $u=x/\|x\|$. Let $r<\|x\|$, the function g is homogeneous, consequently,

$$\begin{aligned} g(x-ru) &= g\left(x\left(1-\frac{r}{\|x\|}\right)\right) = \left(1-\frac{r}{\|x\|}\right)g(x) \text{ and} \\ g(x+ru) &= \left(1+\frac{r}{\|x\|}\right)g(x). \end{aligned}$$

Let $y\in(0,+\infty)^p$ be such that the following inequality occurs componentwise : $x-ru\leqslant y\leqslant x+ru$**.** Because g is componentwise-increasing, we have $g(x-ru)\leqslant g(y)\leqslant g(x+ru)$. More precisely,

$$\forall y\in[x_1-ru_1,x_1+ru_1]\times\cdots\times[x_p-ru_p,x_p+ru_p], |g(y)-g(x)|\leqslant\frac{r}{\|x\|}|g(x)|. \qquad (2.8)$$

Let $\epsilon>0$; one can choose $r_0\geqslant 0$ small enough such that $r_0|g(x)|/\|x\|\leqslant\epsilon$. We set $\eta=$

$r_0 \min\{u_1, \ldots, u_p\}$; thus, the inequality (2.8) gives

$$\|y - x\|_\infty \leqslant \eta \Rightarrow |g(y) - g(x)| \leqslant \epsilon,$$

which proves the continuity of g on $(0, +\infty)^p$. □

Lemma B *The function $f : \delta \in (0, +\infty)^p \mapsto \phi(V_\delta)$ reaches its minimum for at least one element δ^*.*

Proof Let us remind the expression of the function f

$$\forall \delta \in (0, +\infty)^p, f(\delta) = \frac{\lambda_0(\delta)}{\delta_1} \times \cdots \times \frac{\lambda_0(\delta)}{\delta_p}.$$

Since λ_0 is homogeneous, f satisfies the property $\forall k > 0, f(k\delta) = f(\delta)$. Consequently, if the minimum of f over $\mathcal{E} := \{\delta \in (0, +\infty)^p \mid \|\delta\|_\infty = 1\}$ is reached at a point $\delta \in \mathcal{E}$ then f reaches its minimum on the set $\{k\delta \mid k > 0\}$. To prove that the minimum of f over $\mathcal{E}$ cannot be reached for "small δ", we are going 1) to decompose $\mathcal{E}$ in two disjoint sets $\mathcal{E} := A_{\eta_0} \cup B_{\eta_0}$, where

$$A_{\eta_0} := \{\delta \in (0, +\infty)^p \mid \|\delta\|_\infty = 1 \text{ and } \min\{\delta_1, \ldots, \delta_p\} \geq \eta_0\} \text{ and}$$

$$B_{\eta_0} := \{\delta \in (0, +\infty)^p \mid \|\delta\|_\infty = 1 \text{ and } \min\{\delta_1, \ldots, \delta_p\} < \eta_0\}.$$

2) and then to prove that there exists $\eta_0 \in (0, 1)$ and a point δ_A in A_{η_0} such that $f(\delta_A) < \inf_{\delta \in B_{\eta_0}}\{f(\delta)\}$. This will show that $\inf_{\delta \in \mathcal{E}}\{f(\delta)\}$ is equal to $\inf_{\delta \in A_{\eta_0}}\{f(\delta)\}$. The final step of the proof will show that the minimum of f is reached over A_{η_0}.

Let us first build $\eta_0 \in (0, 1)$. For all $i \in [\![1, p]\!]$, let us denote $q_i := \text{se}(\hat{\beta}_i^{\text{mle}}) z_{1-\alpha/2}$ with $z_{1-\alpha/2}$ the $1-\alpha/2$ quantile of a $\mathcal{N}(0, 1)$ distribution. Defined as this, q_i is also the $1-\alpha$ quantile of $|Z_i^{\text{mle}}|$. Notice that $q_i > 0$ because $\text{se}(\hat{\beta}_i^{\text{mle}}) > 0$ and $\alpha \in (0, 1)$. By definition, $\lambda_0(\delta)$ is the $1 - \alpha$ quantile of $\max_{1 \leq i \leq p}\{\delta_i |Z_i^{\text{mle}}|\}$. Consequently, when $\delta \in \mathcal{E}$ we have $\lambda_0(\delta) \geq \min\{q_1, \ldots, q_p\}$ because at least one component of δ is equal to 1. Let us denote $m := \min\{q_1, \ldots, q_p\}$, $\delta_A := (1, \ldots, 1)$ and $\eta_0 := \min\{m^p / f(\delta_A), 1/2\}$.

Let $\delta \in B_{\eta_0}$, because $\delta_1 \times \cdots \times \delta_p < \eta_0$ the following inequality holds

$$f(\delta) = \frac{\lambda_0(\delta)}{\delta_1} \times \cdots \times \frac{\lambda_0(\delta)}{\delta_p} \geq \frac{m}{\delta_1} \times \cdots \times \frac{m}{\delta_p} > \frac{m^p}{\eta_0}.$$

In particular, this shows that $\forall \delta \in B_{\eta_0}, f(\delta) > f(\delta_A)$ consequently, the minimum cannot be reached on B_{η_0}. Because f is continuous on A_{η_0} and A_{η_0} is compact, f reaches its minimum on A_{η_0}. □

The following lemma is a consequence of corollary 3 of Anderson (1955).

Lemma C (Anderson) *Let $V = (V_1, \ldots, V_n)$ and $W = (W_1, \ldots, W_n)$ be centred Gaussian vectors with variance matrices Γ_V and Γ_W, respectively. Assume that the matrix $\Gamma_W - \Gamma_V$ is a positive semidefinite matrix; then,*

$$\forall x \geqslant 0, \mathbb{P}(\max\{|W_1|, \ldots, |W_n|\} \geqslant x) \geqslant \mathbb{P}(\max\{|V_1|, \ldots, |V_n|\} \geqslant x).$$

This inequality implies that $\max\{|W_1|, \ldots, |W_n|\}$ is stochastically greater than $\max\{|V_1|, \ldots, |V_n|\}$.

Proof (Lemma 2.2) For any $U \in G$, the matrix $(UX)^T UX$ is diagonal and $(UX)^T UX = \Delta = \mathrm{diag}(\delta_1, \ldots, \delta_p) = \mathrm{diag}(\delta)$. The difference between the covariance matrices of the Gaussian vectors $(\delta_1 Z_1^{\mathrm{ols}}(U), \ldots, \delta_p Z_p^{\mathrm{ols}}(U)) = \Delta Z^{\mathrm{ols}}(U)$ and $(\delta_1 Z_1^{\mathrm{mle}}, \ldots, \delta_p Z_p^{\mathrm{mle}}) = \Delta Z^{\mathrm{ols}}(V_\delta)$ is semidefinite positive. Indeed, reminding that Σ is the covariance matrix of the maximum likelihood estimator, we obtain that

$$\begin{aligned} \forall x \in \mathbb{R}^p, x^T(\mathrm{var}(\Delta Z^{\mathrm{ols}}(U)) - \mathrm{var}(\Delta Z^{\mathrm{mle}}))x &= (\Delta x)^T(\mathrm{var}(Z^{\mathrm{ols}}(U)) - \Sigma)\Delta x, \\ &= (\Delta x)^T(\mathrm{var}(\hat{\beta}^{\mathrm{ols}}(U)) - \Sigma)\Delta x \geqslant 0. \end{aligned}$$

The last inequality is a consequence of the Gauss-Markov theorem (Rencher and Schaalje, 2008) (page 146). Because λ_0^U and $\lambda_0^{V_\delta}$ are the respective $1-\alpha$ quantiles of $\max\{\delta_1|Z_1^{\mathrm{ols}}(U)|, \ldots, \delta_p|Z_p^{\mathrm{ols}}(U)|\}$ and $\max\{\delta_1|Z_1^{\mathrm{mle}}|, \ldots, \delta_p|Z_p^{\mathrm{mle}}|\}$, the lemma C gives $\lambda_0^U \geqslant \lambda_0^{V_\delta}$. This last inequality gives

$$\phi(V_\delta) = \frac{\lambda_0^{V_\delta}}{\delta_1} \times \cdots \times \frac{\lambda_0^{V_\delta}}{\delta_p} \leq \frac{\lambda_0^U}{\delta_1} \times \cdots \times \frac{\lambda_0^U}{\delta_p} = \phi(U).$$

Finally, using lemma 6.1, the inequality $\phi(V_\delta) \geqslant \phi(V_{\delta^*})$ gives the result. □

Proof (Theorem 2.1) The lemmas 2.1 and 2.2 allow to prove the theorem 2.1. □

Proof (Proposition 2.2) To simplify the computation of the gradients, we consider the following problem which has the same solution as the problem (6.4)

$$\min f(b) = \sum_{i=1}^{p} \ln(b_i) \text{ subject to } F(b) = \mathbb{P}\left(|Z_1^{\mathrm{mle}}|/b_1 \leq 1, \ldots, |Z_p^{\mathrm{mle}}|/b_p \leq 1\right) = 1 - \alpha.$$

Because this problem reaches its minimum at b^*, $\nabla f(b^*)$ is collinear to $\nabla F(b^*)$. Let us set D the matrix $D = \mathrm{diag}(\mathrm{b}_1, \ldots, \mathrm{b}_\mathrm{p})$, we have the following expression for $F(b_1, \ldots, b_p)$

$$\begin{aligned} F(b_1, \ldots, b_p) &= \int_{[-1,1]^p} R \exp\left(-\frac{1}{2} x^T D \Sigma^{-1} D x\right) \det(D) dx, \\ &= \int_{[-1,1]^p} R \exp\left(-\frac{1}{2} x^T D \Sigma^{-1} D x + \ln(\det(D))\right) dx, \end{aligned}$$

with $R = 1/((2\pi)^{p/2}\det(\Sigma)^{1/2})$. Next, the expression of the partial derivative

$$\frac{\partial}{\partial b_i}\left(-\frac{1}{2}x^T D\Sigma^{-1}Dx + \ln(\det(D))\right) = \frac{1}{b_i} - \sum_{j=1}^{p}\Sigma^{-1}_{i,j}x_i x_j b_j,$$

implies that the gradient of F is equal to

$$\begin{aligned}\frac{\partial F}{\partial b_i}(b_1,\ldots,b_p) &= \frac{1}{b_i}F(b_1,\ldots,b_p) - R\sum_{j=1}^{p}\int_{[-1,1]^p}(\Sigma^{-1}_{i,j}x_i x_j b_j)\exp\left(-\frac{1}{2}x^T D\Sigma^{-1}Dx\right)\det(D)dx \\ &= \frac{1-\alpha}{b_i} - R\sum_{j=1}^{p}\int_{[-1,1]^p}(\Sigma^{-1}_{i,j}x_i x_j b_j)\exp\left(-\frac{1}{2}x^T D\Sigma^{-1}Dx\right)\det(D)dx\end{aligned}$$

Thus, $\nabla F(b) = (1-\alpha)\nabla f(b) + v(b)$, where $v(b) \in \mathbb{R}^p$ is the following vector

$$v(b) := \left(\sum_{j=1}^{p}\Sigma^{-1}_{i,j}\int_{[-1,1]^p}x_i x_j b_j^* R\exp\left(-\frac{1}{2}x^T D\Sigma^{-1}Dx\right)\det(D)dx\right)_{1\leqslant i\leqslant p}.$$

Consequently, $\nabla f(b^*)$ and $\nabla F(b^*)$ are collinear if and only if $\nabla f(b^*)$ and $v(b^*)$ are collinear.

$$\begin{aligned}&\exists k \in \mathbb{R} \text{ such that } v(b^*) = k\nabla f(b^*), \\ \Leftrightarrow\ &\forall i \in [\![1,p]\!], \sum_{j=1}^{p}\Sigma^{-1}_{i,j}\int_{[-1,1]^p}x_i b_i^* x_j b_j^* R\exp\left(-\frac{1}{2}x^T D\Sigma^{-1}Dx\right)\det(D)dx = k, \\ \Leftrightarrow\ &\forall i \in [\![1,p]\!], \sum_{j=1}^{p}\Sigma^{-1}_{i,j}\int_{u\in\mathbb{R}^p}u_i u_j\frac{R}{1-\alpha}\exp\left(-\frac{1}{2}u\Sigma^{-1}u\right)\mathbb{1}_{u\in B^*}du = \frac{k}{1-\alpha}. \qquad (2.9)\end{aligned}$$

The expression (2.9) is obtained *via* the change of variables $\forall i \in [\![1,p]\!], u_i = x_i b_i^*$. To conclude, one recognizes that

$$\int_{u\in\mathbb{R}^p}u_i u_j\frac{R}{1-\alpha}\exp\left(-\frac{1}{2}u\Sigma^{-1}u\right)\mathbb{1}_{u\in B^*}du = \mathbb{E}\left(T_i^{b^*}T_j^{b^*}\right) = \mathrm{cov}\left(T_i^{b^*}, T_j^{b^*}\right).$$

Thus the diagonal coefficients of $\Sigma^{-1}\mathrm{var}(T_{b^*})$ are equals to $k/(1-\alpha)$. □

Acknowledgements

The authors are grateful for the real data provided by the following metabolomicians from Toxalim : Cécile Canlet, Laurent Debrauwer and Marie Tremblay-Franco and are grateful to Holger Rauhut for its careful reading. This work is part of the project GMO90+ supported by the grant CHORUS 2101240982 from the Ministry of Ecology, Sustainable Development and Energy in the national research program RiskOGM. Patrick Tardivel is partly supported by a phD fellowship from GMO90+. We also received a grant for the project from the IDEX of Toulouse "Transversalité 2014".

2.10 Commentaires sur la procédure de tests multiples

Dans cette section, nous allons présenter la procédure de tests multiples développée dans Tardivel et al. (2017b) et reprendre la proposition 2.2 avec un point de vue différent. Pour $i \in [\![1, p]\!]$, notons T_i la statistique de test $T_i := \hat{\beta}_i^{\text{mle}}/\text{se}(\hat{\beta}_i^{\text{mle}})$ pour l'hypothèse nulle $\mathcal{H}_i : \beta_i^* = 0$. Notre procédure de tests multiples suggère d'avoir autant de seuils de rejet $s_1, \ldots, s_p$ que d'hypothèses à tester ; l'hypothèse $\mathcal{H}_i$ est rejetée dès que $|T_i| > s_i$. Le FWER est inférieur à α si $\mathbb{P}(\exists i \notin \mathcal{A} \text{ tel que } |T_i| > s_i) \leq \alpha$. Pour garantir l'inégalité précédente, il suffit de choisir des seuils $s_1, \ldots, s_p$ de telle sorte que l'égalité suivante ait lieu

$$\mathbb{P}\left(\bigcap_{1 \leq i \leq p} |T_i| \leq s_i\right) = 1 - \alpha.$$

La région d'acceptation de la procédure de tests multiples est l'hyperrectangle $[-s_1, s_1] \times \cdots \times [-s_p, s_p]$; lorsque $T_1 \in [-s_1, s_1], \ldots, T_p \in [-s_p, s_p]$ aucune hypothèse est rejetée. Intuitivement, plus le volume de la région d'acceptation est petit plus la procédure est puissante. Le seuils optimaux $s_1^*, \ldots, s_p^*$ sont ceux pour lesquelles l'hyperrectangle $[-s_1^*, s_1^*] \times \cdots \times [-s_p^*, s_p^*]$ a un volume minimale. Ces seuils sont obtenus en résolvant le problème d'optimisation sous contrainte suivant

$$\text{minimiser } \prod_{i=1}^{p} s_i \text{ sous la contrainte que } \mathbb{P}\left(\bigcap_{1 \leq i \leq p} |T_i| \leq s_i\right) = 1 - \alpha.$$

Notons T^{s^*} le vecteur aléatoire gaussien tronqué sur l'hyperrectangle $[-s_1^*, s_1^*] \times \cdots \times [-s_p^*, s_p^*]$ de densité

$$f_{T^{s^*}}(u) = \frac{1}{(1-\alpha)\sqrt{(2\pi)^p \det(\Sigma)}} \exp(-u\Sigma^{-1}u) \mathbb{1}_{|u_1| \leq s_1^*} \times \cdots \times \mathbb{1}_{|u_p| \leq s_p^*} du,$$

et $\Sigma = (X^T \Gamma^{-1} X)^{-1}$ la matrice de covariance de maximum de vraisemblance. La proposition 2.2 montre que les coefficient diagonaux de la matrice $\Sigma^{-1}\text{var}(T^{s^*})$ sont tous égaux. En général, les seuils optimaux $s_1^*, \ldots, s_p^*$ ne sont pas tous égaux ainsi, les hypothèses ne sont pas testées avec le même seuil de rejet, ceci est la principale différence entre notre procédure et les procédures classiques (Dunn, 1961; Holm, 1979; Romano and Wolf, 2005).

Bibliographie

Anderson, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proceedings of the American Mathematical Society*, 6(2) :170–176.

Astle, W., De Iorio, M., Richardson, S., Stephens, D., and Ebbels, T. (2012). A bayesian model

of nmr spectra for the deconvolution and quantification of metabolites in complex biological mixtures. *Journal of the American Statistical Association*, 107(500) :1259–1271.

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

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

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*.

Bühlmann, P., Meier, L., and van de Geer, S. (2014). Discussion :"a significance test for the lasso". *Ann. Statist*, 42 :469–477.

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

Dudoit, S. and Van Der Laan, M. J. (2007). *Multiple Testing Procedures with Applications to Genomics*. Springer.

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

G'Sell, M. G., Wager, S., Chouldechova, A., and Tibshirani, R. (2015). Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Sociey : Series B (Statistical Methodology)*, 78(2) :423–444.

Hao, J., Astle, W., De Iorio, M., and Ebbels, T. M. (2012). BATMAN - an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a bayesian model. *Bioinformatics*, 28(15) :2088–2090.

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

Hiriart-Urruty, J.-B. and Lemaréchal, C. (2013). *Convex analysis and minimization algorithms I : fundamentals*, volume 305. Springer Science & Business Media.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2) :65–70.

Janson, L. and Su, W. (2016). Familywise error rate control via knockoffs. *Electronic Journal of Statistics*, 10(1) :960–975.

Jia, J. and Rohe, K. (2012). Preconditioning to comply with the irrepresentable condition. *arXiv preprint arXiv :1208.5584*.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition.

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics*, 42(2) :413–468.

Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of statistics*, 2 :90–102.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3) :1436–1462.

Ravanbakhsh, S., Liu, P., Bjordahl, T. C., Mandal, R., Grant, J. R., Wilson, M., Eisner, R., Sinelnikov, I., Hu, X., Luchinat, C., Greiner, R., and Wishart, D. S. (2015). Accurate, fully-automated NMR spectral profiling for metabolomics. *PLoS ONE*, 10(5) :e0124219.

Rencher, A. C. and Schaalje, G. B. (2008). *Linear models in statistics*. John Wiley & Sons.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469) :94–108.

S̃idàk, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318) :626–633.

Su, W. and Candes, E. (2016). Slope is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3) :1038–1068.

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

Tardivel, P., Servien, R., and Concordet, D. (2017b). A powerful multiple testing procedure in linear Gaussian model. *Submitted*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1) :267–288.

Tulpan, D., Léger, S., Belliveau, L., Culf, A., and Čuperlović-Culf, M. (2011). Metabohunter : an automatic approach for identification of metabolites from ^{1}H-NMR spectra of complex mixtures. *BMC Bioinformatics*, 12(1) :400.

Weljie, A. M., Newton, J., Mercier, P., Carlson, E., and Slupsky, C. M. (2006). Targeted profiling : Quantitative analysis of ^{1}H-NMR metabolomics data. *Analytical Chemistry*, 78(13) :4430–4442.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing : Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7 :2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476) :1418–1429.

Deuxième partie

Application de la procédure de tests multiples à la métabolomique

Chapitre 3

Modélisation d'un spectre de mélange complexe en RMN

Nous avons vu dans la partie précédente (tiré de Tardivel et al. (2017b)) une application peu développée des résultats sur le FWER en métabolomique. Cette application est développée plus en détail dans Tardivel et al. (2017a) qui est le second chapitre de cette partie. Cet article, accepté dans la revue *Metabolomics*, introduit la méthode ASICS, basée sur la procédure de tests multiples décrite dans la première partie de ce manuscrit. Cette nouvelle méthode permet d'identifier et de quantifier des métabolites. Dans l'article de Tardivel et al. (2017a), ASICS est également comparé aux méthodes actuellement utilisées dans la communauté RMN. Cet article étant dédié à un public d'experts en métabolomique, il n'a pas vocation à donner tous les détails sur la modélisation utilisée. Les détails supplémentaires sur la modélisation des spectres sont fournis dans le chapitre suivant.

3.1 Identification et quantification des métabolites

La métabolomique est une science qui s'intéresse à l'identification et la quantification de métabolites (sorte de molécules) que l'on retrouve dans les cellules, les tissus, les fluides biologiques et les organismes. La technique la plus utilisée pour obtenir cette caractérisation est la Résonance Magnétique Nucléaire des protons (RMN). Chaque métabolite possède un spectre RMN qui est caractéristique. Afin d'identifier ces métabolites, les experts utilisent une bibliothèque de spectres de métabolites purs et comparent de visu ces spectres à celui du mélange biologique à analyser. Plus précisément, lorsqu'un expert veut savoir si un métabolite particulier est présent dans un mélange, il vérifie si tous les pics du spectre de ce métabolite se retrouvent dans le spectre du mélange. Cette méthode dépend donc grandement des connaissances de l'expert, notamment du nombre de spectres de métabolites qu'il connaît. Cette identification peut également être rendue délicate par la déformation des spectres (due par exemple à une variation de pH) ou par le chevauchement de certains des pics des métabolites présents dans le

mélange. Voir l'article (Tardivel et al., 2017a) pour plus de détails et des références sur le sujet.

La figure 3.1 est un spectre annoté par des experts en métabolomique.

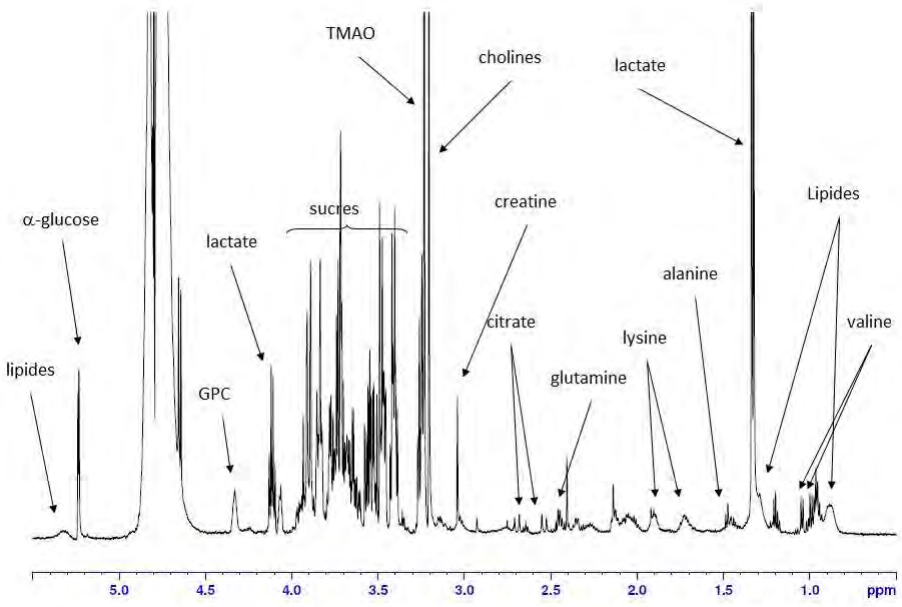


FIGURE 3.1 – Exemple de spectre annoté. On remarque que certains pics de lipides et de valine se superposent.

Le spectre RMN du i^{eme} métabolite pur est représenté par la fonction $f_i : [a, b] \to \mathbb{R}_+$; cette fonction est connue sur une subdivision régulière de l'intervalle $[a, b]$, $a = t_1 \leq t_2 \leq \cdots \leq t_s = b$. Les spectres des métabolites purs ont une aire sous la courbe égale à 1 ainsi, pour tout $i \in [\![1, p]\!]$, la fonction f_i vérifie $\int_a^b f_i(t)dt = 1$. Les conditions d'observations de référence (fréquence, température, PH, ...) des spectres de métabolites purs sont toutes identiques. Dans notre cas, la fréquence de la RMN est de 600.13 MHz, la température et le PH du métabolite pur sont respectivement de 300 K et 7.0.

3.1.1 Le mélange obtenu dans des conditions de référence

Lorsque le mélange est obtenu dans des conditions de référence, le spectre du mélange g, non observé, est une combinaison linéaire des spectres de métabolites purs dont l'expression est la suivante

$$g : t \in [a, b] \mapsto \sum_{i=1}^{p} \alpha_i f_i(t).$$

Les coefficients $\alpha_1, \ldots, \alpha_p$ sont liés aux concentrations des métabolites dans le mélange (*via* le nombre d'atomes d'hydrogène de chaque métabolite). Le spectre du mélange est observé avec du bruit sur la subdivision régulière $a = t_1 \leq \cdots \leq t_s = b$. Le spectre bruité du mélange est

modélisé de la façon suivante

$$\forall j \in [\![1, s]\!], Y(t_j) := \sum_{i=1}^{p} \alpha_i f_i(t_j) + \varepsilon(t_j).$$

Les résidus $\varepsilon(t_1), \ldots, \varepsilon(t_j)$ ne sont pas homoscédastiques, la figure suivante illustre que l'écart-type des résidus a une composante multiplicative.

