Familywise Error Rate Control With a Lasso Estimator

Patrick J.C Tardivel, Rémi Servien, Didier Concordet

Familywise Error Rate Control With a Lasso Estimator

Patrick J.C. Tardivel*, Rémi Servien and Didier Concordet

Toxalim, Université de Toulouse, INRA, ENVT, Toulouse, France.

Abstract

We propose a new method to control the familywise error rate (FWER) in linear Gaussian models. Our method relies on a lasso-type estimator and can be used for all $n \times p$ full-rank design matrices. We provide an explicit and non-asymptotic choice for the lasso tuning parameter that controls the FWER. Numerical experiments highlight the performances 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, Tuning parameter, Metabolomics.

1 Introduction

Let us consider the linear Gaussian model

$$Y = X\beta^* + \varepsilon, \tag{1}$$

where $X = (X_1|\dots|X_p)$ is a $n \times p$ full-rank design matrix, 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 $\mathcal{H}_i$. When a sparse estimator $\hat{\beta}$ of β^* is available, a very simple way to test the hypothesis $\mathcal{H}_i$ is to reject it when $\hat{\beta}_i \neq 0$. The lasso estimator [Tibshirani, 1996] is probably the most popular sparse estimator. It defined by

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

Obviously, the test's performances depend on the estimator $\hat{\beta}(\lambda)$ and, by consequence, on the choice of λ. Meinshausen and Bühlmann [2006], Zhao and Yu [2006], Zou [2006] showed that the irrepresentable condition is

*corresponding author: patrick.tardivel@inra.fr

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 $\mathcal{A}$. Geometrically, this condition means that each variable X_i with $i \notin \mathcal{A}$ is almost orthogonal to the subspace $\text{Vect}\{X_i,\ i \in \mathcal{A}\}$. This condition has been relaxed [Zou, 2006] by considering a consistent estimator of $\mathcal{A}$ based on the adaptive lasso estimator defined by

$$\hat{\beta}^{\text{adapt}}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \left\{ \frac{1}{2} \|Y - X\beta\|^2 + \lambda \sum_{i=1}^{n} \frac{1}{|\tilde{\beta}_i|} |\beta_i| \right\}, \tag{3}$$

where $\tilde{\beta}$ is any consistent estimator of β^*. More precisely, Zou [2006] showed that as soon as $\lambda = Cn^\gamma$, with $C > 0$ and $\gamma \in]0, 1/2[$, the derived estimator $\mathcal{A}(\hat{\beta}^{\text{adapt}}(\lambda))$ is consistent. However, no explicit choice is currently available for C and γ whereas choosing them *a priori* can lead to poor results [Chand, 2012].

In practice, the tuning parameter is often selected using cross-validation. This way of choosing λ can be used for any design matrix X and is implemented in some well-known R packages, such as lars [Efron et al., 2004] or glmnet [Friedman et al., 2010]. Unfortunately, this procedure is unsuitable for the active set estimation [Leng et al., 2006].

Explicit choices of λ have been provided to control the False Discovery Ratio (FDR) [Bogdan et al., 2015, Su and Candes, 2016] or to estimate the active set [Lounici, 2008]. Both works assume that the design matrix is close to an orthogonal matrix (which implies the irrepresentable condition). In addition to this assumption, Lounici [2008] requires that the smallest non-null parameter of β^* is larger than a threshold (beta-min condition) to control the probability of $\{\mathcal{A}(\hat{\beta}(\lambda)) = \mathcal{A}\}$. However, no results are available for a general full rank matrix X.

The λ choice is not an issue for the multiple testing procedures based on the lasso knots $(\hat{\lambda}_i)_{1 \leq i \leq p}$. The lasso knots correspond to values of $\hat{\lambda}$ at which the estimated active set $\mathcal{A}(\hat{\beta}(\hat{\lambda}))$ changes. A first example of a multiple testing procedure based on lasso knots is the covariance test [Lockhart et al., 2014] that tests for a specific knot $(\hat{\lambda}_k)$ the hypothesis $\mathcal{A} \subset \{i \in [\![1,p]\!] \mid \hat{\beta}_i(\hat{\lambda}_k) \neq 0\}$. G'Sell et al. [2015] defined a procedure to test ordered hypotheses $\mathcal{H}_k : \mathcal{A} \subset \{i_1, \ldots, i_{k-1}\}$, where the indices $i_1, \ldots, i_{k-1}$ are such that $\hat{\lambda}_{i_1} \geq \cdots \geq \hat{\lambda}_{i_p}$. This procedure controls the FDR. However, the rejection of the hypothesis $\mathcal{H}_k$ simply indicates that $\mathcal{A}$ is not included in $\{i_1, \ldots, i_{k-1}\}$ but does not provide any information on the set it is included in. Consequently, the set of rejected hypotheses does not allow to properly estimate the active set.

Other multiple testing procedures that use the lasso knots have been developed in Barber and Candès [2015], Janson and Su [2016]. The procedure described in Barber and Candès [2015] controls the FDR while the one given in Janson and Su [2016] controls the $k-$FWER, that is the probability of wrongly rejecting at least k of the true null hypotheses. Both procedures compare knots of the original lasso $(\hat{\lambda}_i)_{1 \leq i \leq p}$ to knockoff lasso knots $(\tilde{\lambda}_i)_{1 \leq i \leq p}$. One can view knots of the knockoff lasso $(\tilde{\lambda}_i)_{1 \leq i \leq p}$ as knots of the lasso when $\forall i \in [\![1,p]\!], \beta_i^* = 0$.

