# A powerful multiple testing procedure in linear Gaussian model

Patrick J C Tardivel, Rémi Servien, Didier Concordet

# A powerful multiple testing procedure in linear Gaussian model

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

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

**Abstract**

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.

## 1 Introduction

Let us consider the linear Gaussian model

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

where $X = (X_1|\dots|X_p)$ is a $n \times p$ design matrix of rank $p$, $\varepsilon$ is a centered Gaussian vector with an invertible variance matrix $\Gamma$, and $\beta^*$ 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

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

*corresponding author: patrick.tardivel@inra.fr

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 $\text{rank}(X) = p$).

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 $\lambda$ 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 $\text{Vect}\{X_i,\ i \in \mathcal{A}\}$. When the design matrix is close to an orthogonal matrix (which implies the irrepresentable condition), an explicit $\lambda$ 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 ($\text{rank(X)} = \text{p}$), 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 $\text{rank}(X)$ 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

$$\forall i \in [\![1, p]\!], \hat{\beta}_i^{U^*}(\lambda) = \ \text{sign}(\hat{\beta}_i^{\text{mle}}) \left( |\hat{\beta}_i^{\text{mle}}| - \lambda/\delta_i^* \right)_+ , \text{where } \hat{\beta}^{\text{mle}} := (X^T \Gamma^{-1} X)^{-1} X^T \Gamma^{-1} Y. \tag{3}$$

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

In contrast, the value $\delta^*$ in (3) giving a multidimensional acceptance region with a minimal volume leads to decision rules where $\mu$ 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 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 $\beta^*$. Let $Z^{\text{ols}}$ denotes a centered Gaussian vector with the same covariance matrix as $\hat{\beta}^{\text{ols}}$, the tuning parameter giving a FWER at level $\alpha$ 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 $\lambda_0$ to control the FWER.

**Proposition 1** *Let $X$ be a $n \times p$ matrix such that $X^T X = \text{diag}(d_1, \ldots, d_p)$ 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 $\mathcal{N}\left(0, (X^T X)^{-1} X^T \Gamma X (X^T X)^{-1}\right)$ distribution. Let $\alpha \in (0, 1)$, if $\lambda_0$ is the $1 - \alpha$ quantile of $\max_{i \in [\![1,p]\!]}\{d_i \times |Z_i^{\text{ols}}|\}$ then,*

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

When the covariance matrix $\Gamma$ is given *a priori*, the distribution of $Z^{\text{ols}}$ is known and $\lambda_0$ can be obtained by simple numerical simulations. In the next section we study the more general case where $X$ has no longer orthogonal columns.

# 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 $(UX)^T UX$ is diagonal. For example the matrix $U := (X^T X)^{-1} X^T$ 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 $\lambda_0$ to control the FWER. To overcome this problem, we propose to apply a linear transformation $U \in G$ to each member of the model (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 $\beta^*$ 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 $\lambda_0^U$ giving a FWER $\alpha$ 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: 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 $\phi$ 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 $\phi$ is minimal and the lasso is a soft thresholded maximum likelihood estimator.

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

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

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

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

*where $\hat{\beta}^{\text{mle}}$ is the maximum likelihood estimator of the model (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 $|\hat{\beta}_i^{\text{mle}}| \geq \lambda/\delta_i^*$ 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 $\delta^*$ of the theorem 1 is not collinear to $1/\text{se}(\hat{\beta}_1^{\text{mle}}), \ldots, 1/\text{se}(\hat{\beta}_p^{\text{mle}})$. Consequently the random variables $\delta_1^*\hat{\beta}_1^{\text{mle}}, \ldots, \delta_p^*\hat{\beta}_p^{\text{mle}}$ have different variances. This remark is the main difference with the classical procedures for which statistical tests $\hat{\beta}_1^{\text{mle}}/\text{se}(\hat{\beta}_1^{\text{mle}}), \ldots, \hat{\beta}_p^{\text{mle}}/\text{se}(\hat{\beta}_p^{\text{mle}})$ are re-scaled to have

unit variance. To provide a mutiple testing procedure which reject $\mathcal{H}_i : \beta_i^* = 0$ as soon as $|\hat{\beta}_i^{\text{mle}}| \geq \lambda/\delta_i^*$ the parameter $\lambda$ have to be chosen as the $1-\alpha$ quantile of $\max\{\delta_1^*|Z_1^{\text{mle}}|, \ldots, \delta_p^*|Z_p^{\text{mle}}|\}$. From now on, we denote $\lambda_0(\delta)$ the $1-\alpha$ quantile of $\max\{\delta_1|Z_1^{\text{mle}}|, \ldots, \delta_p|Z_p^{\text{mle}}|\}$ where $\delta = (\delta_1, \ldots, \delta_p) \in (0, +\infty)^p$.