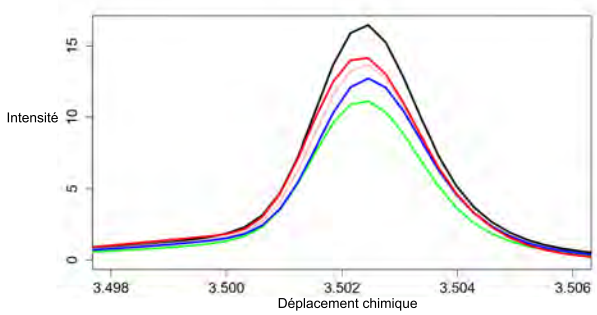


FIGURE 3.2 – Représentation graphique de cinq spectres de glucose pur obtenus dans des conditions identiques. On remarque que l'amplitude du pic varie. Cette variation nous a permis de modéliser le bruit. Plus l'intensité du signal est grande plus le bruit est important. Ceci suggère que le bruit a une composante multiplicative.

Ces observations nous ont permis de modéliser la loi marginales des résidus $\varepsilon(t_1), \ldots, \varepsilon(t_s)$ de la façon suivante $\varepsilon(t_j) \sim \mathcal{N}(0, \sigma_1^2 + \sigma_2^2 g(t_j))$, où σ_1 et σ_2 sont des paramètres connus. La structure de corrélation des résidus sera discutée dans la prochaine section.

3.1.2 Procédure de tests multiples et identification des métabolites

Afin d'identifier les métabolites nous allons tester les hypothèses $\alpha_i = 0$ pour $i \in [\![1, p]\!]$. Le i^{eme} métabolite est identifié lorsque l'hypothèse $\alpha_i = 0$ est rejetée. Les experts en métabolomique souhaiteraient avoir une méthode qui ne fasse aucune mauvaise identification et qui soit capable d'identifier les métabolites ayant une très faible concentration. La FamilyWise Error Rate (FWER) est la probabilité d'avoir un faux positif. Avoir un contrôle de la FWER à un niveau bas nous permettra d'éviter d'avoir de mauvaises identifications. Par ailleurs, avoir une procédure de tests multiples "puissante" nous permettra d'identifier les métabolites ayant des concentrations faibles.

Dans un premier temps nous considérons le modèle simplifié où le spectre du mélange complexe serait obtenu dans des conditions de référence. Dans ces conditions, le spectre du mélange est le vecteur gaussien suivant

$$(Y(t_1), \ldots, Y(t_s)) \sim \mathcal{N}\left(\left(\sum_{i=1}^{p} \alpha_i f_i(t_j)\right)_{1 \leq j \leq s}, \Gamma\right).$$

Pour la loi jointe des résidus, nous avons fait l'hypothèse que les résidus étaient indépendants. Cette hypothèse est simplificatrice en effet, lorsque $|t_j - t_{j'}|$ est presque nulle les résidus $\varepsilon(t_j)$ et $\varepsilon(t_{j'})$ sont corrélés. Sous cette hypothèse d'indépendance, la matrice de covariance des résidus est

$$\Gamma := \operatorname{diag}\left(\sigma_1^2 + \sigma_2^2 \sum_{i=1}^{p} \alpha_i f_i(t_j)\right)_{1 \leq j \leq n}.$$

Les coefficient à estimer $\alpha_1, \ldots, \alpha_p$ étant inconnus, l'expression $\sum_{i=1}^{p} \alpha_i f_i(t_j)$ inconnue peut être remplacée par $Y(t_j)$ qui est connu. Ceci donne donc la matrice de covariance (aléatoire) $\hat{\Gamma} := \operatorname{diag}(\sigma_1^2 + \sigma_2^2 Y(t_j))_{1 \leq j \leq n}$. La procédure de test pour les hypothèses $\mathcal{H}_i : \alpha_i = 0$ avec $i \in [\![1, p]\!]$ est celle décrite dans l'article (Tardivel et al., 2017b). Cette procédure est basée sur l'estimateur

$$\hat{\alpha} := (X^T \hat{\Gamma}^{-1} X)^{-1} X^T \hat{\Gamma}^{-1} Y.$$

Lorsque l'hypothèse $\mathcal{H}_i$ est rejetée, le paramètre α_i (estimé par $\hat{\alpha}_i$) correspondant est significativement non nul.

Les deux raisons qui nous ont motivées à faire l'hypothèse d'indépendance des résidus sont les suivantes :

— Lorsque $\hat{\Gamma}$ est diagonale son inverse est très facile à obtenir, ceci simplifie le calcul de l'estimateur $\hat{\alpha}$.

— Les variances marginales $\operatorname{var}(\hat{\alpha}_1), \ldots, \operatorname{var}(\hat{\alpha}_p)$ ont tendance à être plus grandes lorsque les résidus sont indépendants que lorsqu'ils sont corrélés. Ainsi, en faisant l'hypothèse d'indépendance des résidus, le FWER à tendance à être plus petit que le niveau visé.

3.1.3 Le mélange n'est pas obtenu dans des conditions de référence

Pour avoir un spectre obtenu dans des conditions de référence, il est nécessaire de contrôler plusieurs facteurs dont la fréquence, la température et le PH. En pratique le PH est le facteur le plus difficile à contrôler. Par exemple si le mélange complexe est de l'urine, son PH est compris entre 4.5 et 7.5. Ainsi, un ajustement du PH à 7 est effectué avant l'obtention du spectre du mélange complexe. Malgré cet ajustement, le spectre obtenu n'est pas exactement celui que l'on aurait obtenu dans les conditions de référence (PH 7.0, 300 K, 600.13 MHz,...). La figure 3.3 illustre que la localisation des pics d'un spectre de métabolite peuvent varier suivant les conditions d'observation. Les fonctions déformantes ont été introduites afin de modéliser la variation de la localisation des pics d'un spectre (Veeraraghavan et al., 2006; Wierzbicki et al., 2014). Une fonction déformante sur $[a, b]$ est une fonction $\phi : [a, b] \to [a, b]$ continue, strictement croissante telle que $\phi(a) = a$ et $\phi(b) = b$. Lorsque l'on prend en compte la variation

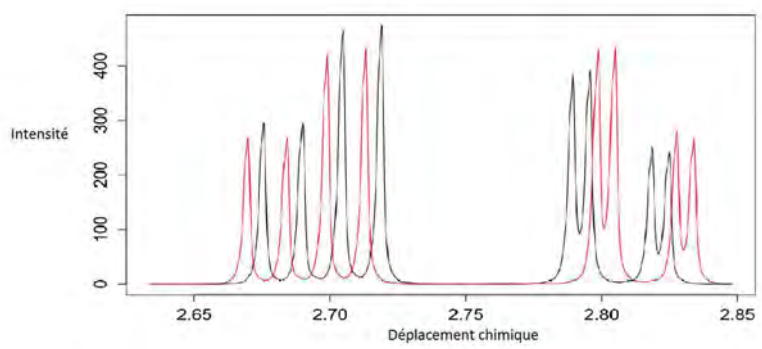


FIGURE 3.3 – Exemple de spectre d'acide aspartic obtenu dans des conditions de température et de PH différentes. L'abscisse de ce graphique représente le déplacement chimique et l'ordonnée représente l'intensité. Entre le spectre noir et le spectre rouge, la taille et la localisation des pics varient.

sur la localisation des pics, on modélise le spectre du mélange de la façon suivante

$$Y(t_j) = \sum_{i=1}^{p} \alpha_i f_i(\phi_i(t_j)) + \varepsilon(t_j), \text{ avec } \varepsilon(t_j) \sim \mathcal{N}\left(0, \sigma_1^2 + \sigma_2^2 \sum_{i=1}^{p} \alpha_i f_i(\phi_i(t_j))\right).$$

La fonction déformante ϕ_i modélise la variation sur la localisation des pics du spectre du i^{eme} métabolite pur par rapport à la localisation des pics du spectre f_i qui est obtenu dans des conditions de référence. Lorsque les fonctions déformantes sont égales à l'identité, on retrouve la modélisation faîte quand le spectre du mélange est obtenu dans les conditions de référence.

En résumé, Y est une combinaison linéaire bruitée des spectres déformés $f_1 \circ \phi_1, \ldots, f_p \circ \phi_p$. Lorsque le mélange n'est pas obtenu dans les conditions de référence (PH 7.0, 300 K, 600.13 MHz,...), les fonctions déformantes $\phi_1, \ldots, \phi_p$ ne sont pas l'identité. Ainsi, avant de chercher à identifier et quantifier les métabolites, il est nécessaire de "corriger" les spectres $f_1, \ldots, f_p$. Dans la partie suivante, nous présentons une étape préliminaire de déformation des spectres. Cet étape "corrige" les spectres $f_1, \ldots, f_p$ en $f_1 \circ \Phi_1^*, \ldots, f_p \circ \Phi_p^*$; les fonctions $\Phi_1^*, \ldots, \Phi_p^*$ sont des approximations des fonctions déformantes $\phi_1, \ldots, \phi_p$. Une fois l'étape de déformation effectuée, l'identification et la quantification des métabolites est faite en utilisant la méthodologie développée dans la section précédente.

3.2 Étape de déformation des spectres

Les spectres RMN nécessitent différentes étapes de pré-traitement comme l'utilisation des séries de Fourier ou la correction de la ligne de base... Pour ces étapes-là, nous avons utilisé des méthodes usuelles qui sont décrites dans Tardivel et al. (2017a). Pour l'étape de déformation

nous avons développé une méthode spécifique qui est décrite en détail dans cette partie.

3.2.1 Fonctions déformantes élémentaires

Afin de proposer une méthode itérative capable d'approximer une fonction déformante quelconque, nous avons introduit les fonctions déformantes élémentaires. La proposition 3.1 montre qu'une fonction déformante quelconque peut être uniformément approchée par une composition de fonction déformantes élémentaires. Un tel résultat justifie que les fonctions déformantes élémentaires sont les bonnes fonctions déformantes à considérer pour la construction d'une méthode itérative.

Définition 3.1 *Soit $I = [c, d]$ un intervalle telle que $I \subset [a, b]$ et $e, f \in I$. On définit la fonction $\psi_I^{e\to f}$*

$$\psi_I^{e\to f}(x) := \begin{cases} \forall x \notin I, \psi_I^{e\to f}(x) = x \\ \forall x \in [c, e], \psi_I^{e\to f}(x) = c + \frac{f-c}{e-c}(x - c) \\ \forall x \in]e, d], \psi_I^{e\to f}(x) = d + \frac{d-f}{d-e}(x - d) \end{cases} .$$

Une telle fonction est appelée fonction déformante élémentaire de l'intervalle $[a, b]$.

Par définition la fonction déformante $\psi_I^{e\to f}$ ne déforme pas en dehors de l'intervalle I. En effet, si g est une fonction définie sur $[a, b]$ alors, $\forall x \in [a, b] \setminus I, g \circ \psi_I^{e\to f}(x) = g(x)$. La figure 3.4 est une illustration de fonction déformante élémentaire

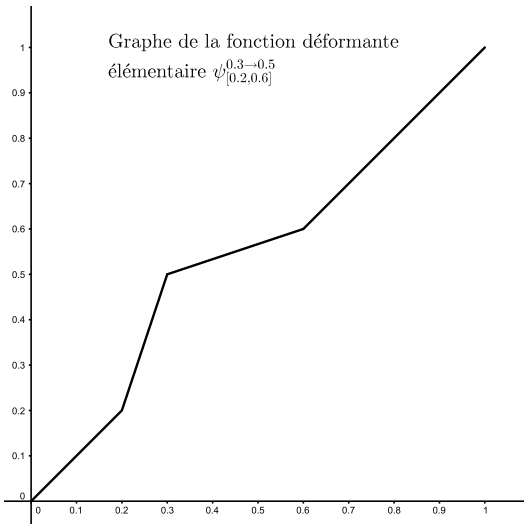


FIGURE 3.4 – Graphe de la fonction déformante élémentaire $\psi_{[0.2,0.6]}^{0.3\to0.5}$ de l'intervalle [0,1] .

Proposition 3.1 *Soit ϕ une fonction déformante de l'intervalle $[a, b]$ alors, pour tout $\epsilon > 0$, il existe un entier $n \geq 1$ et des fonctions déformantes élémentaires $\psi_1, \ldots, \psi_n$ de l'intervalle*

$[a,b]$ *telle que*

$$\sup_{x\in[a,b]}\{|\phi(x)-\psi_1\circ\cdots\circ\psi_n(x)|\}\leq\epsilon$$

Il est facile de démontrer que toute fonction déformante peut être uniformément approchée par une fonction déformante linéaire par morceau. Le lemme suivant montre que toute fonction déformante linéaire par morceau est une composition de fonctions déformantes élémentaires. Ainsi, ce lemme implique la proposition précédente.

Lemme 3.1 *Soient* $n\geq 1$, $a=a_0<a_1<\cdots<a_n<b$, $a=b_0<b_1<\cdots<b_n<b$ *et* $\phi:[a,b]\to[a,b]$ *une fonction telle que* $\forall i\in[\![0,n]\!],\phi(a_i)=b_i$ *et* $\phi(b)=b$ *que l'on prolonge par linéarité en posant*

$$\forall i\in[\![0,n-1]\!],\forall x\in]a_i,a_{i+1}[,\phi(x)=b_i+\frac{b_{i+1}-b_i}{a_{i+1}-a_i}(x-a_i)\text{ et }\forall x\in]a_n,b[,\phi(x)=b_n+\frac{b-b_n}{b-a_n}(x-a_n).$$

alors ϕ *est la composition de fonctions élémentaires suivantes*

$$\phi=\psi_{[a_0,b]}^{a_1\to c_1}\circ\psi_{[a_1,b]}^{a_2\to c_2}\circ\cdots\circ\psi_{[a_{n-1},b]}^{a_n\to c_n}\text{ avec }\forall k\in[\![1,n]\!],c_k=a_{k-1}+\frac{b_k-b_{k-1}}{b-b_{k-1}}(b-a_{k-1})$$

Preuve : Nous allons prouver cette proposition par récurrence. Si $n=1$ alors $\phi=\psi_{[a,b]}^{a_1\to b_1}$, par ailleurs on a

$$c_1=a_0+\frac{b_1-b_0}{b-b_0}(b-a_0)=a+\frac{b_1-a}{b-a}(b-a)=b_1.$$

Ainsi, $\phi=\psi_{[a_0,b]}^{a_1\to c_1}$ donc la proposition est vraie pour $n=1$. Soit $n\geqslant 1$ un entier pour lequel la proposition est vraie montrons que la proposition reste vraie au rang $n+1$.

Soit ϕ une fonction continue croissante affine par morceau telle que $\forall i\in[\![0,n+1]\!],\phi(a_i)=b_i$ et $\phi(b)=b$. On pose ϕ_1 telle que $\forall x\in[a_0,a_n],\phi_1(x)=\phi(x)$ et telle que la restriction de ϕ_1 à l'intervalle $[a_n,b]$ soit affine avec $\phi_1(a_n)=b_n$ et $\phi_1(b)=b$. Ainsi, la fonction ϕ_1 vérifie $\forall i\in[\![0,n]\!],\phi_1(a_i)=b_i$ et $\phi_1(b)=b$. D'après l'hypothèse de récurrence, la fonction ϕ_1 est une composition des fonctions déformantes élémentaires suivantes

$$\phi_1=\psi_{[a_0,b]}^{a_1\to c_1}\circ\psi_{[a_1,b]}^{a_2\to c_2}\circ\cdots\circ\psi_{[a_{n-1},b]}^{a_n\to c_n}.$$

Il reste à vérifier que $\phi=\phi_1\circ\psi_{[a_n,b]}^{a_{n+1}\to c_{n+1}}$. Si $x\leq a_n$ alors,

$$\phi_1\circ\psi_{[a_n,b]}^{a_{n+1}\to c_{n+1}}(x)=\phi_1(x)=\phi(x).$$

Si $x\in[a_n,b]$ alors, la fonction $\phi_1\circ\psi_{[a_n,b]}^{a_{n+1}\to c_{n+1}}$ est continue et affine sur chaque intervalle $[a_n,a_{n+1}]$ et $[a_{n+1},b]$. On vérifie que $\phi_1\circ\psi_{[a_n,b]}^{a_{n+1}\to c_{n+1}}(a_n)=b_n=\phi(a_n)$ et que $\phi_1\circ\psi_{[a_n,b]}^{a_{n+1}\to c_{n+1}}(b)=b=\phi(b)$

enfin on a

$$\begin{aligned}\phi_1 \circ \psi_{[a_n,b]}^{a_{n+1}\to c_{n+1}}(a_{n+1}) &= \phi_1(c_{n+1}),\\ &= \phi_1\left(a_n + \frac{b_{n+1}-b_n}{b-b_n}(b-a_n)\right),\\ &= b_n + \frac{b-b_n}{b-a_n}\frac{b_{n+1}-b_n}{b-b_n}(b-a_n) = b_{n+1} = \phi(a_{n+1}).\end{aligned}$$

Parce que $\phi(x) = \phi_1 \circ \psi_{[a_n,b]}^{a_{n+1}\to c_{n+1}}(x)$ avec $x \in \{a_n, a_{n+1}, b\}$, par linéarité des fonctions ϕ et $\phi_1 \circ \psi_{[a_n,b]}^{a_{n+1}\to c_{n+1}}$ sur chaque intervalle $[a_n, a_{n+1}]$ et $[a_{n+1}, b]$, on en déduit que $\phi = \phi_1 \circ \psi_{[a_n,b]}^{a_{n+1}\to c_{n+1}}$ sur $[a_n, b]$. □

La propriété 3.1 justifie que les fonctions déformantes élémentaires sont adaptées pour le calcul de fonctions déformantes. En revanche, cette propriété ne donne pas de justifications théoriques sur la méthode de calcul des fonctions $\phi_1, \ldots, \phi_p$ que nous allons décrire à la section suivante.

3.2.2 Calcul des fonctions déformantes

Parce que Y est une combinaison linéaire bruitée des spectres déformés $f_1 \circ \phi_1, \ldots, f_p \circ \phi_p$, intuitivement la somme des carrés résiduels $\min_{\alpha_1,\ldots,\alpha_p\in\mathbb{R}} \|Y - \sum_{i=1}^p \alpha_i f_i \circ \phi_i\|^2$ devrait être plus petite que la somme des carrés résiduels $\min_{\alpha_1,\ldots,\alpha_p\in\mathbb{R}} \|Y - \sum_{i=1}^p \alpha_i f_i\|^2$. Cette remarque a motivé la recherche de fonctions déformantes permettant de minimiser la somme des carrés résiduels. Le calcul des fonctions déformantes a été fait de la façon suivante :

On pose

$$(\alpha_1^{(0)}, \ldots, \alpha_p^{(0)}) = \underset{\alpha_1,\ldots,\alpha_p\in\mathbb{R}}{\operatorname{argmin}} \left\| Y - \sum_{i=1}^p \alpha_i f_i \right\|^2.$$

Pour approximer la fonction ϕ_1, on minimise l'expression suivante

$$\min_{\alpha_1,\Phi_1} \left\| Y - \left(\alpha_1 f_1 \circ \Phi_1 + \sum_{i=2}^p \alpha_i^{(0)} f_i\right)\right\|^2. \tag{3.1}$$

On pose Φ_1^* une fonction déformante pour laquelle ce minimum est atteint, notons que rien ne justifie que $\Phi_1^* = \phi_1$. C'est lors de la minimisation du problème (3.1) que l'on emploie des fonctions déformantes élémentaires afin d'obtenir numériquement la fonction déformante Φ_1^*. Enfin, on calcule $\left(\alpha_1^{(1)}, \ldots, \alpha_p^{(1)}\right)$ de la façon suivante

$$\left(\alpha_1^{(1)}, \ldots, \alpha_p^{(1)}\right) = \underset{\alpha_1,\ldots,\alpha_p\in\mathbb{R}}{\operatorname{argmin}} \left\| Y - \left(\alpha_1 f_1 \circ \Phi_1^* + \sum_{i=2}^p \alpha_i f_i\right)\right\|^2.$$

Soit $i < p$, à l'issue de l'étape i, on dispose des coefficients $\left(\alpha_1^{(i)}, \ldots, \alpha_p^{(i)}\right)$ ainsi que des fonctions déformantes $\Phi_1^*, \ldots, \Phi_i^*$. À l'étape $i+1$, on calcule la fonction déformante Φ_{i+1}^* et les coefficients

$\left(\alpha_1^{(i+1)}, \ldots, \alpha_p^{(i+1)}\right)$ de la façon suivante

$$\min_{\alpha_{i+1}, \Phi_{i+1}} \left\| Y - \left(\sum_{j=1}^{i} \alpha_j^{(i)} f_j \circ \Phi_j^* + \alpha_{i+1} f_{i+1} \circ \Phi_{i+1} + \sum_{j=i+2}^{p} \alpha_j^{(i)} f_j \right) \right\|^2 .$$

On pose Φ_{i+1}^* une fonction déformante pour laquel ce minimum est atteint puis on effectue le calcul

$$(\alpha_1^{(i+1)}, \ldots, \alpha_p^{(i+1)}) = \underset{\alpha_1, \ldots, \alpha_p \in \mathbb{R}}{\operatorname{argmin}} \left\| Y - \left(\sum_{j=1}^{i+1} \alpha_j f_j \circ \Phi_j^* + \sum_{j=i+2}^{p} \alpha_i f_i \right) \right\|^2 .$$

La figure 3.5 illustre cette étape de déformation des spectres. Empiriquement, nous avons observé que le pré-traitement qui "corrige" les spectres $f_1, \ldots, f_p$ en $f_1 \circ \Phi_1^*, \ldots, f_p \circ \Phi_p^*$ permettait d'améliorer nettement l'identification et la quantification des métabolites. L'étape de déformation des spectres est l'étape la plus longue de la méthode ASICS. Le temps de calcul est raisonnable, lorsqu'un mélange complexe est analysé à l'aide de 175 spectres, cette étape prend au maximum 1 minute.

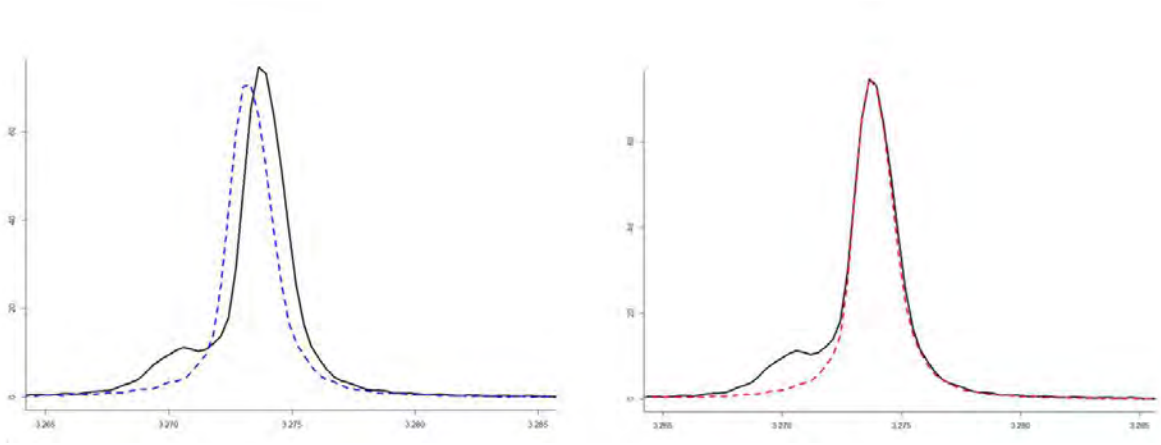


FIGURE 3.5 – Sur la gauche, en ligne pleine, une partie du spectre de l'urine synthétique et en ligne pointillé, le principal pic du spectre de la creatinine. Sur la droite, en pointillé, le spectre de la creatinine après déformation.

Le chapitre suivant présente la méthode ASICS.

Bibliographie

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

Tardivel, P., Servien, R., and Concordet, D. (2017b). A powerful multiple testing procedure in linear Gaussian model. *Submitted*.

Veeraraghavan, A., Chellappa, R., and Roy-Chowdhury, A. K. (2006). The function space of an activity. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 959–968. IEEE.

Wierzbicki, M. R., Guo, L.-B., Du, Q.-T., and Guo, W. (2014). Sparse semiparametric nonlinear model with application to chromatographic fingerprints. *Journal of the American Statistical Association*, 109(508) :1339–1349.

Chapitre 4

ASICS : an automatic method for identification and quantification of metabolites in complex 1D ^{1}H NMR spectra

Ce chapitre est l'article publié dans la revue *Metabolomics*. Le code R de la méthode ASICS ainsi que les ressources supplémentaires de l'article sont disponibles en ligne sur le site de la revue ou sur le site de dépôt en ligne HAL.

Metabolomics (2017) 13:109
DOI 10.1007/s11306-017-1244-5

ORIGINAL ARTICLE

ASICS: an automatic method for identification and quantification of metabolites in complex 1D ^{1}H NMR spectra

Patrick J. C. Tardivel[1] · Cécile Canlet[1,2] · Gaëlle Lefort[3] · Marie Tremblay-Franco[1,2] · Laurent Debrauwer[1,2] · Didier Concordet[1] · Rémi Servien[1]

Received: 10 April 2017 / Accepted: 3 August 2017

Abstract
Introduction Experiments in metabolomics rely on the identification and quantification of metabolites in complex biological mixtures. This remains one of the major challenges in NMR/mass spectrometry analysis of metabolic profiles. These features are mandatory to 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.
Objectives In this article we propose a method, named ASICS, based on a strong statistical theory that handles automatically the metabolites identification and quantification in proton NMR spectra.
Methods A statistical linear model is built to explain a complex spectrum using a library containing pure metabolite spectra. This model can handle local or global chemical shift variations due to experimental conditions using a warping function. A statistical lasso-type estimator identifies and quantifies the metabolites in the complex spectrum. This estimator shows good statistical properties and handles peak overlapping issues.
Results The performances of the method were investigated on known mixtures (such as synthetic urine) and on plasma datasets from duck and human. Results show noteworthy performances, outperforming current existing methods.
Conclusion ASICS is a completely automated procedure to identify and quantify metabolites in ^{1}H NMR spectra of biological mixtures. It will enable empowering NMR-based metabolomics by quickly and accurately helping experts to obtain metabolic profiles.

