Non-asymptotic active set properties of lasso-type estimators in small-dimension

Patrick J.C Tardivel, Rémi Servien, Didier Concordet

Non-asymptotic active set properties of lasso-type estimators in small-dimension

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

Abstract

We propose to estimate the active set associated with a standard linear Gaussian model when the design matrix is a $n \times p$ full-rank matrix (thus, $n \geqslant p$). Asymptotic results are available for lasso-type estimators when the irrepresentable condition holds. In this paper, we give non-asymptotic results on the estimation of the active set in small dimension by providing an explicit tuning parameter choice without this condition. Both theoretical and numerical arguments illustrate the benefits of our approach. An application to the detection of metabolites in metabolomic data is provided.

Keywords: Lasso, Adaptive lasso, Active set estimation, Tuning parameter.

1 Introduction

Let us consider the linear Gaussian model

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

where $X = (X_1|\dots|X_p)$ is an $n \times p$ full-rank design matrix with $n \geqslant p$, ε is a Gaussian vector with an invertible variance matrix Γ, and β^* is an unknown parameter. We want to estimate the set $\mathcal{A} = \{i \in [\![1, p]\!] \mid \beta_i^* \neq 0\}$ of relevant variables, which is called the active set. A natural

*corresponding author: patrick.tardivel@toulouse.inra.fr

way to estimate $\mathcal{A}$ is to consider

$$\mathcal{A}(\hat{\beta}) = \{i \in [\![1, p]\!] \mid \hat{\beta}_i \neq 0\} \tag{2}$$

where $\hat{\beta}$ is any sparse estimator of β^*. The lasso estimator (Tibshirani, 1996) is probably the most popular sparse estimator. It defined by

$$\hat{\beta}^{\text{lasso}}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2} \|Y - X\beta\|^2 + \lambda \|\beta\|_1 \right\}. \tag{3}$$

The consistency of $\hat{\beta}^{\text{lasso}}(\lambda_n)$ does not imply that of $\mathcal{A}(\hat{\beta}^{\text{lasso}}(\lambda_n))$. Meinshausen and Bühlmann (2006); Zhao and Yu (2006); Zou (2006) pointed out that the irrepresentable condition is an almost necessary and sufficient condition to build a consistent active set estimator from the lasso estimator. Geometrically, this condition means that each variable X_i with $i \notin \mathcal{A}$ is almost orthogonal with the subspace $\text{Vect}\{X_i,\ i \in \mathcal{A}\}$. Because this condition is not often met in practice, Zou (2006) relaxed it by proposing a consistent estimator of $\mathcal{A}$ based on the adaptive lasso estimator defined by

$$\hat{\beta}^{\text{adapt}}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{2} \|Y - X\beta\|^2 + \lambda \sum_{i=1}^{n} \frac{1}{|\tilde{\beta}_i|} |\beta_i| \right\} \tag{4}$$

where $\tilde{\beta}$ is an *a priori* known consistent estimator of β^*. Zou (2006) showed that, as soon as $\lambda = Cn^\gamma$, with $C > 0$ and $\gamma \in]0, 1/2[$, the derived estimator $\mathcal{A}(\hat{\beta}^{\text{adapt}})$ is consistent. Some methods have been proposed to select the tuning parameter λ appropriately based, for example, on an extended regularized information criterion (Hui et al., 2015) or variable selection stability (Sun et al., 2013). But, they only provide asymptotic results. For a given n, no explicit choice is currently available for the tuning parameter whereas choosing *a priori* values for C and γ can lead to poor results (Chand, 2012). In practice, the tuning parameter is often selected using cross-validation. This method can be applied in a general setting (for any design matrix X) and is implemented in some well-known R packages, such as lars (Efron et al., 2004) and glmnet (Friedman et al., 2010). Unfortunately, as shown by Leng et al. (2006), this procedure is unsuitable for the active set estimation.

Recently, a new method referred to as covariance test (Lockhart et al., 2014) used the entire lasso solution path to provide a set containing the active set. This method tests if the active set $\mathcal{A}$ is contained in the current lasso model (i.e. $\mathcal{A} \subset \{i \in [\![1,p]\!] \mid \hat{\beta}_i^{\text{lasso}}(\lambda) \neq 0\}$ where λ is a knot of the lasso solution path). G'Sell et al. (2015) applied this approach to define 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 given by the lasso solution path. This method controls the false discovery rate of the sequential procedure. However, the rejection of the hypothesis $\mathcal{H}_k$ does not give a set containing $\mathcal{A}$. Consequently, the set of rejected hypotheses does not allow to estimate the active set properly.

Finally, Lounici (2008) provided an explicit choice for the tuning parameter λ and gave non-asymptotic results for controlling the probability of $\{\hat{\mathcal{A}} = \mathcal{A}\}$ under the assumptions that the design matrix is close to an orthogonal matrix (which implies the irrepresentable condition) and the smallest non-null parameter of β^* is larger than a threshold (beta-min condition). All of these results were obtained in the high dimension framework. If these results can be used in a small-dimension setting ($n \geq p$), they could be improved by using weakened assumptions.

Because the irrepresentable condition is not met in practice, because no guidelines are available for the choice of λ, this article proposes an explicit non-asymptotic choice of λ that guarantees, whatever $n \geq p$, non-asymptotic properties for a lasso-type estimator of the active set $\hat{\mathcal{A}}$ without requiring any condition on X. More precisely, we propose to give an explicit value for λ such that, up to a predetermined set $\mathcal{E} \subset \mathcal{A}$ (the smallest possible), the events $\{\hat{\mathcal{A}}(\lambda) \subset \mathcal{A}\}$ and $\{\mathcal{A} \setminus \mathcal{E} \subset \hat{\mathcal{A}}(\lambda)\}$ both occur with large probabilities.

This article is organised as follows. In section 2, we study the particular case in which the design matrix has orthogonal columns (i.e. $X^T X$ is diagonal), whereas section 3 addresses the general case in which X is a full-rank design matrix. A naive and common method to perform active set estimation in the small-dimension is to compute the maximum likelihood estimator to test the nullity of each component of β^*. The set of rejected hypotheses provides an estimator $\hat{\mathcal{A}}^{\text{mle}}$, and the event $\{\hat{\mathcal{A}}^{\text{mle}} \subset \mathcal{A}\}$ is then controlled via a Bonferroni-type procedure that can be quite approximate. In section 4, we show that our tuning parameter selection provides an estimator that outperforms both Lounici's and the maximum likelihood estimators of the active set. Section 5 is devoted to simulation experiments. In section 6, we focus on the analysis of metabolomic data that motivated this work.

2 Orthogonal-columns case

When the design matrix X of the Gaussian linear model (1) has orthogonal columns (i.e. X^TX is diagonal), the irrepresentable condition holds. In this setting, the probability of $\{\hat{\mathcal{A}} \subset \mathcal{A}\}$ is easy to compute. The results obtained in this section will be adapted to the general case where X^TX is no longer diagonal. Let $\hat{\beta}^{\text{pen}}$ denotes indistinctively $\hat{\beta}^{\text{lasso}}$ or $\hat{\beta}^{\text{adapt}}$ (respectively defined in equations (3) and (4)). The active set estimator $\mathcal{A}(\hat{\beta}^{\text{pen}}(\lambda)) \triangleq \hat{\mathcal{A}}^{\text{pen}}(\lambda)$ (defined in (2)) depends on a tuning parameter λ. The next proposition shows how to choose λ such that $\{\hat{\mathcal{A}}^{\text{pen}}(\lambda) \subset \mathcal{A}\}$ occurs with a controlled probability.

Proposition 1 *Let $d_1, \dots, d_p$ be the diagonal coefficients of X^TX, Z^{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, and T^{lasso} (resp. T^{adapt}) be defined by*

$$T^{\text{lasso}} = \max_{i\in[\![1,p]\!]}\{d_i \times |Z_i^{\text{ols}}|\} \text{ (resp. } T^{\text{adapt}} = \max_{i\in[\![1,p]\!]}\{d_i \times (Z_i^{\text{ols}})^2\}).$$

The $1-\alpha$ quantile of the T^{pen} distribution is denoted λ_0^{pen} (T^{pen} indistinctly denotes T^{adapt} or T^{lasso}). If $\lambda \geqslant \lambda_0^{\text{pen}}$, the following inequality holds:

$$\mathbb{P}(\hat{\mathcal{A}}^{\text{pen}}(\lambda) \subset \mathcal{A}) \geqslant 1-\alpha. \tag{5}$$

This proposition guarantees with high probability that all elements of $\hat{\mathcal{A}}^{\text{pen}}$ belong to the active set. However, there can exist some $i \in [\![1,p]\!]$ such that $\beta_i^* \neq 0$ and that do not belong to $\hat{\mathcal{A}}^{\text{pen}}$. Intuitively, large $|\beta_i^*|$ are easy to detect, so the probability of the event $\{i \in \hat{\mathcal{A}}^{\text{pen}}\}$ is high. The next proposition gives a precise meaning to how large $|\beta_i^*|$ should be for detection.

Proposition 2 *Let $\text{se}(\hat{\beta}_{\text{i}}^{\text{ols}})$ and $z_{1-\eta/p}$ be the standard error of $\hat{\beta}_i^{\text{ols}}$ and the $1-\eta/p$ quantile of a $\mathcal{N}(0,1)$ distribution, respectively. We will call detection thresholds the quantities c_i^{lasso} and c_i^{adapt}, which are respectively defined as*

$$c_i^{\text{lasso}} = \lambda_0^{\text{lasso}}/d_i + \text{se}(\hat{\beta}_{\text{i}}^{\text{ols}})\text{z}_{1-\eta/\text{p}},\ \text{c}_{\text{i}}^{\text{adapt}} = \sqrt{\lambda_0^{\text{adapt}}/\text{d}_{\text{i}}} + \text{se}(\hat{\beta}_{\text{i}}^{\text{ols}})\text{z}_{1-\eta/\text{p}}.$$

If $\mathcal{E}^{\text{pen}}$ is the set $\mathcal{E}^{\text{pen}} = \{i \in \mathcal{A} \mid |\beta_i^*| \leqslant c_i^{\text{pen}}\}$ (c_i^{pen} indistinctly denotes c_i^{lasso} or c_i^{adapt}), then

the following inequality holds:

$$\mathbb{P}(\mathcal{A} \setminus \mathcal{E}^{\text{pen}} \subset \hat{\mathcal{A}}^{\text{pen}}(\lambda_0^{\text{pen}})) \geqslant 1 - \eta. \tag{6}$$

A direct consequence of (6) is that if $\mathcal{A} \bigcap \mathcal{E}^{\text{pen}} = \emptyset$, then

$$\mathbb{P}(\mathcal{A} \subset \hat{\mathcal{A}}^{\text{pen}}(\lambda_0^{\text{pen}})) \geqslant 1 - \eta.$$

In other words, when all non-null $|\beta_i^*|$ are sufficiently large, the active set is contained in $\hat{\mathcal{A}}^{\text{pen}}(\lambda_0^{\text{pen}})$ with a high probability. Note that the detection thresholds are deterministic numbers that depend only on the design matrix and Γ. Consequently, when the variance matrix Γ is known *a priori*, a detection threshold can be computed for each column of X before the data analysis. In contrast, in their book, Bühlmann and van de Geer (2011) proposed the same detection threshold for all columns of X (beta-min condition). Using a similar idea, by setting $c = \max\{c_1^{\text{pen}}, \ldots, c_p^{\text{pen}}\}$, we obtain a single detection threshold that can be used for all columns of X. Because c_i^{pen} can be quite different from one another, this single detection threshold can be very approximate and should be used with care.