To manage the previous multiple testing procedure based on the maximum likelihood estimator, the keystone is to compute the optimal parameter $\delta^*$. The next section deals with this issue.

# 4 A new procedure based on the maximum likelihood estimator

The theorem 1 does not explain how to get such an optimal parameter $\delta^*$. 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 $\delta^*$ 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 $\Sigma$ be the covariance matrix of the maximum likelihood estimator and let $Z^{\text{mle}}$ be distributed according to $\mathcal{N}(0_{\mathbb{R}^p}, \Sigma)$. The rectangular parallelepiped $B^*$ has the smallest volume among rectangular parallelepiped $B$ such that $P\left(Z^{\text{mle}} \in B\right) = 1-\alpha$. This is a constraint optimization problem whose solutions are stationary points of the Lagrangian. The condition given in the following proposition should hold for $B^*$.

**Proposition 2** *Let $b^* = \left(b_1^*, \ldots, b_p^*\right)$ be a solution of the following optimisation problem*

$$\min \prod_{i=1}^{p} b_i \text{ subject to } \mathbb{P}\left(|Z_1^{\text{mle}}| \leq b_1, \ldots, |Z_p^{\text{mle}}| \leq b_p\right) = 1-\alpha. \tag{6}$$

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

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

*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 $\Sigma$, all the diagonal coefficients of $\Sigma^{-1}\text{Var}(T^{b^*})$ would be equal, indicating that $b^*$ is a solution of (6). Because the diagonal terms of $\text{var}(T^{b^*})$ are always smaller than the diagonal terms of $\Sigma$, $\text{var}(T^{b^*})$ cannot be equal to $\Sigma$. 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 truncated Gaussian variable $Z^{\text{mle}}$ 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 $Z^{\text{mle}}$. 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 $\delta^*$ 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 $\hat{\beta}_1^{\mathrm{mle}},\dots,\hat{\beta}_p^{\mathrm{mle}}$ are independent thus, $\Sigma$ is the diagonal matrix $\mathrm{diag}(\mathrm{var}(\hat{\beta}_1^{\mathrm{mle}}),\dots,\mathrm{var}(\hat{\beta}_1^{\mathrm{mle}}))$. From Proposition 2, the vector $b^*$ must satisfy

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

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

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

**3) In the block diagonal equicorrelated case :** the covariance matrix $\Sigma$ of $\hat{\beta}^{\mathrm{mle}}$ is the following block diagonal matrix $\mathrm{diag}(M(1,\rho),M(1,\rho'))$ 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 $\Sigma^{-1}$ is block diagonal with $\Sigma^{-1} = \mathrm{diag}(M(a,b),M(a',b'))$. 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 $\delta_1^*\hat{\beta}_1^{\mathrm{mle}},\dots,\delta_p^*\hat{\beta}_p^{\mathrm{mle}}$ 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 $\Sigma$, the diagonal coefficients of $\Sigma$ are equal to 1. Indeed, one can check that $(b_1^*/\sqrt{\Sigma_{1,1}},\dots,b_p^*/\sqrt{\Sigma_{p,p}})$ is the solution of the following problem

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

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 (1), namely $\Sigma := (X^T\Gamma X)^{-1}$.

2. The parameter $\delta^* \in (0, +\infty)^p$ is obtain by solving the problem (6). This optimal parameter must satisfies the relation $\Sigma^{-1}\text{var}(T^{b^*})$ given in the proposition 2.

3. One compute $\lambda_0(\delta^*)$ which is the $1-\alpha$ quantile of the random variable $\{\delta_1^*|Z_1^{\text{mle}}|, \ldots, \delta_p^*|Z_p^{\text{mle}}|\}$. The quantile $\lambda_0(\delta^*)$ is computed numerically using a large number of realizations of $Z^{\text{mle}}$ distributed according to $\mathcal{N}(0, \Sigma)$.