This article proposes a multiple testing procedure that uses a lasso type estimator for which the tuning parameter λ controls the FWER *via* an explicit non-asymptotic choice of λ. This procedure does not require

any condition on the design matrix X.

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), whereas Section 3 addresses the general case where X is a full-rank design matrix. Section 4 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 5 details the analysis of metabolomic data that motivated this work.

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

Proposition 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. If λ_0 is the $1-\alpha$ quantile of $\max_{i\in[\![1,p]\!]}\{d_i \times |Z_i^{\text{ols}}|\}$ then,

$$\mathbb{P}(\forall i \notin \mathcal{A}, \hat{\beta}_i(\lambda_0) = 0) \leq 1 - \alpha. \tag{4}$$

When the covariance matrix Γ is given *a priori*, the distribution of Z^{ols} is known and λ_0 can be obtained by numerical simulations. In the next section we are going to adapt the Proposition 1 to the more general case where X has no longer orthogonal columns.

3 General case

In this section, we assume that the design matrix X is a full rank matrix. Let us consider the set G of applications that orthogonalise X. In other terms, if $U \in G$, the matrix $(UX)^T UX$ is diagonal. 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 of a linear transformation $U \in G$ to each member of the model (1). This leads to the new linear Gaussian model

$$\tilde{Y} = \tilde{X}\beta^* + \tilde{\varepsilon} \text{ with } \tilde{Y} = UY, \tilde{X} = UX \text{ and } \tilde{\varepsilon} = U\varepsilon. \tag{5}$$

Because $\tilde{X}$ has orthogonal columns, it is possible to use the Proposition 1 of the previous section. For all $\lambda \geq 0$, the lasso estimator of β^* is $\hat{\beta}^U(\lambda) = \left(\text{sign}(\hat{\beta}_i^{\text{ols}}(U)) \left(|\hat{\beta}_i^{\text{ols}}(U)| - \lambda/d_i(U) \right)_+ \right)_{1 \leq i \leq p}$ and 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 (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. 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 $\beta^* = 0$,

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

The minimization of ϕ thus leads to minimize the volume of the rectangular parallelepiped $\left[-\frac{\lambda_0^U}{d_1(U)}, \frac{\lambda_0^U}{d_1(U)} \right] \times \cdots \times \left[-\frac{\lambda_0^U}{d_p(U)}, \frac{\lambda_0^U}{d_p(U)} \right]$ among those that have a level $1-\alpha$. The following theorem shows that it is possible to pick a transformation U^* for which ϕ is minimal.

Theorem 1 *There exists a linear transformation $U^* \in G$, such that*

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

The previous theorem gives the existence of U^* but does not guarantee its uniqueness. The building of an optimal U^* can be performed in two steps. First, because we want a small 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 (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 1 *Let $\delta \in]0,+\infty[^p$ and P_δ be an invertible $n \times n$ matrix such that*

$$P_\delta X = \begin{pmatrix} \Delta \\ 0 \end{pmatrix}, \text{ with } \Delta = diag(\sqrt{\delta_1}, \ldots, \sqrt{\delta_p}) \text{ and } 0 \text{ the null matrix.}$$

Let M be a $p \times n$ matrix defined by

$$M = ((P_\delta X)^T (P_\delta \Gamma P_\delta^T)^{-1} P_\delta X)^{-1} (P_\delta X)^T (P_\delta \Gamma P_\delta^T)^{-1} = \begin{pmatrix} M_1 & M_2 \end{pmatrix},$$

where M_1 and M_2 are $p \times p$ and a $p \times (n-p)$ matrices, respectively. Finally, let us consider the $n \times n$ matrix V_δ defined by

$$V_\delta = \begin{pmatrix} Id_p & \Delta M_2 \\ 0 & 0 \end{pmatrix} P_\delta.$$

Then, for all $\delta \in (0,\infty)^p$, the matrix V_δ belongs to G, and $\hat{\beta}^{\text{ols}}(V_\delta)$ has the same variance as the the maximum likelihood estimator of (1).

As explained in the introduction, the lasso estimator requires a "nearly" orthogonal design matrix X to get a consistent the active set estimator. Since most multiple test procedures are based on the lasso estimator, it is always a good idea to apply a linear transformation V_δ in (5) before using it. Lemma 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 (5) is to apply the linear transformation $\Gamma^{-1/2}$. However, and contrarily to the V_δ transformations, the obtained design matrix does not have orthogonal columns.

As an example for Lemma 1, let us set $\Gamma = Id_4$ 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 $P_{(1,1)}$ and $V_{(1,1)}$ satisfying Lemma 1 is

$$P_{(1,1)} := \begin{pmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.5 & -0.5 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix} \text{ and } V_{(1,1)} := \begin{pmatrix} 0.5 & 1/6 & -1/6 & 1/6 \\ 0.5 & -1/6 & 1/6 & -1/6 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

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 efficient

$$\tilde{X} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}^T \text{ and } \text{Var}\left(\hat{\beta}^{\text{ols}}(V_{(1,1)})\right) = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T V V^T \tilde{X}(\tilde{X}^T X)^{-1} = (X^T X)^{-1} = \begin{pmatrix} 1/3 & 1/6 \\ 1/6 & 1/3 \end{pmatrix}.$$

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 *Set*

$$U^* = V_{\delta^*} \text{ with } \delta^* = \underset{\delta \in]0,+\infty[^p}{\text{arginf}} \phi(V_\delta), \tag{6}$$

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

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

As shown in the proof, there always exists at least a vector $\delta^* \in]0,+\infty[^p$ such that the infimum is reached. Consequently, Theorem 1 holds for $U^* = V_{\delta^*}$.