3 General case

In this section, we no longer assume that the design matrix X has orthogonal columns. In this general setting, the lasso or adaptive lasso estimators do not have a closed form. Consequently, it becomes difficult to choose a tuning parameter λ and find a set $\mathcal{E}^{\text{pen}}$ that ensure that the events $\{\hat{\mathcal{A}}^{\text{pen}}(\lambda) \subset \mathcal{A}\}$ and $\{\mathcal{A} \setminus \mathcal{E}^{\text{pen}} \subset \hat{\mathcal{A}}^{\text{pen}}\}$ occur with a controlled probability. To overcome this difficulty, we propose application of a linear transformation $U \in G$ that orthogonalises the matrix X (i.e. $(UX)^T UX$ is diagonal) to each member of the model (1). This leads to the new linear Gaussian model

$$\tilde{Y} = \tilde{X}\beta^* + \tilde{\varepsilon} \text{ with } \tilde{Y} = UY, \tilde{X} = UX \text{ and } \tilde{\varepsilon} = U\varepsilon. \tag{7}$$

One can notice that the lasso estimator of the model (7)

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

is different from the lasso estimator of the model (1).

Although the active set estimator provided by $\hat{\beta}^{\text{lasso}}$ is not consistent as soon as the irrepresentable condition on the design matrix X is not verified, this condition is no longer necessary to build a consistent active set estimator with $\hat{\beta}_U^{\text{lasso}}$.

For all $U \in G$, $\hat{\beta}^{\text{ols}}(U)$ denotes the ordinary least squares estimator of β^* once the transformation U has been applied; namely, $\hat{\beta}^{\text{ols}}(U) = \left[\tilde{X}^T \tilde{X}\right]^{-1} \tilde{X}^T \tilde{Y}$. Because $\tilde{X}$ has orthogonal columns, it is possible to use the propositions of the previous section. More precisely, the propositions 1 and 2 provide a tuning parameter $\lambda_0^{\text{pen}}(U)$ and detection thresholds $c_i^{\text{pen}}(U)$ that controlled the probabilities of the events $\{\hat{\mathcal{A}}^{\text{pen}}(\lambda_0^{\text{pen}}(U)) \subset \mathcal{A}\}$ and $\{i \in \hat{\mathcal{A}}(\lambda_0^{\text{pen}}(U))\}$, respectively. The set G of linear transformations that orthogonalise X is large. Among these, we will select a transformation for which the detection thresholds $c^{\text{pen}}(U)$ are as small as possible. When the detection thresholds $c^{\text{pen}}(U)$ become small, the cardinality of $\mathcal{E}_U^{\text{pen}}$ decreases, and the cardinality of the set of elements at the least detected $\mathcal{A} \setminus \mathcal{E}_U^{\text{pen}}$ consequently increases. Because $c^{\text{pen}}(U)$ is a vector (there is a detection threshold for each column of X), we need a norm ϕ that indicates how small $c^{\text{pen}}(U)$ is. From now, ϕ will denote a L^q norm ($q > 0$) on $\mathbb{R}^p$. The theorem 1 shows that it is possible to pick a transformation U_ϕ for which $\phi(c^{\text{pen}}(U_\phi))$ is minimal.

Theorem 1 *There exists a linear transformation $U_\phi \in G$, a tuning parameter $\lambda_0^{\text{pen}}(U_\phi)$ and a set $\mathcal{E}_{U_\phi}^{\text{pen}}$ such that*
1) $\mathbb{P}(\hat{\mathcal{A}}^{\text{pen}}(\lambda_0^{\text{pen}}(U_\phi)) \subset \mathcal{A}\}) \geqslant 1 - \alpha$ and
2) $\mathbb{P}(\mathcal{A} \setminus \mathcal{E}_{U_\phi}^{\text{pen}} \subset \hat{\mathcal{A}}^{\text{pen}}(\lambda_0^{\text{pen}}(U_\phi)) \geqslant 1 - \eta$.
Moreover, $c^{\text{pen}}(U_\phi)$ is minimal for the norm ϕ.

The linear transformation U_ϕ depends on the penalized estimator, α and η. We choose to simplify the notation by not writing these dependencies.

The previous theorem gives the existence of U_ϕ. The following two lemmas are the main steps of its construction. Because U_ϕ should provide small detection thresholds, let us first recall

their expressions,

$$c_i^{\text{lasso}} = \lambda_0^{\text{lasso}}/d_i + se(\hat{\beta}_i^{ols})z_{1-\eta/p} \text{ and } c_i^{\text{adapt}} = \sqrt{\lambda_0^{\text{adapt}}/d_i} + se(\hat{\beta}_i^{ols})z_{1-\eta/p}.$$

Notice that $c^{\text{pen}}(U)$ increases when the variance matrix of $\hat{\beta}^{\text{ols}}(U)$ is large. Indeed, the term $se(\hat{\beta}_i^{ols}(U))z_{1-\eta/p}$ increases with the standard error of $\hat{\beta}_i^{\text{ols}}(U)$. Furthermore $\lambda_0^{\text{pen}}(U)$ is inflated when the variance matrix of $\hat{\beta}^{\text{ols}}(U)$ is large. Lemma 1 exhibits linear transformations V_δ that orthogonalise the design matrix X and for which the estimator $\hat{\beta}^{ols}(V_\delta)$ has a small variance (it is efficient).

Lemma 1 *Let $\delta \in]0, +\infty[^p$ and P_δ be a $n \times n$ 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

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

where M_1 and M_2 are $p \times p$ and a $p \times (n-p)$ matrices, respectively. The matrix V_δ is an $n \times n$ matrix defined by

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

V_δ then belongs to G, and $\hat{\beta}^{\text{ols}}(V_\delta)$ is equal to the maximum likelihood estimator $\hat{\beta}^{\text{mle}}$ of (1).

If the linear transformation U in (7) is equal to V_δ, we obtain a tuning parameter $\lambda_0^{\text{pen}}(V_\delta)$ and detection thresholds $c^{\text{pen}}(V_\delta)$, both of which have a closed form. More precisely, if Z^{mle} is a centred Gaussian random variable with the same variance matrix as the maximum likelihood estimator $\hat{\beta}^{\text{mle}}$,

$$Z^{\text{mle}} \sim \mathcal{N}(0, (X^T \Gamma^{-1} X)^{-1}), \tag{8}$$

the tuning parameters $\lambda_0^{\text{lasso}}(V_\delta)$ and $\lambda_0^{\text{adapt}}(V_\delta)$ are respectively defined as the $1-\alpha$ quantiles of the distributions of $\max\{\delta_1|Z_1^{\text{mle}}|, \ldots, \delta_p|Z_p^{\text{mle}}|\}$ and $\max\{\delta_1(Z_1^{\text{mle}})^2, \ldots, \delta_p(Z_p^{\text{mle}})^2\}$. Further-

more, if $se(\hat{\beta}_i^{\text{mle}})$ is the standard errors of $\hat{\beta}_i^{\text{mle}}$, threshold detection $c_i^{\text{lasso}}(V_\delta)$ and $c_i^{\text{adapt}}(V_\delta)$ are equal to

$$c_i^{\text{lasso}}(V_\delta) = \lambda_0^{\text{lasso}}(V_\delta)/\delta_i + se(\hat{\beta}_i^{\text{mle}})z_{1-\eta/p} \text{ and } c_i^{\text{adapt}}(V_\delta) = \sqrt{\lambda_0^{\text{adapt}}(V_\delta))/\delta_i} + se(\hat{\beta}_i^{\text{mle}})z_{1-\eta/p}.$$

The expression of $c^{\text{pen}}(V_\delta)$ is easy to optimise with respect to δ and only requires simulation of the Gaussian vector Z^{mle}. This optimisation allows minimal detection thresholds $c^{\text{pen}}(U_\phi)$ to be obtained for the norm ϕ. The next lemma proves that theorem 1 holds for the linear transformation U_ϕ.

Lemma 2 *Set*

$$U_\phi = V_{\delta^*} \text{ with } \delta^* = \underset{\delta \in]0,+\infty[^p}{\text{arginf}} \ \phi(c^{\text{pen}}(V_\delta)). \tag{9}$$

Then,

$$\forall U \in G, \phi(c^{\text{pen}}(U_\phi)) \leqslant \phi(c^{\text{pen}}(U)).$$

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

In the particular case where ϕ is the supremum norm, the next proposition shows that the components of $c^{\text{pen}}(U_\infty)$ are equal. By optimizing c^{pen} for this norm, we obtain a single detection threshold (the same for all columns of X).

Proposition 3 *Let V_{δ^*} as in lemma 2 be an element for which $\|c^{\text{pen}}(V_{\delta^*})\|_\infty$ is minimal. Assume that η, defined in proposition 2, is such that $\eta/p < 1/2$; then we have*

$$\forall U \in G, \|c^{\text{pen}}(V_{\delta^*})\|_\infty \leqslant \|c^{\text{pen}}(U)\|_\infty \Leftrightarrow c_1^{pen}(V_{\delta^*}) = \cdots = c_p^{\text{pen}}(V_{\delta^*}).$$

The assumption $\eta/p < 1/2$ is not restrictive. It simply ensures that the detection thresholds are positive. The result given by theorem 1 depends on the penalized estimator used. *A priori*, one might think that the estimators $\hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}(U_\phi^{\text{lasso}}))$ and $\hat{\mathcal{A}}^{\text{adapt}}(\lambda_0^{\text{adapt}}(U_\phi^{\text{adapt}}))$ given by this theorem are different. However, the next proposition shows that the distributions of these two estimators are equal.

Proposition 4 *Let δ^* be an element of $]0,+\infty[^p$ for which $\phi(c^{\text{lasso}}(V_{\delta^*}))$ is minimal. Define $\zeta^* = ((\delta_1^*)^2, \ldots, (\delta_p^*)^2)$; then, we have the two results that $\phi(c^{\text{adapt}}(V_{\zeta^*}))$ is minimal and the distribution of $\hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}(V_{\delta^*}))$ is equal to that of $\hat{\mathcal{A}}^{\text{adapt}}(\lambda_0^{\text{adapt}}(V_{\zeta^*}))$.*

Consequently, there is no gain from using an adaptive lasso instead of a lasso estimator to estimate the active set. Thus, in the following, we restrict ourself to $\hat{\mathcal{A}}^{\text{lasso}}$.

