Multiple Hypotheses Testing For Variable Selection

F. Rohart*

UMR 5219, Institut de Mathématiques de Toulouse, INSA de Toulouse, 135 Avenue de Rangueil, 31077 Toulouse cedex 4, France

UMR 444 Laboratoire de Génétique Cellulaire, INRA Toulouse, 31320 Castanet Tolosan cedex, France

Abstract

Many methods have been developed to estimate the set of relevant variables in a sparse linear model $Y = X\beta + \epsilon$ where the dimension p of β can be much higher than the length n of Y. Here we propose two new methods based on multiple hypotheses testing, either for ordered or non-ordered variables. Our procedures are inspired by the testing procedure proposed by Baraud et al. [2]. The new procedures are proved to be powerful under some conditions on the data and their properties are non asymptotic. They gave better results in estimating the set of relevant variables than both the False Discovery Rate (FDR) and the Lasso, both in the common case ($p < n$) and in the high-dimensional case ($p \geq n$).

Keywords: model selection, FDR, Lasso, Bolasso, multiple hypotheses testing, high-dimension

1. Introduction

Recent technologies have provided scientists with a new kind of data; very high-dimensional data, especially with high-throughput DNA/RNA chips in biology. Unravelling the relevant variables -genes for example- underlying an observation is a well known problem in statistics and is still one of the current major challenges. Indeed, with a large number of variables there is often a desire to select a smaller subset that not only fits as well as the full set of variables, but also contains the more important ones. Discovering the relevant variables leads to higher prediction accuracy, an important criterion in variable selection.

Many methods have been developed to estimate the set of relevant variables in the linear model $Y = X\beta + \epsilon$ where the dimension p of β can be much higher than the length n of Y. In particular, a lot of model selection methods have been developed based on a penalized criterion. The mostly known is probably the Lasso that had been presented by Tibshirani [14]; l^1 penalization of the least squares estimate which shrinks to zero some irrelevant coefficients, hence an estimation of the set of relevant variables. A lot of studies have been conducted on the Lasso and many results are available; e.g. consistency of the Lasso in high-dimensional linear regression [16], sparsity oracle inequalities [5] and variable selection in high-dimensional graphs with the Lasso [12]. The Lasso has several variants such as an adaptative Lasso [9], a bootstrap Lasso [1] or a

*Corresponding author
Email address: florian.rohart@gmail.com (F. Rohart)

Group Lasso [8]. A l^1 penalization has also been used in the Sparse-PLS, which induces a limited number of variables in each PLS direction; see Tenenhaus [13] for an introduction on PLS, and Lê Cao et al. [11] for further details on Sparse-PLS. Other kinds of penalization have also been used, such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), two methods based on a likelihood penalization on the number of variables included in the model. Despite that the major portion of model selection methods was developed to perform in low dimension, some of them apply in the high-dimensional case. There is still some others that were actually developed to be powerful when p is higher than n, such as the Dantzig selector [7]. Nonetheless a recent paper shows that under a sparsity condition on the linear model, the Dantzig selector and the Lasso exhibit similar behavior [4]. Nevertheless, penalization criterion is not the only way to perform model selection. For instance, the False Discovery Rate (FDR) procedure, developed in the context of multiple hypotheses testing by Benjamini and Hochberg [3], was used in variable selection by Bunea et al. [6]. This procedure has been extended to high-dimensional analysis and is presently used in biology for QTL research and transcriptome analysis; a p-value is calculated for each variable X_i from the regression of Y onto that variable and selection is performed through an adjusted threshold.

Most of the selection methods cited above give quite good results when p is lower than n, but the results get worst as p grows larger than n. In the context of this paper, variable selection when p is much higher than n, those methods are disappointing and unsatisfactory. Moreover, most of theoretical results are only asymptotic and only prove consistency of the estimators. Non asymptotic results are more sought since small samples are usual in practice. Concerning methods using a penalized criterion such as AIC, BIC or any other penalization on the likelihood, another major drawback is of computational nature. Indeed, a search through all the 2^p possible spaces may be needed and this search is as complex as p grows.

This paper deals with the problem of recovering the set of relevant variables in a sparse linear model when p can be lower or far higher than n. We consider the regression model:

$$Y = X\beta + \epsilon \tag{1}$$

where Y is the observation of length n, $X = (X_1, ..., X_p)$ is the matrix of p variables, β is an unknown vector of $\mathbb{R}^p$, ϵ a Gaussian vector with i.i.d. components, $\epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$ where I_n is the identity matrix of $\mathbb{R}^n$, and σ some unknown positive quantity. We set $J = \{j, \beta_j \neq 0\}$ and $|J| = k_0$. We denote $\beta_J = (\beta_j)_{j \in J}$. Let $\mu = E(Y) = X\beta$ and $\mathbb{P}_\mu$ the distribution of Y obeying to model (1).