Keywords Metabolomics · Nuclear magnetic resonance · Identification of metabolites · Quantification of metabolites · NIST plasma

1 Introduction

The development of new technologies has enabled the growth of the omics as a new science field. This refers to a field of biology focused basically on the study of the genome (genomics), the transcriptome (transcriptomics), the proteome (proteomics) or the metabolome (metabolomics) and their modulation by various stimuli. As a common trait, these different approaches produce very large datasets. Consequently, metabolomics experiments are conducted without any hypotheses on the discriminant metabolites to assess the differences between trial groups. Indeed, the whole characterization of the data would lead to intractable computational problems. Nevertheless, efficiency of metabolomics experiments relies on the identification and quantification of metabolites in complex biological mixtures (Blow 2008; Nicholson and Lindon 2008). One of the major challenges in NMR/mass spectrometry analysis of metabolic profiles

Electronic supplementary material The online version of this article (doi:10.1007/s11306-017-1244-5) contains supplementary material, which is available to authorized users.

✉ Patrick J. C. Tardivel
patrick.tardivel@inra.fr

[1] Toxalim, Université de Toulouse, INRA, ENVT, INP-Purpan, UPS, 31027 Toulouse, France

[2] Axiom Platform, MetaToul-MetaboHUB, National Infrastructure for Metabolomics and Fluxomics, 31027 Toulouse, France

[3] GenPhySE, Université de Toulouse, INRA, ENVT, Castanet Tolosan, France

Published online: 18 August 2017

remains the automatic metabolite identification from spectral data (Pontoizeau et al. 2010). Concerning proton NMR spectroscopy, each generated spectrum is usually first divided into intervals called buckets (De Meyer et al. 2008; 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. Today, this identification is mainly manually carried out by an expert, based on his knowledge and on direct comparisons with known metabolite spectra. This identification is tedious, time consuming and expert dependent (Tredwell et al. 2011). Furthermore, some problems, such as peak overlapping, warped spectra due to experimental variations or the high number of possible metabolites for a given chemical shift are very usual in complex mixtures and make identification very challenging. Recently, some automatic methods have been proposed for metabolite identification (see Ravanbakhsh et al. (2015) or Alonso et al. (2015) for a complete review) but none can be considered as an unanimous gold-standard. These methods could allow the use of metabolomics in a standard way using a priori formulated hypotheses on the metabolites by providing an automatic characterization of any complex 1D ^{1}H NMR spectrum.

This article proposes a new method called ASICS (Automatic Statistical Identification in Complex Spectra). ASICS works relying on a library of pure metabolites spectra. The identification of metabolites is performed by comparing the spectrum of the mixture with spectra of the library. These comparisons are carried out using a statistical theory with established statistical properties (Tardivel et al. 2017). ASICS handles experimental problems such as the baseline correction or the variation of chemical shifts. This method is very fast, very competitive with the methods cited above and could help NMR experts in the analysis of complex mixtures. The R code is available as online resource and could be used with any Bruker NMR file of a complex mixture.

2 Materials and methods

2.1 Sample preparation and NMR spectroscopy

A known mixture containing five metabolites in close concentrations and displaying close proton NMR signals was first prepared in replicate (n = 5) to assess the performances of the method. Mean concentrations were 10.11 mM galactose, 4.86 mM GABA (γ-aminobutyric acid), 5.22 mM acetic acid, 20.10 mM L-lysine and 9.97 mM L-tryptophane. The samples were prepared in deuterated water phosphate buffer (pH 7.0).

Performances of ASICS were also assessed on the basis of a homemade synthetic urine sample (CDC 2010) prepared in ultrapure water (see Table S4 and Table S5 for details), duck plasma and a reference human plasma (NIST SRM1950). Details on the duck plasma analysis can be found in Theron et al. (2011) or in Bonnefont et al. (2014) and details on the plasma NIST SRM1950 are gathered as online resource 2.

For NMR analysis, 500 µl of this synthetic urine sample was mixed with 200 µl of phosphate buffer (pH 7.0) prepared in deuterated water and containing 1 mM TSP. The mixture was vortexed, centrifuged at 5000 g for 10 min at 4 °C and 600 µl of supernatant were transferred into a 5 mm NMR tube.

The 1D ^{1}H NMR spectra of 175 reference compounds were collected to build the spectral library (Table S2 in online resource 3). These compounds have been prepared at the concentration of 20 mM in phosphate buffer (0.2 M; pH 7.0) prepared in D_2O/H_2O in a 70:30 ratio (v/v).

All NMR spectra were recorded at 300 K using a Bruker Avance III HD spectrometer (^{1}H frequency: 600.13 MHz, Bruker, Germany) with a 5 mm CQPCI cryoprobe.

^{1}H NMR spectra of synthetic urine sample and reference compounds were recorded using the Noesypr1d NMR sequence for the suppression of water resonance, with a mixing time of 100 ms. A total of 128 transients were collected into 32k data points using a spectral width of 20 ppm, a relaxation delay of 2 s and an acquisition time of 1.36 s. Prior to Fourier Transformation, an exponential line broadening function of 0.3 Hz was applied to the FID.

All spectra were phase and baseline corrected using the Topspin v3.2 software (Bruker, Germany) and were calibrated to TSP signal (δ 0 ppm). We apply BaselineCorrector (Wang et al. 2013) for automatically estimating the baselines of the different spectra.

2.2 Modelling the spectrum of the complex mixture

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 $\phi: I \rightarrow I$ 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, $f \circ \phi$ 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

$$g = \sum_{1 \le i \le p} \alpha_i f_i \circ \phi_i + \varepsilon,$$

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

$$\varepsilon = \sqrt{\sum_{1 \le i \le p} \alpha_i f_i \circ \phi_i}\,\varepsilon_1 + \varepsilon_2\ ,$$

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.

In this model, g is observed, the spectra $f_1, \ldots, f_p$ are known, the α_i's are unknown parameters, the ϕ_i are unknown warping functions, and the noise ε is unobserved.

2.3 Cleansing step

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.

2.4 Identification of metabolites in a complex mixture

The ith metabolite is considered as identified in the complex mixture when its coefficient α_i is greater than zero. The identification of the metabolites belonging to a complex mixture relies on the estimation of the active set A defined as follows

$$A = \left\{ i \in \{1, \ldots, p\} \text{ such that } \alpha_i \neq 0 \right\}.$$

If a sparse estimator (estimator whose some components are exactly zero) $\hat{\alpha} = \left(\hat{\alpha}_1, \ldots, \hat{\alpha}_p\right)$ of $\left(\alpha_1, \ldots, \alpha_p\right)$ was available, the active set could be estimated as

$$A(\hat{\alpha}) = \left\{ i \in \{1, \ldots, p\} \text{ such that } \hat{\alpha}_i \neq 0 \right\}.$$

However, the warping functions $\phi_1, \ldots, \phi_p$ need to be known to obtain a sparse estimator of $\alpha_1, \ldots, \alpha_p$. To solve this problem, ASICS proceeds in two stages.

During the first stage, the warping functions are successively estimated using non sparse estimates of $\left(\alpha_1, \ldots, \alpha_p\right)$. At the beginning of the kth step of this first stage, the estimates of the first k − 1 warping functions $\phi_1^{(1)}, \ldots, \phi_{k-1}^{(k-1)}$ and non-sparse 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 $\alpha_i^{(k-1)}$ indicates the step at which the estimate was obtained.

The kth warping function is estimated by solving the following optimization problem

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

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

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

Figure 1 provides an illustration of this 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 estimated warping functions are used at the second stage to derive lasso-type sparse estimates of $(\alpha_1, \ldots, \alpha_p)$ (Tibshirani 1996; Bühlmann and Geer 2011) by minimizing in $\alpha_1, \ldots, \alpha_p$ the following expression

$$\left\| U\left(g - \sum_{1 \le i \le p} \alpha_i f_i \circ \phi_i^{(i)} \right) \right\|^2 + \lambda \sum_{1 \le i \le p} |\alpha_i|,$$

where U is a linear transformation. This estimation gives a sparse estimation of the proportions $\hat{\alpha}$ and thus an estimation of the active set $A(\hat{\alpha})$. When $\lambda = 0$, the so-called least squares estimator that is not sparse is obtained. Conversely, when the parameter λ is too large, all coefficients $\alpha_1, \ldots, \alpha_p$ become equal to zero. As a consequence the choice of λ plays an important role on the properties of the estimator $A(\hat{\alpha})$. This choice, as well as that of the linear transformation U, is documented in Tardivel et al. (2017) to obtain an estimator of the active set showing good statistical properties. These properties guarantee a reliable identification of metabolites by controlling the two possible sources of errors: identify a metabolite that is not present in the complex mixture (false positive) or not identify a metabolite that is present (false negative). All these properties are based on an identifiability assumption: we assume that the library is identifiable up to a warping function *i.e.*, in the library, there is no weighted sum of two (or more) spectra of pure metabolites (up to a warping function) that could result in a spectrum which could correspond to another metabolite. Assuming this assumption is met, all the properties of our lasso-type estimator applied and the signal overlap of the different metabolites can be handled by ASICS.

2.5 Quantification of the metabolites

Lasso-type estimators of the parameters $\alpha_1, \ldots, \alpha_p$ are known to be biased (Hastie et al. 2009). For this reason, the quantification of metabolites is performed with a least squares method limited to the metabolites identified (i.e. with $\hat{\alpha}_i$ greater than zero) at the previous step. The quantification of the metabolites is obtained through the relative concentrations that could be easily computed form the coefficients $\alpha_1, \ldots, \alpha_p$ and the numbers of hydrogen atoms of the metabolites. The concentrations obtained are given relatively to the largest one. There is no maximum bound to the measured concentration and, according to the different experimentations, ASICS will be able to detect a relative concentration of approximately 1%. Regarding absolute concentrations, and taking into account our equipment and our experimental conditions for data acquisition, metabolites at concentrations lower than 1 μM will be not detected (signal/noise ratio too low) whereas metabolites at concentrations higher than 1 M will result in a broadening of NMR signals, increasing signal overlapping. Any additional reference compound with known concentration is sufficient to obtain the absolute quantities.

This whole procedure has been implemented on a R freeware code and all further results have been obtained using a classical personal computer and the R 3.2.2 version. All presented results could be computed using the code reported as online resource 1. The input parameters are the following: the complex mixture, the exclusion areas (by default [5.1;4.5] ppm to delete the water peak) and the maximum

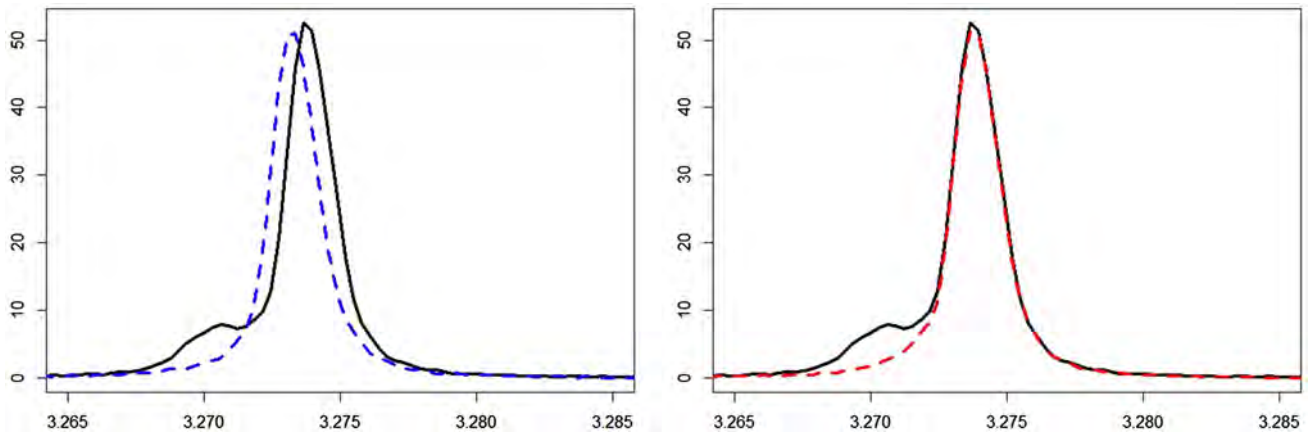


Fig. 1 On the *left in solid line*, the main peak of the creatinine in the spectrum of the synthetic urine. In *dotted line*, the same peak observed on the spectrum of the creatinine before the warping stage. On the *right in dotted line*, the main peak of the creatine spectrum observed after the warping stage

variation M allowed (by default 0.02 ppm). The output delivers the identified metabolites together with their estimated relative concentrations as well as both the estimated and real mixture spectra. The spectrum of the synthetic urine is gathered with an explanation on how to use the code. The interested user could also easily modify or replace our library by a personal/customized one. For a more user-friendly interface, ASICS is also implemented in Galaxy, a dedicated interface for metabolomic data treatment workflows (Guitton et al. 2017).

The accuracy measure reported on the following results has been defined in Ravanbakhsh et al. (2015) by ratio of correct labels (true positives plus true negatives) to the library size.

3 Results and discussion

3.1 Assessment on known mixtures

ASICS was firstly assessed on known mixtures. From the five metabolites mixture reported in Table 1, ASICS always identified the 5 metabolites that are actually present. However, it provided some false positives (between 8 and 11 depending on the considered replicate) yielding an accuracy measure ranging from 96.6 to 98.2%. The quantification results of ASICS on the known mixtures are reported in Table 1.

Due to the quantification of some non-present compounds, the mean estimated proportions were slightly below the real one. Indeed, the false positive compounds were all quantified below 3.3% with respect to the lysine concentration. ASICS thus proved to be robust for the whole spectrum preparation and processing as the final results are not very sensitive to these bias.

3.2 Validation using comparisons with dosages

To validate ASICS quantifications, data available from previous works (Theron et al. 2011; Bonnefont et al. 2014) carried out on duck plasma were used. For two plasmatic metabolites, namely glucose and lactate, concentration ratios provided by ASICS were compared to those obtained by Theron et al. (2011) using a validated enzymatic method. Results presented in Fig. 2 show that the two determination methods are well correlated with a correlation of 0.81.

Indeed, a linear regression applied on this dataset leads to the equation $y = -0.15 + 0.98x$. The intercept and the slope are not statistically different from their expected value (resp. 0 and 1). This good correlation validates the order of magnitude of the quantitative information obtained using ASICS.

Table 1 Results of ASICS on the known mixtures

	Mean of the five estimated proportions	Maximum of the five estimated proportions	Minimum of the five estimated proportions	Real proportion
Lysine	1	1	1	1
Galactose	0.409	0.434	0.392	0.503
Tryptophane	0.446	0.462	0.430	0.496
Acetic acid	0.156	0.125	0.180	0.259
GABA	0.248	0.278	0.237	0.242

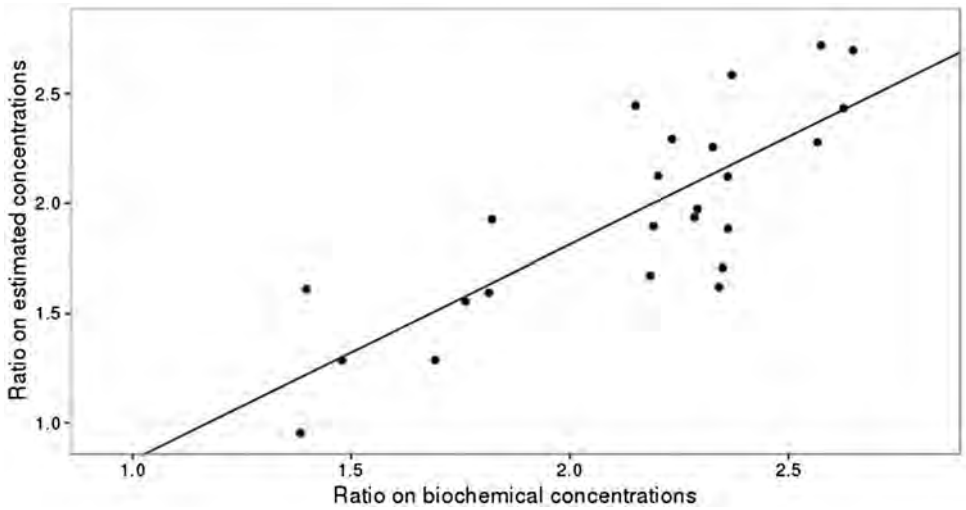


Fig. 2 Glucose and lactate concentration ratios obtained by dosage (*x-axis*) and by using ASICS in the NMR spectrum (*y-axis*) on 24 ducks. The linear regression is *plotted in straight line*

3.3 Comparison with other methods

ASICS was compared to other current methods available for the analysis of complex mixtures NMR spectra (i.e. MetaboHunter, Batman, Bayesil and Chenomx). 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:

a. Synthetic urine containing salts to simulate a typical urine sample with known concentrations of metabolites;
b. Biological human plasma sample (NIST SRM1950 plasma): a reference plasma sample already annotated by NMR experts (Simón-Manso et al. 2013).

3.3.1 Synthetic urine

For the synthetic urine sample, the $10-0.5$ ppm spectral range was used, excluding the region between 6.5 and 4.5 ppm which include the very intense water and urea signals.

In order to accelerate spectra processing with BATMAN, its library was reduced to only 147 metabolites that were also present in our library (Table S3). The library of Bayesil does not contain trimethylamine-*N*-oxide and trigonelline. All the methods have been ran using the default parameter settings as a new user would proceed. The results of identifications are presented in Table 2.

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 by 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. As indicated in Table 2, 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.

Four metabolites (namely ascorbic acid, L-glutamine, malonic acid and formic acid) were not identified by ASICS. This can be due to different reasons. The missing of ascorbic acid is probably due to an experimental problem: its corresponding peaks are not present in the spectrum and this metabolite was identified neither by Bayesil nor by Chenomx. One can assume that the ascorbic acid has been degraded as it is known to be an easily oxidisable metabolite. The L-glutamine was only identified by Bayesil with an unrealistic quantification. ASICS is missing this compound likely because its related signals are located in a range of the spectrum displaying many signals and thus, they may have been falsely attributed to other metabolites. For malonic acid, this can be attributed to acidic hydrogen–deuterium exchange occurring in deuterated water (D_2O). Indeed, ^1H NMR spectra of malonic acid and urine sample have been obtained with different proportions of D_2O, namely 70% for the pure compound, and 30% for the urine sample. In the spectrum of malonic acid acquired in 70% D_2O, a triplet is observed at 3.11 ppm corresponding to the CHD signal, together with a singlet at 3.13 ppm corresponding to the CH_2

Table 2 Comparison of the identification 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	2 min 38 s
MetaboHunter	4	51	17	795	92	867	<1 min
Batman	21	125	0	1	18	147	74 h
Bayesil	12	17	7	53	73	89	10 min 48 s
Chenomx	15	48	6	269	54	338	<1 min

signal (proportion 56/44 respectively). In the urine sample, the triplet signal is very weak compared to the singlet (proportion 14/86), due to the lower exchange rate, explaining why ASICS was unable to identify this metabolite. This underlines the fact that, for this kind of compound, a great attention has to be paid to this phenomenon and, that ideally, the library may include a spectrum for each potential matrix. Finally, ASICS did not identify formic acid whereas the other methods did. This highlights one of the limitations of our method: since it relies on a lasso-type estimator (and, by consequence, mean square thresholded estimator), it uses the area under the curve assigned to each metabolite, which is proportional to the concentration of the metabolite and to its number of hydrogen atoms. Therefore, a metabolite bearing a single non-exchangeable hydrogen atom, such as formic acid, displays a smaller area under the curve than another metabolite at the same concentration but with a higher number of hydrogen atoms. This may explain why formic acid was not identified by ASICS whereas other metabolites (e.g. L-tyrosine with seven hydrogen atoms) were identified due to their greater area under the curve.

Performances of the various tested software were also compared in term of quantification. Results are summarized in Table 3.

The quantifications provided by ASICS or by Chenomx both fit quite well the order of magnitude of the real proportion of the different metabolites. Conversely, some quantification results are very far from the real proportion for Bayesil (citric acid, L-glutamine …) and for BATMAN (citric acid, trimethylamine-*N*-oxide …).

The same results were obtained using the synthetic urine without salts, highlighting the robustness of ASICS (data not shown). The Bruker file of the synthetic urine spectrum is included in online resource 1 and, thus, all the presented results can be easily recovered.

The above results suggest that ASICS represents the best trade-off between method accuracy (for both identification and quantification) and computational time. However, these results have to be analyzed with caution. First, depending on the experimental conditions, it is likely that a given method cannot be always better than others. Second, each method runs with its own specific library, which produces a bias in the comparison. Then, each method has been computed with default parameters and it is likely that, spending time to set these parameters, the performances of the different methods may be improved.

Table 3 Comparison of the relative quantification of the four methods on the synthetic urine

Compound	Real proportion	ASICS's proportion	Bayesil's proportion	Chenomx's proportion (automatic fit)	BATMAN's proportion
Creatinine	1	1	1	1	1
Citric acid	0.434	0.693	12.38	Not identified	0.089
Hippuric acid	0.338	0.344	Not identified	0.312	0.072
Trimethylamine-*N*-oxide	0.286	0.439	Not in library	0.311	1.311
Ascorbic acid	0.156	Not identified	Not identified	Not identified	0.568
Malonic acid	0.073	Not identified	Not identified	0.015	0.058
Ethanolamine	0.062	0.044	Not identified	Not identified	0.259
L-Lysine	0.044	0.076	0.512	Not identified	0.214
Dimethylamine	0.047	0.054	0.079	Not identified	0.025
Betaine	0.042	0.053	0.246	0.055	0.754
L-Alanine	0.042	0.056	0.219	0.045	0.230
D-Glucose	0.041	0.059	0.705	0.046	0.023
Guanidinoacetic acid	0.033	0.052	Not identified	0.022	0.226
L-Carnitine	0.033	0.045	0.159	0.023	0.029
L-Glutamine	0.032	Not identified	4.100	Not identified	0.342
Acetic acid	0.032	0.031	Not identified	0.035	0.129
Glycine	0.031	0.036	0.089	0.032	0.304
Lactic acid	0.028	0.027	0.307	0.018	0.025
Trigonelline	0.026	0.011	Not in library	0.018	0.044
Formic acid	0.017	Not identified	0.006	0.029	0.007
L-Tyrosine	0.012	0.024	Not identified	0.014	0.571

3.3.2 NIST plasma

The NIST plasma sample is of particular interest since it represents a real biological sample and it has been extensively studied and characterized by several teams, making available several results on metabolites identification.

As the composition of the NIST plasma is still an open-question, it cannot be used to assess the superiority of any method. Nevertheless, it could be interesting to compare the different results to highlight the potential benefits of these automatic approaches. From the NIST plasma sample, 27 compounds were identified by an NMR expert at level 1 using the 1D ^{1}H and 2D NMR spectra of plasma NIST sample and reference compounds. All the main compounds identified by the experts were also identified by ASICS whereas it is not the case for the other methods. The quantification of these compounds by ASICS provides an accurate order of magnitude. BAYESIL identified 44 compounds (20 in common with the expert at level 1), Chenomx identified 78 compounds (17 in common with the expert at level 1) whereas Simón-Manso et al. (2013) identified 39 compounds in filtered plasma (21 in common with the expert at level 1). In addition to the 21 compounds common with Simón-Manso et al. (2013), ASICS allowed identifying L-serine and GPC that were further confirmed by the NMR experts at level 1 using ^{1}H and ^{13}C chemical shifts compared with reference compounds. Furthermore, ASICS also pointed out 21 other compounds that were not identified at level 1 by the NMR expert. These compounds could be false positives (i.e. not present in the NIST plasma) or new detections. Some of these compounds (TMAO, L-ornithine and pantothenic acid) have been detected by ASICS and Chenomx but not by the expert at level 1, and so further work is required to assess the potential presence of these compounds. All these results show that these automatic approaches constitute a helpful tool for NMR experts, but have to be used with a careful control.

4 Conclusion

In this article we propose a method able to identify and quantify metabolites in a complex mixture of NMR 1D ^{1}H spectra. The warping strategy implemented in ASICS could deal with local modifications (including translations or more complex geometrical transformations) of the complex mixture spectra. ASICS proved to be helpful to save time for NMR experts, providing a useful method for the use of metabolomics in a standard way using a priori hypotheses on metabolites. However, ASICS also obviously still has limitations. For example, the method for correcting the baseline is likely to provide poor results in spectrum areas displaying a high number of peaks. Theoretically, identification and quantification methods require the library to contain all spectra of metabolites contained in the mixture. In practice, it is not possible to make such an assumption and a library containing the main metabolites spectra in terms of concentration yields satisfactory results for our method. Then, the estimation of warping functions could depend on the order in which they are estimated. In theory, it would be better to estimate all of them simultaneously, but this computation cannot be carried out within a reasonable time. The quantification step seems to be sensitive to the variability of the NMR spectrum and only provides an order of magnitude for the metabolites concentrations. Additionally, like all other automatic methods, metabolites with overlapping single resonances (for example formic acid, acetic acid and succinate) would potentially be difficult to identify and to quantify. The complex mixture spectrum also needs to be recorded at the same pH as the library spectra to reduce the potential variations of chemical shift. As pointed out using the synthetic urine sample, NMR spectra of metabolites that are very sensitive to proton exchanges with deuterated water (e.g. malonic acid), also need careful attention. Recording the spectra in the same conditions for both standard compounds and complex mixture would help overcoming potential differences between the two spectra.

Nevertheless, the study of examples of known compositions and comparisons with other methods show that this method provides better results than other existing methods. ASICS was also tested on two real matrices, providing good results on duck plasma and on the NIST plasma sample. Furthermore, ASICS is completely automated, freely available and the library of metabolites may easily be upgraded or replaced by an interested researcher. Finally, we hope that ASICS will enable the wealth of new applications of NMR by quickly and accurately help NMR experts for the study of metabolic profiles.