The study of the asymptotic properties of lasso-type active set estimators is not the main purpose of this article. Nevertheless, it is interesting to show that these estimators behave properly when n tends to $+\infty$ and p is fixed as shown in the following proposition. Remind that the estimator $\hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}(U_\phi))$ depends on α and n.

Proposition 5 *There exists a sequence $(\alpha_n)_{n\in\mathbb{N}^*}$ that converges to zero such that*

$$\lim_{n\to+\infty} \mathbb{P}(\hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}(U_\phi)) = \mathcal{A}) = 1.$$

4 Comparison with other active set estimators

In this part the letters z and q denotes respectively quantiles of a $\mathcal{N}(0,1)$ and χ_1^2 distributions.

4.1 Comparison with Lounici's estimator

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

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

He proved that the event $\{\hat{\mathcal{A}}^L = \mathcal{A}\}$ has a controlled probability when the Gram matrix $\frac{1}{n}X^TX$ is close to the identity, the noise ε is Gaussian, and the smallest non-null parameter $|\beta_i^*|$ is sufficiently large. To enable comparison of $\hat{\mathcal{A}}^{\text{L}}$ and $\hat{\mathcal{A}}^{\text{lasso}}$, we assume that X is orthogonal ($X^TX = Id_p$). In this setting, the detection thresholds given in Proposition 2 are all equal to $c^{\text{lasso}} = \sigma\left(\sqrt{q_{\sqrt[p]{1-\alpha}}} + z_{1-\eta/p}\right)$. If λ_0^{lasso} is chosen as in the proposition 1, we have

$$\min_{i\in\mathcal{A}}\{\beta_i^*\} \geqslant c^{\text{lasso}} \Rightarrow \mathbb{P}(\hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}) = \mathcal{A}) \geqslant 1-\alpha-\eta,$$

whereas in the same setting, Lounici gave

$$\min_{i\in\mathcal{A}}\{\beta_i^*\} \geqslant c^{\mathrm{L}} \Rightarrow \mathbb{P}(\hat{\mathcal{A}}^{\mathrm{L}} = \mathcal{A}) \geqslant 1 - p^{1-A^2/8},$$

where $c^{\mathrm{L}} = 3A\sigma\sqrt{\log(p)}$ and $A \geqslant 2\sqrt{2}$.

If c^{L} and c^{lasso} are both chosen such that $\mathbb{P}(\hat{\mathcal{A}}^{\mathrm{L}} = \mathcal{A}) \geqslant 1-\alpha$ and $\mathbb{P}(\hat{\mathcal{A}}^{\mathrm{lasso}} = \mathcal{A}) \geqslant 1-\alpha$, they have the same order of magnitude $\sigma\sqrt{\log(p)}$, but for any p, c^{lasso} is smaller than c^{L}, as illustrated by Table 1.

p	10	20	50	100
c^{L}/σ	18.20	19.53	21.15	22.30
$c^{\mathrm{lasso}}/\sigma$	5.37	5.82	6.37	6.76

Table 1: This table provides a numerical comparison of c^{L} and c^{lasso}. We chose $\alpha = \eta = 0.025$ and A such that $1 - p^{1-A^2/8} = 0.95$. These values ensure that $\mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) \geqslant 0.95$. This table shows that for any $p \geqslant 1$, c^{lasso} is smaller than c^{L}.

The main explanation of the observed difference between c^{lasso} and c^{L} relies on the choice of the tuning parameter. Indeed, the tuning parameter $\lambda_0^{\mathrm{lasso}}$ is the $1-\alpha$ quantile of $\max\{|Z_1^{\mathrm{ols}}|, \ldots, |Z_p^{\mathrm{ols}}|\}$, whereas Lounici's tuning parameter bounds above the $1-\alpha$ quantile of $2\max\{|Z_1^{\mathrm{ols}}|, \ldots, |Z_p^{\mathrm{ols}}|\}$. It results that $\hat{\mathcal{A}}^{\mathrm{L}} \subseteq \hat{\mathcal{A}}^{\mathrm{lasso}}$, implying that $c^{\mathrm{L}} \geqslant c^{\mathrm{lasso}}$.

4.2 Comparison with the maximum likelihood estimator

Using the maximum likelihood estimator, one can build a test that rejects the null hypothesis $\beta_i^* = 0$ when $|\hat{\beta}_i^{\mathrm{mle}}|$ is larger than a threshold t_i. The set of rejected hypotheses provides an estimator $\hat{\mathcal{A}}^{\mathrm{mle}}$ of $\mathcal{A}$ defined by

$$\hat{\mathcal{A}}^{\mathrm{mle}} = \{i \in [\![1, p]\!] \mid |\hat{\beta}_i^{\mathrm{mle}}| > t_i\}.$$

This estimator depends on the thresholds $t_1, \ldots, t_p$ that can be chosen as follows. If there is at least one false rejection (if there exists an integer i for which $\beta_i^* = 0$ and $|\hat{\beta}_i^{\mathrm{mle}}| > t_i$), then $\hat{\mathcal{A}}^{\mathrm{mle}} \not\subset \mathcal{A}$. Thus, to ensure that the event $\{\hat{\mathcal{A}}^{\mathrm{mle}} \subset \mathcal{A}\}$ holds with a large probability, it is necessary to control the family-wise error rate (FWER) (Lehmann and Romano, 2005). When the $\hat{\beta}_i^{\mathrm{mle}}$ values are independent, a Šidák procedure gives exact control of the FWER (Dudoit

and Van Der Laan, 2007), whereas in all other cases, a Bonferroni procedure provides only approximate control. The following proposition compares the distributions of $\hat{\mathcal{A}}^{\text{lasso}}$ and $\hat{\mathcal{A}}^{\text{mle}}$ in a simplified case of independence.

Proposition 6 *Assume that* $\hat{\beta}^{\text{mle}} \sim \mathcal{N}(\beta^*, \sigma^2 Id_p)$*, and let us set* $t_1 = \cdots = t_p = \sigma\sqrt{q_{\sqrt[p]{1-\alpha}}}$ *and* V_{δ^*} *as defined in proposition 3. Then,* $\mathbb{P}(\{\hat{\mathcal{A}}^{\text{mle}} \subset \mathcal{A}\}) \geqslant 1-\alpha$*. Moreover, the distributions of* $\hat{\mathcal{A}}^{\text{mle}}$ *and* $\hat{\mathcal{A}}^{\text{lasso}}(\lambda^{\text{lasso}}(V_{\delta^*}))$ *are equal.*

As shown hereafter, when the components of the maximum likelihood estimator are no longer independent, $\hat{\mathcal{A}}^{\text{lasso}}(\lambda^{\text{lasso}}(U_\phi))$ has better performance than $\hat{\mathcal{A}}^{\text{mle}}$. Let us recall that the proposition 1 ensures that $\mathbb{P}(\hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}(U_\phi)) \subset \mathcal{A}) \geqslant 1-\alpha$ and the proposition 2 gives a set $\mathcal{E}^{\text{lasso}}$ such that $\mathbb{P}(\mathcal{A} \setminus \mathcal{E}^{\text{lasso}} \subset \hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}(U_\phi)) \leqslant 1-\eta$. It is possible to define an estimator $\hat{\mathcal{A}}^{\text{mle}}$ with similar characteristics to $\hat{\mathcal{A}}^{\text{lasso}}(\lambda^{\text{lasso}}(U_\phi))$. Indeed, a Bonferroni procedure yields thresholds $t_1, \ldots, t_p$ such that $\{\hat{\mathcal{A}}^{\text{mle}} \subset \mathcal{A}\}$ occurs with a probability of at least $1-\alpha$. Furthermore, in a similar manner as $\mathcal{E}^{\text{lasso}}$, it is possible to build $\mathcal{E}^{\text{mle}}$ such that $\mathbb{P}(\mathcal{A} \setminus \mathcal{E}^{\text{mle}} \subset \hat{\mathcal{A}}^{\text{mle}}) \geqslant 1-\eta$. The following proposition compares the detection thresholds associated to $\mathcal{E}^{\text{mle}}$ with that associated to $\mathcal{E}^{\text{lasso}}$.

Proposition 7 *If we set* $t_1 = se(\hat{\beta}_1^{\text{mle}}) z_{1-\alpha/2p}, \ldots, t_p = se(\hat{\beta}_p^{\text{mle}}) z_{1-\alpha/2p}$ *then*

$$\mathbb{P}(\hat{\mathcal{A}}^{\text{mle}} \subset \mathcal{A}) \geqslant 1-\alpha.$$

Let us define $\mathcal{E}^{\text{mle}} = \{i \in [\![1,p]\!] \mid |\beta_i^*| \leqslant c_i^{\text{mle}}\}$ *with* $c_i^{\text{mle}} = t_i + se(\hat{\beta}_i^{\text{mle}}) z_{1-\eta/p}$ *then*

$$\mathbb{P}(\mathcal{A} \setminus \mathcal{E}^{\text{mle}} \subset \hat{\mathcal{A}}^{\text{mle}}) \geqslant 1-\eta.$$

Furthermore, let $U_\phi \in G$ *be the linear transformation for which* $\phi(c^{\text{lasso}}(U_\phi))$ *is minimal then,* $\phi(c^{\text{lasso}}(U_\phi)) \leqslant \phi(c^{\text{mle}})$.

Because the detection thresholds $c^{\text{lasso}}(U_\phi)$ and c^{mle} measure the performance of $\hat{\mathcal{A}}^{\text{lasso}}(\lambda^{\text{lasso}}(U_\phi))$ and $\hat{\mathcal{A}}^{\text{mle}}$, respectively, the estimator $\hat{\mathcal{A}}^{\text{lasso}}(\lambda^{\text{lasso}}(U_\phi))$ is better than $\hat{\mathcal{A}}^{\text{mle}}$. Heuristically, the performance of $\hat{\mathcal{A}}^{\text{lasso}}$ is even better than that of $\hat{\mathcal{A}}^{\text{mle}}$ when the components of $\hat{\beta}^{\text{mle}}$ (and thus those of Z^{mle} in (8)) are correlated or when p is large.

5 Simulation experiments

The simulations given in subsection 5.1 illustrate the theoretical results obtained in the previous section. The numerical comparison of the optimal threshold detection for several norms is performed in subsection 5.2. In subsection 5.3, it is shown that the detection thresholds give exact control of $\mathbb{P}(i \in \hat{\mathcal{A}}^{\text{lasso}})$.