4. The multiple testing procedure rejects the null hypothesis $\mathcal{H}_i : \beta_i^* = 0$ when $|\hat{\beta}_i^{\text{mle}}| > \lambda_0(\delta^*)/\delta_i^*$. 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.

# 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 $p$.

## 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 $\alpha$ and improves the Bonferroni procedure since the cutoff $\alpha/(p-i+1)$ associated to the hypothesis $\mathcal{H}_{s(i)}$ is smaller than $\alpha/p$.

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 $T := (T_1, \dots, T_p)$. 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 $\mathcal{H}_{r(1)}, \dots, \mathcal{H}_{r(p)}$ 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 $\max\{|Z_{r(s)}|, \dots, |Z_{r(p)}|\}$.

For the numerical experiments, we performed 1000 simulations. The covariance matrix $\Sigma$ of the maximum likelihood estimator is $\Sigma := \mathrm{diag}(M(1,\rho), Id_{500})$, where $M(1,\rho)$ and $Id_{500}$ are both $500 \times 500$ 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 $\delta^*$ of the lemma 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 $\delta^*$ were used to derive $\lambda_0(\delta^*)$ 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 $\rho$ . 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

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

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.

## 5.2 Comparison with Lounici's estimator

Lounici [2008] used a thresholded lasso estimator $\hat{\beta}^{\mathrm{th}}$ to build the following estimator of $\mathcal{A}$:

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

He proved that the event $\{\mathcal{A}(\hat{\beta}_i^{\mathrm{th}}(\lambda_L)) = \mathcal{A}\}$ has a controlled probability when the design matrix $X$ is close to

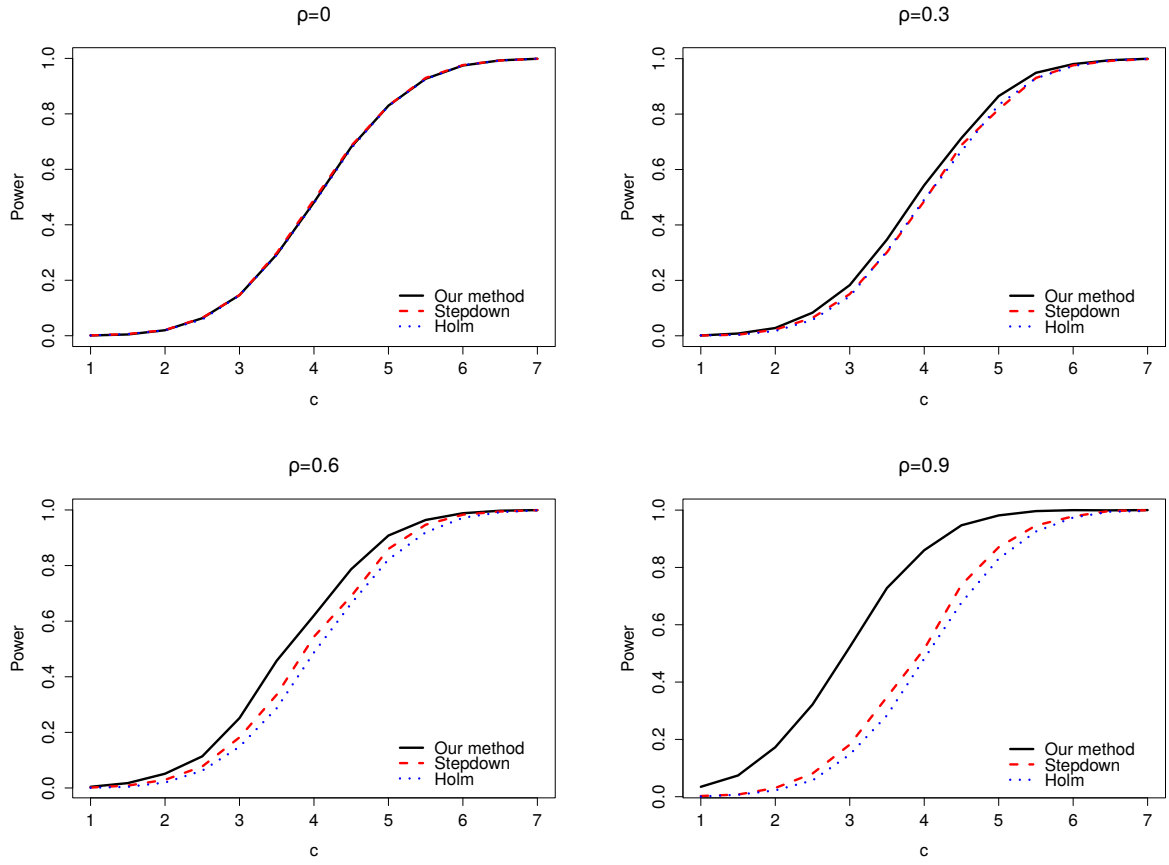


Figure 1: This figure shows the power our multiple testing procedure, the power of multiple testing procedures generic stepdown and the power of Holm's procedure. When $\rho = 0$, the three procedures have approximately the same power. When $\rho$ increases, the difference between the power of our procedure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