But this lemma does not explain how to get such a δ^*. We did not manage to obtain a closed form of it. However some simple remarks could help its numerical computation. First, because for all $t > 0$, the application $t \mapsto \phi(V_{t\delta})$ is constant, 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 $b_1 = \lambda_0^{V_\delta}/\delta_1, \ldots, b_p = \lambda_0^{V_\delta}/\delta_p$ (resp. $b_1^* = \lambda_0^{V_\delta^*}/\delta_1^*, \ldots, b_p^* = \lambda_0^{V_\delta^*}/\delta_p^*$) and consider the rectangular parallelepiped $B = [-b_1, b_1] \times \ldots \times [-b_p, b_p]$ (resp. $B^* = [-b_1^*, b_1^*] \times \cdots \times [-b_p^*, b_p^*]$). Let Σ the covariance matrix of the maximum likelihood estimator, the centered Gaussian random variable $Z^{\text{opt}} = \hat{\beta}^{ols}(V_\delta) - \beta^*$ is distributed according to $\mathcal{N}(0_{\mathbb{R}^p}, \Sigma)$. The rectangular parallelepiped B^* has the smallest volume among rectangular parallelepiped B such that $P\left(Z^{\text{opt}} \in B\right) = 1 - \alpha$. 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 *Let $b^* = \left(b_1^*, \ldots, b_p^*\right)$ be a solution of the following optimisation problem*

$$\min \prod_{i=1}^{p} b_i \text{ subject to } \mathbb{P}\left(|Z_1^{\text{opt}}| \leq b_1, \ldots, |Z_p^{\text{opt}}| \leq b_p\right) = 1 - \alpha. \tag{7}$$

Then, if T_{b^*} denotes the random vector $T_{b^*} := \left(Z_1^{\text{opt}}\mathbb{1}_{\{|Z_1^{\text{opt}}|\leq b_1^*\}}, \ldots, Z_p^{\text{opt}}\mathbb{1}_{\{|Z_p^{\text{opt}}|\leq b_p^*\}}\right)$, 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 (7). 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 can be intuitively interpreted. The optimal (with respect to the volume) rectangular parallelepiped should be such that the covariance of the Gaussian variable Z^{opt} 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^{opt}. 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 transformation V_{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 : we set $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. From Proposition 2, the vector b^* must satisfy

$$\frac{1}{\sigma_1^2}\mathbb{E}\left((Z_1^{\text{opt}})^2\mathbb{1}_{\{-b_1^*\leq Z_1^{\text{opt}}\leq b_1^*\}}\right) = \cdots = \frac{1}{\sigma_p^2}\mathbb{E}\left((Z_p^{\text{opt}})^2\mathbb{1}_{\{-b_p^*\leq Z_p^{\text{opt}}\leq b_p^*\}}\right).$$

One deduces that $b_1^* = \sigma_1, \ldots, b_p^* = \sigma_p$. Consequently, the only vector $\delta^* = (\delta_1^*, \ldots, \delta_p^*)$ such that $\|\delta^*\|_\infty = 1$ and for which $(1/\delta_1^*, \ldots, 1/\delta_p^*)$ is collinear to b^* satisfies $\delta_1^* = \sigma_{i_0}/\sigma_1, \ldots, \delta_p^* = \sigma_{i_0}/\sigma_p$, with $\sigma_{i_0} = \max\{\sigma_1, \ldots, \sigma_p\}$.

2) In the equicorrelated case : we set $\Sigma = M(1, \rho)$, consequently, $\Sigma^{-1} = M(a, b)$ 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.

3) In the equicorrelated and independent case : let Σ be the block diagonal matrix $\text{diag}(M(1, \rho), Id_{p-s})$ with $M(1, \rho)$ a $s \times s$ matrix. It follows that Σ^{-1} is block diagonal with $\Sigma^{-1} = \text{diag}(M(a, b), Id_{p-s})$. If we set $\delta_1^* = \cdots = \delta_s^* = k_1$ and $\delta_{s+1}^* = \cdots = \delta_p^* = k_2$, one deduces that $\text{Var}(T_{b^*})$ is block diagonal with $\text{Var}(T_{b^*}) = \text{diag}(M(c, d), eId_{p-s})$ for some c, d, e. Consequently, whatever k_1 and k_2, the s first diagonal coefficients of $\Sigma^{-1}\text{Var}(T_{b^*})$ are equal and the $p - s$ last diagonal coefficients of $\Sigma^{-1}\text{Var}(T_{b^*})$ are equal. It remains to tune k_1 and k_2 such that all the diagonal coefficients of $\Sigma^{-1}\text{Var}(T_{b^*})$ become equal.

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 $(b_1^*/\sigma_1, \ldots, b_p^*/\sigma_p)$ is the solution of the following problem

$$\min \prod_{i=1}^{p} b_i \text{ subject to } \mathbb{P}\left(|Z_1^{\text{opt}}|/\sigma_1 \leq b_1, \ldots, |Z_p^{\text{opt}}|/\sigma_p \leq b_p\right) = 1 - \alpha.$$

4 Comparison with other multiple testing procedure

We developed a multiple testing procedure that controls the FWER *via* the tuning parameter of the lasso estimator. In this section, we compare its performances to the one 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 perform 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.