5.1 Numerical comparison with $\hat{\mathcal{A}}^{\text{mle}}$

In this subsection, we numerically compare $\mathbb{P}(\hat{\mathcal{A}}^{\text{lasso}} = \mathcal{A})$ and $\mathbb{P}(\hat{\mathcal{A}}^{\text{mle}} = \mathcal{A})$. The tuning parameter λ_0^{lasso} and the thresholds $t_1, \ldots, t_p$ were chosen according to proposition 1 and section 4.2 to guarantee that the event $\{\hat{\mathcal{A}} \subset \mathcal{A}\}$ holds with a probability greater than 0.95. We set $p = 10$ or $p = 100$; for all $i \leqslant p$, $\hat{\beta}_i^{\text{mle}} \sim \mathcal{N}(\beta_i^*, \sigma^2)$, and for all $i \neq j$, $\text{corr}(\hat{\beta}_i^{\text{mle}}, \hat{\beta}_j^{\text{mle}}) = \rho$ and $\beta_1^* = \cdots = \beta_5^* = c > 0$; $\beta_6^* = \cdots = \beta_p^* = 0$. Figure 1 represents the curve of

$$\frac{\mathbb{P}_c(\mathcal{A} = \hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}))}{\mathbb{P}_c(\mathcal{A} = \hat{\mathcal{A}}^{\text{mle}})}$$

as a function of c/σ.

This figure illustrates that the ratio $\mathbb{P}_c(\hat{\mathcal{A}}^{\text{lasso}} = \mathcal{A})/\mathbb{P}(\hat{\mathcal{A}}^{\text{mle}} = \mathcal{A})$ increases with increasing p and ρ. For $p = 10$ (resp. $p = 100$), $\mathbb{P}_c(\hat{\mathcal{A}}^{\text{lasso}} = \mathcal{A}) \geqslant \mathbb{P}(\hat{\mathcal{A}}^{\text{mle}} = \mathcal{A})$ when $c/\sigma \leqslant 5$ (resp. $c/\sigma \leqslant 5.5$), whereas the reverse inequality holds elsewhere. Consequently, for any p and ρ, when $|\beta_i^*| \leqslant 5\text{se}(\hat{\beta}_{\text{i}}^{\text{mle}})$, $\hat{\mathcal{A}}^{\text{lasso}}$ better estimates $\mathcal{A}$ than $\hat{\mathcal{A}}^{\text{mle}}$.

When $c/\sigma \geqslant 5$, $\hat{\mathcal{A}}^{\text{mle}}$ is better. This is not surprising because

$$\lim_{c \to +\infty} \frac{\mathbb{P}_c(\hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}) = \mathcal{A})}{\mathbb{P}_c(\hat{\mathcal{A}}^{\text{mle}} = \mathcal{A})} = \lim_{c \to +\infty} \frac{\mathbb{P}_c(\hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}) \subset \mathcal{A})}{\mathbb{P}_c(\hat{\mathcal{A}}^{\text{mle}} \subset \mathcal{A})}.$$

This latter ratio (which, in fact, does not depend on c) is less than 1 because $\mathbb{P}(\hat{\mathcal{A}}^{\text{lasso}} \subset \mathcal{A})$ is closer to the nominal probability (0.95) than $\mathbb{P}(\hat{\mathcal{A}}^{\text{mle}} \subset \mathcal{A})$. This performance of $\hat{\mathcal{A}}^{\text{lasso}}$ simply reflects that $\hat{\mathcal{A}}^{\text{lasso}}$ has better control of $\mathbb{P}(\hat{\mathcal{A}} \subset \mathcal{A})$ than $\hat{\mathcal{A}}^{\text{mle}}$.

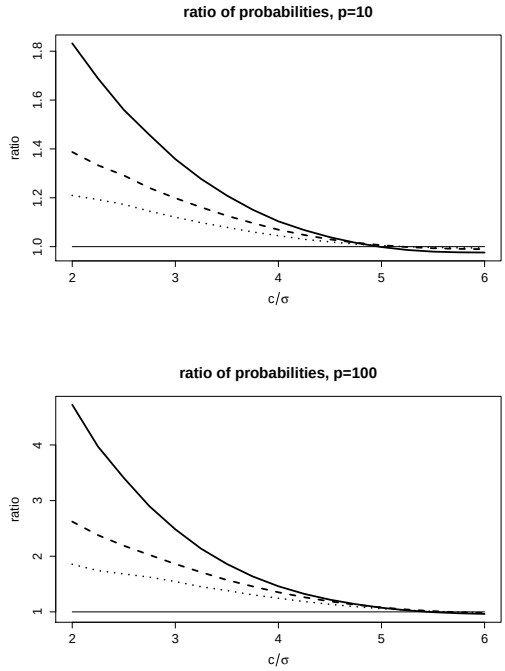

Figure 1: These figures represent the curves of $\mathbb{P}_c(\mathcal{A} = \hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}))/\mathbb{P}_c(\mathcal{A} = \hat{\mathcal{A}}^{\text{mle}})$ as a function of c/σ for $\rho = 0.9$ (solid line), $\rho = 0.7$ (dashed line) and $\rho = 0.5$ (dotted line). One can observe that this ratio increases with increasing number of columns p and with the correlation ρ. When the coefficient c/σ goes to infinity, the ratio converges to $\mathbb{P}_c(\hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}) \subset \mathcal{A})/\mathbb{P}_c(\hat{\mathcal{A}}^{\text{mle}} \subset \mathcal{A})$.

5.2 A single or several detection thresholds?

From proposition 3, optimisation of the detection thresholds c^{lasso} for the supremum norm leads to a single threshold (i.e. the threshold is the same for each column of X). If the optimisation of c^{lasso} is performed with another norm, several detection thresholds are obtained. In this section, we study the influence of the standard errors of $\hat{\beta}_i^{\text{mle}}$ on the optimal thresholds for the L^1 and L^∞ norms. Recall that these thresholds are defined by

$$c^{\text{lasso}}(U_i) = c^{lasso}(V_{\delta^*}^*), \text{ with } \delta^* = \underset{\delta \in]0,+\infty[^p}{\text{arginf}} \|c^{\text{lasso}}(V_\delta)\|_i, \text{ with } i \in \{1, \infty\}.$$

We used 100,000 realizations of the random vector Z^{mle} defined by (8) to compute $c^{\text{lasso}}(V_\delta)$. For simplicity, we assumed that $\hat{\beta}^{\text{mle}}$ has independent components. We first assumed that $\text{var}(\hat{\beta}^{\text{mle}}) = \sigma^2 Id_p$. In this case, the optimal thresholds for the L^1 norm are equal to optimal thresholds for

the supremum norm,

$$\forall i \in [\![1, 10]\!], \frac{c_i^{\text{lasso}}(U_\infty)}{\sigma} = \frac{c^{\text{lasso}}(U_1)}{\sigma} = 4.44.$$

This equality is no longer true when $\text{var}(\hat{\beta}^{\text{mle}}) = \sigma^2 \text{diag}(1, \ldots, 10)$. The optimal thresholds for the L_1 norm are

$$\frac{c^{\text{lasso}}(U_1)}{\sigma} = (5.39, 6.24, 7.81, 8.86, 9.82, 10.72, 11.70, 12.32, 13.18, 13.61),$$

whereas those that are optimal for the supremum norm are

$$\forall i \in [\![1, 10]\!], \frac{c_i^{\text{lasso}}(U_\infty)}{\sigma} = 12.29.$$

In this second setting, one notes that the optimal threshold for the supremum norm is slightly less than the maximum of optimal thresholds for the L^1 norm. Because in practice the standard errors of $\hat{\beta}^{\text{mle}}$ are unequal, it is preferable to optimise the L^1 norms of the thresholds.

5.3 Lasso estimation of the active set

The detection thresholds should ensure that if $\beta_i^* \geqslant c_i^{\text{lasso}}$, $\mathbb{P}(i \in \hat{\mathcal{A}}^{\text{lasso}}) \geqslant 1-\eta/p$. However, λ_0^{lasso} as defined in proposition 1 should guarantee that $\mathbb{P}(\hat{\mathcal{A}}^{\text{lasso}} \subset \mathcal{A}) \geqslant 1-\alpha$. We will evaluate whether these latter probabilities are close to their targets. In this simulation, we set $n = 100, p = 10$; Γ is equal to

$$\forall i, j \in [\![1, 100]\!], \Gamma_{i,j} = \sigma^2 \rho^{|i-j|}, \text{ with } \rho = 0.5.$$

The design matrix X was chosen such that

$$(X^T X)_{i,j} \begin{cases} 1 \text{ if } i = j, \\ 0.5 \text{ if } |i - j| = 1, \\ 0 \text{ in the other cases.} \end{cases}$$

Table 2 shows that for the i$^{\text{th}}$ variable: β_i^*, c_i^{lasso} and $\mathbb{P}(i \in \hat{\mathcal{A}}^{\text{lasso}})$. These numerical results were obtained assuming $\alpha = \eta/p = 0.05$.

i	β_i^*/σ	$c_i^{\text{lasso}}/\sigma$	$P(i \in \hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}(U_1)))$
1	5.65	8.47	0.563
2	7.82	9.39	0.824
3	10.95	10.95	0.950
4	13.38	11.15	0.994
5	16.92	11.28	0.999
6	0	11.34	0.010
7	0	10.82	0.012
8	0	10.45	0.008
9	0	9.28	0.007
10	0	8.35	0.003

Table 2: As soon as $|\beta_i^*|$ is greater than the corresponding threshold detection, $\mathbb{P}(i \in \hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}(U_1)) \geqslant 0.95$. The parameter $\beta_3^* = c_3^{\text{lasso}}$; thus, $\mathbb{P}(3 \in \hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}(U_1))) = 0.95$. Note that the sum of the probabilities $\mathbb{P}(i \in \hat{\mathcal{A}}^{\text{lasso}})$ over the variables that do not belong to $\mathcal{A}$ (i.e. 6, 7, 8, 9, 10) is less than 0.05.

The set $\mathcal{E}^{\text{lasso}}$ of variables that are difficult to detect is comprised of those for which $0 < \beta_i^* < c_i^{\text{lasso}}$ that is $\{1, 2\}$. Note that for these variables, $\mathbb{P}(i \in \hat{\mathcal{A}}^{\text{lasso}}) < 0.95$. When λ_0^{lasso} is chosen as described in proposition 1-i.e. without knowing the cardinality of $\mathcal{A}$-we have

$$\mathbb{P}(\hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}(U_1)) \subset \{1, \ldots, 5\}) = 0.969 \geqslant 0.95,$$

whereas

$$\mathbb{P}(\mathcal{A} \setminus \mathcal{E}^{\text{lasso}} \subset \hat{\mathcal{A}}^{\text{lasso}}) = 0.943 \text{ and } \mathbb{P}(\hat{\mathcal{A}}^{\text{lasso}} = \mathcal{A}) = 0.389.$$

6 Application in metabolomics: detection of metabolites

Metabolomics is the science concerned with the detection of metabolites (small molecules) in biological mixtures (e.g. blood and urine). The most common technique for performing such characterization is proton nuclear magnetic resonance (NMR). Each metabolite generates a characteristic resonance signature in the NMR spectra with an intensity proportional to its concentration in the mixture. The number of peaks generated by a metabolite and their locations and ratio of heights are reproducible and uniquely determined: each metabolite has its own signature in the spectra. Each signature spectrum of each metabolite can be stored in a library that could contain hundreds of spectra. One of the major challenges in NMR analysis of metabolic

profiles remains to be automatic metabolite assignment from spectra. To identify metabolites, experts use spectra of pure metabolites and manually compare these spectra to the spectrum of the biological mixture under analysis. Such a method is time-consuming and requires domain-specific knowledge. Furthermore, complex biological mixtures can contain hundreds or thousands of metabolites, which can result in highly overlapping peaks. Figure 2 gives an example of an annotated spectrum of a mixture.