Acknowledgements This work is part of the project GMO90+ (Grant CHORUS 2101240982) from the French Ministry of Ecology, Sustainable Development and Energy within the national research program RiskOGM. Patrick Tardivel is partially supported by a PhD fellowship from GMO90+. The IDEX of Toulouse "Transversalité 2014" is thanked for its support to this project. The authors also thank the French National Infrastructure of Metabolomics and Fluxomics (MetaboHUB-ANR-11-INBS-0010) for their support. The authors thank Alyssa Bouville and Roselyne Gautier for help in the sample preparation and NMR analyses.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest regarding this work.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Alonso, A., Marsal, S., & Julià, A. (2015). Analytical methods in untargeted metabolomics: State of the art in 2015. *Frontiers in Bioengineering and Biotechnology, 3*, 23.

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

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

Blow, N. (2008). Metabolomics: Biochemistry's new look. *Nature, 455*(7213), 697–700.

Bonnefont, C. M., Guerra, A., Théron, L., Molette, C., Canlet, C., & Fernandez, X. (2014). Metabolomic study of fatty livers in ducks: Identification by ^{1}H-NMR of metabolic markers associated with technological quality. *Poultry Science, 93*(6), 1542–1552.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer: New York.

CDC (Center for Disease Control and Prevention). (2010). *Bisphenol A and other environmental phenols and Parabens in urine*. https://www.cdc.gov/nchs/data/nhanes/nhanes_07_08/eph_e_met_phenols_parabens.pdf.

De Meyer, T., Sinnaeve, D., Van Gasse, B., Tsiporkova, E., Rietzschel, E. R., De Buyzere, M. L., et al. (2008). NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Analytical Chemistry, 80*, 3783–3790.

Guitton, Y., Tremblay-Franco, M., Le Corguillé, G., Martin, J.-F., Pétéra, M., Roger-Mele, P., et al. (2017). Create, run, share, publish, and reference your LC-MS, GC-MS, and NMR data analysis workflows with Workflow4Metabolomics 3.0, the Galaxy online e-infrastructure for metabolomics. *International Journal of Biochemistry and Cell Biology*. doi:10.1016/j.biocel.2017.07.002.

Hao, J., Astle, W., De Iorio, M., & Ebbels, T. M. D. (2012). BATMAN—an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. *Bioinformatics, 28*(15), 2088–2090.

Hao, J., Liebeke, M., Astle, W., De Iorio, M., Bundy, J. G., & Ebbels, T. M. D. (2014). Bayesian deconvolution and quantification of metabolites in complex 1D NMR spectra using BATMAN. *Nature Protocols, 9*, 1416–1427.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer Series in Statistics.

Nicholson, J. K., & Lindon, J. C. (2008). Systems biology: Metabonomics. *Nature, 455*(7216), 1054–1056.

Pontoizeau, C., Herrmann, T., Toulhoat, P., Elena-Herrmann, B., & Emsley, L. (2010). Targeted projection NMR spectroscopy for unambiguous metabolic profiling of complex mixtures. *Magnetic Resonance in Chemistry, 48*(9), 727–733.

Ravanbakhsh, S., Liu, P., Bjordahl, T. C., Mandal, R., Grant, J. R., Wilson, M., et al. (2015). Accurate, fully-automated NMR spectral profiling for metabolomics. *PLoS ONE, 10*(5), e0124219.

Simón-Manso, Y., Lowenthal, M. S., Kilpatrick, L. E., Sampson, M. L., Telu, K. H., Rudnick, P. A., et al. (2013). 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.

Tardivel, P.J., Servien, R., & Concordet, D. (2017). *A powerful mutiple testing procedure in linear Gaussian model*. https://hal.archives-ouvertes.fr/hal-01322077.

Theron, L., Fernandez, X., Marty-Gasset, N., Pichereaux, C., Rossignol, M., Chambon, C., et al. (2011). Identification by proteomic analysis of early post-mortem markers involved in the variability in fat loss during cooking of mule duck "foie gras". (2011). *Journal of Agricultural and Food Chemistry, 59*, 12617–12628.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58*(1), 267–288.

Tredwell, G. D., Behrends, V., Geier, F. M., Liebeke, M., & Bundy, J. G. (2011). Between-person comparison of metabolite fitting for NMR-based quantitative metabolomics. *Analytical Chemistry, 83*(22), 8683–8687.

Tulpan, D., Léger, S., Belliveau, L., Culf, A., & Čuperlović-Culf, M. (2011). MetaboHunter: an automatic approach for identification of metabolites from 1 H-NMR spectra of complex mixtures. *BMC Bioinformatics, 12*(1), 1.

Wang, K. C., Wang, S. Y., Kuo, C. H., & Tseng, Y. J. (2013). Distribution-based classification method for baseline correction of metabolomic 1D proton nuclear magnetic resonance spectra. *Analytical Chemistry, 85*(2), 1231–1239.

Weljie, A. M., Newton, J., Mercier, P., Carlson, E., & Slupsky, C. M. (2006). Targeted profiling: Quantitative analysis of ^{1}H NMR metabolomics data. *Analytical Chemistry, 78*(13), 4430–4442.

Troisième partie

Sparsest representations of a vector in a family spanning $\mathbb{R}^n$

Chapitre 5

Sparsest representations of the expected value of a linear model response

First, I am going to briefly introduce the works Perrot-Dockès et al. (2017a,b) who analysed sap coming from three varieties of tree. Contrarily to us, the metabolites of sap tree are known ; they only need to quantify the concentration of each metabolite. To measure these concentrations, they have used a liquid chromatography-mass spectrometry. Such a technique is only used for quantification but not for identification.

5.1 High-dimensional data in metabolomics

Perrot-Dockès et al. aimed to find the metabolites for which the concentrations are significantly different among these three varieties of tree. For the variety $v \in V$, the i^{th} measurement of the metabolite $j \in [\![1, p]\!]$ was modelled as follows : $Y_{v,i}^{(j)} = \mu_v^{(j)} + \varepsilon_{v,i}^{(j)}$, where $Y_{v,i}^{(j)}$ is observed, the expected value $\mu_v^{(j)}$ and the residual term $\varepsilon_{v,i}^{(j)}$ are unknowns ($p = 1000$ in this application). Let n be the total number of sap samples analysed ($n = 30$ in this application), the following modification of the previous model allows to obtain a sparse parameter $\tilde{\mu}_v^{(j)}$ defined hereafter

$$\tilde{Y}_{v,i}^{(j)} = Y_{v,i}^{(j)} - \frac{1}{n}\sum_{v,i} Y_{v,i}^{(j)} = \mu_v^{(j)} + \varepsilon_{v,i}^{(j)} - \frac{1}{n}\sum_{v,i}\left(\mu_v^{(j)} + \varepsilon_{v,i}^{(j)}\right) = \tilde{\mu}_v^{(j)} + \tilde{\epsilon}_{v,i}^{(j)}.$$

When the expected value of the concentration of the metabolite j does not depend on the variety $v \in V$ we have $\forall v \in V, \tilde{\mu}_v^{(j)} = 0$, namely

$$\text{If there exists } c \in \mathbb{R} \text{ such that } \forall v \in V, \mu_v^{(j)} = c \text{ then } \forall v \in V, \tilde{\mu}_v^{(j)} = 0.$$

Perrot-Dockès et al. assumed that for the most of element $j \in [\![1,p]\!]$, the concentration of metabolite j did not depend on the variety implying thus $\forall v \in V, \tilde{\mu}_v^{(j)} = 0$. Related to the sparsity of the parameter $\tilde{\mu} := (\tilde{\mu}_v^{(j)})_{1\leq j\leq p, v\in V}$, they also assumed that the irrepresentable condition held for $\tilde{\mu}$. Under this assumption, with a tuning parameter λ_n correctly chosen, they proved the sign consistency of the lasso estimator $\hat{\mu}(\lambda_n)$ namely $\lim_{n\to+\infty} \mathbb{P}(\text{sign}(\hat{\mu}(\lambda_n)) = \text{sign}(\tilde{\mu})) = 1$. As a conclusion, the works of Perrot-Dockès et al. illustrate that the estimation of a sparse parameter in the high-dimensional linear model is a challenging issue in metabolomics.

5.2 High-dimensional linear model

Let Y be a Gaussian vector of $\mathbb{R}^n$ distributed according to $\mathcal{N}(m, \sigma^2 Id_n)$ and let $X_1, \ldots, X_p \in \mathbb{R}^n$ be fixed explicative variables. Assume that Y is observed, σ^2 is known and m is unknown. In the linear model, the classical assumption is $m \in \text{vect}(X_1, \ldots, X_p)$. There are two possibilities :

1. The family $X_1, \ldots, X_p$ is linearly independent. In this case, there exists a unique parameter $\beta^* \in \mathbb{R}^p$ such that $m = \beta_1^* X_1 + \cdots + \beta_p^* X_p$. This expression allows to rewrite Y as $Y = X\beta^* + \varepsilon$ with X a $n \times p$ matrix whose columns are $X_1, \ldots, X_p$ and ε a centered Gaussian vector distributed according to $\mathcal{N}(0, \sigma^2 Id_n)$.
2. The family $X_1, \ldots, X_p$ is not linearly independent, such situation always occurs when $p > n$. The parameter β^* given previously is not unique. In a lot of work, the unknown expected value m is written as $m = X\beta^*$; additional assumptions are assumed to avoid the problem of the identifiability of β^*. For example, these assumptions are the following
   — The cardinal $\text{card}\{i \in [\![1,p]\!] \mid \beta_i^* \neq 0\}$ is very small (Bickel et al., 2009; Lounici, 2008)
   — The parameter β^* satisfies the irrepresentable condition (Ollier and Viallon, 2017; Perrot-Dockès et al., 2017b).

Instead to assume $m = X\beta^*$ and try to estimate the parameter β^*, Meinshausen proposes to estimate a l^1 sparse representation of m defined by

$$\underset{\beta\in\mathbb{R}^p}{\text{argmin}} \sum_{i=1}^{p} |\beta_i| \text{ under the constraint that } X\beta = m. \tag{5.1}$$

Without any other conditions, the set of solutions of (5.1) can be empty or can have several elements. However, as soon as the columns of X span $\mathbb{R}^n$, this equation admits at least a solution and a unique solution when the general position condition holds for X.

The convex problem (5.1) is usually used to recover the sparsest representation of m in X that is the solution of the following intractable problem (5.2)

$$\underset{\beta\in\mathbb{R}^p}{\text{argmin}} \, \text{card}\{i \in [\![1,p]\!] \mid \beta_i \neq 0\} \text{ under the constraint that } X\beta = m. \tag{5.2}$$

As explained by Meinshausen, under some conditions (null space property (Donoho and Elad, 2003; Gribonval and Nielsen, 2003), restricted isometry property (Candes, 2008)...) these problems (5.1) and (5.2) have the same solution. In the noiseless case when $\sigma = 0$ implying thus $Y = m$, the problem (5.1) is convex consequently, its solutions could be efficiently obtained. Contrarily to the problem (5.1), the problem (5.2) is not convex. Even in the noiseless case, solving (5.2) remains a challenging issue.

I think that improvement of methods to solve the problem (5.2) in the noiseless case could provide better methods to estimate the solutions of (5.2) when σ is no longer null.

As we have already said, there are conditions for which the solutions of the problem (5.1), called l^0 minimization, are the same as the solutions of the problem (5.2) called l^1 minimization. We give the bibliography of these conditions in the following section. In this section, we are going to use the same notations as those used in the Tardivel et al. (2017).

5.3 l^0 minimization and l^1 minimization

Let $d_1, \ldots, d_p$ be a family of vectors of $\mathbb{R}^n$ such that $\text{vect}\{d_1, \ldots, d_p\} = \mathbb{R}^n$. When $p > n$ this family is not linearly independent and consequently the set of representations of a vector $y \in \mathbb{R}^n$ in the family $d_1, \ldots, d_p$ defined hereafter

$$\{x \in \mathbb{R}^p \mid x_1 d_1 + \cdots + x_p d_p = y\}$$

is infinite. Let us denote D the matrix whose columns are $d_1, \ldots, d_p$ and $\|x\|_0 := \text{card}\{i \in [\![1, p]\!] \mid x_i \neq 0\}$, the sparsest representations are the solutions of the following problem

$$\underset{x \in \mathbb{R}^p}{\text{argmin}} \|x\|_0 \text{ subject to } Dx = y. \tag{$\mathcal{P}_0$}$$

This problem is called the l^0 minimization. An intuitive method to solve this problem is to compute $\tilde{x} = \tilde{D}^{-1} y$ for each $n \times n$ invertible submatrices $\tilde{D}$ of D. The elements $\tilde{x}$ for which $\|\tilde{x}\|_0$ is minimal provide the solutions of $\mathcal{P}_0$. Because there are $\binom{p}{n}$ $n \times n$ submatrices, this intuitive method is only tractable when n and p are together very small. To deal with the case when p is large or n is large, other approaches such as the l^1 minimization have been developed to obtain the sparsest representations. The l^1 minimization also called basis pursuit (Donoho and Elad, 2003) is the following problem

$$\underset{x \in \mathbb{R}^p}{\text{argmin}} \|x\|_1 \text{ subject to } Dx = y. \tag{$\mathcal{P}_1$}$$

This problem is convex, one can rewrite it as a linear programming problem (Foucart and Rauhut (2013) page 63). Even if n or p are large, linear programming problems are efficiently solved by using, for example, the R package lpSolve. Now, we give the bibliography of conditions

insuring a representation x of the vector y in the family $d_1, \ldots, d_p$ is a solution of $\mathcal{P}_0$, $\mathcal{P}_1$ or simultaneously $\mathcal{P}_0$ and $\mathcal{P}_1$.

5.3.1 Conditions to have a solution for $\mathcal{P}_0$

In this section, we provide conditions so that a representation be the only one sparsest representation. The first condition relies on the spark of the matrix D.

Definition 5.1 *The spark of the matrix D is defined as*

$$\text{spark}(D) := \min_{h \in \ker(D), h \neq 0} \|h\|_0.$$

This definition allows to state the following proposition (Donoho and Elad, 2003; Gribonval and Nielsen, 2003).

Proposition 5.1 (spark condition) *Let x be a representation of y in D such that $\|x\|_0 <$ $\text{spark}(D)/2$ then x is the unique solution of $\mathcal{P}_0$.*

The computation of the $\text{spark}(D)$ is intractable. However when the unique representation property (URP) holds for the matrix D then $\text{spark}(D) = n + 1$. The following definition and the following proposition are given in Woodworth and Chartrand (2016).

Definition 5.2 *The URP condition holds for the matrix D if for all subset I with* $\text{card}(I) \leq n$ *the family $(d_i)_{i \in I}$ is linearly independent.*

The proposition below shows that the URP is generic for a matrix D.

Proposition 5.2 *Let D be a random matrix with a continuous distribution onto the $n \times p$ matrix with $n \leq p$ then almost surely the URP holds.*

This proposition shows that the URP is a very weak condition. As a consequence, in practice, when a representation x satisfies $\|x\|_0 < (n+1)/2$, this representation is the unique solution of $\mathcal{P}_0$.

5.3.2 Conditions to have solution for $\mathcal{P}_1$

In this section, we provide conditions so that a representation be the solution of $\mathcal{P}_1$. The following proposition is a result of Daubechies et al. (2010), this proposition provides a characterisation of the basis pursuit solutions.

Proposition 5.3 (characterisation of the basis pursuit solution) *Let x be a representation of y in D, then :*

1. *The representation x is a solution of $\mathcal{P}_1$ if and only if*

$$\forall h \in \ker(D), h \neq 0, \left| \sum_{i \in \text{supp}(x)} \text{sign}(x_i) h_i \right| \leq \sum_{i \notin \text{supp}(x)} |h_i|.$$

2. *The representation x is the unique solution of $\mathcal{P}_1$ if and only if*

$$\forall h \in \ker(D), h \neq 0, \left| \sum_{i \in \text{supp}(x)} \text{sign}(x_i) h_i \right| < \sum_{i \notin \text{supp}(x)} |h_i|.$$

The null space property (Donoho and Elad, 2003; Gribonval and Nielsen, 2003) is then a sufficient condition, derived from the previous proposition, so that a representation x to be the unique solution of $\mathcal{P}_1$.

Proposition 5.4 (Null space property) *Let x be a representation of the vector y in D. If the following inequality holds*

$$\forall h \in \ker(D), h \neq 0, \sum_{i \in \text{supp}(x)} |h_i| < \sum_{i \notin \text{supp}(x)} |h_i|$$

then x is the unique solution of $\mathcal{P}_1$.

The irrepresentable condition is also a well known condition to have the sign consistency of the lasso estimator (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Zou, 2006). In the proposition, we show that if the irrepresentable condition holds for a representation x then x is the unique solution of $\mathcal{P}_1$.

Proposition 5.5 (Irrepresentable condition) *Let x be a representation of y in D such that the family $(d_i)_{i \in \text{supp}(x)}$ is linearly independent. Let S be the set $S := \text{supp}(x)$, let us denote D_S and D_{S^c} be respectively the matrices whose columns are $(d_i)_{i \in S}$ and $(d_i)_{i \notin S}$. If the following inequality occurs*

$$\left\| D_{S^c}^T D_S (D_S^T D_S)^{-1} \text{sign}(x_S) \right\|_\infty < 1$$

then x is the unique solution of $\mathcal{P}_1$.

The inequality $\left\| D_{S^c}^T D_S (D_S^T D_S)^{-1} \text{sign}(x_S) \right\|_\infty < 1$ is called the irrepresentable condition. In the book of Bühlmann and van de Geer (2011) or in the article of van de Geer and Bühlmann (2009), it is proved that the irrepresentable condition implies the compatibility condition (see the article (van de Geer and Bühlmann, 2009) for the definition). Furthermore, as explained by Meinshausen (2015) and proved by Raskutti et al. (2010), the null space property is a particular case of the compatibility condition. So, I guess that the result of the proposition 5.5 is at least known for some statisticians. However, to my knowledge, a straightforward proof of the proposition 5.5 has never been given.

Proof of the proposition 5.5 : Let h_S and h_{S^c} be respectively the vectors $(h_i)_{i\in S}$ and $(h_i)_{i\notin S}$. For all $h \in \ker(X), h \neq 0$ we have

$$\begin{aligned}\left|\sum_{i\in\text{supp}(x)} \text{sign}(x_i)h_i\right| - \sum_{i\notin\text{supp}(x)} |h_i| &= |h_S^T sign(x_S)| - \|h_{S^c}\|_1 \\ &= |h_S^T D_S^T D_S (D_S^T D_S)^{-1} sign(x_S)| - \|h_{S^c}\|_1.\end{aligned}$$

Because $0 = Dh = D_S h_S + D_{S^c} h_{S^c}$, one deduces that

$$\begin{aligned}\left|\sum_{i\in\text{supp}(x)} \text{sign}(x_i)h_i\right| - \sum_{i\notin\text{supp}(x)} |h_i| &= |-h_{S^c}^T D_{S^c}^T D_S (D_S^T D_S)^{-1} sign(x_S)| - \|h_{S^c}\|_1, \\ &\leq \|h_{S^c}\|_1 \|D_{S^c}^T D_S (D_S^T D_S)^{-1} sign(x_S)\|_\infty - \|h_{S^c}\|_1, \\ &< \|h_{S^c}\|_1 - \|h_{S^c}\|_1 = 0.\end{aligned}$$

The characterisation of the basis pursuit solution given in the proposition 5.3 allows to conclude that x is the unique solution of $\mathcal{P}_1$. □

If x is a representation of y in D for which a weaker condition than the irrepresentable condition holds then, x is a unique solution of $\mathcal{P}_1$. An exhaustive list of conditions implying the irrepresentable condition is given in the article of van de Geer and Bühlmann (2009) or in the book of Bühlmann and van de Geer (2011) page 177.

5.3.3 Conditions to have the same solution for both $\mathcal{P}_1$ and $\mathcal{P}_0$

In this section, we give conditions so that a representation x be simultaneously the solution of $\mathcal{P}_0$ and $\mathcal{P}_1$. The first condition relies on the mutual coherence.

Definition 5.3 *Let us assume that for all $i \in [\![1, p]\!]$, $\|d_i\|_2 = 1$. The coherence of the matrix D is*

$$M = \max_{i\neq j} \langle d_i, d_j \rangle.$$

The coherence of the matrix D is easy to compute. As a consequence, the mutual coherence condition is easily checkable.

Proposition 5.6 (mutual coherence condition) *Let us assume that columns $(d_i)_{1\leq i\leq p}$ are such that $\forall i \in [\![1, p]\!]$, $\|d_i\|_2 = 1$. Let x be a representation of y in D such that $\|x\|_0 < (1+1/M)/2$ then x is both the unique solution of $\mathcal{P}_0$ and $\mathcal{P}_1$*

The proof of the previous proposition is given in Donoho and Elad (2003) or Gribonval and Nielsen (2003). In practice this previous result is useful when the bound $(1 + 1/M)/2$ is large. This bound is large when the coherence M is small meaning that the family $d_1, \ldots, d_p$ is close

to an orthogonal family. However, when $p > n$, this family is not orthogonal thus M is not equal to zero. More precisely, the Welch bound is a lower bound for M. The proof of the theorem 5.1 is given in Sustik et al. (2007) or in the book of Foucart and Rauhut (2013).

Theorem 5.1 (Welch bound) *Let D be a $n \times p$ matrix with $n < p$. Let $d_1, \dots, d_p$ be the columns of D such that for all $i \in [\![1, p]\!]$, $\|d_i\|_2 = 1$. A lower bound of the coherence M of the matrix D is the following*

$$M \geq \sqrt{\frac{p-n}{n(p-1)}}.$$

The equality occurs if and only if $d_1, \dots, d_p$ form a tightframe namely

$$\exists c \geq 0 \text{ such that } \forall i, j \in [\![1, p]\!], i \neq j, |\langle d_i, d_j \rangle| = c.$$

As a consequence of the theorem 5.1, when $p \gg n$, the bound $(1 + 1/M)/2$ given in the mutual coherence condition is smaller than $(1 + \sqrt{n})/2$.

Other conditions which insure that a representation x is simulatenously a solution of $\mathcal{P}_0$ and $\mathcal{P}_1$ relies on the restrictive isometry constant.

Definition 5.4 *Let $s \in [\![1, p]\!]$, the restrictive isometry constant δ_s is defined as follow*

$$\delta_s := \inf \left\{ l \in \mathbb{R}_+ \mid \forall x \in \mathbb{R}^p \text{ such that } \|x\|_0 \leq s, \text{ we have } (1-l)\|x\|^2 \leq \|Dx\|^2 \leq (1+l)\|x\|^2 \right\}.$$

If $\delta_s = 0$ then each family of at most s columns of D would be orthogonal. Of course, when $p > n$ the family $(d_i)_{1 \leq i \leq p}$ is not orthogonal but a small δ_s indicates that $(d_i)_{1 \leq i \leq p}$ is close to an orthogonal family. There are a lot of results which relie on the restricted isometry constant, we do not pretend to provide an exhaustive list of all these results. The propositions 5.7 and 5.8 are given respectively in the articles of Candes (2008) and Cai and Zhang (2013).

Proposition 5.7 (restricted isometry condition) *Let x be a representation such that $\|x\|_0 \leq s$. If $\delta_{2s} \leq \sqrt{2} - 1$ then x is the unique solution of these both problems $\mathcal{P}_0$ and $\mathcal{P}_1$.*

Proposition 5.8 (restricted isometry condition) *Let x be a representation such that $\|x\|_0 \leq s$ with $s \geq 2$. If $\delta_s < 1/3$ then x is the unique solution of these both problems $\mathcal{P}_0$ and $\mathcal{P}_1$.*

5.3.4 l^α minimization with $\alpha \in (0, 1]$

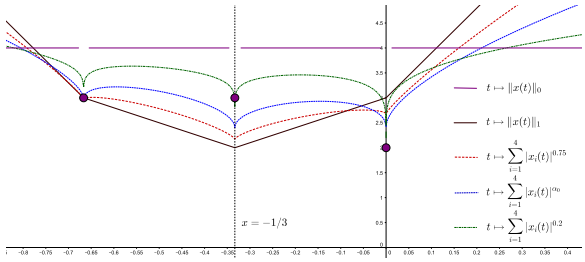
Although a lot of conditions insure that a representation is simultaneously the solution of $\mathcal{P}_0$ and $\mathcal{P}_1$, these problems are not equivalent. Hereafter, an example in which the solution of $\mathcal{P}_1$ and $\mathcal{P}_0$ are different. Let us define the matrix D and the vector y as follows

$$D := \begin{pmatrix} 2 & -1 & -1 & -1 \\ -1 & 2 & 1 & -1 \\ -1 & 1 & 2 & 2 \end{pmatrix} \text{ and } y = D \begin{pmatrix} 0 \\ 1 \\ -2 \\ 0 \end{pmatrix} = \begin{pmatrix} -3 \\ 4 \\ 4 \end{pmatrix}.$$

The affine space $Dx = y$ is the set $\{(t, 1+3t, -2-3t, 2t), t \in \mathbb{R}\}$. Let us denote $x(t) := (t, 1+3t, -2-3t, 2t)$, by minimizing respectively the functions $t \in \mathbb{R} \mapsto \|x(t)\|_0$ and $t \in \mathbb{R} \mapsto \|x(t)\|_1$ we obtain that the unique solution of $\mathcal{P}_0$ is $x(0) = (0, 1, -2, 0)$ while, the unique solution of $\mathcal{P}_1$ is $x(-1/3) = (-1/3, 0, -1, -2/3)$. In this same example, we are going to illustrate that the minimization of the l^α "norm" with $\alpha \in (0, 1]$ allows to recover the sparsest solution. The minimization of the l^α "norm" is the following problem

$$\underset{x \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{p} |x_i|^\alpha \text{ subject to } Dx = y. \qquad (\mathcal{P}_\alpha)$$

The figure 5.1 illustrates a situation in which the solution of $\mathcal{P}_\alpha$ is the unique sparsest representation as soon as α is small enough.



value of α	solution of $\mathcal{P}_\alpha$
$0 < \alpha < \alpha_0$	$(0, 1, -2, 0)$
$\alpha = \alpha_0 \approx 0.493$	$(0, 1, -2, 0), (-\frac{1}{3}, 0, -1, -\frac{2}{3})$
$\alpha_0 < \alpha < 1$	$(-\frac{1}{3}, 0, -1, -\frac{2}{3})$

FIGURE 5.1 – The left figure illustrates the graphics of the functions $t \mapsto \|x(t)\|_0$ and $t \mapsto \sum_{i=1}^{4} |x_i(t)|^\alpha$ with $\alpha \in \{0.2, \alpha_0, 0.75, 1\}$. Minimizing these functions allows to obtain the solutions of $\mathcal{P}_0, \mathcal{P}_{0.2}, \mathcal{P}_{\alpha_0}, \mathcal{P}_{0.75}$ and $\mathcal{P}_1$. The table in the right provides the solution of $\mathcal{P}_\alpha$. When $\alpha > \alpha_0$, the solution of $\mathcal{P}_\alpha$ is $x(-1/3) = (-1/3, 0, -1, -2/3)$ (the same solution as $\mathcal{P}_1$) while, when $\alpha < \alpha_0$ the solution of $\mathcal{P}_\alpha$ is $x(0) = (0, 1, -2, 0)$. Consequently, when $\alpha < \alpha_0$ the minimization of $\mathcal{P}_\alpha$ provides the sparsest representation.