4.1 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 with $n = 2500$, $p = 1000$. The matrix X and σ^2 were chosen so that the covariance matrix $\Sigma := \sigma^2(X^TX)^{-1}$ is a block diagonal matrix ; $\Sigma = \text{diag}(M(1,\rho), Id_{500})$ and $M(1,\rho)$ 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 $\rho \in \{0, 0.3, 0.6, 0.9\}$. We applied a linear transformation V_{δ^*} provided by the Lemma 1 in (5), with $\delta_1^* = \cdots = \delta_{500}^* = k_1$ 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 $\lambda_0^{V_{\delta^*}}$ giving a FWER less that $\alpha = 0.05$. In figure 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

$$\frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}}\mathbb{E}_c\left(\mathbb{1}_{\{\hat{\beta}_i^{V_{\delta^*}}\left(\lambda_0^{V_{\delta^*}}\right)\neq 0\}}\right), \frac{1}{|\mathcal{A}|}\sum_{s(i)\in\mathcal{A}}\mathbb{E}_c\left(\prod_{j=1}^{i}\mathbb{1}_{\{P_{s(j)}\leq\frac{\alpha}{p+1-j}\}}\right) \text{ and } \frac{1}{|\mathcal{A}|}\sum_{s(i)\in\mathcal{A}}\mathbb{E}_c\left(\prod_{j=1}^{i}\mathbb{1}_{\{t_{r(j)}\leq|T_{r(j)}|\}}\right).$$

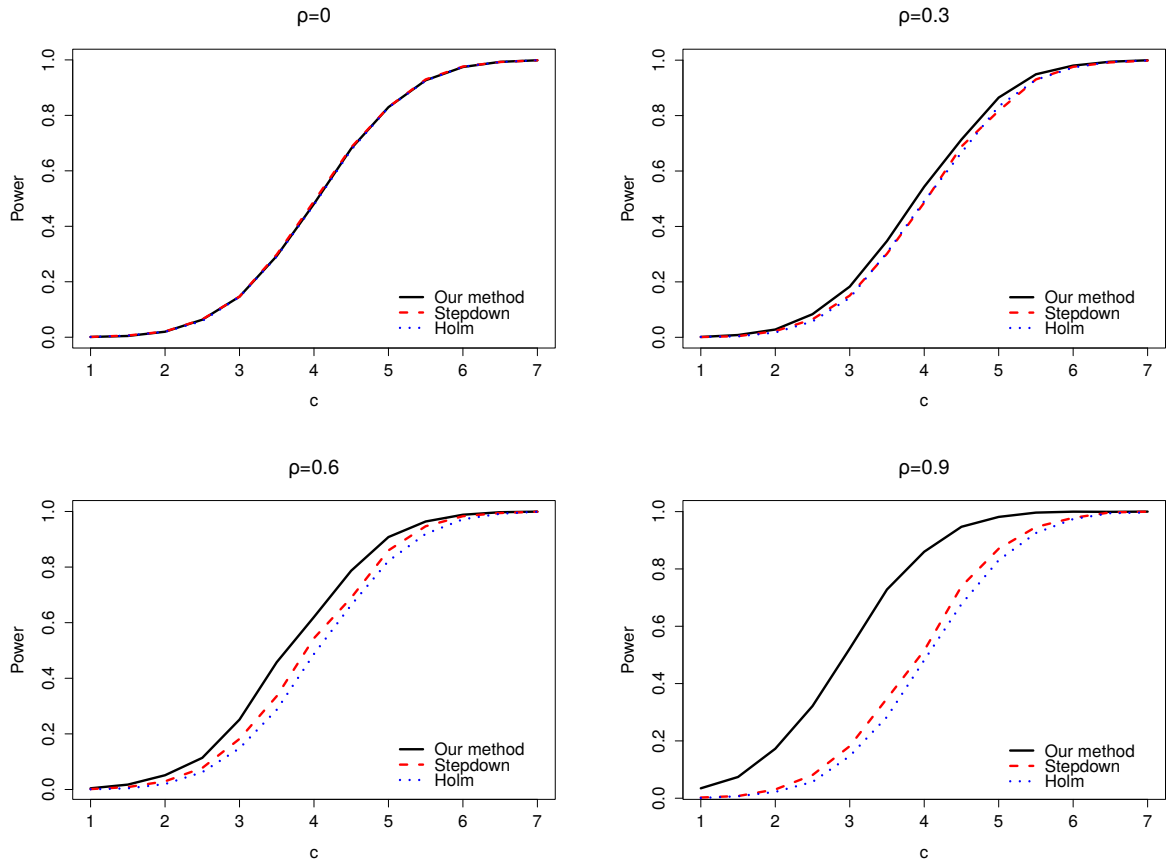

Figure 1: This figure shows the power our multiple testing procedure, the power of multiple testing procedures generic stepdown and the power of Holm's procedure. When $\rho = 0$, the three procedures have approximately the same power. When ρ increases, the difference between the power of our procedure 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 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 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.

4.2 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

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

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 $\mathbb{P}_c(\mathcal{A}(\hat{\beta}(\lambda_0)) = \mathcal{A})$ and $\mathbb{P}_c(\mathcal{A}(\hat{\beta}_i^{\text{th}}(\lambda_L)) = \mathcal{A})$ are represented in figure 2.

The main explanation of the observed difference between $\mathbb{P}_c(\mathcal{A}(\hat{\beta}(\lambda_0)) = \mathcal{A})$ and $\mathbb{P}_c((\hat{\beta}_i^{\text{th}}(\lambda_L)) = \mathcal{A})$ relies on the choice of the tuning parameter. Indeed, the tuning parameter λ_0 is the $1 - \alpha$ quantile of $\max\{|Z_1^{\text{opt}}|, \ldots, |Z_p^{\text{opt}}|\}$, whereas Lounici's tuning parameter λ_L bounds above the $1 - \alpha$ quantile of $2\max\{|Z_1^{\text{opt}}|, \ldots, |Z_p^{\text{opt}}|\}$. With our multiple testing procedure, the probability of no false discovery is $\mathbb{P}(\forall i \in [\![21, 1000]\!] \mid \hat{\beta}_i(\lambda_0) = 0)$ is exactly equal to 0.9510. As one can notice in figure 2, when the all the parameters β_i^* in the active set increase, *ie* when c increases, the probability $P_c(\mathcal{A}(\hat{\beta}(\lambda_0)) = \mathcal{A})$ does not go to 1. This is because, when there is at least one false discovery, 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.