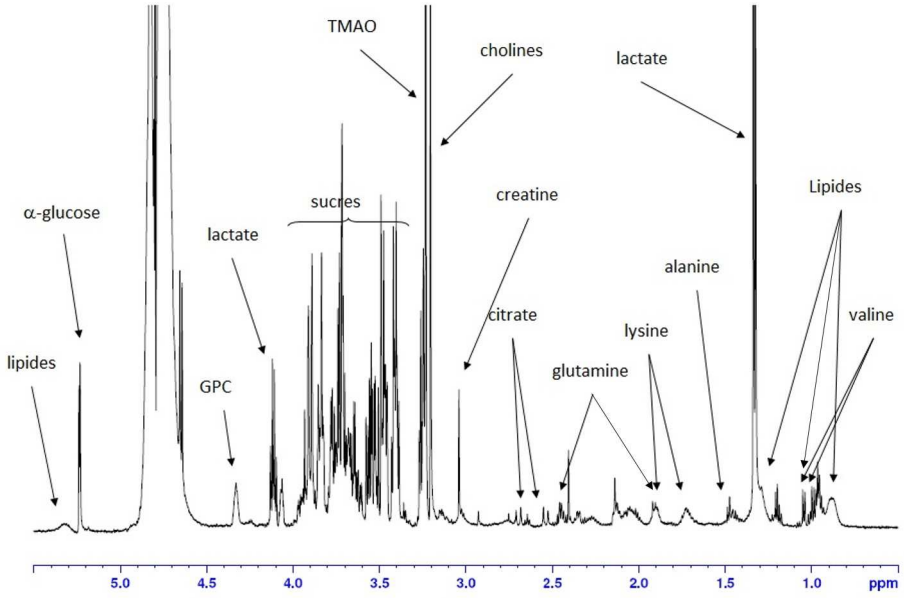

Figure 2: Example of an annotated mixture spectrum. There are overlaps between peaks of lipides and valine and between the peaks of glutamine and lysine.

Recently, automatic methods have been proposed. For example, Metabohunter (Tulpan et al., 2011) is very fast, but the statistical interpretation of the scoring function associated with each metabolite is difficult. Other methods, such as BATMAN (Astle et al., 2012; Hao et al., 2012), Mercier et al. (2011) or Zheng et al. (2011), are based on a modelling using a Lorentzian shape and a Bayesian strategy. Nevertheless, they are time-consuming and thus cannot be applied to a large library of metabolites, and 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.

Let us denote $\mathcal{A}$ as the set of metabolites in the mixture and $\hat{\mathcal{A}}$ as the set of detected metabolites. The event $\{\hat{\mathcal{A}} \subset \mathcal{A}\}$ implies that there is no false detection; thus, the method must control the probability of false detections. Furthermore, if $\mathcal{E} \subset \mathcal{A}$ is the set of undetected metabolites, the main objective is to provide a set $\mathcal{E}$ as small as possible. As proved in the previous sections, the estimator $\hat{\mathcal{A}}^{\text{lasso}}$ can fulfil these conditions and provide us with a good estimate for the active set $\mathcal{A}$.

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 metabolites $\{f_i\}_{1 \leqslant i \leqslant p}$ (with $\int_T f_i(t)dt = 1$) that can be found in a mixture. This family of p spectra is assumed to be linearly independent. In a first approximation, the observed spectrum of the mixture Y can be modelled as a discretized noisy convex combination of the pure spectra:

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

The random vector $(\varepsilon_1, \ldots, \varepsilon_n)$ is a standard Gaussian $\mathcal{N}(0, \sigma^2 Id_n)$. The variance σ^2 is estimated using several observations of a metabolite spectrum. Recall that the objective is to estimate the active set $\mathcal{A} = \{i \in [\![1, p]\!] \mid \beta_i^* \neq 0\}$.

6.2 Real dataset

The method for the detection of metabolites was performed on a known mixture. The metabolomicians 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 3.

The 6 metabolites that are present in the complex mixture are detected, including those with small proportions. There is no false detection because the 30 other metabolites are not detected. The detection thresholds are very different from one metabolite to another. They are strongly impacted by two characteristics of the metabolite spectrum: first, the height of the peaks. If, among all peaks of a spectrum, there is one large peak, the detection threshold would be lower,

Metabolites	Detected by $\hat{\mathcal{A}}^{\text{lasso}}$	Actual proportions	c^{lasso}
Choline chloride	Yes	0.545	0.011
Creatinine	Yes	0.209	0.011
Benzoic acid	Yes	0.086	0.018
L-Proline	Yes	0.069	0.034
D-Glucose	Yes	0.060	0.036
L-Phenylalanine	Yes	0.029	0.025
30 other metabolites	No	0	$[0.010; 0.034]$

Table 3: This table presents the results for the 36 metabolites of the library. The metabolites detected using $\hat{\mathcal{A}}^{\text{lasso}}$ are presented in the first column. The actual proportions of each metabolite are presented in the second column. The detection thresholds, calculated using proposition 2 with $\eta/p = 0.05$, are listed in the last column.

and this metabolite would be easier to detect. This is the case for the choline chloride spectrum but not for that of D-glucose, which is composed of many small peaks. Second, the detection threshold decreases with increasing number of peaks that do not overlap with others. Because the whole procedure is quite fast, lasting only a few seconds, it could be easily applied to a library containing several hundred metabolites.

7 Conclusions

In this article, we proposed a lasso-type estimation of the active set $\hat{\mathcal{A}}^{\text{pen}}$ in the small-dimensional setting. When the design matrix has orthogonal columns, we gave a tuning parameter λ_0^{pen} and a set $\mathcal{E}^{\text{pen}}$ such that the events $\{\hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{pen}}) \subset \mathcal{A}\}$ and $\{\mathcal{A} \setminus \mathcal{E}^{\text{pen}} \subset \hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{pen}})\}$ occur with a controlled probability. 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. We obtained good performance for the estimator $\hat{\mathcal{A}}^{\text{pen}}(\lambda_0^{\text{pen}}(U))$ based on $\hat{\beta}^{\text{ols}}(U)$. However, this estimator is no longer available in the high-dimensional setting. Of course, the least squares estimator and orthogonal design matrix are specific to the small-dimensional setting. However, the methods developed in this paper could most likely be combined with the results of Lounici to relax the well-known irrepresentable condition and improve active set estimation in the high-dimensional setting.

8 Appendix

Proof (Proposition 1) When the design matrix is orthogonal (i.e. $X^T X = Id_p$) or has orthogonal columns, both the lasso and adaptive lasso estimators have explicit expressions (Tibshirani, 1996; Hastie et al., 2009; Bühlmann and van de Geer, 2011) given by

$$\hat{\beta}_i^{\text{adapt}}(\lambda) = sgn(\hat{\beta}_i^{\text{ols}}) \left(|\hat{\beta}_i^{\text{ols}}| - \frac{\lambda}{d_i |\hat{\beta}_i^{\text{ols}}|} \right)_+ \text{ and } \hat{\beta}_i^{\text{lasso}}(\lambda) = sgn(\hat{\beta}_i^{\text{ols}}) \left(|\hat{\beta}_i^{\text{ols}}| - \frac{\lambda}{d_i} \right)_+ . \quad (10)$$

When $i \notin \mathcal{A}$, $\beta_i^* = 0$, the Gaussian vector $(\hat{\beta}_i^{\text{ols}})_{i \notin \mathcal{A}}$ has the same distribution as $(Z_i^{\text{ols}})_{i \notin \mathcal{A}}$. Therefore, we obtain that for the adaptive lasso,

$$\begin{aligned}
\mathbb{P}\left(\hat{\mathcal{A}}^{\text{pen}}(\lambda) \subset \mathcal{A}\right) &= \mathbb{P}\left(\forall i \notin \mathcal{A}, \hat{\beta}_i^{\text{pen}}(\lambda) = 0\right) \\
&= \mathbb{P}\left(\forall i \notin \mathcal{A}, |\hat{\beta}_i^{\text{ols}}| - \frac{\lambda_0^{\text{adapt}}}{d_i |\hat{\beta}_i^{\text{ols}}|} \leqslant 0\right), \\
&= \mathbb{P}\left(\forall i \notin \mathcal{A}, (Z_i^{\text{ols}})^2 \times d_i \leqslant \lambda_0^{\text{adapt}}\right), \\
&\geqslant \mathbb{P}\left(\forall i \in [\![1, p]\!], (Z_i^{\text{ols}})^2 \times d_i \leqslant \lambda_0^{\text{adapt}}\right), \\
&\geqslant \mathbb{P}\left(T^{\text{adapt}} \leqslant \lambda_0^{\text{adapt}}\right) = 1 - \alpha.
\end{aligned}$$

Using the same arguments, the same result holds for the usual lasso. □

Proof (Proposition 2) As in the proof of proposition 1, we provide only the proof for the adaptive lasso because the same arguments can be used for the lasso. From the form of the adaptive lasso solution given in (10), we derive

$$\hat{\beta}_i^{\text{adapt}}(\lambda_0^{\text{adapt}}) \neq 0 \Leftrightarrow d_i \times (\hat{\beta}_i^{\text{ols}})^2 > \lambda_0^{\text{adapt}}.$$

If we set $Z^{\text{ols}} = \hat{\beta}^{\text{ols}} - \beta^*$, we obtain

$$\begin{aligned}
\hat{\beta}_i^{\text{adapt}}(\lambda_0^{\text{adapt}}) \neq 0 &\Leftrightarrow d_i \times (Z_i^{\text{ols}} + \beta_i^*)^2 > \lambda_0^{\text{adapt}}, \\
&\Leftrightarrow \left(\beta_i^* + Z_i^{\text{ols}} - \sqrt{\lambda_0^{\text{adapt}}/d_i}\right)\left(\beta_i^* + Z_i^{\text{ols}} + \sqrt{\lambda_0^{\text{adapt}}/d_i}\right) > 0.
\end{aligned}$$