The figure 5.1 illustrates a situation in which the minimization of the l^α "norm" provides a sparsest representation as soon as α is small enough. In Tardivel et al. (2017), we are going to study the minimization of the l^α "norm". The problem $\mathcal{P}_0$ could have several solutions. In this article, we prove that without any condition, the minimization of the l^α "norm" with α small enough provides at least one solution of the problem $\mathcal{P}_0$.

Chapitre 6

Sparsest representations and approximations of a high-dimensional linear system

Patrick J.C. Tardivel [1], Rémi Servien and Didier Concordet
Toxalim, Université de Toulouse, INRA, ENVT, Toulouse, France.

summary : In a high-dimensional linear system of equations, constrained l^1 minimization methods such as the basis pursuit or the lasso are often used to recover one of the sparsest representations or approximations of the system. The null space property is a sufficient and "almost" necessary condition to recover a sparsest representation with the basis pursuit. Unfortunately, this property can not be easily checked. On the other hand, the mutual coherence or the restricted isometry property are checkable sufficient conditions insuring the basis pursuit to recover one of the sparsest representation. Because both of these conditions are too strong, they are hardly met in practice. Even with these conditions, to our knowledge, there is no theoretical result insuring that the lasso solution is one of the sparsest approximations. In this article, we study a novel constrained problem that gives, without any condition, one of the sparsest representations or approximations. To solve this problem, we provide a numerical method and we prove its convergence. Numerical experiments show that this approach gives better results than both the basis pursuit problem and the reweighted l^1 minimization problem.

Keywords : Basis pursuit, Lasso, Sparsest representations, Sparsest approximations.

1. corresponding author : patrick.tardivel@inra.fr

6.1 Introduction

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 $\|y - (x_1 d_1 + \cdots + x_p d_p)\|^2 \leq \epsilon$. 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

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

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 $(d_j)_{1 \leq j \leq p}$. Many applications are related to the resolution of this problem. For example there are applications concerning tomography (Burger et al., 2016; Liu and Gao, 2016; Prieto and Dorn, 2016) or the radar (Baraniuk and Steeghs, 2007; Herman and Strohmer, 2009).

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

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

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 Donoho and Elad (2003); Donoho et al. (2006); Gribonval and Nielsen (2003). 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 Cohen et al. (2009); Donoho and Elad (2003); Donoho et al. (2006); Gribonval and Nielsen (2003) is probably the most known condition. However, as pointed out by Tillmann et al. Tillmann and Pfetsch (2014), this condition is uncheckable. Another condition is the restricted isometry property detailed in Cai and Zhang (2013); Candes (2008); Candes et al. (2006); Candes and Tao (2005); Foucart and Rauhut (2013). This condition is not easy to use because the computation of the restricted isometry constant is intractable c.f. 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 $\mathcal{P}_1$ do 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

$$\text{argmin} \sum_{1 \leqslant i \leqslant p} \ln(1+|x_i|/\delta) \text{ subject to } Dx = y. \tag{6.1}$$

An iterative method converging to a stationary point of the problem (6.1) is provided in Lobo et al. (2007). 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 (6.1) are also solutions of $\mathcal{P}_0$ and the choice of δ plays a major role on the performances of the method.

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 ; see Bertsekas (1999) (chapter 5.3) for more details.

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 Bunea et al. (2007); Donoho et al. (2006); Dossal (2012); van de Geer (2008); van de Geer and Bühlmann (2009). Among these conditions (for an exhaustive list, see Bühlmann and van de Geer (2011) page 177), the two most known are probably the irrepresentable condition Meinshausen and Bühlmann (2006); Zhao and Yu (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 article 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). To obtain such solutions, we define and solve the following problem

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

We provide functions $f_\alpha : \mathbb{R}_+ \to \mathbb{R}$, depending on a parameter $\alpha > 0$, guaranteeing without any condition that

— when $\epsilon = 0$, there exists α_0 such that whatever $0 < \alpha \leq \alpha_0$, the previous problem is "almost equivalent" to $\mathcal{P}_0$ since $S_{f_\alpha}^0 \subset S_0$,

— when $\epsilon > 0$, $S_{f_\alpha}^\epsilon$ becomes arbitrary close to S_0^ϵ when α converges to 0.

This article is organized as follows. In section 2, we study the case $\epsilon = 0$. We prove that there exists α_0 such that, whatever $\alpha \leq \alpha_0$, each element of $S_{f_\alpha}^0$ is a solution of $\mathcal{P}_0$ and that a Maximisation Minimisation (MM) method provides an iterative sequence which converges to a local minimum of $\mathcal{P}_0$. Section 3 is dedicated to the case $\epsilon > 0$. We prove that $S_{f_\alpha}^\epsilon$ becomes arbitrary close to the set S_0^ϵ when α converges to 0 and we give necessary conditions that must satisfy the limit points of the iterative sequence provided by the MM method. We also exhibit a subset of S_0^ϵ that fulfilled these necessary conditions. The section 4 is devoted to simulations. Numerical experiments show that this approach gives better results to recover one of the sparsest representations than both the basis pursuit problem $\mathcal{P}_1$ and the reweighted l^1 minimization problem.

6.2 A sparsest representation

As already explained, solving $\mathcal{P}_0$ is difficult. Replacing the l^0 “norm" by a l^1 norm leads to the problem $\mathcal{P}_1$ which provides sparse solutions. However, the conditions guaranteeing that a solution of $\mathcal{P}_1$ is also a solution of $\mathcal{P}_0$ are unverifiable. 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. 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).

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

where $\|x\|_\alpha = (\sum_{i=1}^p |x_i|^\alpha)^{1/\alpha}$ is the l^α "norm" of the vector x. The study of this problem has been the subject of an abundant literature Chartrand (2007); Foucart and Lai (2009); Gribonval and Nielsen (2003, 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 Gribonval and Nielsen

(2003, 2007) or the restricted isometry property Chartrand (2007); Foucart and Lai (2009); Lai (2010); Sun (2012) hold. As for the basis pursuit, these conditions are uncheckable.

In this section we show that there exists $\alpha_0 > 0$ such that the solutions of $\mathcal{P}_\alpha$ are also solutions of $\mathcal{P}_0$ as soon as $\alpha < \alpha_0$. When $\alpha < 1$, the function $x = (x_1, \ldots, x_p) \mapsto \|x\|_\alpha$ is a concave function on each domain of the form $I_1 \times \cdots \times I_p$, with $I_k = (-\infty, 0]$ or $I_k = [0, +\infty)$. Solving $\mathcal{P}_\alpha$ leads to minimize a locally concave function on a convex set. This is not a convex optimization problem. In this respect, we propose in this section a numerical method to solve it. 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". A geometric interpretation linking the shape of the spheres of the l^α "norm" to the sparseness of the solutions of $\mathcal{P}_\alpha$ is given in Hastie et al. (2009). In the theorem 6.1, we focus on the following problem

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

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 6.1 *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 (see Sun et al. (2013)). However, 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 6.1 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.

In the theorem 6.1, we made relatively weak assumptions on the f_α functions. Indeed, a function f_α for which the properties of the theorem 6.1 hold can be not derivable on $(0, +\infty)$ or not continuous in 0. 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) popularized in statistics by the EM algorithm Dempster et al. (1977). 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.

In a similar way as in Candes et al. (2008); Lobo et al. (2007), 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

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

Note that without any other consideration, nothing guarantees that $x^{(k+1)}$ is unique. The general position condition for D (as defined in Tibshirani (2013)) is a sufficient condition for the uniqueness of $x^{(k+1)}$ Rosset et al. (2004). The general position condition is very weak. Indeed, when D is a random matrix with a continuous distribution on the set of the $n \times p$ matrix, the general position condition holds almost surely Tibshirani (2013). Consequently, in practice, the uniqueness of the basis pursuit solution always holds.

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

$$x^{(k+1)} := \operatorname{argmin} \sum_{1\leqslant i\leqslant p} f'_\alpha(|x_i^{(k)}| + \Delta)|x_i| \text{ subject to } Dx = y. \tag{6.2}$$

The theoretical results justifying the introduction of Δ are provided in the theorem 6.2 and proposition 6.1.

Theorem 6.2 *For every $x^{(0)} \in \mathbb{R}^p$, for every $\Delta > 0$, there exists an integer k_0 such that $\forall k \geqslant k_0$, the sequence $x^{(k)}$ defined in (6.2) is so that $x^{(k)} = x^{(k_0)}$.*

A similar theorem that deals only with the convergence of the iterative method in the special case where $f_\alpha(x) = \log(1 + x/\alpha)$ already denoted as (6.1) is given in Lobo et al. (2007). This theorem shows that the iterative sequence converges onto a stationary point of the problem

$\min \sum_{1\leqslant i\leqslant p} \log(1+|x_i|/\alpha)$ subject to $Dx = y$ which is not *a priori* a local minimum of $\mathcal{P}_0$. Moreover, the proposed proof in Lobo et al. (2007) seems incorrect because even for a bounded sequence, the fact that $\lim_{k\to+\infty} x_i^{(k+1)} - x_i^{(k)} = 0$ does not imply the convergence of $(x_i^{(k)})_{k\in\mathbb{N}}$. The proposition 6.1 states the limit of the sequence $(x^{(k)})_{k\in\mathbb{N}}$ defined in (6.2) is a local minimum of the problem $\mathcal{P}_0$.

Proposition 6.1 *Let $(x^{(k)})_{k\in\mathbb{N}}$ be the sequence defined in (6.2) 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 $\|x\|_0 > \|l\|_0$.*

The limit l given in the previous proposition depends on $x^{(0)} \in \mathbb{R}^p$ and $\Delta > 0$. In Section 6.4 we discuss the choice of the initial point $x^{(0)}$ and we propose to test different values for Δ in order to keep the local minimum having the lowest l^0 "norm".

6.3 Sparsest $\epsilon-$approximations

In the previous section, we obtained one of the sparsest representations of y by solving the problem $\mathcal{P}_{f_\alpha}$ instead of $\mathcal{P}_0$ with α small enough. Similarly, 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 6.3 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 $\mathcal{P}_0^\epsilon$. This justifies to solve $\mathcal{P}_{f_\alpha}^\epsilon$ instead of $\mathcal{P}_0^\epsilon$. There are situations in which solving $\mathcal{P}_{f_\alpha}^\epsilon$, with a small enough α, gives one of the sparsest approximations. However, there are situations in which it is not the case. Unfortunately, we do not have any general criterion separating these two cases. This is the reason why, we propose the following theorem that states that the solutions of $\mathcal{P}_{f_\alpha}^\epsilon$ are arbitrarily close to S_0^ϵ. 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 6.3 *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 6.3 hold for the families of functions $f_\alpha : x \in \mathbb{R}_+ \mapsto x/(\alpha + x)$ and $f_\alpha : x \in \mathbb{R}_+ \mapsto \arctan(x/\alpha)$. The figure 6.1 illustrates this result in two different cases. In the first case, with a small enough α, the problem $\mathcal{P}^\epsilon_{f_\alpha}$ captures one of the sparsest approximations. In the second case, whatever $\alpha > 0$, the solution of the problem $\mathcal{P}^\epsilon_{f_\alpha}$ is not one of the sparsest approximations but stays close to S^ϵ_0.

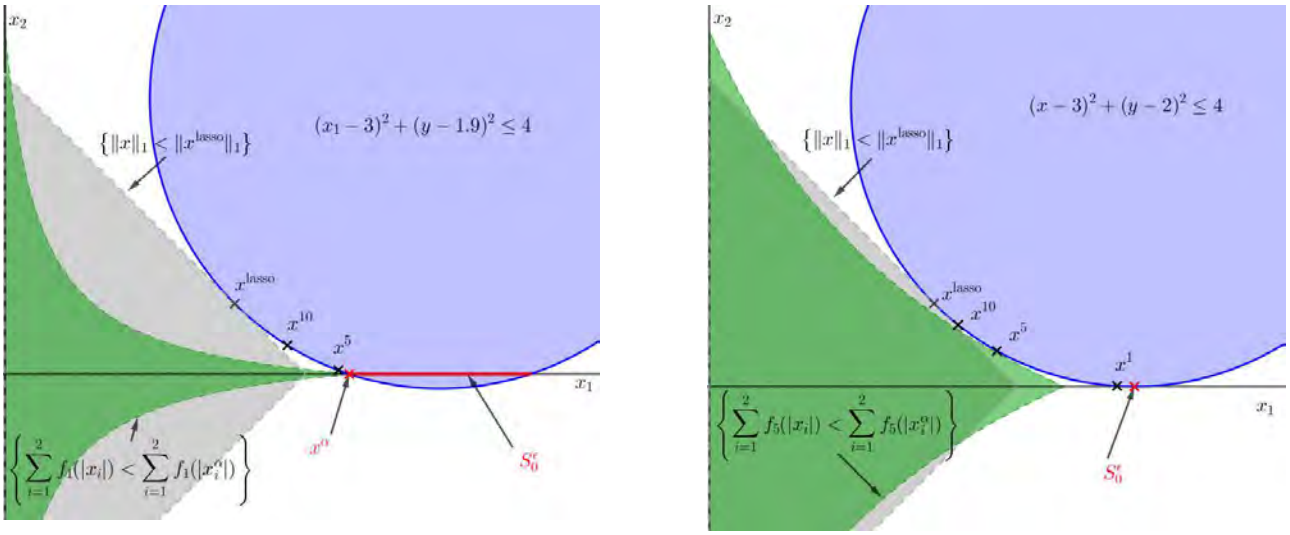


FIGURE 6.1 – Let f_α be the function $f_\alpha : x \in \mathbb{R}_+ \mapsto x/(x+\alpha)$ with $\alpha > 0$. On the left, we represent the solution of the problem argmin $\sum_{i=1}^2 f_\alpha(|x_i|)$ subject to $(x_1-3)^2+(x_2-1.9)^2 \leq 4$ for several values of α and the solution of the lasso problem argmin $\sum_{i=1}^2 |x_i|$ subject to $(x_1-3)^2+(x_2-1.9)^2 \leq 4$ denoted x^{lasso}. The points x^{10}, x^5 and x^α are the solutions of the first problem when $\alpha = 10$, $\alpha = 5$ and $\alpha \leq \alpha_0$ with $\alpha_0 \approx 4.5$. Geometrically, x^α and x^{lasso} are respectively the unique solution of the first problem with $\alpha = 1$ and of the lasso problem because the "open balls" $\{\sum_{i=1}^2 f_1(|x_i|) < \sum_{i=1}^2 f_1(|x_i^\alpha|)\}$ (in green) and $\{\|x\|_1 < \|x^{\text{lasso}}\|_1\}$ (in grey) do not share any point with the constraint set $(x_1-3)^2+(x_2-1.9)^2 \leq 4$ (in blue). Note that when $\alpha \leq \alpha_0$, the first problem catches an element x^α of S^ϵ_0 (in red). On the right, we represent the solution of the lasso problem and the solutions x^{10}, x^5, x^1 of the problem argmin $\sum_{i=1}^2 f_\alpha(|x_i|)$ subject to $(x_1-3)^2+(x_2-2)^2 \leq 4$ when $\alpha = 10, \alpha = 5$ and $\alpha = 1$. In addition we draw the "open balls" $\{\sum_{i=1}^2 f_5(|x_i|) < \sum_{i=1}^2 f_5(|x_i^5|)\}$ (in green) and $\{\|x\|_1 < \|x^{\text{lasso}}\|_1\}$ (in grey). When α is small the solution is close to S^ϵ_0. However, one can prove that whatever $\alpha > 0$, this second problem never catches exactly an element of S^ϵ_0.

In the previous section, we have seen that a MM method provides a sequence (6.2) which is stationary from a certain rank onto a local minimum of the problem $\mathcal{P}_0$. To solve the problem $\mathcal{P}^\epsilon_{f_\alpha}$, one uses the same MM method as in (6.2) 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

$$x^{(k+1)} := \text{argmin} \sum_{1\leqslant i\leqslant p} f'_\alpha(|x_i^{(k)}| + \Delta)|x_i| \text{ subject to } \|y - Dx\|^2 \leq \epsilon. \tag{6.3}$$

Similarly to the basis pursuit problem, the lasso problem (6.3) does not always have an unique

solution. However, the general position condition for D is sufficient to insure the uniqueness of the lasso solution Rosset et al. (2004); Tibshirani (2013). As already explained, the general position condition is very weak Tibshirani (2013) and, in practice, the uniqueness of the lasso solution always occurs.

In the theorem 6.4, we prove that the sequence $(x^{(k)})_{k\in\mathbb{N}}$, as defined in (6.3), is bounded that is, when k is large enough, $x^{(k)}$ is close to a limit point. The theorem 6.4 shows that the optimality conditions hold for the limit points of the sequence $(x^{(k)})_{k\in\mathbb{N}}$.

Theorem 6.4 *Let $y \in \mathbb{R}^p$ such that $\|y\|^2 > \epsilon$. Let $(f_\alpha)_{\alpha>0}$ be a family of increasing, concave and two times differentiable functions defined on $(0, +\infty)$ such that $\forall \alpha > 0, f'_\alpha$ is convex and*

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

Then :

1. *The sequence $(x^{(k)})_{k\in\mathbb{N}}$ described in (6.3) is bounded.*
2. *For any limit point $\tilde{x}$ of the sequence $(x^{(k)})_{k\in\mathbb{N}}$, we have*
   - *i) The vector $\tilde{x}$ is on the boundary of the constraints' set thus, $\|y - D\tilde{x}\|^2 = \epsilon$.*
   - *ii) The family of D matrix columns $(d_i)_{i\in \mathrm{supp}(\tilde{x})}$ is linearly independent.*
   - *iii) The vectors $(d_i^T(y - D\tilde{x}))_{i\in\mathrm{supp}(\tilde{x})}$ and $(f'_\alpha(|\tilde{x}_i| + \Delta))_{i\in\mathrm{supp}(\tilde{x})}$ are collinear.*

When $\|y\|^2 \leq \epsilon$ then 0 is the unique solution of the problem $\mathcal{P}^\epsilon_{f_\alpha}$ and for any $k \geq 0$ we have $x^{(k)} = 0$. In particular when $\|y\|^2 < \epsilon$, the condition i) is not met. As for the theorem 6.3, the assumptions on f_α given in theorem 6.4 hold for the function $f_\alpha : x \in \mathbb{R}_+ \mapsto x/(\alpha + x)$. The points for which the properties i), ii) and iii) hold are kind of "critical points" of the problem $\mathcal{P}^\epsilon_{f_\alpha}$. The properties i), ii), iii) described in the previous theorem are verified at all points x^α of $S^\epsilon_{f_\alpha}$.

Actually, a proof similar to the proof of the lemma 6.9 shows that x^α is on the boundary of the constraint $\|y - Dx\|^2 \leq \epsilon$. Consequently, the property i) holds for x^α.

By the lemma 6.1, the family $(d_i)_{i\in\mathrm{supp}(x^\alpha)}$ is linearly independent thus property ii) holds.

Finally, because x^α is a solution of the problem $\mathcal{P}^\epsilon_{f_\alpha}$, $(x^\alpha)_{i\in\mathrm{supp}(x^\alpha)}$ is also a solution of the problem

$$\text{argmin} \sum_{i\in\mathrm{supp}(x^\alpha)} f_\alpha(|x_i|) \text{ subject to } \|y-\tilde{D}x\|^2 \leq \epsilon \text{ where } \tilde{D} \text{ is the matrix with columns } (d_i)_{i\in\mathrm{supp}(x^\alpha)}. \tag{6.4}$$

Consequently $(x^\alpha)_{\mathrm{supp}(x^\alpha)}$ is a stationary point of a Lagrangian function (Lange (2004) page 71, Boyd and Vandenberghe (2004) page 243) implying thus the property iii) to hold with $\Delta = 0$.

The previous remark and the theorem 6.3 have a nice geometric interpretation illustrated on figure 6.2 for $p = 3$ and $n = 2$.

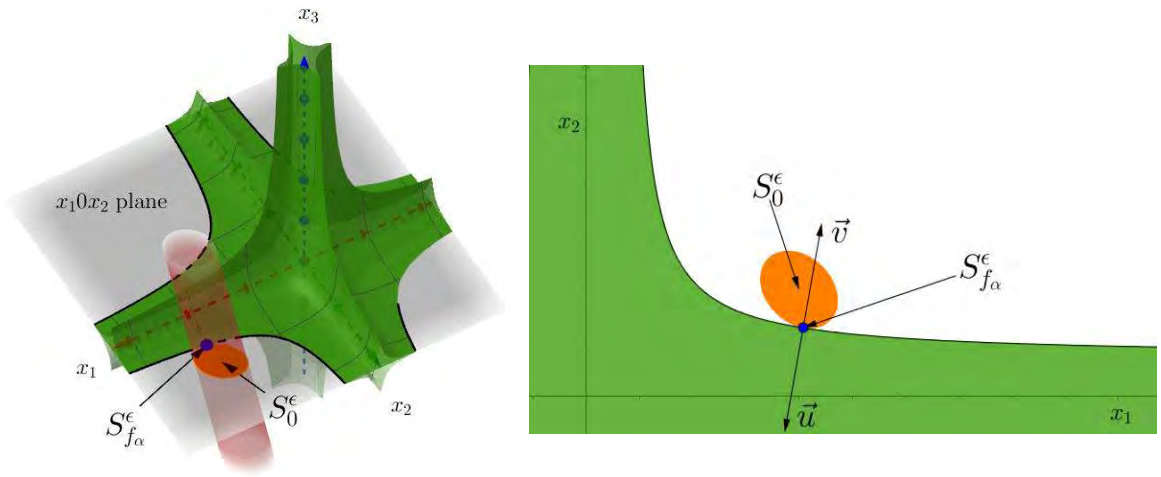


FIGURE 6.2 – In the left panel the set of constraints $\|y - Dx\|^2 \leq \epsilon$ (in orange) and the "ball" $\sum_{i=1}^3 f_\alpha(|x_i|) \leq R$ (in green) are represented. The radius R is the smallest positive number for which the cylinder $\|y - Dx\|^2 \leq \epsilon$ and the "ball" $\sum_{i=1}^3 f_\alpha(|x_i|) \leq R$ share at least one common point. The set S_0^ϵ is a union of three ellipsoids which are the intersection of the cylinder $\|y - Dx\|^2 \leq \epsilon$ with the planes $x_1 0 x_2, x_1 0 x_3$ and $x_2 0 x_3$. To keep this illustration understandable, we only plot the intersection of the cylinder $\|y - Dx\|^2 \leq \epsilon$ and the plane $x_1 0 x_2$. The set $S^\epsilon_{f_\alpha} = \{x^\alpha\}$, represented as a blue point in the left figure, is a singleton of S_0^ϵ. This illustrates theorem 6.3 showing that whatever $\eta > 0$ $S^\epsilon_{f_\alpha} \subset G_\eta$. In the right panel, we focus on the intersection of the cylinder $\|y - Dx\|^2 \leq \epsilon$ and the intersection of the "ball" $\sum_{i=1}^3 f_\alpha(|x_i|) \leq R$ with the plane $x_1 0 x_2$. The vectors $\vec{u} = \left(-d_i^T(y - Dx^\alpha)\right)_{1\leq i\leq 2} = \left(\frac{\partial \|y-Dx\|^2}{\partial x_i}(x^\alpha)\right)_{1\leq i\leq 2}$ and $\vec{v} = (\ \mathrm{sign}(x_i) f'_\alpha(|x_i^\alpha|))_{1\leq i\leq 2} = \left(\frac{\partial \sum_{i=1}^3 f_\alpha(|x_i|)}{\partial x_i}(x^\alpha)\right)_{1\leq i\leq 2}$ represent respectively the normalized normal vectors to the ellipsoid and the "ball". Note that the solution x^α of the problem (6.4) is i) on the boundary of the cylinder ii) completely included in the plane $(x_1 0 x_2)$, and iii) that at this point, the normal vectors $\vec{u}$ and $\vec{v}$ are collinear.

Because for each element x^α in $S^\epsilon_{f_\alpha}$, the property iii) holds with $\Delta = 0$, this value of Δ could appear as the ideal value. It is not the case. Indeed, if we define the set L^α by

$$L^\alpha := \underset{x\in S_0^\epsilon}{\mathrm{argmin}} \sum_{i=1}^p f_\alpha(|x_i| + \Delta), \tag{6.5}$$

for an arbitrary $\Delta > 0$, the proposition 6.2 shows that L^α is a set of "critical points" such that $L^\alpha \subset S_0^\epsilon$. Consequently, whatever Δ, when $x^{(0)}$ is well chosen, one can expect that for k large enough, $x^{(k)}$ is close to the set L^α.