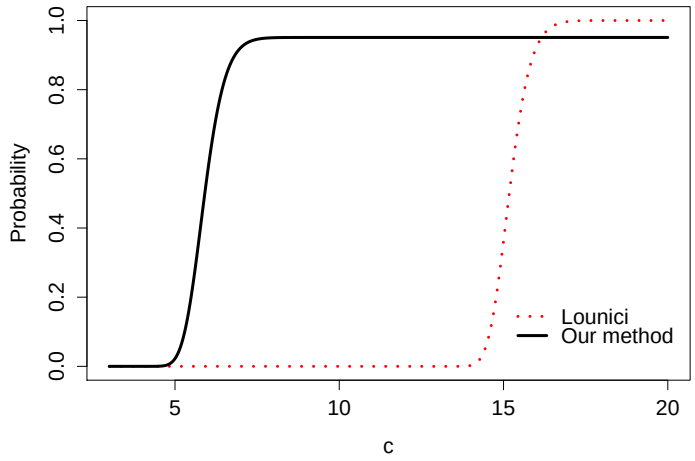

Figure 2: This figure represents the probabilities to recover the active set with Lounici's method ($\mathbb{P}_c(\hat{\mathcal{A}}^{\mathrm{L}}(\lambda_L) = \mathcal{A})$) in red dotted line and with our method ($\mathbb{P}_c(\mathcal{A}(\hat{\beta}(\lambda_0)) = \mathcal{A})$) in black plain line. 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 $\mathbb{P}_c(\mathcal{A}(\hat{\beta}(\lambda_0)) = \mathcal{A}) \approx 0.95$.

4.3 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 $\lambda \in \mathbb{R}_+ \mapsto \tilde{\beta}(\lambda)$ 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 variable $\hat{\lambda}_j = \sup\{\lambda \mid \hat{\beta}_j(\lambda) \neq 0\}$ and $\tilde{\lambda}_j = \sup\{\lambda \mid \tilde{\beta}_j(\lambda) \neq 0\}$ that are called knots of the lasso solution path. When, $|\beta_i^*| >> 0$, 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 $X^T X$ must be equal to 1), we can not determine such a matrix and a standard error $\sigma > 0$ such that $\sigma^2 (X^T X)^{-1} = \mathrm{diag}(M(1, \rho), Id_{500})$. 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 linear transformation V_{δ^*} provided by the Lemma 1 in (5) was used with $\delta^* = (1, \ldots, 1)$. Then, the tuning parameter $\lambda_0^{V_{\delta^*}}$ was determined to obtain a FWER equal to 0.5.

The power of each multiple testing procedure is represented in the figure 3. The power is the average proportion of true discoveries; the expression of the power for our procedure and the Janson's procedure are respectively equal to

$$\frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}} \mathbb{E}_c\left(\mathbb{1}_{\{\hat{\beta}_i^{V_{\delta^*}}(\lambda_0^{V_{\delta^*}})\neq 0\}}\right) \text{ and } \frac{1}{|\mathcal{A}|}\sum_{s(i)\in\mathcal{A}} \mathbb{E}_c\left(\mathbb{1}_{\{\sum_{j=1}^{i}\chi_{\rho(j)}<v\}}\right).$$

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 $\mathbb{P}_c(\exists i \notin \mathcal{A} \mid \hat{\beta}^{V_{\delta^*}}(\lambda_0^{V_{\delta^*}}) \neq 0) = \mathbb{P}(|Z_i^{\text{opt}}| > \lambda_0^{V_{\delta^*}}/\delta_i^*)$. This probability does not depend from c, we obtained 0.462, 0.477, 0.482 0.495 when the correlation ρ were respectively equal to $\rho = 0, \rho = 0.3, \rho = 0.6$ and $\rho = 0.9$. The figure 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.