an orthogonal matrix up to a multiplicative constant, the noise $\varepsilon$ 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 $\lambda_L$ 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.

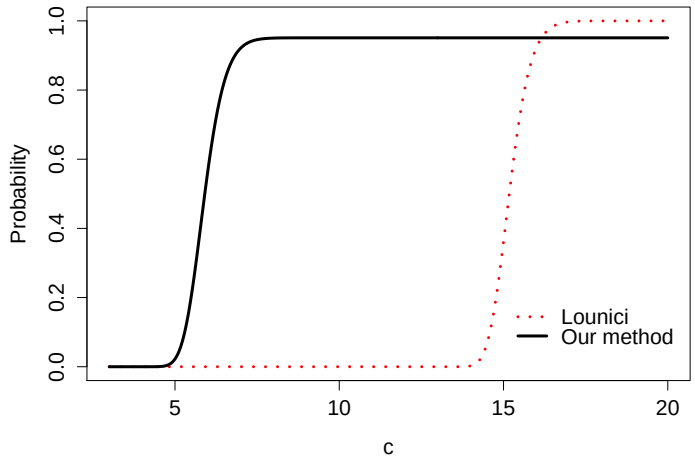


Figure 2: This figure represents the probabilities to recover the active set with Lounici's method ( $\mathbb{P}_c(\hat{\mathcal{A}}^{\text{L}}(\lambda_L) = \mathcal{A})$) in red dotted line and with our method ($\mathbb{P}_c(\{i \in [\![1,p]\!] \mid |\hat{\beta}_i^{\text{mle}}| > \lambda_0(\delta^*)/\delta_i^*\} = \mathcal{A})$) 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, $\mathbb{P}_c((\hat{\beta}_i^{\text{th}}(\lambda_L)) = \mathcal{A}) \approx 1$ and $\mathbb{P}_c(\{i \in [\![1,p]\!] \mid |\hat{\beta}_i^{\text{mle}}| > \lambda_0(\delta^*)/\delta_i^*\} = \mathcal{A}) \approx 0.95$.

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 $\max\{|Z_1^{\text{mle}}|, \dots, |Z_p^{\text{mle}}|\}$ ($\delta^* = (1, \dots, 1)$), whereas Lounici's tuning parameter $\lambda_L$ bounds above the $1-\alpha$ quantile of $2\max\{|Z_1^{\text{mle}}|, \dots, |Z_p^{\text{mle}}|\}$. 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, when the all the parameters $\beta_i^*$ in the active set increase, *ie* when $c$ increases, the probability $\mathbb{P}_c(\{i \in [\![1,p]\!] \mid |\hat{\beta}_i^{\text{mle}}| > \lambda_0(\delta^*)/\delta_i^*\} = \mathcal{A})$ 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.

### 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

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

where the design matrix $X_{\text{KO}} = [X, \tilde{X}]$ 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 $X^T X$ must be equal to 1), we can not determine such a matrix and a standard error $\sigma > 0$ such that $\sigma^2 (X^T X)^{-1} = \text{diag}(M(1, \rho), Id_{500})$. Indeed, diagonal coefficients of $M^{-1}(1, \rho)$ are not equal to 1 when $\rho \neq 0$. Consequently, whatever $\sigma > 0$, the matrix $X^T X = \sigma^2 \text{diag}(M^{-1}(1, \rho), Id_{500})$ 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 $\Sigma = \sigma^2 (X^T X)^{-1} = M(1, \rho)$. Different values of $\rho$ 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 $\delta^*$ provided by the Lemma 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 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 $\rho$ 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