Because $\beta_i^* + Z_i^{\text{ols}} - \sqrt{\lambda_0^{\text{adapt}}/d_i} \leqslant \beta_i^* + Z_i^{\text{ols}} + \sqrt{\lambda_0^{\text{adapt}}/d_i}$, one deduces that $\beta_i^* \geqslant \sqrt{\lambda_0^{\text{adapt}}/d_i} - Z_i^{\text{ols}}$ is a sufficient condition for $\hat{\beta}_i^{\text{adapt}}(\lambda_0^{\text{adapt}}) \neq 0$. Furthermore, because c_i^{adapt} is defined as the $1 - \eta/p$ quantile of the distribution of $\sqrt{\lambda_0^{\text{adapt}}/d_i} - Z_i^{\text{ols}}$, if $|\beta_i^*| \geqslant c_i^{\text{adapt}}$, we have

$$\mathbb{P}(\hat{\beta}_i^{\text{adapt}}(\lambda_0^{\text{adapt}}) \neq 0) = \mathbb{P}(i \in \hat{\mathcal{A}}^{\text{adapt}}(\lambda_0^{\text{adapt}})) \geqslant 1 - \eta/p.$$

Finally, a Bonferroni procedure leads to

$$\mathbb{P}(\forall i \in \mathcal{A} \setminus \mathcal{E}^{\text{adapt}}, i \in \hat{\mathcal{A}}^{\text{adapt}}(\lambda_0^{\text{adapt}})) = \mathbb{P}(\mathcal{A} \setminus \mathcal{E}^{\text{adapt}} \subset \hat{\mathcal{A}}^{\text{adapt}}(\lambda_0^{\text{adapt}})) \geqslant 1 - \eta,$$

which is the announced result. □

Proof (Lemma 1) Let us assume that the linear transformation U in (7) is equal to P_δ. Thus, we have

$$\tilde{Y} = \tilde{X}\beta^* + \tilde{\varepsilon}, \text{ with } \tilde{Y} = P_\delta Y, \tilde{X} = P_\delta X \text{ and } \tilde{\varepsilon} = P_\delta \varepsilon.$$

Because the variance of $\tilde{\varepsilon}$ is equal to $P_\delta \Gamma P_\delta^T$, the maximum likelihood estimator of the model (1) is

$$\begin{aligned}
\hat{\beta}^{\text{mle}} &= (\tilde{X}^T (P_\delta \Gamma P_\delta^T)^{-1} \tilde{X})^{-1} \tilde{X} (P_\delta \Gamma P_\delta^T)^{-1} \tilde{Y} \\
&= ((P_\delta X)^T (P_\delta \Gamma P_\delta^T)^{-1} P_\delta X)^{-1} (P_\delta X)^T (P_\delta \Gamma P_\delta^T)^{-1} P_\delta Y, \\
&= \begin{pmatrix} M_1 & M_2 \end{pmatrix} P_\delta Y.
\end{aligned}$$

Furthermore, the ordinary least squares estimator $\hat{\beta}^{\text{ols}}(V_\delta)$ is

$$\hat{\beta}^{\text{ols}}(V_\delta) = ((V_\delta X)^T (V_\delta X))^{-1} (V_\delta X)^T V_\delta Y.$$

Let us now check that $\hat{\beta}^{\text{mle}} = \hat{\beta}^{\text{ols}}(V_\delta) = \begin{pmatrix} \Delta^{-1} & M_2 \end{pmatrix} P_\delta Y$. If we denote by $(e_1, \ldots, e_n)$ and $(b_1, \ldots, b_p)$ the canonical basis of $\mathbb{R}^n$ and $\mathbb{R}^p$, we have

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

and, as a result, $\left(\Delta^{-1} \quad M_2\right) P_\delta Y = \hat{\beta}^{\text{mle}}$. It is straightforward to show that $V_\delta X = \left(\Delta \quad 0\right)^T$, thus $V_\delta \in G$. Moreover, we have

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

which gives the result □

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

$$\delta \in]0, +\infty[^p \mapsto \phi(c^{\text{pen}}(V_\delta))$$

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

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

1. The function $\delta \in]0, +\infty[^p \mapsto \lambda_0^{\text{pen}}(\delta)$ is homogeneous:

$$\forall k > 0, \forall \delta \in]0, +\infty[^p, \lambda_0^{\text{pen}}(k\delta) = k\lambda_0^{\text{pen}}(\delta).$$

2. The function $\delta \in]0, +\infty[^p \mapsto \lambda_0^{\text{pen}}(\delta)$ is componentwise-increasing:

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

The following lemma provides the continuity of the function $\delta \in]0, +\infty[^p \mapsto \lambda_0^{\text{pen}}(\delta)$.

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

$r \geqslant 0$ such that $x - ru \in]0, +\infty[^p$. The function g is homogeneous; thus,

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

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

$$\forall y \in [x_1 - ru_1, x_1 + ru_1] \times \cdots \times [x_p - ru_p, x_p + ru_p], |g(y) - g(x)| \leqslant \frac{r}{\|x\|} |g(x)|. \qquad (11)$$

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 (11) 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 *Let ϕ be a L^q norm on $\mathbb{R}^p$ with $q > 0$; the function*

$$f : \delta \in]0, +\infty[^p \mapsto \phi(c^{\text{pen}}(V_\delta))$$

attains its minimum for at least one element δ^*.

Proof Let us recall the expression of the function f

$$f : \delta \in]0, +\infty[^p \mapsto \phi\left(\frac{\lambda_0^{\text{pen}}(\delta)}{\delta_1} + se(\hat{\beta}_i^{\text{mle}}) z_{1-\eta/p}, \ldots, \frac{\lambda_0^{\text{pen}}(\delta)}{\delta_p} + se(\hat{\beta}_i^{\text{mle}}) z_{1-\eta/p}\right).$$

Because f is homogeneous, one deduces 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, $\bar{f}$ attains 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

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

hence, the result follows. □

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

Lemma C (Anderson) *Let $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ be centred Gaussian vectors with variance matrices Γ_X and Γ_Y, respectively. Assume that the matrix $\Gamma_Y - \Gamma_X$ is a positive semidefinite matrix; then,*

$$\forall x \geqslant 0, \mathbb{P}(\max\{|Y_1|^\gamma, \ldots, |Y_n|^\gamma\} \geqslant x) \geqslant \mathbb{P}(\max\{|X_1|^\gamma, \ldots, |X_n|^\gamma\} \geqslant x), \text{ with } \gamma \in \{1, 2\}.$$

This inequality implies that $\max\{|Y_1|^\gamma, \ldots, |Y_n|^\gamma\}$ is stochastically greater than $\max\{|X_1|^\gamma, \ldots, |X_n|^\gamma\}$.

Proof (Lemma 2) For any $U \in G$, the matrix $(UX)^T UX$ is diagonal and $(UX)^T UX = \Delta = \text{diag}(\delta_1, \ldots, \delta_p) = \text{diag}(\delta)$. The difference between the variance 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 $(\delta_1 Z_1^{\text{mle}}, \ldots, \delta_p Z_p^{\text{mle}}) = \Delta Z^{\text{mle}}$ is semidefinite positive. Indeed, we obtain that

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

The last inequality is a consequence of the Gauss-Markov theorem (Rencher and Schaalje, 2008) (page 146). Because $\lambda_0^{\text{pen}}(U)$ and $\lambda_0^{\text{pen}}(V_\delta)$ 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 $\lambda_0^{\text{pen}}(U) \geqslant$

$\lambda_0^{\text{pen}}(V_\delta)$. Furthermore, the inequalities $\forall i \in [\![1,p]\!], \text{se}(\hat{\beta}_{\text{i}}^{\text{mle}}) \leqslant \text{se}(\hat{\beta}_{\text{i}}^{\text{ols}}(\text{U}))$ lead to

$$\forall i \in [\![1,p]\!], c_i^{\text{pen}}(V_{\delta^*}) = \frac{\lambda_0^{\text{pen}}(V_\delta)}{\delta_i} + se(\hat{\beta}_i^{\text{mle}}) z_{1-\eta/p} \leqslant c_i^{\text{pen}}(U) = \frac{\lambda_0^{\text{pen}}(U)}{\delta_i} + se(\hat{\beta}_i^{\text{ols}}(U)) z_{1-\eta/p}.$$

Because $c^{\text{pen}}(V_\delta)$ is componentwise smaller than $c^{\text{pen}}(U)$, one deduces that $\phi(c^{\text{pen}}(U)) \geqslant \phi(c^{\text{pen}}(V_\delta))$. Finally, using lemma B, the inequality $\phi(c^{\text{pen}}(V_\delta)) \geqslant \phi(c^{\text{pen}}(V_{\delta^*}))$ gives the result. □

Proof (Proposition 3) The proposition can be shown by proving that

$$\forall \delta \in]0,+\infty[^p, \|c^{\text{pen}}(V_{\delta^*})\|_\infty \leqslant \|c^{\text{pen}}(V_\delta)\|_\infty \Leftrightarrow c_1^{\text{pen}}(V_{\delta^*}) = \cdots = c_p^{\text{pen}}(V_{\delta^*}).$$

Here, we denote $c^{\text{pen}}(V_\delta) = c^{\text{pen}}(\delta) = c^{\text{pen}}(\delta_1, \ldots, \delta_p)$. Assume that $\delta^* \in]0,+\infty[^p$ is such that $c_1^{\text{pen}}(\delta^*) = \cdots = c_p^{\text{pen}}(\delta^*)$. We will prove that $\|c^{\text{pen}}(\delta^*)\|_\infty \leqslant \|c^{\text{pen}}(\delta)\|_\infty$ for any $\delta \neq \delta^*$. For this purpose, we denote

$$k = \max_{i \in [\![1,p]\!]} \{\delta_i^*/\delta_i\}.$$

There exists $i_0 \in [\![1,p]\!]$ such that $k = \delta_{i_0}^*/\delta_{i_0}$. We have

$$\begin{aligned} c_{i_0}^{\text{pen}}(\delta_1, \ldots, \delta_p) &= \sqrt{\frac{\lambda_0^{\text{pen}}(\delta_1, \ldots, \delta_p)}{\delta_{i_0}}} + se(\hat{\beta}_{i_0}^{\text{mle}}) z_{1-\eta/p}, \\ &= \sqrt{\frac{\lambda_0^{\text{pen}}(k\delta_1, \ldots, k\delta_p)}{k\delta_{i_0}}} + se(\hat{\beta}_{i_0}^{\text{mle}}) z_{1-\eta/p}. \end{aligned}$$

Because $(k\delta_1, \ldots, k\delta_p)$ is componentwise-greater than $(\delta_1^*, \ldots, \delta_p^*)$, from the componentwise-increasing property of λ_0^{pen}, one deduces that

$$\begin{aligned} c_{i_0}^{\text{pen}}(\delta_1, \ldots, \delta_p) &\geqslant \sqrt{\frac{\lambda_0^{\text{pen}}(\delta_1^*, \ldots, \delta_p^*)}{\delta_{i_0}^*}} + se(\hat{\beta}_{i_0}^{\text{mle}}) z_{1-\eta/p}, \\ &\geqslant c_{i_0}^{\text{pen}}(\delta_1^*, \ldots, \delta_p^*) = \|c^{\text{pen}}(\delta_1^*, \ldots, \delta_p^*)\|_\infty. \end{aligned}$$

This shows that $\|c^{\text{pen}}(\delta_1, \ldots, \delta_p)\|_\infty \geqslant \|c^{\text{pen}}(\delta_1^*, \ldots, \delta_p^*)\|_\infty$.