5 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

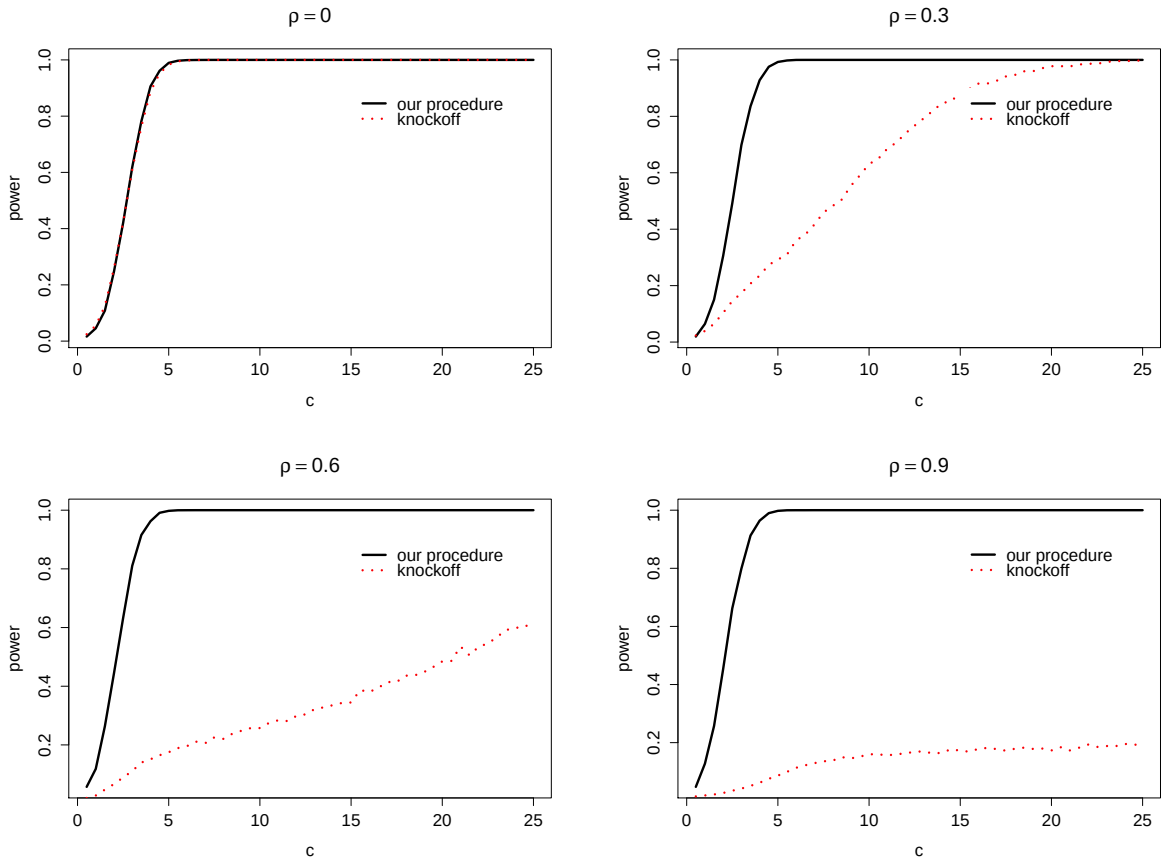

Figure 3: In this figure, we compared the power our multiple testing procedure with the power of the knockoff multiple testing procedure. Each point is an average of 1000 simulations. In the case where $\rho = 0$, components of $\hat{\beta}^{\text{opt}}$ are independent and two procedures have approximately the same power. In the case where $\hat{\beta}^{\text{opt}}$ have equi-correlated components, our procedure is more powerful.

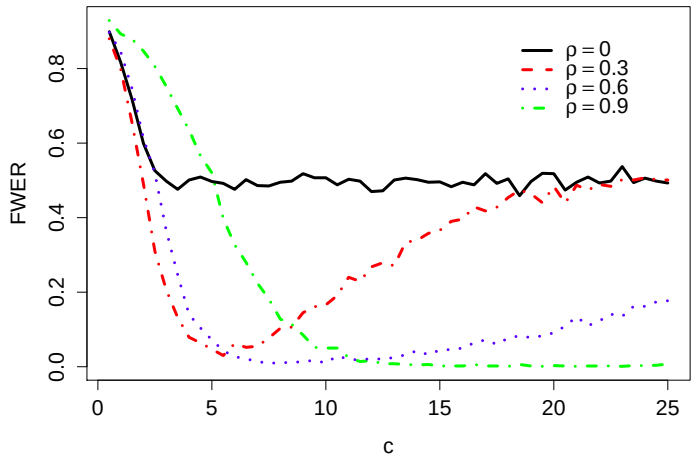

Figure 4: In this figure, we have computed the FWER level of the knockoff procedure for all $c > 0$. When non-null parameters are small (i.e c is small), the FWER level is not well controlled. When c is large enough, except in the independent case, the FWER level is largely smaller than its nominal value 0.5. Each point is an average of 1000 simulations.

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 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 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.

5.1 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

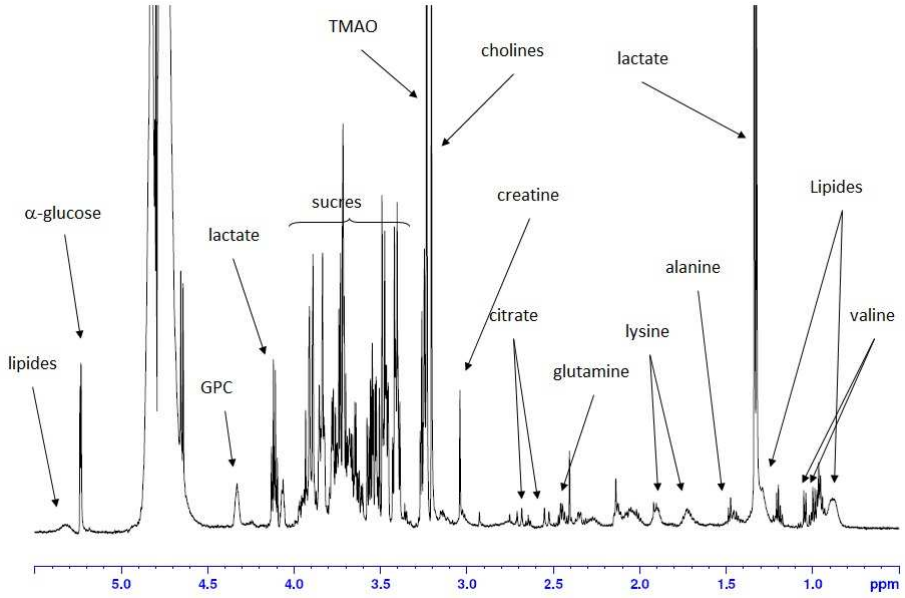

Figure 5: Example of an annotated mixture spectrum. There are overlaps between peaks of lipides and valine and between the peaks of glutamine and lysine.

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.

5.2 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.

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: 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. [2017].