The proposition 6.2 shows that every element of L^α satisfies the property i), ii) and iii).

Proposition 6.2 *Let $y \in \mathbb{R}^p$ such that $\|y\|^2 > \epsilon$. Let x^α be an arbitrary element of L^α. Then, the three following properties hold for x^α.*

i) The vector x^α is on the boundary of the constraint thus, $\|y - Dx^\alpha\|^2 = \epsilon$.

ii) The family $(d_i)_{i\in \mathrm{supp}(x^\alpha)}$ is linearly independent.

iii) The vectors $(d_i^T(y - Dx^\alpha))_{i\in\text{supp}(x^\alpha)}$ *and* $(f'_\alpha(|x_i^\alpha| + \Delta))_{i\in\text{supp}(x^\alpha)}$ *are collinear.*

6.4 Numerical experiments

In the previous section, we developed a new method able to recover at least one solution of $\mathcal{P}_0$ or $\mathcal{P}_0^\epsilon$. Currently, the basis pursuit $\mathcal{P}_1$ is the reference method to recover a solution of $\mathcal{P}_0$. An alternative to the basis pursuit is the reweighted l^1 minimization Candes et al. (2008). In this section, 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 (6.2).

$$\begin{aligned} \text{Let } a : \; = \; & \text{argmin} \sum_{1\leqslant i\leqslant p} f'_\alpha(|x_i^{(k)}| + \Delta)|x_i| \text{ subject to } Dx = y \\ & \text{and set } \begin{cases} x^{(k+1)} = a \text{ if } \|a\|_0 \leq \|x^{(k)}\|_0 \\ x^{(k+1)} = x^{(k)} \text{ otherwise} \end{cases} . \end{aligned} \tag{6.6}$$

As for the sequence given in (6.2), when k is large enough, the sequence (6.6) is stationary onto a point l. As defined in (6.6) 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 $\|l\|_0 \leq \|x^{\text{bp}}\|_0$. 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. The following section provides a smart initial point $x^{(0)}$.

6.4.1 Choice of the initial point $x^{(0)}$

Because the MM algorithm converges to a local minimum of $\mathcal{P}_0$, the choice of its initial point is critical. Candès et al Candes et al. (2008) took the solution of problem $\mathcal{P}_1$ as the initial point for the iterative sequence (6.2). 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. This intuition is confirmed by the right panel of the figure 6.3 which illustrates that x^{bp} catch easily the largest components of $\tilde{x}$.

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

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

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

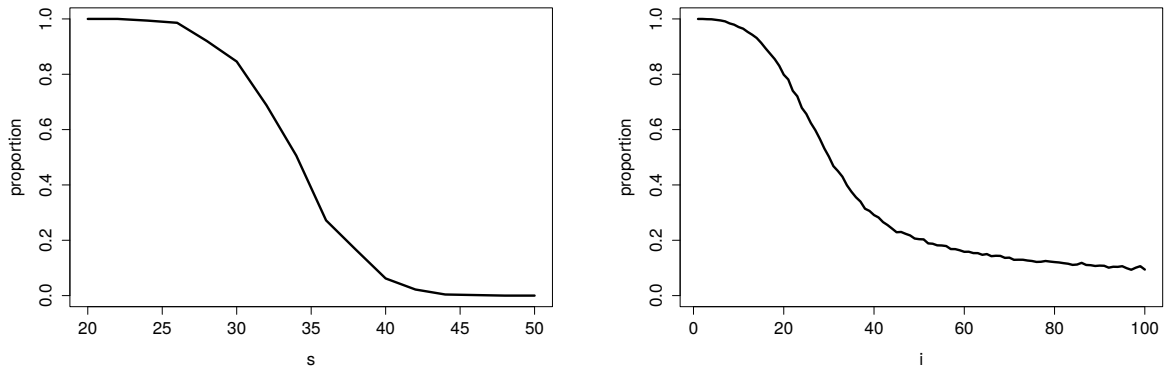


FIGURE 6.3 – In this figure, $\tilde{x}$ is a random vector such that $\text{supp}(\tilde{x})=[\![1,s]\!]$, with $s\in\{20,22,\dots,50\}$ and $|\tilde{x}_1|\geq\cdots\geq|\tilde{x}_s|$. For every $s\in\{20,22,\dots,50\}$, a sample of 500 families $\mathcal{D}=\{d_1,\dots,d_{256}\}$ and 500 observations of the random vectors $\tilde{x}$ have been simulated. For each family and observation of $\tilde{x}$, we compute the solution x^{bp} of the basis pursuit problem $\mathcal{P}_1$. On the left panel, we have the representation of the proportion of times when $x^{\text{bp}}=\tilde{x}$ as a function of s. One notices that when $s\geq 45$, the event $x^{\text{bp}}=\tilde{x}$ is never observed. In the right panel, we set $s=50$ and r is a permutation of $[\![1,100]\!]$ such that $|x^{\text{bp}}_{r(1)}|\geq\cdots\geq|x^{\text{bp}}_{r(100)}|$ (by lemma 6.3, $\text{Card}(\text{supp}(x^{\text{bp}}))\leq 100$). For each $i\in[\![1,100]\!]$ in the x-axis, the y-axis represents the proportion of times for which $r(i)\in\text{supp}(\tilde{x})$. Note that largest components of x^{bp} are elements of $\text{supp}(\tilde{x})$.

The figure 6.4 gives an algorithm which describes how to choose $x^{(0)}$. The input of the

algorithm 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

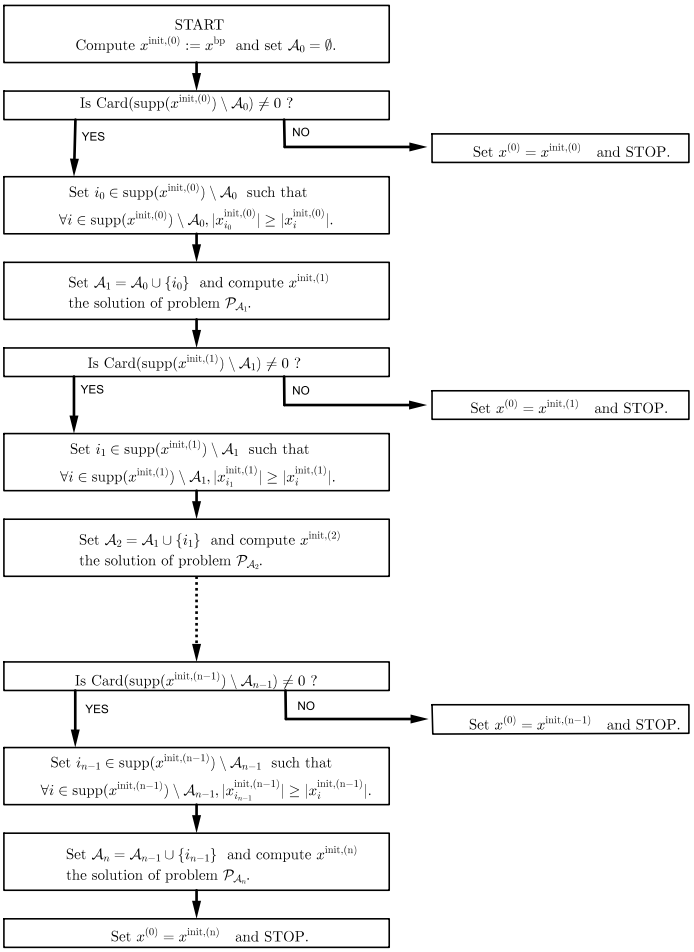


FIGURE 6.4 – In this figure, we give the different steps of the algorithm to obtain the initial point $x^{(0)}$.

the problems $\mathcal{P}_{\mathcal{A}_1}, \mathcal{P}_{\mathcal{A}_2}, \ldots$ should be increasingly close to $\tilde{x}$. 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. When at the j^{th} iteration $\text{Card}(\text{supp}(x^{\text{init},(\text{j})}) \setminus \mathcal{A}_j) = 0$, it is not possible to find an element i_j to construct the set $\mathcal{A}_{j+1}$ and the algorithm stops.

6.4.2 Comparisons

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 Candes et al. (2008) and the solution given by our method as defined by (6.6). The reweighted l^1 solution is the limit of the

sequence $(x^{\mathrm{l1,(k)}})_{k\in\mathbb{N}}$ defined by $x^{\mathrm{l1},(0)} = x^{\mathrm{bp}}$ and

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

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 (6.6) was computed using the algorithm described previously. After 8 iterations, we keep the sparsest solution among the one obtained with $\Delta \in \{0.01, 0.1, 0.5, 1, 2, 4\}$.

The figure 6.5 shows the performances of the basis pursuit, the reweighted l^1 minimization and our method.

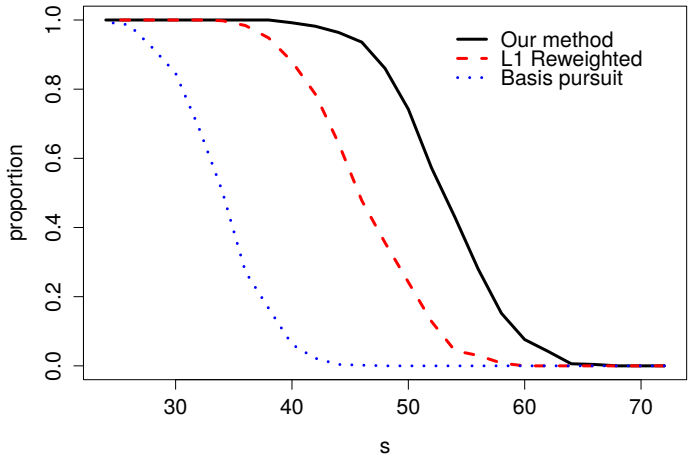


FIGURE 6.5 – The performances of the three competing methods are represented by the proportions of realisations of the events $x^{\mathrm{bp}} = \tilde{x}$, $x^{\mathrm{l1},(8)} = \tilde{x}$ and $x^{(8)} = \tilde{x}$ as a function of the number of non null components of $\tilde{x}$ denoted s. One notices 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 6.5 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.

Additional simulations are performed when D is a partial random circulant matrix. This simulations are given in appendix.

6.5 Conclusion

In this article, we studied the problems $\mathcal{P}_{f_\alpha}$ and $\mathcal{P}^{\epsilon}_{f_\alpha}$ which recover respectively one of the sparsest representations or one of the sparsest approximations of a high-dimensional linear system. Theoretical results are proved and a MM method is then used to solve these problems. Numerical experiments highligh the performances of our method compared to the basis pursuit and the reweighted l^1 minimization ones. 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 of the expectation of y.

6.6 Appendix 1 : Proofs

6.6.1 Proof of the theorem 6.1

By construction, the function to be minimized in the problem $\mathcal{P}_{f_\alpha}$ converges pointwise to the l^0 "norm" when α goes to 0. As the l^0 norm is not continuous, this convergence can not be uniform onto $\mathbb{R}^p$. However, a straightforward consequence of the lemma 6.1 is that the number of possible solutions of the problem $\mathcal{P}_{f_\alpha}$ is finite and the convergence of $\sum_{i=1}^p f_\alpha(|x_i|)$ to $\|x\|_0$ is therefore uniform onto this finite set. The proof of theorem 6.1 is based on this uniform convergence.

Lemma 6.1 *Let f_α be a function defined on $\mathbb{R}_+$ strictly increasing and strictly concave such that*

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

Denote x^α a solution of the problem $\mathcal{P}_{f_\alpha}$ (resp. $\mathcal{P}^{\epsilon}_{f_\alpha}$) then the family $(d_i)_{i \in \mathrm{supp}(x^\alpha)}$ is linearly independent.

Proof : Let us assume that the family $(d_i)_{i \in \mathrm{supp}(x^\alpha)}$ is not linearly independent. There exist coefficients $(\gamma_i)_{i \in \mathrm{supp}(x^\alpha)}$ not simultaneously null such that

$$\sum_{i \in \mathrm{supp}(x^\alpha)} \gamma_i d_i = \vec{0}.$$

To provide a contradiction, we are going to show that$\sum_{i=1}^p f_\alpha(|x_i^\alpha|)$ is no longer minimal. That is, there exists an admissible point z so that $\sum_{i=1}^p f_\alpha(|z_i|) < \sum_{i=1}^p f_\alpha(|x_i^\alpha|)$. Let us define $\{i_1, \ldots, i_s\} := \{i \in \mathrm{supp}(x^\alpha) \mid \gamma_i \neq 0\}$, the set of non-null components of γ. We are looking for z among the admissible points $x(t)$ defined by

$$\forall t \in \mathbb{R}, x_i(t) = x_i^\alpha + t\gamma_i \text{ if } i \in \{i_1, \ldots, i_s\} \text{ and } x_i(t) = x_i^\alpha \text{ otherwise.}$$

For all $i \in \{i_1, \dots, i_s\}$, let us denote $t_i = -x_i^\alpha/\gamma_i$. Without loss of generality, we assume that $t_{i_1} \leqslant \cdots \leqslant t_{i_s}$. The function $t \in \mathbb{R} \mapsto f_\alpha(|x_i(t)|)$ is strictly decreasing and strictly concave on $(-\infty, t_i]$ and strictly increasing and strictly concave on $[t_i, +\infty)$ when $i \in \{i_1, \dots, i_s\}$.

Assume that $0 \notin [t_{i_1}, t_{i_s}]$; because each function $t \in \mathbb{R} \mapsto f_\alpha(|x_i(t)|)$ with $i \in \{i_1, \dots, i_s\}$ is strictly decreasing on $(-\infty, t_i]$ (resp. strictly increasing on $[t_i, +\infty)$), one deduces that $t \in \mathbb{R} \mapsto \sum_{i=1}^p f_\alpha(|x_i(t)|)$ is strictly decreasing on $(-\infty, t_{i_1}]$ (resp. strictly increasing on $[t_{i_s}, +\infty)$). These monotony results imply that

$$\sum_{i=1}^p f_\alpha(|x_i(0)|) = \sum_{i=1}^p f_\alpha(|x_i^\alpha|) > \min\left\{\sum_{i=1}^p f_\alpha(|x(t_{i_1})|), \sum_{i=1}^p f_\alpha(|x(t_{i_s})|)\right\},$$

which provides a contradiction for the minimality of $\sum_{i=1}^p f_\alpha(|x_i^\alpha|)$.

Assume that $0 \in [t_{i_1}, t_{i_s}]$ then, there exists i_k such that $0 \in (t_{i_k}, t_{i_{k+1}})$ (note that t_{i_k} and $t_{i_{k+1}}$ are not null). Because each function $t \in \mathbb{R} \mapsto f_\alpha(|x_i(t)|)$ with $i \in \{i_1, \dots, i_s\}$ is strictly concave on $[t_{i_k}, t_{i_{k+1}}]$, one deduces that $t \in \mathbb{R} \mapsto \sum_{i=1}^p f_\alpha(|x(t)|)$ is also strictly concave on $[t_{i_k}, t_{i_{k+1}}]$. Consequently, the restriction of the function $t \in \mathbb{R} \mapsto \sum_{i=1}^p f_\alpha(|x(t)|)$ to the set $[t_{i_k}, t_{i_{k+1}}]$ reaches its minimum at t_{i_k} or $t_{i_{k+1}}$ and nowhere else. This concavity result implies that

$$\sum_{i=1}^p f_\alpha(|x_i(0)|) = \sum_{i=1}^p f_\alpha(|x_i^\alpha|) > \min\left\{\sum_{i=1}^p f_\alpha(|x_i(t_{i_k})|), \sum_{i=1}^p f_\alpha(|x_i(t_{i_{k+1}})|)\right\},$$

which provides a contradiction for the minimality of $\sum_{i=1}^p f_\alpha(|x_i^\alpha|)$. □

We now consider the set $\mathcal{E}$ of subsets $I \subset [\![1, p]\!]$ such that

— The family $(d_i)_{i \in I}$ is linearly independent.

— $y \in Vect(d_i)_{i \in I}$.

Given a subset $I \in \mathcal{E}$, let x_I be the unique vector such that $\text{supp}(x_I) = I$ and $Dx_I = y$. Let us introduce $S := \{x_I, I \in \mathcal{E}\}$. As $\mathcal{E}$ is finite, this set of vectors is finite.

Whatever the function f_α satisfying the properties of the lemma 6.1, the lemma 6.1 shows that the family $(d_i)_{i \in \text{supp}(x^\alpha)}$ is linearly independent. As x^α is admissible, $y \in Vect(d_i)_{i \in \text{supp}(x^\alpha)}$. It follows that for all $x^\alpha \in S_{f_\alpha}$, $x^\alpha \in S$; that is $S_{f_\alpha} \subset S$. The next lemma shows that the solutions of the problem $\mathcal{P}_0$ are also included in S.

Lemma 6.2 *The set S_0 of solutions of $\mathcal{P}_0$ satisfies $S_0 \subset S$.*

Proof : Let x^* be a solution of $\mathcal{P}_0$, we have $Dx^* = y$. To show that $x^* \in S$, it remains to prove that the family $(d_i)_{i \in \text{supp}(x^*)}$ is linearly independent. Suppose that this family is not linearly independent then there exist coefficients $(\gamma_i)_{i \in \text{supp}(x^*)}$ not simultaneously null such that

$$\sum_{i \in \text{supp}(x^*)} \gamma_i d_i = \vec{0}.$$

To provide a contradiction for the minimality of $\|x^*\|_0$, we are going to prove that there exists an admissible point z such that $\|z\|_0 < \|x^*\|_0$. We are looking for z among admissible points $x(t)$ defined by

$$\forall t \in \mathbb{R}, x_i(t) = x_i^* + t\gamma_i \text{ if } i \in \text{supp}(x^*) \text{ and } x_i(t) = x_i^* = 0 \text{ otherwise.}$$

By construction, we have $\forall t \in \mathbb{R}, \text{supp}(x(t)) \subset \text{supp}(x^*)$. To conclude this proof, we have to find $t_0 \in \mathbb{R}$ for which the inclusion is strict. Let $i_0 \in \text{supp}(x^*)$ such that $\gamma_{i_0} \neq 0$ and define $t_0 = -x_{i_0}^*/\gamma_{i_0}$. The i_0^{th} component of $x(t_0)$ is null. Consequently, $\|x(t_0)\|_0 < \|x^*\|_0$ which provides a contradiction to the fact that x^* is a solution of $\mathcal{P}_0$. □

Proof of theorem 6.1 : By the lemma 6.1 and 6.2, we have $S_{f_\alpha} \subset S$ and $S_0 \subset S$. If the elements of $S \setminus S_0$ are not solution of $\mathcal{P}_{f_\alpha}$, one deduces that $S_{f_\alpha} \subset S_0$. Let x and x^* be respectively an arbitrary element of $S \setminus S_0$ and of S_0. A straightforward consequence of the inequality $\sum_{i=1}^p f_\alpha(|x_i|) > \sum_{i=1}^p f_\alpha(|x_i^*|)$ is that x is not a solution of $\mathcal{P}_{f_\alpha}$. We are going prove that this inequality holds when α is small enough. We have that $\forall x \in S \setminus S_0$,

$$\sum_{i=1}^p f_\alpha(|x_i|) - \sum_{i=1}^p f_\alpha(|x_i^*|) = \sum_{i=1}^p f_\alpha(|x_i|) - \|x\|_0 + \|x\|_0 - \|x^*\|_0 + \|x^*\|_0 - \sum_{i=1}^p f_\alpha(|x_i^*|).$$

Because x is not a solution of $\mathcal{P}_0$ contrarily to x^*, one has $\|x\|_0 - \|x^*\|_0 \geqslant 1$. Furthermore, the uniform convergence of $\sum_{i=1}^p f_\alpha(|x_i|)$ to $\|x\|_0$ onto the set S gives $\alpha_0 > 0$ such that

$$\forall \alpha \in (0, \alpha_0), \forall x \in S, \left| \sum_{i=1}^p f_\alpha(|x_i|) - \|x\|_0 \right| < 1/2.$$

Consequently, one obtains

$$\forall \alpha \in (0, \alpha_0), \forall x \in S \setminus S_0, \sum_{i=1}^p f_\alpha(|x_i|) > \sum_{i=1}^p f_\alpha(|x_i^*|).$$

Thus, as soon as $\alpha < \alpha_0$, the solution of $\mathcal{P}_{f_\alpha}$ satisfies $S_{f_\alpha} \subset S_0$ □

6.6.2 Proof of the theorem 6.2 and of the proposition 6.1

The main consequence of lemma 6.3, is that the iterative sequence $(x^{(k)})_{k\geq 1}$ provided by the MM method (6.2) satisfies $\forall k \geq 1, x^{(k)} \in S$. Because S is a finite set, this result is useful for the proof of the theorem 6.2.

Lemma 6.3 *Let us denote*

$$S_\omega := \text{argmin} \sum_{i=1}^p w_i |x_i| \text{ subject to } y = Dx, \text{with } \forall i \in [\![1, p]\!], \omega_i > 0 \tag{6.7}$$

and

$$S_\omega^\epsilon := \operatorname{argmin} \sum_{i=1}^{p} w_i |x_i| \text{ subject to } \|y - Dx\|_2^2 \leqslant \epsilon, \text{with } \forall i \in [\![1, p]\!], \omega_i > 0. \tag{6.8}$$

Then, there exists an element $x^\omega \in S_\omega$ (resp. $x^\omega \in S_\omega^\epsilon$) such that the family $(d_i)_{i \in \text{supp}(x^\omega)}$ is linearly independent.

Proof : When the set S_ω (resp. S_ω^ϵ) is not a singleton, we set x^ω an element of S_ω (resp. S_ω^ϵ) with a minimal l^0 norm. Assume that $(d_i)_{i \in \text{supp}(x^\omega)}$ is not linearly independent. There exist coefficients $(\gamma_i)_{i \in \text{supp}(x)}$ not simultaneously null such that $\sum_{i \in \text{supp}(x^\omega)} \gamma_i d_i = \vec{0}$. Let us set $\mathcal{A}' := \{i \in \text{supp}(x^\omega) \text{ such that } \gamma_i \neq 0\}$. One defines the admissible $x(t)$ of the problem (6.7) (resp. (6.8)) as follows

$$x_i(t) := \begin{cases} x_i^\omega + t\gamma_i \text{ if } i \in \mathcal{A}', \\ x_i^\omega \text{ otherwise.} \end{cases}$$

By definition, the point $x(t)$ satisfies $\text{supp}(x(t)) \subset \text{supp}(x^\omega)$. To provide a contradiction for the minimality of the l^0 "norm" of the solution x^ω, we could build an element $x(t_0) \in S_\omega$ (resp. S_ω^ϵ) with a strictly lower l^0 "norm".

Let f be the function $\forall t \in \mathbb{R}, f(t) := \sum_{i=1}^p w_i |x_i(t)|$. This function is equal to $f(t) = \sum_{i \in \mathcal{A}'} \omega_i |x_i + t\gamma_i| + \sum_{i \notin \mathcal{A}'} \omega_i |x_i|$. The minimum of f is reached on the set $\{-x_i/\gamma_i\}_{i \in \mathcal{A}'}$. If $t_0 := -x_{i_0}/\gamma_{i_0}$, with $i_0 \in \mathcal{A}'$, is a value for which the minimum of f is reached, one sees that $x_{i_0}(t_0) = 0$. This shows $\|x(t_0)\|_0 < \|x^\omega\|_0$ and $x(t_0)$ is an admissible point for which $\sum_{i=1}^p \omega_i |x_i(t_0)| \leq \sum_{i=1}^p \omega_i |x_i(0)| = \sum_{i=1}^p \omega_i |x_i^\omega|$. Consequently, $x(t_0)$ is point of S_ω (resp. S_ω^ϵ) with a strictly smaller l^0 "norm" than the one of x^ω which contradicts the minimality of $\|x^\omega\|_0$. □

Remind that for each $k \geq 1$, $x^{(k)}$ defined in (6.2) is the solution of a weighted basis pursuit problem. We have already noted that in practice weighted basis pursuit problem admits a unique solution. Consequently, by the lemma 6.3 the family $(d_i)_{i \in \text{supp}(x^{(k)})}$ is linearly independent and, on the other hand, $y = Dx^{(k)}$ which implies that $x^{(k)} \in S$.

Proof of theorem 6.2 : The MM method for the function $x \in \mathbb{R}^p \mapsto \sum_{1 \leqslant i \leqslant p} f_\alpha(|x_i| + \Delta)$ provides the sequence $(x^{(k)})_{k \geq 0}$ defined in (6.2). In the following, we prove that the sequence $(u_k)_{k \in \mathbb{N}}$ with $u_k := \sum_{1 \leqslant i \leqslant p} f_\alpha(|x_i^{(k)}| + \Delta)$ is stationary.

For $k \geq 1$, the vector $x^{(k)}$ is a solution of a weighted basis pursuit problem. Consequently, the lemma 6.3 insures that $x^{(k)} \in S$. Since S is a finite set, the sequence $(u_k)_{k \leq 1}$ can only take a finite number of values

$$\forall k \in \mathbb{N}^*, u_k \in \left\{ \sum_{1 \leqslant i \leqslant p} f_\alpha(|x_i^I| + \Delta), I \in \mathcal{E} \right\}.$$

If we show that the sequence $(u_k)_{k \in \mathbb{N}}$ is decreasing that implies its stationary for a large enough k. We follow the proof given in Hunter and Lange (2004); Lange (2004). Remind that $x^{(k+1)}$ is

defined as follow

$$x^{(k+1)} := \operatorname{argmin} \sum_{1\leqslant i\leqslant p} f_\alpha(|x_i^{(k)}| + \Delta) + f'_\alpha(|x_i^{(k)}| + \Delta)(|x_i| - |x_i^{(k)}|).$$

Let us set $L_{x^{(k)}}(x) := \sum_{1\leqslant i\leqslant p} f_\alpha(|x_i^{(k)}| + \Delta) + f'_\alpha(|x_i^{(k)}| + \Delta)(|x_i| - |x_i^{(k)}|)$. The concavity of the function $x \in \mathbb{R} \mapsto f_\alpha(x + \Delta)$ on $\mathbb{R}_+$ implies that

$$\forall x \in \mathbb{R}^p, \sum_{1\leqslant i\leqslant p} f_\alpha(|x_i| + \Delta) \leqslant L_{x^{(k)}}(x).$$

Because, the minimum of $L_{x^{(k)}}(x)$ is reached at $x^{(k+1)}$, one obtains the following property

$$u_{k+1} = \sum_{1\leqslant i\leqslant p} f_\alpha(|x_i^{(k+1)}| + \Delta) \leqslant L_{x^{(k)}}(x^{(k+1)}) \leqslant L_{x^{(k)}}(x^{(k)}) = \sum_{1\leqslant i\leqslant p} f_\alpha(|x_i^{(k)}| + \Delta) = u_k.$$

Since the sequence $(u_k)_{k\in\mathbb{N}}$ is decreasing, there exists $k_0 \geq 0$ such that $(u_k)_{k\in\mathbb{N}}$ is stationary for $k \geq k_0$.