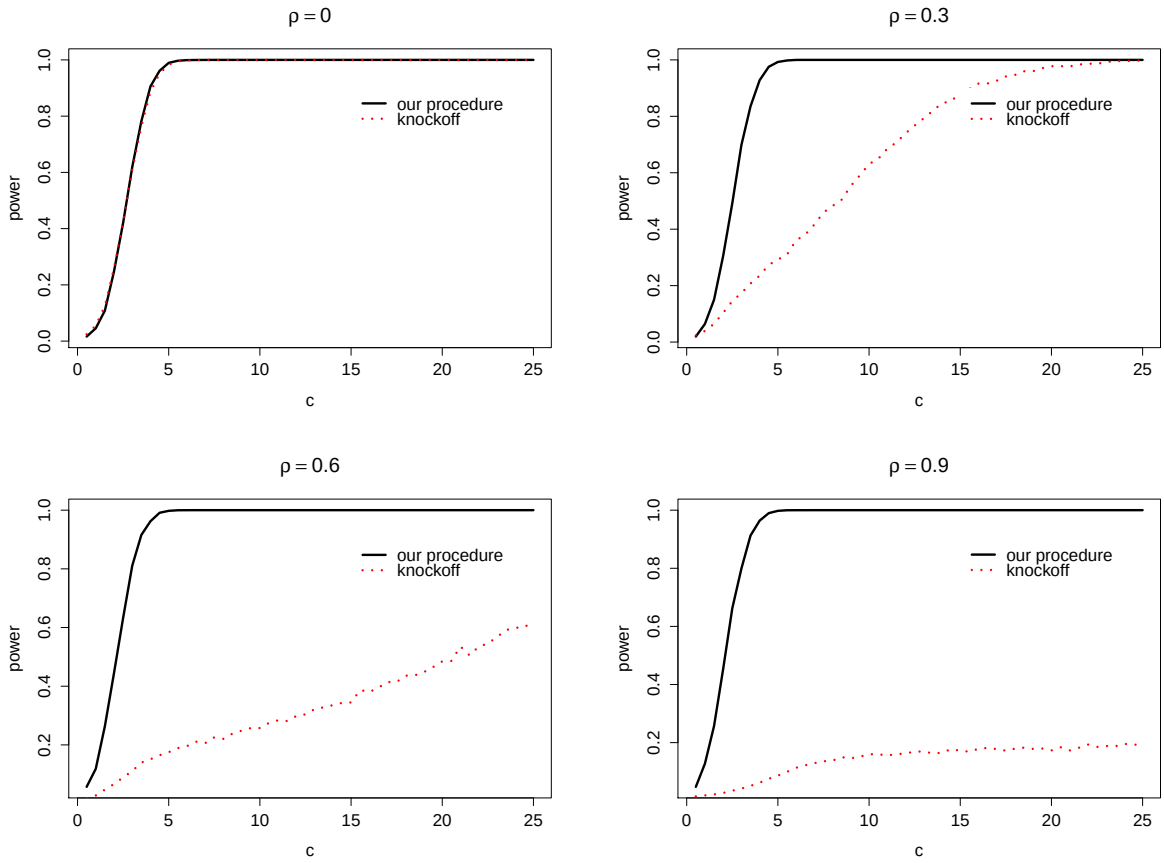


Figure 3: In this figure, we compared the power our multiple testing procedure with the power of the knockoff multiple testing procedure. Each point is an average of 1000 simulations. In the case where $\rho = 0$, components of $\hat{\beta}^{\text{mle}}$ are independent and two procedures have approximately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

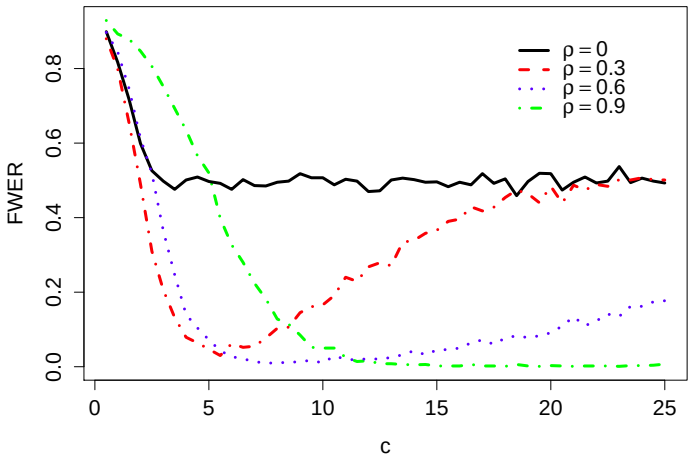


Figure 4: In this figure, we have computed the FWER level of the knockoff procedure for all $c > 0$. When non-null parameters are small (i.e $c$ is small), the FWER level is not well controlled. When $c$ is large enough, except in the independent case, the FWER level is largely smaller than its nominal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

knockoff multiple testing procedure does not control the FWER at a level 0.5 for small values of $c$.