6 Conclusions

In this article, we proposed a multiple testing procedure that rejects the hypothesis $\beta_i^* = 0$ when a lasso-type estimator $\hat{\beta}_i(\lambda)$ is not null. When the design matrix X of the Gaussian linear model (1) has orthogonal columns, we gave a tuning parameter λ_0 such that the multiple testing procedure controls the FWER at a level α. When X has no longer orthogonal columns, the keystone of the paper is to apply a linear transformation U to each member of the model (1) such that U orthogonalises X and for which the estimator $\hat{\beta}^{\text{ols}}(U)$ is efficient. We then applied the results from the orthogonal case to the non-orthogonal one. Numerical comparisons illustrate the benefit of our procedure compare to the state-of-the-art procedures that control the FWER. In a future work, we will explore a stepdown procedure adapted to the procedure given in this paper that could increase the power.

7 Appendix

Proof (Proposition 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 is 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 1) It is straightforward to show that $V_\delta X = \begin{pmatrix}\Delta & 0\end{pmatrix}^T$. This implies that $V_\delta \in G$. We are going to show that $\hat{\beta}^{\text{ols}}(V_\delta) = MP_\delta Y$. Indeed,

$$\begin{aligned}\hat{\beta}^{\text{ols}}(V_\delta) &= ((V_\delta X)^T (V_\delta X))^{-1} (V_\delta X)^T V_\delta Y, \\ &= \Delta^{-2} \begin{pmatrix}\Delta & 0\end{pmatrix} V_\delta Y, \\ &= \begin{pmatrix}\Delta^{-1} & M_2\end{pmatrix} P_\delta Y.\end{aligned}$$

It remains to show that $\Delta^{-1} = M_1$. For the next calculus, let us introduce $(e_1, \ldots, e_n)$ and $(f_1, \ldots, f_p)$ the canonical basis of $\mathbb{R}^n$ and $\mathbb{R}^p$.

$$\begin{aligned}\forall i \in [\![1,p]\!], M_1 f_i &= Me_i, \\ &= ((P_\delta X)^T (P_\delta \Gamma P_\delta^T)^{-1} P_\delta X)^{-1} (P_\delta X)^T (P_\delta \Gamma P_\delta^T)^{-1} e_i, \\ &= ((P_\delta X)^T (P_\delta \Gamma P_\delta^T)^{-1} P_\delta X)^{-1} (P_\delta X)^T (P_\delta \Gamma P_\delta^T)^{-1} \frac{1}{\sqrt{\delta_i}} P_\delta X f_i = \frac{1}{\sqrt{\delta_i}} f_i.\end{aligned}$$

Since $\hat{\beta}^{\text{ols}}(V_\delta) = MP_\delta Y$, one deduces that

$$\begin{aligned}\text{Var}(\hat{\beta}^{\text{ols}}(V_\delta)) &= MP_\delta \Gamma P_\delta^T M^T, \\ &= ((P_\delta X)^T (P_\delta \Gamma P_\delta^T)^{-1} P_\delta X)^{-1}, \\ &= (X^T P_\delta^T (P_\delta^T)^{-1} \Gamma^{-1} P_\delta^{-1} P_\delta X)^{-1}, \\ &= (X^T \Gamma^{-1} X)^{-1}.\end{aligned}$$

One can recognize the covariance matrix of the maximum likelihood estimator. □

The proof of lemma 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.

In the following, we denote $\lambda_0^{V_\delta} = \lambda_0(\delta)$, with $\delta \in]0, +\infty[^p$. It is straightforward to show that λ_0 given in the proposition 1 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 $x = (x_1, \ldots, x_p) \in]0, +\infty[^p$, 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,

$$\forall y \in [x_1 - ru_1, x_1 + ru_1] \times \cdots \times [x_p - ru_p, x_p + ru_p], |g(y) - g(x)| \leqslant \frac{r}{\|x\|}|g(x)|. \tag{8}$$

Let $\epsilon \geqslant 0$; 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 (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 recall the expression of the function f

$$f(\delta) = \frac{\lambda_0(\delta)}{\delta_1} \times \cdots \times \frac{\lambda_0(\delta)}{\delta_p}.$$

Since λ_0 is homogeneous, f satisfy the property $\forall k > 0, f(k\delta) = f(\delta)$. This property implies that if the restriction of f onto the unit sphere reaches its minimum, then f has a global minimum on $]0, +\infty[^p$. We

denote $S_\infty(1)$ as the unit sphere of $\mathbb{R}^p$ for the supremum norm. Using Lemma A, we obtain that f is continuous; moreover, the restriction of f onto the set $]0,+\infty[^p \cap S_\infty(1)$ can be extended by continuity to $[0,+\infty[^p \cap S_\infty(1)$ by setting

$$\bar{f} : \delta \in [0,+\infty[^p \cap S_\infty(1) \begin{cases} f(\delta) \text{ if } \delta \in]0,+\infty[^p \cap S_\infty(1) \\ +\infty \text{ if } \exists i \in [\![1,p]\!] \text{ such that } \delta_i = 0. \end{cases}$$

The function $\bar{f}$ is continuous on the compact set $[0,+\infty[^p \cap S_\infty(1)$; thus, $\bar{f}$ reaches its minimum at δ^*. The minimum of the function $\bar{f}$ is finite, so one deduces that $\delta^* \in]0,+\infty[^p \cap S_\infty(1)$. Finally, we obtain

$$\forall \delta \in]0,+\infty[^p, \phi(V_\delta) \geqslant \phi(V_{\delta^*}).$$