Conversely, assume that $\delta^* \in]0,+\infty[^p$ such that $\|c^{\text{pen}}(\delta^*)\|_\infty$ is minimal, and assume that

the inequality $c_1^{\text{pen}}(\delta^*) = \cdots = c_p^{\text{pen}}(\delta^*)$ does not hold. We set

$$I_0 = \{i \in [\![1, p]\!] \mid c_i(\delta^*) = \|c(\delta^*)\|_\infty\}.$$

Because $I_0 \subsetneq [\![1, p]\!]$, one can choose $\epsilon > 0$ such that

$$\forall i \notin I_0, (1+\epsilon)c_i(\delta^*) < \|c(\delta^*)\|_\infty.$$

We define δ^0 as the parameter

$$\delta^0 := \begin{cases} \delta_i^0 = (1+\epsilon)\delta_i^* \text{ if } i \in I_0, \\ \delta_i^0 = \delta_i^* \text{ if } i \notin I_0. \end{cases}$$

If $i \in I_0$, we have

$$c_i^{\text{pen}}(\delta^0) = \sqrt{\frac{\lambda_0^{\text{pen}}(\delta^0)}{\delta_i^0}} + se(\hat{\beta}_{i_0}^{\text{mle}}) z_{1-\eta/p}.$$

Because λ_0^{pen} is strictly componentwise-increasing, one deduces that

$$\begin{aligned} c_i^{\text{pen}}(\delta^0) &< \sqrt{\frac{\lambda_0^{\text{pen}}((1+\epsilon)\delta^*)}{(1+\epsilon)\delta_i^*}} + se(\hat{\beta}_{i_0}^{\text{mle}}) z_{1-\eta/p}, \\ &< c_i^{\text{pen}}(\delta^*) = \|c^{\text{pen}}(\delta^*)\|_\infty. \end{aligned}$$

If $i \notin I_0$, we have

$$\begin{aligned} c_i^{\text{pen}}(\delta^0) &= \sqrt{\frac{\lambda_0^{\text{pen}}(\delta^0)}{\delta_i^0}} + se(\hat{\beta}_{i_0}^{\text{mle}}) z_{1-\eta/p}, \\ &= \sqrt{\frac{\lambda_0^{\text{pen}}(\delta^0)}{\delta_i^*}} + se(\hat{\beta}_{i_0}^{\text{mle}}) z_{1-\eta/p}, \\ &< \sqrt{\frac{\lambda_0^{\text{pen}}((1+\epsilon)\delta^*)}{\delta_i^*}} + se(\hat{\beta}_{i_0}^{\text{mle}}) z_{1-\eta/p}. \end{aligned}$$

Because the function λ_0 is homogeneous, we obtain

$$\begin{aligned} c_i^{\text{pen}}(\delta^0) &< \sqrt{1+\epsilon}\sqrt{\frac{\lambda_0^{\text{pen}}(\delta^*)}{\delta_i^*}} + se(\hat{\beta}_{i_0}^{\text{mle}})z_{1-\eta/p}, \\ &< (1+\epsilon)\left(\sqrt{\frac{\lambda_0^{\text{pen}}(\delta^*)}{\delta_i^*}} + se(\hat{\beta}_{i_0}^{\text{mle}})z_{1-\eta/p}\right), \\ &< (1+\epsilon)c_i(\delta^*) < \|c(\delta^*)\|_\infty. \end{aligned}$$

Therefore, $\|c^{\text{pen}}(\delta^0)\|_\infty < \|c^{\text{pen}}(\delta^*)\|_\infty$, which results in a contradiction. □

Proof (Proposition 4) Let $\delta \in]0,+\infty[^p$, and let us define $\zeta = (\zeta_1,\ldots,\zeta_p) = (\delta_1^2,\ldots,\delta_p^2)$. We have

$$\begin{aligned} \{E = \hat{\mathcal{A}}^{\text{adapt}}(\lambda_0^{\text{adapt}}(V_\zeta))\} &= \left\{\bigcap_{i\in E}\{i \in \hat{\mathcal{A}}^{\text{adapt}}(\lambda_0^{\text{adapt}}(V_\zeta))\}\right\}\bigcap\left\{\bigcap_{i\notin E}\{i \notin \hat{\mathcal{A}}^{\text{adapt}}(\lambda_0^{\text{adapt}}(V_\zeta))\}\right\}, \\ &= \left\{\bigcap_{i\in E}\{\hat{\beta}_i^{\text{adapt}}(\lambda_0^{\text{adapt}}(V_\zeta)) \neq 0\}\right\}\bigcap\left\{\bigcap_{i\notin E}\{\hat{\beta}_i^{\text{adapt}}(\lambda_0^{\text{adapt}}(V_\zeta)) = 0\}\right\}. \end{aligned}$$

The closed form of the adaptive lasso estimator given in (10) gives that

$$\begin{aligned} \{E = \hat{\mathcal{A}}^{\text{adapt}}(\lambda_0^{\text{adapt}}(V_\zeta))\} &= \left\{\bigcap_{i\in E}\{\zeta_i(\hat{\beta}_i^{\text{mle}})^2 > \lambda_0^{\text{adapt}}(V_\zeta)\}\right\}\bigcap\left\{\bigcap_{i\notin E}\{\zeta_i(\hat{\beta}_i^{\text{mle}})^2 \leqslant \lambda_0^{\text{adapt}}(V_\zeta)\}\right\}, \\ &= \left\{\bigcap_{i\in E}\{\delta_i|\hat{\beta}_i^{\text{mle}}| > \sqrt{\lambda_0^{\text{adapt}}(V_\zeta)}\}\right\}\bigcap\left\{\bigcap_{i\notin E}\{\delta_i|\hat{\beta}_i^{\text{mle}}| \leqslant \sqrt{\lambda_0^{\text{adapt}}(V_\zeta)}\}\right\}. \end{aligned}$$

Furthermore, we have

$$\{E = \hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}(V_\delta))\} = \left\{\bigcap_{i\in E}\left\{\delta_i|\hat{\beta}_i^{\text{mle}}| > \lambda_0^{\text{lasso}}(V_\delta)\right\}\right\} \cap \left\{\bigcap_{i\notin E}\left\{\delta_i|\hat{\beta}_i^{\text{mle}}| \leqslant \lambda_0^{\text{lasso}}(V_\delta)\right\}\right\}.$$

Thus, the distributions of $\hat{\mathcal{A}}^{\text{adapt}}(\lambda_0^{\text{adapt}}(V_\zeta))$ and $\hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}(V_\delta))$ are equal if and only if $\lambda_0^{\text{lasso}}(V_\delta) = \sqrt{\lambda_0^{\text{adapt}}(V_\zeta)}$. We have that

$$\begin{aligned} 1-\alpha &= \mathbb{P}(\max\{\zeta_1(Z_1^{\text{mle}})^2,\ldots,\zeta_p(Z_p^{\text{mle}})^2\} \leqslant \lambda_0^{\text{adapt}}), \\ &= \mathbb{P}((\max\{\delta_1|Z_1^{\text{mle}}|,\ldots,\delta_p|Z_p^{\text{mle}}|\})^2 \leqslant \lambda_0^{\text{adapt}}), \\ &= \mathbb{P}\left(\max\{\delta_1|Z_1^{\text{mle}}|,\ldots,\delta_p|Z_p^{\text{mle}}|\} \leqslant \sqrt{\lambda_0^{\text{adapt}}}\right). \end{aligned}$$

Because $\mathbb{P}\left(\max\{\delta_1|Z_1^{\rm mle}|,\ldots,\delta_p|Z_p^{\rm mle}|\}\leqslant\lambda_0^{\rm lasso}\right)=1-\alpha$, one deduces that $\lambda_0^{\rm lasso}=\sqrt{\lambda_0^{\rm adapt}}$. This leads to $c^{\rm lasso}(V_\delta)=c^{\rm adapt}(V_\zeta)$. Thus, if δ^* is an element for which $\phi(c^{\rm lasso}(V_{\delta^*}))$ is minimal, then $\phi(c^{\rm adapt}(V_{\zeta^*}))$ is also minimal. □

Proof (Proposition 5) In this proof, the dependencies in α and n are provided in order to facilitate understanding. So, instead of $\hat{\beta}^{\rm mle}$, $\lambda_0^{\rm lasso}(U)$ and $\hat{\mathcal{A}}^{\rm lasso}(\lambda_0^{\rm lasso}(U))$ we denote respectively $\hat{\beta}^n,\lambda_{1-\alpha}(U)$ and $\hat{\mathcal{A}}^n(\lambda_{1-\alpha}(U))$. We have $\mathbb{P}(\hat{\mathcal{A}}^n(\lambda_{1-\alpha_n}(U_\phi))\subset\mathcal{A})\geqslant 1-\alpha_n$. Since α_n converges toward 0, one deduces that $\mathbb{P}(\hat{\mathcal{A}}^n(\lambda_{1-\alpha_n}(U_\phi))\subset\mathcal{A})\underset{n\to+\infty}{\longrightarrow}1$. It remains to check that $\mathbb{P}(\mathcal{A}\subset\hat{\mathcal{A}}^n(\lambda_{1-\alpha_n}(U_\phi)))\underset{n\to+\infty}{\longrightarrow}1$. We have the following equalities

$$\begin{aligned}\mathbb{P}(\mathcal{A}\subset\hat{\mathcal{A}}^n(\lambda_{1-\alpha_n}(U_\phi))) &= \mathbb{P}(\forall i\in\mathcal{A},\hat{\beta}_i^{\rm lasso}(\lambda_{1-\alpha_n}(U_\phi))\neq 0),\\ &= \mathbb{P}\left(\forall i\in\mathcal{A},sgn(\hat{\beta}_i^n)\left(|\hat{\beta}_i^n|-\lambda_{1-\alpha_n}(V_{\delta^*})/\delta_i^*\right)_+\neq 0\right),\\ &= \mathbb{P}\left(\forall i\in\mathcal{A},|\hat{\beta}_i^n|\geqslant\lambda_{1-\alpha_n}(V_{\delta^*})/\delta_i^*\right).\end{aligned}$$

For all $i\in\mathcal{A}$, we have that $|\hat{\beta}_i^n|\underset{n\to+\infty}{\longrightarrow}|\beta_i^*|>0$. Thus, to prove that $\mathbb{P}\left(\forall i\in\mathcal{A},|\hat{\beta}_i^n|\geqslant\lambda_{1-\alpha_n}(V_{\delta^*})/\delta_i^*\right)\underset{n\to+\infty}{\longrightarrow}1$, it suffices to prove

$$\left(\frac{\lambda_{1-\alpha_n}(V_{\delta^*})}{\delta_1^*},\ldots,\frac{\lambda_{1-\alpha_n}(V_{\delta^*})}{\delta_p^*}\right)\underset{n\to+\infty}{\longrightarrow}0_{\mathbb{R}^p}\Leftrightarrow\phi\left(\frac{\lambda_{1-\alpha_n}(V_{\delta^*})}{\delta_1^*},\ldots,\frac{\lambda_{1-\alpha_n}(V_{\delta^*})}{\delta_p^*}\right)\underset{n\to+\infty}{\longrightarrow}0. \quad (12)$$