# 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 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

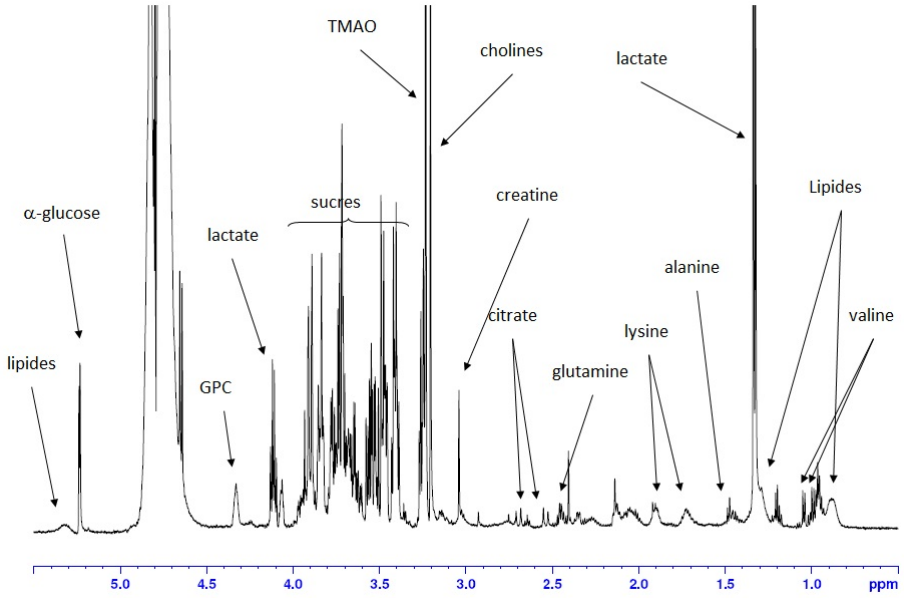


Figure 5: Example of an annotated mixture spectrum. There are overlaps between peaks of lipides and valine and between the peaks of glutamine and lysine.

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.

## 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 $\sigma^2$ is estimated using several observations of a metabolite spectrum.

### 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.

| 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] where our procedure is compared to the existing ones on more complex datasets.

## 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 (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 $|\hat{\beta}_i^{\mathrm{mle}}| > \lambda_0(\delta^*)/\delta_i^*$. The parameter $\delta^*$ is the optimal one given in proposition 2 and $\lambda_0(\delta^*)$ is the $1-\alpha$ quantile of $\max\{\delta_1^*|Z_1^{\mathrm{mle}}|, \ldots, \delta_p^*|Z_p^{\mathrm{mle}}|\}$. The keystone of this procedure is to compute the optimal parameter $\delta^*$, an exact computation of $\delta^*$ 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 $\delta^*$ 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 $\delta^*$. 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.

# 8 Appendix 1 : construction of the matrix $U^*$

The theorem 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 $\lambda_0^U$, 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 $P$ 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 $V_\delta$ 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 $V_\delta$ belongs to $G$, and $\hat{\beta}^{\text{ols}}(V_\delta) = \hat{\beta}^{\text{mle}}$, where $\beta^{\text{mle}}$ is the maximum likelihood estimator of the model (1).*

The matrix $P$ given in the lemma 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 1 evidences $V_\delta$ 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 $\Gamma^{-1/2}$. 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 $V_\delta$ 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 et al. [2015] which relax the irrepresentable condition. When the rank of $X$ is $p$, $FX$ is orthogonal thus $F \in G$. However contrarily to the $V_\delta$ transformations, the estimator $\hat{\beta}^{\text{ols}}(F)$ is not efficient.

As an example for Lemma 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 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 $\phi$.

**Lemma 2** *Set*

$$U^* = V_{\delta^*} \text{ with } \delta^* = \underset{\delta\in]0,+\infty[^p}{\text{arginf}}\ \phi(V_\delta), \tag{7}$$

*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 1 holds for $U^* = V_{\delta^*}$.