The result follows. □

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) 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 $(\delta_1 Z_1^{\mathrm{opt}}, \ldots, \delta_p Z_p^{\mathrm{opt}}) = \Delta Z^{\mathrm{ols}}(V_\delta)$ is semidefinite positive. Indeed, we obtain that

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

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 $\max\{\delta_1|Z_1^{\mathrm{opt}}|, \ldots, \delta_p|Z_p^{\mathrm{opt}}|\}$, 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 B, the inequality $\phi(V_\delta) \geqslant \phi(V_{\delta^*})$ gives the result. □

Proof (Proposition 2) To simplify the computation of the gradients, we consider the following problem

which has the same solution as the problem (7)

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

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

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

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

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

implies that the gradient of F is equal to

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

Thus $\nabla f(b^*)$ and $\nabla F(b^*)$ are collinear if and only if the components of the vector

$$\left(\sum_{j=1}^{p} \Sigma_{i,j}^{-1} \int_{[-1,1]^p} x_i b_i^* x_j b_j^* \exp\left(-\frac{1}{2} x^T D\Sigma^{-1} Dx\right) \det(D) dx\right)_{1 \leqslant i \leqslant p} \tag{9}$$

are equal. To conclude, one recognizes that

$$\int_{[-1,1]^p} R x_i b_i^* x_j b_j^* \exp\left(-\frac{1}{2} x^T D\Sigma^{-1} Dx\right) \det(D) dx = \mathbb{E}\left(Z_i^{\text{opt}} \mathbb{1}_{\{|Z_i^{\text{opt}}| \leq b_i^*\}} Z_j^{\text{opt}} \mathbb{1}_{\{|Z_j^{\text{opt}}| \leq b_j^*\}}\right)$$

In the previous equality, the second term is the i, j coefficient of the matrix $\text{Var}(T_{b^*})$. Since components of (9) are equals, one deduces that the diagonal coefficients of $\Sigma^{-1}\text{Var}(T_{b^*})$ are equals. □

Acknowledgements

The authors are grateful for the real data provided by the following metabolomicians from Toxalim: Cécile Canlet, Laurent Debrauwer and Marie Tremblay-Franco. 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".

References

Theodore W Anderson. 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, 1955.

William Astle, Maria De Iorio, Sylvia Richardson, David Stephens, and Timothy Ebbels. 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, 2012.

Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. The Annals of Statistics, 43(5):2055–2085, 2015.

Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J Candès. Slope - adaptive variable selection via convex optimization. The Annals of Applied Statistics, 9(3):1103–1140, 2015.

Peter Bühlmann and Sara van de Geer. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, 2011. ISBN 3642201911, 9783642201912.

Sohail Chand. On tuning parameter selection of lasso-type methods - A Monte Carlo study. In Proceedings of 2012 9th International Bhurban Conference on Applied Sciences Technology (IBCAST), pages 120–129, 2012.

Sandrine Dudoit and Mark J Van Der Laan. Multiple Testing Procedures with Applications to Genomics. Springer, 2007.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. The Annals of Statistics, 32(2):407–499, 2004.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1):1–22, 2010.

Max Grazier G'Sell, Stefan Wager, Alexandra Chouldechova, and Robert Tibshirani. Sequential selection procedures and false discovery rate control. Journal of the Royal Statistical Sociey: Series B (Statistical Methodology), 78(2):423–444, 2015.

Jie Hao, William Astle, Maria De Iorio, and Timothy MD Ebbels. BATMAN - an R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a bayesian model. Bioinformatics, 28(15):2088–2090, 2012.

Trevoir Hastie, Rob Tibshirani, and Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer Series in Statistics. Springer, 2009. ISBN 9780387848587.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. Convex Analysis and Minimization Algorithms I: Fundamentals, volume 305. Springer Science & Business Media, 2013.

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

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

Erich L. Lehmann and Joseph P. Romano. Testing Statistical Hypotheses. Springer Texts in Statistics. Springer, New York, third edition, 2005.

Chenlei Leng, Yi Lin, and Grace Wahba. A note on the lasso and related procedures in model selection. Statistica Sinica, 16(4):1273–1284, 2006.

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

Karim Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. Electronic Journal of Statistics, 2:90–102, 2008.

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

Siamak Ravanbakhsh, Philip Liu, Trent C. Bjordahl, Rupasri Mandal, Jason R. Grant, Michael Wilson, Roman Eisner, Igor Sinelnikov, Xiaoyu Hu, Claudio Luchinat, Russell Greiner, and David S Wishart. Accurate, fully-automated NMR spectral profiling for metabolomics. PLoS ONE, 10(5):e0124219, 2015.

Alvin C Rencher and G Bruce Schaalje. Linear Models in Statistics. John Wiley & Sons, 2008.

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

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

Patrick J.C. Tardivel, Cécile Canlet, Gaëlle Lefort, Marie Tremblay-Franco, Laurent Debrauwer, Didier Concordet, and Rémi Servien. ASICS: an automatic method for identification and quantification of metabolites in NMR 1D ^{1}H spectra. Submitted, 2017.

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

Dan Tulpan, Serge Léger, Luc Belliveau, Adrian Culf, and Miroslava Čuperlović-Culf. Metabohunter: an automatic approach for identification of metabolites from ^{1}H-NMR spectra of complex mixtures. BMC Bioinformatics, 12(1):400, 2011.

Aalim M. Weljie, Jack Newton, Pascal Mercier, Erin Carlson, and Carolyn M. Slupsky. Targeted profiling: quantitative analysis of ^{1}H-NMR metabolomics data. Analytical Chemistry, 78(13):4430–4442, 2006.

Peng Zhao and Bin Yu. On model selection consistency of lasso. The Journal of Machine Learning Research, 7:2541–2563, 2006.

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