The strict concavity of the function $x \in \mathbb{R}_+ \mapsto f(x + \Delta)$ implies that

$$f_\alpha(|x_i^{(k_0+1)}| + \Delta) \leq f_\alpha(|x_i^{(k_0)}| + \Delta) + f'_\alpha(|x_i^{(k_0)}| + \Delta)(|x_i^{(k_0+1)}| - |x_i^{(k_0)}|),$$

with a strict inequality when $|x_i^{(k_0+1)}| \neq |x_i^{(k_0)}|$. Thus, if there exists $i_0 \in [\![1, p]\!]$ such that $|x_{i_0}^{(k_0+1)}| \neq |x_{i_0}^{(k_0)}|$, $u_{k_0+1} < L_{x^{(k_0)}}(x^{(k_0+1)}) \leq u_{k_0}$ which provides a contradiction for the stationary of the sequence $(u_k)_{k\in\mathbb{N}}$. Consequently, we have

$$\forall i \in [\![1, p]\!], |x_i^{(k_0+1)}| = |x_i^{(k_0)}|.$$

This equality gives that $\operatorname{supp}(x^{(k_0)}) = \operatorname{supp}(x^{(k_0+1)})$. Because $x^{(k_0)}$ and $x^{(k_0+1)}$ are admissible points,

$$\sum_{i\in\operatorname{supp}(x^{(k_0)})} x_i^{(k_0)} d_i = \sum_{i\in\operatorname{supp}(x^{(k_0)})} x_i^{(k_0+1)} d_i.$$

Finally, the lemma 6.3 implies that the family $(d_i)_{i\in\operatorname{supp}(x^{(k_0)})}$ is linearly independent. One deduces that $x^{(k_0)} = x^{(k_0+1)}$. A straightforward consequence is that the sequence $(x^{(k)})_{k\in\mathbb{N}}$ is stationary when $k \geq k_0$. □

Proof of proposition 6.1 : Remind that l is the limit of the sequence $x^{(k)}$ given in (6.2). Let us defined $r := \min\{|l_i|, i \in \operatorname{supp}(l)\}$. One can check that $\forall x \in B_\infty(l, r)$ we have $x_i \neq 0$ once $l_i \neq 0$. Consequently, $\operatorname{supp}(l) \subset \operatorname{supp}(x)$. Assume $\operatorname{supp}(x) = \operatorname{supp}(l)$. Since $Dx = Dl$, one

deduces that

$$\sum_{i\in\text{supp}(l)} x_i d_i = \sum_{i\in\text{supp}(l)} l_i d_i.$$

Since the family $(d_i)_{i\in\text{supp}(l)}$ is linearly independent, one deduces that $x = l$. Consequently, $\forall x \in B_\infty(l, r)$ such that $x \neq l$, we have $\text{supp}(l) \subsetneq \text{supp}(x)$ thus, $\|l\|_0 < \|x\|_0$. □

6.6.3 Proof of the theorem 6.3

By the lemma 6.1, for any x^* in $S^\epsilon_{f_\alpha}$, the family $(d_i)_{i\in x^*}$ is linearly independent. Moreover, x^* is an admissible point, thus $\|y - Dx^*\|^2 \leq \epsilon$. Consequently, $x^* \in \bigcup_{I\in\mathcal{I}} E_I$, where

$$\mathcal{I} := \{I \subset [\![1, p]\!] \mid (d_i)_{i\in I} \text{ is linearly independent } \} \text{ and}$$

$$E_I := \{x \in \mathbb{R}^p \mid \text{supp}(x) \subset I \text{ and } \|y - Dx\|^2 \leq \epsilon\}.$$

Let us denotes $E := \bigcup_{I\in\mathcal{I}} E_I$.

Lemma 6.4 *The set E is compact.*

Proof : Let us denote $\bar{x} \in \mathbb{R}^p$ with $\text{supp}(\bar{x}) \subset I$ such that $D\bar{x}$ is the orthogonal projection of y onto the space $\text{Vect}(d_i)_{i\in I}$. If $\|y - D\bar{x}\|^2 > \epsilon$ then the set E_I is empty. Otherwise,

$$E_I = \{x \in \mathbb{R}^p \mid \text{supp}(x) \subset I \text{ and } \|D(x - \bar{x})\|^2 \leq \epsilon'\}, \text{ with } \epsilon' = \epsilon - \|y - D\bar{x}\|^2.$$

Since $\text{supp}(x) \subset I$ and $\text{supp}(\bar{x}) \subset I$, one shows that

$$\|D(x - \bar{x})\|^2 = \|D_S(x_I - \bar{x}_I)\|^2,$$

with $x_I := (x_i)_{i\in I}$, $\bar{x}_I := (\bar{x}_i)_{i\in I}$ and D_I is matrix whose columns are $(d_i)_{i\in S}$. Because the family $(d_i)_{i\in I}$ is linearly independent, the Gram matrix $D_I^T D_I$ is invertible thus, $\|D_I(x_I - \bar{x}_I)\|^2 \leq \epsilon'$ is an ellipsoid of $\mathbb{R}^{\text{Card(I)}}$. Therefore, E_I is a compact. Consequently, the finite union of compact set $\bigcup_{I\in\mathcal{I}} E_I$ is a compact set. □

In the lemma 6.5 and the theorem 6.3, we denote $s_0 := \min \|x\|$ subject to $\|y - Dx\|^2 \leq \epsilon$.

Lemma 6.5 *For $\eta > 0$, let us denote G_η the open set $G_\eta = \bigcup_{x\in S^\epsilon_0} B(x, \eta)$. The function*

$$F_\alpha : x \in E \setminus G_\eta \mapsto \min\left\{s_0 + 1, \sum_{i=1}^{p} f_\alpha(|x_i|)\right\}$$

converges uniformly to the function $F : x \in E \setminus G_\eta \mapsto s_0 + 1$ when α converges to 0.

Proof : Let $(\alpha_n)_{n\in\mathbb{N}}$ be a decreasing sequence converging toward 0. Because $f_\alpha \geq f_{\alpha'}$ once $\alpha \leq \alpha'$, $(F_{\alpha_n})_{n\in\mathbb{N}}$ is a monotonic sequence of continuous functions. Furthermore, on the compact set

$E\backslash G_\eta$, this sequence converges pointwise toward the continuous function $F : x \in E\backslash G_\eta \mapsto s_0+1$. Consequently, the Dini's theorem gives the uniform convergence of $(F_{\alpha_n})_{n\in\mathbb{N}}$. Therefore, for all $\delta > 0$, there exists n_0 such that

$$\forall n \geq n_0, \sup_{x\in E\backslash G_\eta} \{|F_{\alpha_n}(x) - s_0 - 1|\} \leq \delta.$$

Finally, if $\alpha \leq \alpha_{n_0}$, for all $x \in E \setminus G_\eta$ we have the following inequalities

$$-\delta \leq F_{\alpha_{n_0}}(x) - s_0 - 1 \leq F_\alpha(x) - s_0 - 1 \leq 0.$$

Consequently, one obtains

$$\sup_{x\in E\backslash G_\eta} \{|F_\alpha(x) - s_0 - 1|\} \leq \sup_{x\in E\backslash G_\eta} \left\{|F_{\alpha_{n_0}}(x) - s_0 - 1|\right\} \leq \delta,$$

which shows the uniform convergence. □

Proof of theorem 6.3 : Let x^* be an arbitrary element of S_0^ϵ, we are going to prove that for $\alpha > 0$ small enough,

$$\forall x \in E \setminus G_\eta, \sum_{i=1}^{p} f_\alpha(|x_i|) > \sum_{i=1}^{p} f_\alpha(|x_i^*|). \tag{6.9}$$

If the inequality (6.9) holds then $S_{f_\alpha}^\epsilon \subset G_\eta$. Actually, by definition, $S_{f_\alpha}^\epsilon \subset E$ and by the inequality (6.9), the elements of $E\backslash G_\eta$ are not solution of $\mathcal{P}_{f_\alpha}^\epsilon$. The convergence of $\sum_{i=1}^{p} f_\alpha(|x_i^*|)$ toward s_0 once α converges to 0 implies that

$$\exists \alpha_1 > 0 \text{ such that } \forall \alpha \leq \alpha_1, \sum_{i=1}^{p} f_\alpha(|x_i^*|) < s_0 + 1/2.$$

The uniform convergence given in the previous lemma 6.5 implies that

$$\exists \alpha_2, \forall \alpha \leq \alpha_2, \forall x \in E \setminus G_\eta \min \left\{ s_0 + 1, \sum_{i=1}^{p} f_\alpha(|x_i|) \right\} > s_0 + 1/2.$$

Finally, if we set $\alpha_0 = \min\{\alpha_1, \alpha_2\}$, we have

$$\forall \alpha \leq \alpha_0, \forall x \in E \setminus G_\eta, \min \left\{ s_0 + 1, \sum_{i=1}^{p} f_\alpha(|x_i|) \right\} - \sum_{i=1}^{p} f_\alpha(|x_i^*|) > 0,$$

which implies

$$\forall \alpha \leq \alpha_0, \forall x \in E \setminus G_\eta, \sum_{i=1}^{p} f_\alpha(|x_i|) > \sum_{i=1}^{p} f_\alpha(|x_i^*|).$$

□

6.6.4 Proof of the theorem 6.4 and of the proposition 6.2

Let $(x^{(\phi(k))})_{k\geq 0}$ be a subsequence of $x^{(k)}$ (defined in 6.3) that converges to $\tilde{x}$. The lemmas 6.6, 6.7 and 6.8 are used to prove that the sequence $(x^{(\phi(k)+1)})_{k\geq 0}$ has the same limit as $(x^{(\phi(k))})_{k\geq 0}$.

Lemma 6.6 *Let $f : \mathbb{R}_+ \to \mathbb{R}$ be an striclty increasing, strictly concave and two times differentiable function such that f' is convex then,*

$$\forall \eta > 0, \exists \epsilon > 0 \text{ such that } \forall a \in [0, a_0], \forall b \in \mathbb{R}_+, |a-b| > \eta \Rightarrow f(a) + f'(a)(b-a) - f(b) > \epsilon. \tag{6.10}$$

Proof : Let us defined the function $g_{a_0}(h)$ as follows

$$\forall h \geq 0, g_{a_0}(h) := f(a_0) + f'(a_0)h - f(a_0 + h).$$

We are going to prove that (6.10) holds when $\epsilon = g_{a_0}(\eta)$. In a first step, let us prove that $f(a) + f'(a)(b-a) - f(b) \geq g_{a_0}(|b-a|)$. We set $t = b - a$, the convexity of f' gives

$$\frac{\partial}{\partial a}\left(f(a) + f'(a)|t| - f(a+|t|)\right) = f'(a) + f''(a)|t| - f'(a+|t|) \leq 0.$$

The concavity of f gives

$$f(a) + f'(a)t - f(a+t) \geq f(a) + f'(a)|t| - f(a+|t|).$$

Indeed, when $t \geq 0$, the result is obvious otherwise, when $t < 0$, we have $t = -|t|$, the previous inequality is a consequence of the next one

$$\frac{f(a) - f(a-|t|)}{|t|} \geq f'(a) \geq \frac{f(a+|t|) - f(a)}{|t|}$$

From these inequalities, one deduces that

$$f(a)+f'(a)t-f(a+t) \geq f(a)+f'(a)|t|-f(a+|t|) \geq f(a_0)+f'(a_0)|t|-f(a_0+|t|) = g_{a_0}(|b-a|).$$

The function f' is strictly decreasing (because f is strictly concave) consequently $\forall h > 0, g'_{a_0}(h) = f'(a_0) - f'(a_0 + h) > 0$ thus, g is strictly increasing. Since $g_{a_0}(0) = 0$, we have $\epsilon := g_{a_0}(\eta) > 0$. Finally, if $|b-a| > \eta$ we have

$$f(a) + f'(a)(b-a) - f(b) \geq g_{a_0}(|b-a|) > g_{a_0}(\eta) = \epsilon.$$

□

In the following, we denote $|x| := (|x_i|)_{1\leq i\leq p}$ with $x \in \mathbb{R}^p$.

Lemma 6.7 *The sequence $(x^{(k)})_{k\in\mathbb{N}}$ described in (6.3) satisfies*

$$\lim_{k\to+\infty} d_\infty(|x^{(k+1)}|, |x^{(k)}|) = 0$$

Proof : Let us define the sequence $(u_k)_{k\in\mathbb{N}}$ with $u_k := \sum_{1\leq i\leq p} f_\alpha(|x_i^{(k)}| + \Delta)$. The convergence of this sequence is given in the proof of the theorem 6.2.

Assume that $d_\infty(|x^{(k+1)}|, |x^{(k)}|)$ does not converge to 0, we have

$$\exists \eta > 0, \forall K \geq 0, \exists k_0 \geq K \text{ such that } d_\infty(|x^{(k_0+1)}|, |x^{(k_0)}|) \geq \eta.$$

If $d_\infty(|x^{(k_0+1)}|, |x^{(k_0)}|) \geq \eta$ then, there exists $i_0 \in [\![1, p]\!]$ such that $\left||x_{i_0}^{(k_0+1)}| - |x_{i_0}^{(k_0)}|\right| \geq \eta$. Because the sequence $(x^{(k)})_{k\in\mathbb{N}}$ is bounded (proof 1 of the theorem 6.4), there exists $a_0 \geq 0$ such that $\forall k \in \mathbb{N}, \|x^{(k)}\|_\infty \leq a_0$. By the lemma 6.6 we have

$$\exists \epsilon > 0 \text{ such that } f_\alpha(|x_{i_0}^{(k_0)}| + \Delta) + f'_\alpha(|x_{i_0}^{(k_0)}| + \Delta)(|x_{i_0}^{(k_0+1)}| - |x_{i_0}^{(k_0)}|) - f_\alpha(|x_{i_0}^{(k_0+1)}| + \Delta) \geq \epsilon.$$

Furthermore the concavity of f_α implies that

$$\forall i \neq i_0, f_\alpha(|x_i^{(k_0)}| + \Delta) + f'_\alpha(|x_i^{(k_0)}| + \Delta)(|x_i^{(k_0+1)}| - |x_i^{(k_0)}|) - f_\alpha(|x_i^{(k_0+1)}| + \Delta) \geq 0.$$

These two inequalities imply that

$$u_{k_0+1} + \epsilon = \sum_{i=1}^{p} f_\alpha(|x_i^{(k_0+1)}| + \Delta) + \epsilon \leq \sum_{i=1}^{p} f_\alpha(|x_i^{(k_0)}| + \Delta) + f'_\alpha(|x_i^{(k_0)}| + \Delta)(|x_i^{(k_0+1)}| - |x_i^{(k_0)}|)$$

Furthermore, by definition of $x^{(k_0+1)}$, we have

$$\sum_{i=1}^{p} f_\alpha(|x_i^{(k_0)}| + \Delta) + f'_\alpha(|x_i^{(k_0)}| + \Delta)(|x_i^{(k_0+1)}| - |x_i^{(k_0)}|) \leq \sum_{i=1}^{p} f_\alpha(|x_i^{(k_0)}| + \Delta) = u_{k_0}.$$

The previous inequality implies that

$$\forall K, \exists k_0 \geq K \text{ such that } |u_{k_0+1} - u_{k_0}| \geq \epsilon.$$

The last inequality provides a contradiction for the convergence of the sequence $(u_k)_{k\in\mathbb{N}}$. □

Lemma 6.8 *Let $x^{(\phi(k))}$ be a subsequence of $(x^{(k)})_{k\in\mathbb{N}}$ that converges toward $\tilde{x}$ then, the sequence $(x^{(\phi(k)+1)})_{k\in\mathbb{N}}$ converges toward $\tilde{x}$.*

Proof : The proof 1) in the theorem 6.4 shows that the sequence $(x^{(k)})_{k\in\mathbb{N}}$ is bounded. Consequently, $(x^{(\phi(k)+1)})_{k\in\mathbb{N}}$ is bounded too. To prove that the bounded sequence $(x^{(\phi(k)+1)})_{k\in\mathbb{N}}$ converges to $\tilde{x}$, it is sufficient to show that $\tilde{x}$ is the only limit point of this sequence. Let

$(x^{(\phi(\psi(k))+1)})_{k\in\mathbb{N}}$ be a converging subsequence such that

$$\lim_{k\to+\infty} x^{(\phi(\psi(k))+1)} = \tilde{x}^1, \text{ with } \tilde{x}_1 \neq \tilde{x}.$$

By the lemma 6.7, we have $\lim_{k\to+\infty} d_\infty(|x^{(\phi(\psi(k))+1)}|, |x^{(\phi(\psi(k)))}|) = 0$. Since $\lim_{k\to+\infty} x^{(\phi(\psi(k)))} = \tilde{x}$, one deduces that $|\tilde{x}| = |\tilde{x}^1|$. Let us define $\tilde{x}^2$ as $\tilde{x}^2 := (\tilde{x}^1 + \tilde{x})/2$. Because

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

we have

$$\sum_{i=1}^{p} f'_\alpha(|x_i^{(\phi(\psi(k)))}| + \Delta)|x_i^{(\phi(\psi(k))+1)}| \leq \sum_{i=1}^{p} f'_\alpha(|x_i^{(\phi(\psi(k)))}| + \Delta)(|\tilde{x}_i^2|).$$

Taking the limit in the previous expression, one obtains

$$\sum_{1\leq i\leq p} f'_\alpha(|\tilde{x}_i| + \Delta)|\tilde{x}_i^1| \leq \sum_{1\leq i\leq p} f'_\alpha(|\tilde{x}_i| + \Delta)|\tilde{x}_i^2|. \tag{6.11}$$

On the other hand, $\text{supp}(\tilde{x}^2) = \{i \in \text{supp}(\tilde{x}^1) \mid \tilde{x}_i = \tilde{x}_i^1\}$, which implies that $\text{supp}(\tilde{x}^2) \subsetneq \text{supp}(\tilde{x}^1)$ and $\forall i \in \text{supp}(\tilde{x}^2), \tilde{x}_i^2 = \tilde{x}_i^1$. Consequently, we have

$$\sum_{1\leq i\leq p} f'_\alpha(|\tilde{x}_i|+\Delta)|\tilde{x}_i^1| > \sum_{i\in\text{supp}(\tilde{x}^2)} f'_\alpha(|\tilde{x}_i|+\Delta)|\tilde{x}_i^1| = \sum_{i\in\text{supp}(\tilde{x}^2)} f'_\alpha(|\tilde{x}_i|+\Delta)|\tilde{x}_i^2| = \sum_{1\leq i\leq p} f'_\alpha(|\tilde{x}_i|+\Delta)|\tilde{x}_i^2|. \tag{6.12}$$

The inequality (6.12) provides a contradiction with the inequality (6.11). Therefore, the only limit point of the bounded sequence $(x^{(\phi(k)+1)})_{k\in\mathbb{N}}$ is $\tilde{x}$. □

Lemma 6.9 *Let x^ω be a solution of the weighted lasso problem*

$$\text{argmin} \sum_{i=1}^{p} w_i|x_i| \text{ subject to } \|y - Dx\|^2 \leqslant \epsilon, \text{with } \forall i \in [\![1, p]\!], \omega_i > 0. \tag{6.13}$$

Furthermore, let us assume that $\|y\|^2 > \epsilon$ then, $\|y - Dx^\omega\|^2 = \epsilon$.

Proof : Let us assume that $\|y - Dx^\omega\|^2 < \epsilon$. Consider the points $x(t)$ defined by

$$\forall i \in [\![1, p]\!], x_i(t) = sign(x_i^\omega)(|x_i^\omega| - t)_+, \text{ where } (a)_+ = \max\{a, 0\}.$$

One can check that $\|x(t) - x^\omega\|_\infty \leq t$. Because the set $\{x \in \mathbb{R}^p \mid \|y - Dx\|^2 < \epsilon\}$ is an open set, there exists $t_0 > 0$ small enough such that $\|y - Dx(t_0)\|^2 < \epsilon$. Finally, we have

$$\forall i \notin \text{supp}(x^\omega), |x_i(t_0)| = |x_i^\omega| = 0 \text{ and } \forall i \in \text{supp}(x^\omega), |x_i(t_0)| < |x_i^\omega|.$$

Because $\vec{0}$ is not an admissible point, one has $x^\omega \neq \vec{0}$. Consequently, we have the following

inequality.

$$\sum_{i=1}^{p} w_i|x_i(t_0)| < \sum_{i=1}^{p} w_i|x_i^{\omega}|.$$

Such a result provides a contradiction for the minimality of $\sum_{i=1}^{p} \omega_i|x_i^{\omega}|$. □

Proof of theorem 6.4 :

1) For any $k \geq 1$, $x^{(k)}$ is the solution of a weighted lasso. By lemma 6.3, the family $(d_i)_{i\in \text{supp}(x^{(k)})}$ is linearly independent. Consequently, $\forall k \geq 1, x^{(k)} \in E$, where E is the set given in the lemma 6.4. Because E is a compact set of $\mathbb{R}^p$, one deduces that $(x^{(k)})_{k\in\mathbb{N}}$ is bounded.

2-i) Because $\lim_{k\to+\infty} x^{(\phi(k))} = \tilde{x}$, there exists k_0 such that

$$\forall k \geq k_0, \text{supp}(\tilde{x}) \subset \text{supp}(x^{(\phi(k))}).$$

Since by lemma 6.3 $(d_i)_{i\in\text{supp}(x^{(k_0)})}$ is linearly independent, one deduces that $(d_i)_{i\in\text{supp}(\tilde{x})}$ is linearly independent.

2-ii) For any $k \geq 1$, $x^{(k)}$ is the solution of a weighted lasso with positive weights and $\|y\|^2 > \epsilon$. Consequently from the lemma 6.9, for all $k \geq 1$, $\|y - Dx^{(k)}\|^2 = \epsilon$. Because the set $\{x \in \mathbb{R}^p \mid \|y - Dx\|^2 = \epsilon\}$ is a closed set, one deduces that the limit point $\tilde{x}$ satisfies $\|y - D\tilde{x}\|^2 = \epsilon$.

2-iii) By definition of $x^{(k)}$ we have

$$x^{(\phi(k)+1)} := \text{argmin} \sum_{i=1}^{p} f'_{\alpha}(|x_i^{(\phi(k))}| + \Delta)|x_i| \text{ subject to } \|y - Dx\|_2^2 \leqslant \epsilon.$$

According to Bertsekas (1999) (chapter 5.3), there exists $\lambda \geq 0$ such that

$$x^{(\phi(k)+1)} := \text{argmin} f'_{\alpha}(|x_i^{(\phi(k))}| + \Delta)|x_i| + \lambda\|y - Dx\|_2^2.$$

Consequently, the subdifferential of the previous expression evaluated in $x^{(\phi(n)+1)}$ contains the null vector

$$0 \in \partial\text{pen}(x^{(\phi(k)+1)}) - \lambda D^T(y - Dx^{(\phi(k)+1)}), \tag{6.14}$$

with $\partial\text{pen}(x^{(\phi(k)+1)}) = C_1 \times \cdots \times C_p$, where

$$C_i := \begin{cases} \left[-f'_{\alpha}(|x_i^{(\phi(k))}| + \Delta), f'_{\alpha}(|x_i^{(\phi(k))}| + \Delta)\right] \text{ if } x_i^{(\phi(k)+1)} = 0 \\ \text{sign}(x_i^{\phi(k)+1}) f'_{\alpha}(|x_i^{(\phi(k))}|) \text{ otherwise} \end{cases}.$$

Since $\lim_{n\to+\infty} x^{(\phi(k))} = \lim_{k\to+\infty} x^{(\phi(k)+1)} = \tilde{x}$, the vectors $(x^{(\phi(k)+1)})_{i\in\text{supp}(\tilde{x})}$ and $(\tilde{x})_{i\in\text{supp}(\tilde{x})}$ have the same sign for k large enough. Moreover, since f'_{α} is continuous, by taking the limit in (6.14), we see that the vectors $(d_i^T(y - D\tilde{x}))_{i\in\text{supp}(\tilde{x})}$ and $(\text{sign}(\tilde{x}_i)f'_{\alpha}(\tilde{x}_i))_{i\in\text{supp}(\tilde{x})}$ are collinear.