# 9 Appendix 2 : Proofs

**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 its global minimum. Because the penalty term is a $L^1$ norm, the function $\psi$ is not differentiable everywhere. However, as $\psi$ is a convex function, it has a subdifferential. To find where the global minimum of $\psi$ 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 $\psi$ 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 $\lambda_0$. 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)** The matrix $V_\delta$ 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 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 $\delta^*$. In the second step, we prove that the linear transformation $V_{\delta^*}$ 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 $\lambda_0$ 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,

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

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 (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 $\delta^*$.*

**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 $\lambda_0$ 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 $\delta$", 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 $\delta_A$ in $A_{\eta_0}$ 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 $A_{\eta_0}$.

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 $\delta$ 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 $B_{\eta_0}$. Because $f$ is continuous on $A_{\eta_0}$ and $A_{\eta_0}$ is compact, f reaches its minimum on $A_{\eta_0}$. □

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 $\Gamma_V$ and $\Gamma_W$, 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 $(UX)^T UX = \Delta = \text{diag}(\delta_1, \ldots, \delta_p) =$

diag($\delta$). The difference between the covariance matrices of the Gaussian vectors $(\delta_1 Z_1^{\text{ols}}(U), \ldots, \delta_p Z_p^{\text{ols}}(U)) = \Delta Z^{\text{ols}}(U)$ and $(\delta_1 Z_1^{\text{mle}}, \ldots, \delta_p Z_p^{\text{mle}}) = \Delta Z^{\text{ols}}(V_\delta)$ is semidefinite positive. Indeed, reminding that $\Sigma$ is the covariance matrix of the maximum likelihood estimator, we obtain that

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

The last inequality is a consequence of the Gauss-Markov theorem [Rencher and Schaalje, 2008] (page 146). Because $\lambda_0^U$ and $\lambda_0^{V_\delta}$ are the respective $1-\alpha$ quantiles of $\max\{\delta_1|Z_1^{\text{ols}}(U)|, \ldots, \delta_p|Z_p^{\text{ols}}(U)|\}$ and $\max\{\delta_1|Z_1^{\text{mle}}|, \ldots, \delta_p|Z_p^{\text{mle}}|\}$, 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 B, the inequality $\phi(V_\delta) \geqslant \phi(V_{\delta^*})$ gives the result. □

**Proof (Theorem 1)** The lemmas 1 and 2 allow to prove the theorem 1. □

**Proof (Proposition 2)** To simplify the computation of the gradients, we consider the following problem which has the same solution as the problem (6)

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

Because this problem reaches its minimum at $b^*$, $\nabla f(b^*)$ is collinear to $\nabla F(b^*)$. Let us set $D$ the matrix $D = \text{diag}(\text{b}_1, \ldots, \text{b}_\text{p})$, 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 $R = 1/((2\pi)^{p/2}\det(\Sigma)^{1/2})$. Next, the expression of the partial derivative

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

implies that the gradient of $F$ is equal to

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

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

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

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

$$\begin{aligned}
& \exists k \in \mathbb{R} \text{ such that } v(b^*) = k \nabla f(b^*), \\
\Leftrightarrow \quad & \forall i \in [\![1,p]\!], \sum_{j=1}^{p} \Sigma^{-1}_{i,j} \int_{[-1,1]^p} x_i b^*_i x_j b^*_j R \exp\left(-\frac{1}{2} x^T D \Sigma^{-1} D x\right) \det(D) dx = k, \\
\Leftrightarrow \quad & \forall i \in [\![1,p]\!], \sum_{j=1}^{p} \Sigma^{-1}_{i,j} \int_{u \in \mathbb{R}^p} u_i u_j \frac{R}{1-\alpha} \exp\left(-\frac{1}{2} u \Sigma^{-1} u\right) \mathbb{1}_{u \in B^*} du = \frac{k}{1-\alpha}. \qquad (9)
\end{aligned}$$

The expression (9) is obtained *via* the change of variables $\forall i \in [\![1,p]\!], u_i = x_i b^*_i$. To conclude, one recognizes that

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

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".


## 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.

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

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

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.

Jinzhu Jia, Karl Rohe, et al. Preconditioning the lasso for sign consistency. Electronic Journal of Statistics, 9 (1):1150–1172, 2015.

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

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 complex 1D $^{1}$H NMR spectra. Metabolomics, 13(10):109, 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.