Let us set $\delta=(\delta_1,\ldots,\delta_p)=(1/se(\hat{\beta}_1^n),\ldots,1/se(\hat{\beta}_p^n))$, the optimality of the linear transformation $U_\phi=V_{\delta^*}$ gives

$$\delta\in]0,+\infty[^p,\phi\left(\frac{\lambda_{1-\alpha_n}(V_{\delta^*})}{\delta_1^*},\ldots,\frac{\lambda_{1-\alpha_n}(V_{\delta^*})}{\delta_p^*}\right)\leqslant\phi\left(\frac{\lambda_{1-\alpha_n}(V_\delta)}{\delta_1},\ldots,\frac{\lambda_{1-\alpha_n}(V_\delta)}{\delta_p}\right).$$

Thus, the limit given in (12) holds if

$$\lim_{n\to+\infty}\phi\left(\frac{\lambda_{1-\alpha_n}(V_\delta)}{\delta_1},\ldots,\frac{\lambda_{1-\alpha_n}(V_\delta)}{\delta_p}\right)=0. \quad (13)$$

The tuning parameter $\lambda_{1-\alpha_n}(V_\delta)$ is the $1-\alpha_n$ quantile of $\max\{|Z_1^{\rm mle}|/se(\hat{\beta}_1^{\rm mle}),\ldots,|Z_p^{\rm mle}|/se(\hat{\beta}_p^{\rm mle})\}$. One notices that $|Z_i^{\rm mle}|/se(\hat{\beta}_i^{\rm mle})$ has a folded standard Gaussian distribution. So, classical inequality yields to $\lambda_{1-\alpha_n}(V_\delta)\leqslant\sqrt{2\ln(p/\alpha_n)}$. If

we set $\alpha_n = 1/n$, one obtains

$$\lambda_{1-\alpha_n}(V_\delta)/\delta_i \leqslant se(\hat{\beta}_i^n) \times \sqrt{2\ln(pn)}.$$

Since $\hat{\beta}^n$ has a rate of convergence of $1/\sqrt{n}$, one deduces that the limits given in (12) and (13) hold together. Therefore, $\lim_{n\to+\infty} \mathbb{P}(\mathcal{A} \subset \hat{\mathcal{A}}^n(\lambda_{1-\alpha_n}(V_\phi))) = 1$ that completes the proof. □

Proof (Proposition 6) Let us recall that the Gaussian vector $(\hat{\beta}_i^{\text{mle}})_{i\notin\mathcal{A}}$ has the same distribution as $(Z_i^{\text{mle}})_{i\notin\mathcal{A}}$ defined in (8). Thus, we have

$$\begin{aligned}
\mathbb{P}(\hat{\mathcal{A}}^{\text{mle}} \subset \mathcal{A}) &= \mathbb{P}(\forall i \notin \mathcal{A}, |\hat{\beta}_i^{\text{mle}}| \leqslant \sigma\sqrt{q_{\sqrt[p]{1-\alpha}}}), \\
&= \mathbb{P}(\forall i \notin \mathcal{A}, |Z_i^{\text{mle}}| \leqslant \sigma\sqrt{q_{\sqrt[p]{1-\alpha}}}), \\
&\geq \mathbb{P}(\forall i \in [\![1,p]\!], |Z_i^{\text{mle}}| \leqslant \sigma\sqrt{q_{\sqrt[p]{1-\alpha}}}).
\end{aligned}$$

Because $\sigma\sqrt{q_{\sqrt[p]{1-\alpha}}}$ is the $1-\alpha$ quantile of $\max\{|Z_1^{\text{mle}}|,\ldots,|Z_p^{\text{mle}}|\}$, one deduces that $\mathbb{P}(\hat{\mathcal{A}}^{\text{mle}} \subset \mathcal{A}) \geq 1-\alpha$.

The norm $\|c^{\text{lasso}}(V_{\delta^*})\|_\infty$ reaches a minimum for a linear transformation V_{δ^*} for which $\delta^* = (1,\ldots,1)$. Thus, the tuning parameter $\lambda_0^{\text{lasso}}(V_{\delta^*})$ is the $1-\alpha$ quantile of $\{|Z_1^{\text{mle}}|,\ldots,|Z_p^{\text{mle}}|\}$, and $\lambda_0^{\text{lasso}}(V_\delta^*) = \sigma\sqrt{q_{\sqrt[p]{1-\alpha}}}$.

For all $E \subset [\![1,p]\!]$, we have

$$\begin{aligned}
\{E = \hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}(V_{\delta^*}))\} &= \left\{\bigcap_{i\in E}\{i \in \hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}(V_{\delta^*}))\}\right\} \bigcap \left\{\bigcap_{i\notin E}\{i \notin \hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}(V_{\delta^*}))\}\right\}, \\
&= \left\{\bigcap_{i\in E}\{\hat{\beta}_i^{\text{lasso}}(\lambda_0^{\text{lasso}}(V_{\delta^*})) \neq 0\}\right\} \bigcap \left\{\bigcap_{i\notin E}\{\hat{\beta}_i^{\text{lasso}}(\lambda_0^{\text{lasso}}(V_{\delta^*})) = 0\}\right\}.
\end{aligned}$$

The closed form of the lasso estimator given by (10) allows us to write

$$\begin{aligned}
\{E = \hat{\mathcal{A}}^{\text{lasso}}(\lambda_0^{\text{lasso}}(V_{\delta^*}))\} &= \left\{\bigcap_{i\in E}\{|\hat{\beta}_i^{\text{mle}}| > \sigma\sqrt{q_{\sqrt[p]{1-\alpha}}}\}\right\} \bigcap \left\{\bigcap_{i\notin E}\{|\hat{\beta}_i^{\text{mle}}| \leqslant \sigma\sqrt{q_{\sqrt[p]{1-\alpha}}}\}\right\}, \\
&= \{E = \hat{\mathcal{A}}^{\text{mle}}\}
\end{aligned}$$

which proves the proposition. □

Proof (Proposition 7) Let us define δ as follow

$$\delta = (\delta_1, \ldots, \delta_p) = \left(\frac{1}{se(\hat{\beta}_1^{mle})}, \ldots, \frac{1}{se(\hat{\beta}_p^{mle})} \right).$$

We are going to prove that

$$\forall i \in [\![1, p]\!], c_i^{\text{lasso}}(V_\delta) \leqslant c_i^{\text{mle}}.$$

The previous inequality implies that $\phi(c^{\text{lasso}}(V_\delta)) \leqslant \phi(c^{\text{mle}})$.

Let $k_{1-\alpha}$ be the $1-\alpha$ quantile of the random variable $\max\{\delta_1|\mathrm{Z}_1^{\text{mle}}|, \ldots, \delta_{\mathrm{p}}|\mathrm{Z}_{\mathrm{p}}^{\text{mle}}|\}$. Since for any integer i, the random variable $\delta_i Z_i^{\text{mle}}$ has a standard Gaussian distribution, the Bonferroni inequality gives $k_{1-\alpha} \leqslant z_{1-\alpha/2p}$. Therefore, we have

$$c_i^{lasso}(V_\delta) = se(\hat{\beta}_i^{\text{mle}})(k_{1-\alpha} + z_{1-\eta/p}) \leqslant se(\hat{\beta}_i^{\text{mle}})(z_{1-\alpha/2p} + z_{1-\eta/p}) = c_i^{\text{mle}}.$$

Since $\phi(c^{\text{lasso}}(U_\phi))$ is smaller than $\phi(c^{\text{lasso}}(V_\delta))$, one deduces that $\phi(c^{\text{lasso}}(U_\phi)) \leqslant \phi(c^{\text{mle}})$. □

The previous proof shows that the difference between $\phi(c^{\text{lasso}}(U_\phi))$ and $\phi(c^{\text{mle}})$ increases when the difference between $k_{1-\alpha}$ and $z_{1-\alpha/2p}$ becomes large. Furthermore, one can notice that when the components of $\hat{\beta}^{\text{mle}}$ are correlated, the distribution of $\max\{\delta_1|\mathrm{Z}_1^{\text{mle}}|, \ldots, \delta_{\mathrm{p}}|\mathrm{Z}_{\mathrm{p}}^{\text{mle}}|\}$ is approximately equals to that of a folded standard Gaussian. Thus, in this case $k_{1-\alpha} \approx z_{1-\alpha/2}$ that is much smaller than $z_{1-\alpha/2p}$ especially when p is large. One deduces that $\hat{\mathcal{A}}^{\text{lasso}}$ is much better than $\hat{\mathcal{A}}^{\text{mle}}$ in this particular case.

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 Ministry of Ecology, Sustainable Development and Energy in the national research program Risk'OGM. We also received a grant for the project from the IDEX of Toulouse " Transversalité 2014 "

References

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

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

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

Chand, S. (2012). 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), pp. 120–129.

Dudoit, S. and M. J. Van Der Laan (2007). Multiple testing procedures with applications to genomics. Springer.

Efron, B., T. Hastie, I. Johnstone, R. Tibshirani, et al. (2004). Least angle regression. The Annals of statistics 32(2), 407–499.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. Journal of statistical software 33(1), 1–22.

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

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

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

Hui, F. K. C., D. I. Warton, and S. D. Foster (2015). Tuning parameter selection for the adaptive lasso using eric. Journal of the American Statistical Association 110(509), 262–269.

Lehmann, E. L. and J. P. Romano (2005). Testing statistical hypotheses (Third ed.). Springer Texts in Statistics. New York: Springer.

Leng, C., Y. Lin, and G. Wahba (2006). A note on the lasso and related procedures in model selection. Statistica Sinica 16(4), 1273–1284.

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

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

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

Mercier, P., M. J. Lewis, D. Chang, D. Baker, and D. S. Wishart (2011). Towards automatic metabolomic profiling of high-resolution one-dimensional proton nmr spectra. Journal of biomolecular NMR 49(3-4), 307–323.

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

Sun, W., J. Wang, and Y. Fang (2013). Consistent Selection of Tuning Parameters via Variable Selection Stability. Journal of Machine Learning Research 14, 3419–3440.

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

Tulpan, D., S. Léger, L. Belliveau, A. Culf, and M. Čuperlović-Culf (2011). Metabohunter: an automatic approach for identification of metabolites from 1h-nmr spectra of complex mixtures. BMC bioinformatics 12(1), 400.

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

Zheng, C., S. Zhang, S. Ragg, D. Raftery, and O. Vitek (2011). Identification and quantification of metabolites in 1h nmr spectra by bayesian model selection. Bioinformatics 27(12), 1637–1644.

Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association 101(476), 1418–1429.