Proof of proposition 6.2 :

i) The proof of this part is exactly the same as the one provided in lemma 6.9.

ii) The proof of this part is exactly the same as the one provided in lemma 6.2.

iii) The vector $x^{\alpha}_{\text{supp}(x^{\alpha})} := (x^{\alpha}_i)_{i\in\text{supp}(x^{\alpha})}$ is a solution of the problem

$$\text{argmin} \sum_{i\in\text{supp}(x^{\alpha})} f_{\alpha}(|x_i|+\Delta) \text{ subject to } \|y-\tilde{D}x\|_2^2 \leqslant \epsilon, \text{ where } \tilde{D} \text{ has columns } (d_i)_{i\in\text{supp}(x^{\alpha})}. \tag{6.15}$$

Indeed, assume that $x^{\alpha}_{\text{supp}(x^{\alpha})}$ is not a solution of the previous problem, then there exists $\bar{x} \in \mathbb{R}^{\text{Card(supp(x}^{\alpha}))}$ such that

$$\|y-\tilde{D}\bar{x}\|_2^2 \leqslant \epsilon \text{ and } \sum_{i\in\text{supp}(x^{\alpha})} f_{\alpha}(|\bar{x}_i|+\Delta) < \sum_{i\in\text{supp}(x^{\alpha})} f_{\alpha}(|x^{\alpha}_i|+\Delta).$$

Let us set $x' \in \mathbb{R}^p$ such that $x'_i := \bar{x}_i$ if $i \in \text{supp}(x^{\alpha})$ and $x'_i := 0$ otherwise. By definition of x' we have $\|x'\|_0 \leq \|x^{\alpha}\|_0$. On the other hand, since $\tilde{D}\bar{x} = Dx'$ we have $\|y-Dx'\|^2 \leq \epsilon$ therefore $x' \in S_0^{\epsilon}$. Let us show that $\sum_{i=1}^p f_{\alpha}(|x'_i|+\Delta) < \sum_{i=1}^p f_{\alpha}(|x^{\alpha}_i|+\Delta)$

$$\begin{aligned}\sum_{i=1}^{p} f_{\alpha}(|x'_i|+\Delta) &= \sum_{i\notin\text{supp}(x^{\alpha})} f_{\alpha}(\Delta) + \sum_{i\in\text{supp}(x^{\alpha})} f_{\alpha}(|\bar{x}_i|+\Delta), \\ &< \sum_{i\notin\text{supp}(x^{\alpha})} f_{\alpha}(\Delta) + \sum_{i\in\text{supp}(x^{\alpha})} f_{\alpha}(|x^{\alpha}_i|+\Delta) = \sum_{i=1}^{p} f_{\alpha}(|x^{\alpha}_i|+\Delta).\end{aligned}$$

The previous inequality contradicts that $x^{\alpha} \in L^{\alpha}$. According to Bertsekas (1999) (chapter 5.3), there exists $\lambda \geq 0$ such that $x^{\alpha}_{\text{supp}(x^{\alpha})}$, the solution of (6.15), is also the solution of the problem

$$\text{argmin} \sum_{i\in\text{supp}(x^{\alpha})} f_{\alpha}(|x_i|+\Delta) + \lambda\|y-\tilde{D}x\|^2, \text{ where } \lambda \geq 0.$$

Because the partial derivatives of $\sum_{i\in\text{supp}(x^{\alpha})} f_{\alpha}(|x_i|+\Delta) + \lambda\|y-\tilde{D}x\|^2$ at $x^{\alpha}_{\text{supp}(x^{\alpha})}$ are null we have

$$\forall i \in \text{supp}(x^{\alpha}),\ \text{sign}(x^{\alpha}_i) f'_{\alpha}(|x^{\alpha}_i|+\Delta) - \lambda d_i^T(y-\tilde{D}x^{\alpha}_{\text{supp}(x^{\alpha})}) = 0.$$

Since $\tilde{D}x^{\alpha}_{\text{supp}(x^{\alpha})} = Dx^{\alpha}$, one obtains that the vectors $(\ \text{sign}(x^{\alpha}_i) f'_{\alpha}(|x^{\alpha}_i|+\Delta))_{i\in\text{supp}(x^{\alpha})}$ and $\left(d_i^T(y-Dx^{\alpha})\right)_{i\in\text{supp}(x^{\alpha})}$ are colinear. □

6.7 Appendix 2 : Simulations with partial random circulant matrices

We use the same setting of simulation than the one given in the section 6.4 except that here D is a partial random circulant matrix as defined in Rauhut (2010). First, before to introduce D, let us define a random circulant matrix Φ. Let $b_0, \ldots, b_{p-1}$ be i.i.d Rademacher random

variables and define the circulant random matrix Φ as follows

$$\Phi = \begin{pmatrix} b_0 & b_1 & \dots & \dots & b_{p-1} \\ b_{p-1} & b_0 & b_1 & \dots & b_{p-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & b_1 \\ b_1 & \dots & \dots & b_{p-1} & b_0 \end{pmatrix}.$$

Let I be a random set independent from $b_0 \dots, b_{p-1}$ having a uniform distribution on combinations of n elements among p (thus $\text{card}(I) = n$). Let us define the $n \times p$ partial random circulant matrix D as follows

$$D = (d_1 | \dots | d_p) = (\Phi_{i,j}/\sqrt{n})_{i \in I, j \in [\![1,p]\!]}.$$

The columns of D are normalized so that $\|d_1\|_2 = \dots = \|d_p\|_2 = 1$. We choose $\tilde{x}$ with $\text{supp}(\tilde{x})$ as in section 6.4 and we let $y = D\tilde{x}$.

6.7.1 Comparisons

For each observation of D and $\tilde{x}$, we compute the basis pursuit solution (denoted x^{bp}) of $\mathcal{P}_1$, the reweighted l^1 minimization solution and the solution given by our method as defined in (6.6). 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 our method was computed using the algorithm described in the figure 6.4. After 8 iterations, we kept the sparsest solution among the one obtained with $\Delta \in \{0.01, 0.1, 0.5, 1, 2, 4\}$.

The figure 6.6 shows the performances of the basis pursuit, the reweighted l^1 minimization and our method.

The performances of these three methods namely the basis pursuit, the l^1 reweighted and our method are similar to the performances given in the section 6.4 namely on figure 6.5. The next subsection suggests that these similar performances are due to the proximity between the kernel of the matrices D and $\tilde{D}$, when n is large.

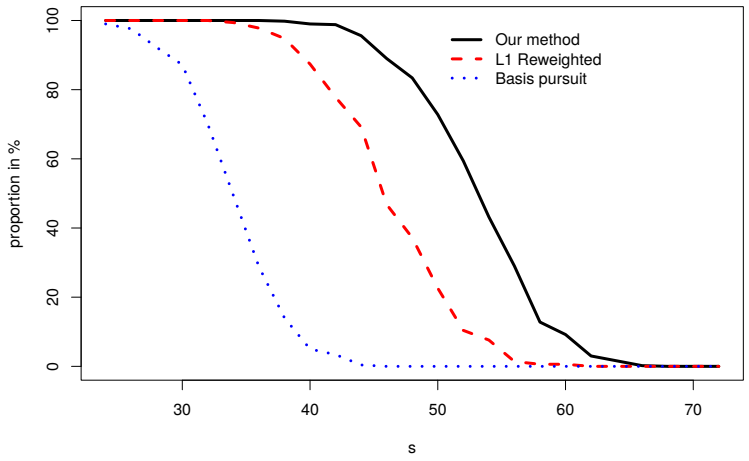


FIGURE 6.6 – The performances of the basis pursuit, l^1 reweighted and our method are represented by the proportions of realisations of the events $x^{\text{bp}} = \tilde{x}$, $x^{\text{l1},(8)} = \tilde{x}$ and $x^{(8)} = \tilde{x}$ as a function of the number of non null components of $\tilde{x}$ denoted s.

6.7.2 Comments on these simulations

Remind that $\tilde{D} := (\tilde{d}_1|\dots|\tilde{d}_p)$ where $\tilde{d}_i = \zeta_i/\|\zeta\|_i$ and $\zeta_1, \dots, \zeta_p$ are i.i.d $\mathcal{N}(0, Id_n)$ distributed and that $\tilde{D}$ is a random matrix used in the numerical study of the section 6.4.

Instead to consider random matrices D and $\tilde{D}$, to simplify, we consider $n \times p$ fixed matrices A and B. Let $\omega_1 > 0, \dots, \omega_p > 0$ be positive weights and let $S_\omega(A)$, $S_\omega(B)$ be the solutions of the following weighted basis pursuit problem

$$S_\omega(A) := \operatorname{argmin} \begin{cases} \sum_{i=1}^p \omega_i |x_i| \\ \text{subject to } Ax = A\tilde{x} \end{cases} \quad \text{and } S_\omega(B) := \operatorname{argmin} \begin{cases} \sum_{i=1}^p \omega_i |x_i| \\ \text{subject to } Bx = B\tilde{x} \end{cases}.$$

To recover $\tilde{x}$ using the measurement matrix A (resp. B) with the basis pursuit we compute $S_\omega(A)$ (resp. $S_\omega(B)$) with $w_1 = \dots = w_p = 1$. For the l^1 reweighted method or for our method, the weights $\omega_1, \dots, \omega_p$ are computed iteratively. A similar proof to the one of the proposition 5.3 allows to show that $\ker(A) = \ker(B)$ implies that $S_\omega(A) = S_\omega(B)$. In particular, when $\ker(A) = \ker(B)$, the performances to recover $\tilde{x}$ using the measurement matrix A and the basis pursuit, the l^1 reweighted method or our method are equal to performances to recover $\tilde{x}$ using the measurement matrix B and one of these three methods.

A sufficient condition to prove that $\ker(A) = \ker(B)$ is to show that the Gram matrices $A^T A$ and $B^T B$ are equal. In our simulation study, the matrices D and $\tilde{D}$ are random thus neither $D^T D = \tilde{D}^T \tilde{D}$ nor $\mathcal{L}(D^T D) = \mathcal{L}(\tilde{D}^T \tilde{D})$ hold. However, asymptotically we have the following result

$$\sqrt{n}\langle d_i, d_j \rangle \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1) \text{ and } \sqrt{n}\langle \tilde{d}_i, \tilde{d}_j \rangle \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Consequently, when n is large, the marginal distributions of the random matrices $D^T D$ and $\tilde{D}^T \tilde{D}$ are close. In my opinion, the proximity between these two distributions of $D^T D$ and $\tilde{D}^T \tilde{D}$ could explain that the curves provided in figure 6.5 and 6.6 are similar.

Bibliographie

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

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

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

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

Bunea, F., Tsybakov, A., and Wegkamp, M. (2007). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1 :169–194.

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

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

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

Candes, E. J., Romberg, J. K., and Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8) :1207–1223.

Candes, E. J. and Tao, T. (2005). Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12) :4203–4215.

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

Chartrand, R. (2007). Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Processing Letters*, 14(10) :707–710.

Cohen, A., Dahmen, W., and DeVore, R. (2009). Compressed sensing and best k-term approximation. *Journal of the American Mathematical Society*, 22(1) :211–231.

Daubechies, I., DeVore, R., Fornasier, M., and Güntürk, C. S. (2010). Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics*, 63(1) :1–38.

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

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

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

Dossal, C. (2012). A necessary and sufficient condition for exact sparse recovery by l_1 minimization. *Comptes Rendus Mathematique*, 350(1) :117–120.

Foucart, S. and Lai, M.-J. (2009). Sparsest solutions of underdetermined linear systems via l_q-minimization for $0 < q \leq 1$. *Applied and Computational Harmonic Analysis*, 26(3) :395–407.

Foucart, S. and Rauhut, H. (2013). *A mathematical introduction to compressive sensing*. Springer.

Gribonval, R. and Nielsen, M. (2003). Sparse representations in unions of bases. *IEEE Transactions on Information Theory*, 49(12) :3320–3325.

Gribonval, R. and Nielsen, M. (2007). Highly sparse representations from dictionaries are unique and independent of the sparseness measure. *Applied and Computational Harmonic Analysis*, 22(3) :335–355.

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

Herman, M. A. and Strohmer, T. (2009). High-resolution radar via compressed sensing. *IEEE transactions on signal processing*, 57(6) :2275–2284.

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, 58(1) :30–37.

Lai, M.-J. (2010). On sparse solutions of underdetermined linear systems. *Journal of Concrete and Applicable Mathematics*, 8(2) :296–327.

Lange, K. (2004). Elementary optimization. In *Optimization*, pages 1–17. Springer.

Liu, J. and Gao, H. (2016). Material reconstruction for spectral computed tomography with detector response function. *Inverse Problems*, 32(11) :114001.

Lobo, M. S., Fazel, M., and Boyd, S. (2007). Portfolio optimization with linear and fixed transaction costs. *Annals of Operations Research*, 152(1) :341–365.

Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electronic Journal of statistics*, 2 :90–102.

Meinshausen, N. (2015). 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.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3) :1436–1462.

Ollier, E. and Viallon, V. (2017). Regression modelling on stratified data with the lasso. *Biometrika*, 104(1) :83–96.

Perrot-Dockès, M., Lévy-Leduc, C., Chiquet, J., Sansonnet, L., Brégère, M., Étienne, M.-P., Robin, S., and Genta-Jouve, G. (2017a). A multivariate variable selection approach for analyzing lc-ms metabolomics data. *arXiv preprint arXiv :1704.00076*.

Perrot-Dockès, M., Lévy-Leduc, C., Sansonnet, L., and Chiquet, J. (2017b). Variable selection in multivariate linear models with high-dimensional covariance matrix estimation. *arXiv preprint arXiv :1707.04145*.

Prieto, K. and Dorn, O. (2016). Sparsity and level set regularization for diffuse optical tomography using a transport model in 2d. *Inverse Problems*, 33(1) :014001.

Raskutti, G., Wainwright, M. J., and Yu, B. (2010). Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(Aug) :2241–2259.

Rauhut, H. (2010). Compressive sensing and structured random matrices. *Theoretical foundations and numerical methods for sparse recovery*, 9 :1–92.

Rosset, S., Zhu, J., and Hastie, T. (2004). Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5 :941–973.

Sun, Q. (2012). Recovery of sparsest signals via l_q-minimization. *Applied and Computational Harmonic Analysis*, 32(3) :329–341.

Sun, W., Wang, J., and Fang, Y. (2013). Consistent Selection of Tuning Parameters via Variable Selection Stability. *Journal of Machine Learning Research*, 14 :3419–3440.

Sustik, M. A., Tropp, J. A., Dhillon, I. S., and Heath, R. W. (2007). On the existence of equiangular tight frames. *Linear Algebra and its applications*, 426(2) :619–635.

Tardivel, P., Servien, R., and Concordet, D. (2017). Sparsest representations and approximations of a high-dimensional linear system. *Submitted*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1) :267–288.

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7 :1456–1490.

Tillmann, A. M. and Pfetsch, M. E. (2014). 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.

van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2) :614–645.

van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3 :1360–1392.

Woodworth, J. and Chartrand, R. (2016). Compressed sensing recovery via nonconvex shrinkage penalties. *Inverse Problems*, 32(7) :075004.

Zhang, Z., Xu, Y., Yang, J., Li, X., and Zhang, D. (2015). A survey of sparse representation : algorithms and applications. *IEEE Access*, 3 :490–530.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7 :2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476) :1418–1429.

Conclusion et perspectives

Ce manuscrit de thèse a été écrit à partir des trois articles suivants

1. A powerful multiple testing procedure in linear Gaussian model (Tardivel et al., 2017b) actuellement soumis,
2. ASICS : an automatic method for identification and quantification of metabolites in complex 1D ^{1}H NMR spectra (Tardivel et al., 2017a) accepté dans *Metabolomics*,
3. Sparsest representations and approximations of a high-dimensional linear system (Tardivel et al., 2017c) actuellement soumis.

Dans le cadre du modèle linéaire gaussien, l'article 1) propose une nouvelle procédure de tests multiples contrôlant le FWER. Dans l'article 2), le spectre d'un mélange complexe est modélisé par la réponse d'un modèle linéaire gaussien. La procédure de tests multiples développée dans l'article 1) permet de tester les hypothèses nulles $\beta_i = 0$ avec $i \in [\![1, p]\!]$ où β_i représente la concentration du i^{eme} métabolite dans le mélange. Dans la troisième partie de ce manuscrit, dans le cadre du modèle linéaire gaussien en grande dimension, nous avons vu qu'il y avait un enjeu à estimer la représentation plus parcimonieuse de l'espérance $m \in \mathbb{R}^n$ d'un vecteur gaussien Y dans une famille génératrice $X_1, \ldots, X_p$ de $\mathbb{R}^n$. Dans le cas non bruité, lorsque $Y = m$, l'article 3) montre qu'une représentation la plus parcimonieuse de m est obtenue *via* la minimisation de la "norme" l^α avec α suffisamment petit.

Certains des résultats obtenus dans ces articles peuvent être améliorés alors que d'autres soulèvent de nouvelles questions. Nous allons lister dans les paragraphes suivants les pistes de recherche dégagées par ces résultats.

Dans l'article 1), dans le cadre du modèle linéaire gaussien, nous avons proposé une nouvelle procédure de tests multiples construite à partir du maximum de vraisemblance $\hat{\beta}^{\text{mle}}$ permettant de tester la nullité des paramètres $\beta_i^* = 0$ pour $i \in [\![1, p]\!]$. Dans les procédures classiques qui contrôlent le FWER, un seuil s, le même pour toutes les hypothèses, est calculé. L'hypothèse nulle $\beta_i^* = 0$ est rejetée dès que $|\hat{\beta}_i^{\text{mle}}|/\text{se}(\hat{\beta}_i^{\text{mle}}) \geq s$. L'innovation par rapport à ces procédures classiques est d'avoir autant de seuils que d'hypothèses à tester. Parmi tous les seuils possibles qui contrôlent le FWER à un niveau fixé, les seuils optimaux $s_1^*, \ldots, s_p^*$ sont ceux pour lesquels le produit $s_1^* \times \cdots \times s_p^*$ est minimal. Nous avons montré que l'utilisation de ces seuils $s_1^*, \ldots, s_p^*$ à la place d'un unique seuil s commun à toutes les hypothèses permet un gain de puissance

moyenne. Bien que nous ayons décrit comment obtenir ces seuils dans certains cas particuliers, leur calcul reste difficile. Une méthode numérique permettant un calcul rapide et précis de ces seuils est un enjeu pour que cette procédure de tests multiples devienne largement utilisée.

Dans ce travail, les résidus sont gaussiens et la matrice de covariance est connue. Ainsi, les écart-types $\mathrm{se}(\hat{\beta}_1^{\mathrm{mle}}), \ldots, \mathrm{se}(\hat{\beta}_p^{\mathrm{mle}})$ sont connus. Lorsque ces écart-types ne sont plus connus mais estimés, les statistiques de test $\hat{\beta}_i^{\mathrm{mle}}/\hat{\mathrm{se}}(\hat{\beta}_i^{\mathrm{mle}})$ suivent, sous l'hypothèse nulle, une loi de student. Il serait intéressant de pouvoir déterminer les seuils optimaux dans ce cadre. Cette étude pourrait de plus être étendue à un cadre beaucoup plus large que le cas gaussien. Pour tout $i \in [\![1, p]\!]$, considérons l'hypothèse nulle $\mathcal{H}_i : \theta_i \in \Theta_i$ avec $\Theta_i \subset \mathbb{R}$ et T_i une statistique de test associée à cette hypothèse. On rejette l'hypothèse nulle dès que $T_i \in R_i$ avec R_i la région de rejet. Pour contrôler le FWER au niveau $\alpha \in [0, 1]$, il suffit que les régions de rejet $R_1, \ldots, R_p$ satisfassent l'inégalité suivante

$$\sup_{\theta_1 \in \Theta_1, \ldots, \theta_p \in \Theta_p} \mathbb{P}_{\theta_1, \ldots, \theta_p}(\exists i \in [\![1, p]\!] \text{ tel que } T_i \in R_i) \leq \alpha.$$

Soit λ la mesure de Lebesgue sur $\mathbb{R}$, nous souhaiterions trouver une procédure de tests multiples pour laquelle le volume de la région d'acceptation $\lambda(\mathbb{R} \setminus R_1) \times \cdots \times \lambda(\mathbb{R} \setminus R_p)$ soit minimale.

Concernant l'article 2), il y a une limite à l'utilisation de la méthode ASICS : lorsque la bibliothèque ne contient pas certains spectres de métabolites présents dans le mélange complexe, les estimateurs $\hat{\alpha}_1, \ldots, \hat{\alpha}_p$ sont biaisés. Ces biais dépendent des spectres non contenus dans la bibliothèque et par conséquent ne peuvent pas être corrigés. Ainsi, dans l'idéal, la bibliothèque devrait contenir tous les spectres des métabolites ayant une concentration non-nulle dans le mélange. En ce qui concerne l'étape de déformation, aucune garantie théorique n'est donnée sur la façon dont les spectres $f_1, \ldots, f_p$ sont "corrigés" en $f_1 \circ \Phi_1^*, \ldots, f_1 \circ \Phi_1^*$. Dans un cadre non bruité, la façon dont les spectres "corrigés" $f_1 \circ \Phi_1^*, \ldots, f_1 \circ \Phi_1^*$ sont obtenus mériterait d'être étudiée.

Malgré ces limitations, les premiers retours sur la méthode ASICS sont encourageants, néanmoins, la convivialité de cette méthode peut être améliorée. Pour le moment, l'analyse d'un spectre de mélange complexe prends environ 2 min 30. Parce que les experts en métabolomique n'analysent pas, en général, un spectre de mélange complexe mais une centaine de spectres de mélanges complexes, il y a un enjeu à diminuer le temps de calcul d'ASICS. Une première façon d'améliorer ASICS serait donc d'implémenter une méthode permettant de calculer rapidement et précisément les seuils optimaux discutés précédemment par exemple en utilisant le langage C++ à la place du logiciel R. Enfin, les spectres doivent être saisis puis traités un par un, il n'est pas possible de les analyser simultanément. Ce détail mériterait d'être pris en compte pour la future version d'ASICS.

Dans l'article 3), nous avons vu qu'une représentation la plus parcimonieuse d'un vecteur $m \in \mathbb{R}^n$ dans une famille génératrice $\{X_1, \ldots, X_p\}$ de $\mathbb{R}^n$ était obtenue *via* la minimisation de la

"norme" l^α avec α suffisamment petit. Lorsque m est l'espérance inconnue d'un vecteur gaussien de loi $\mathcal{N}(m, \sigma^2 Id_n)$, nous souhaiterions construire une région de confiance pour le paramètre inconnu

$$\beta^* := \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{p} |\beta_i|^\alpha \text{ sous la contrainte que } \beta_1 X_1 + \cdots + \beta_p X_p = m.$$

L'écriture précédente suppose l'unicité du paramètre β^* qui peut être garantie sous certaines hypothèses sur la famille $\{X_1, \ldots, X_p\}$. Par exemple, lorsque $\alpha = 1$, la condition générale pour la famille $X_1, \ldots, X_p$ est une condition suffisante (très faible) pour l'unicité de β^* ; lorsque $\alpha < 1$ une hypothèse similaire à la condition générale devrait être formulée. Enfin, l'estimateur plug-in est obtenu en remplaçant m dans l'expression précédente par Y. Un travail futur, consisterait à construire une région de confiance pour β^* à partir de l'estimateur plug-in.

Bibliographie

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

Tardivel, P., Servien, R., and Concordet, D. (2017b). A powerful multiple testing procedure in linear Gaussian model. *Submitted*.

Tardivel, P., Servien, R., and Concordet, D. (2017c). Sparsest representations and approximations of a high-dimensional linear system. *Submitted*.

Résumé

Considérons un vecteur gaussien Y de loi $\mathcal{N}(m, \sigma^2 Id_n)$ et X une matrice non aléatoire de dimension $n \times p$ avec Y observé, m inconnu, σ et X connus. Dans le cadre du modèle linéaire, m est supposé être une combinaison linéaire des colonnes de X. En petite dimension, lorsque $n \geq p$ et que $\ker(X) = 0$, il existe alors un unique paramètre β^* tel que $m = X\beta^*$; on peut alors réécrire Y sous la forme $Y = X\beta^* + \varepsilon$. Dans le cadre du modèle linéaire gaussien en petite dimension, nous construisons une nouvelle procédure de tests multiples contrôlant le FWER pour tester les hypothèses nulles $\beta_i^* = 0$ pour $i \in [\![1, p]\!]$. Cette procédure est appliquée en métabolomique au travers du programme ASICS disponible en ligne. ASICS permet d'identifier et de quantifier les métabolites *via* l'analyse des spectres RMN. En grande dimension, lorsque $n < p$ on a $\ker(X) \neq 0$, ainsi le paramètre β^* décrit précédemment n'est pas unique. Dans le cas non bruité lorsque $\sigma = 0$, impliquant que $Y = m$, nous montrons que les solutions du système linéaire d'équations $Y = X\beta$ ayant un nombre de composantes non nulles minimale s'obtiennent *via* la minimisation de la "norme" l^α avec α suffisament petit.

Mots-clés : Procédure de tests multiples, FWER, Estimateur lasso, Paramètre de régularisation, Minimisation de la norme l^1, Minimisation de la "norme" l^α, Minimisation de la "norme" l^0, Représentation parcimonieuse, Résonance magnétique nucléaire, Identification de métabolites, Quantification de métabolites.

Abstract

Let Y be a Gaussian vector distributed according to $\mathcal{N}(m, \sigma^2 Id_n)$ and X a not random matrix of dimension $n \times p$ with Y observed, m unknown, σ and X known. In the linear model, m is assumed to be a linear combination of the columns of X. In small dimension, when $n \geq p$ and $\ker(X) = 0$, there exists a unique parameter β^* such that $m = X\beta^*$; then we can rewrite Y in the form $Y = X\beta^* + \varepsilon$. In the small-dimensional linear Gaussian model framework, we construct a new multiple testing procedure controlling the FWER to test the null hypotheses $\beta_i^* = 0$ for $i \in [\![1, p]\!]$. This procedure is applied in metabolomics through the freeware ASICS available online. ASICS allows to identify and to quantify metabolites *via* the analyse of RMN spectra. In high dimension, when $n < p$, we have $\ker(X) \neq 0$ consequently the parameter β^* described above is no longer unique. In the noiseless case when $\sigma = 0$, implying thus $Y = m$, we show that the solutions of the linear system of equations $Y = X\beta$ having a minimal number of non-zero components are obtained *via* the l^α minimization with α small enough.

Keywords : Multiple testing procedure, Familywise error rate, Lasso Estimator, Tuning parameter, Basis pursuit, l^α minimization, l^0 minimization, Sparsest representation, Nuclear magnetic resonance, Identification of metabolites, Quantification of metabolites.