The aim of this paper is to estimate J, the set of relevant variables in (1). We distinguish two frameworks. On one hand, the variables $X_1, ..., X_p$ are assumed to be ordered, regardless of Y. We define a powerful procedure for estimating J under some conditions on the data, either when $p \leq n$ or when $p > n$. These properties are non asymptotic. This procedure is a multiple hypotheses testing method based on Baraud et al. [2] which consists of doing several tests to decide whether $E(Y)$ is in V, some linear subspace of $\mathbb{R}^n$, or in a suitable collection of subspaces containing V. On the other hand, the variables are not assumed to be ordered. We provide a procedure to estimate J when σ is known and another similar procedure when σ is unknown. The two procedures are proved to be powerful under some conditions on the data. The properties of the procedures are also non asymptotic.

This paper is organized as follow, in Section 2 we present the first procedure to estimate J,

in the ordered variable case; the non-ordered variable selection is considered in Section 3. A simulation study is provided in Section 4 to compare several variable selection methods.

2. Ordered variable selection

2.1. The case $p < n$

First of all, the common case $p < n$ is considered. The family $(X_i)_{1\leq i\leq p}$ is assumed to be a linearly independent family to which an order is given, regardless of Y. In this section we focus on ordered variables, which means that the relevant variables are supposed to be in the first places, i.e. $J = \{1, .., k_0\}$. Hence an estimation of k_0 gives us an estimation of J. This section focuses on the estimation of k_0.

Let k be a positive integer and let V_k denote a linear subspace of $\mathbb{R}^n$, the following results are based on a test of the null hypothesis "$\mu = E(Y)$ belongs to V_k" against the alternative that it does not. As proposed in Baraud et al. [2], we consider a finite collection of linear subspaces of $V_k^{\perp}$, $\{S_{k,t}, t \in \mathcal{T}\}$, to test the null hypothesis. The index set $\mathcal{T}$ is allowed to depend on the number of observations n, or on the number of parameters p.

Let $\{\alpha_t,\ t \in \mathcal{T}\}$ be a suitable collection of numbers in]0,1[. The testing procedure presented in Baraud et al. [2] consists in doing several Fisher tests of level α_t of the null hypothesis:

$$H_k : \quad \{\mu \in V_k\} \text{ against the alternative } \{\mu \in V_k + S_{k,t}\}.$$

The null hypothesis is rejected if at least one of the Fisher tests does. Our procedure consists in doing successively the tests $(H_k)_{k\geq 0}$ for a suitable collection of linear subspaces $(V_k)_{k\geq 0}$ until the null hypothesis is accepted.

Let us introduce some notations that will be used throughout this section. Note $\|s\|_n^2 = \sum_{i=1}^n s_i^2/n$. For each $k \in \mathbb{N}$, $t \in \mathcal{T}$, we set $V_{k,t} = V_k \oplus S_{k,t}$, and denote by $D_{k,t}$ and $N_{k,t}$ the dimension of $S_{k,t}$ and $V_{k,t}^{\perp}$ respectively. Moreover set Π_V the orthogonal projector onto V for all subspace V. $\bar{F}_{D,N}(u)$ denotes the probability for a Fisher with D and N degrees of freedom to be larger than u. We denote $\forall\ (x, y) \in \mathbb{R}^{n^2} < x, y >_n = \sum_{i=1}^n x_i y_i/n$, and $\forall a \in \mathbb{R}$, $\lfloor a \rfloor$ the integer part of a.

For all $i \in \{1, ..., p\}$, X_i is supposed to be normed to 1: $\forall\ i, < X_i, X_i >_n = 1$.

As the family $(X_i)_{1\leq i\leq p}$ is ordered, a natural choice of the collection V_k is the following: set $\forall 1 \leq k \leq p, V_k = \text{span}(X_1, .., X_k)$ and $V_0 = \{0\}$. With this choice of V_k and as $(X_i)_{1\leq i\leq p}$ is a linearly independent family, we have for all $k \geq 0$, $\dim(V_k) = k$.

For $k \in \{0, ..., p-1\}$, let $t_{max}^k = \lfloor log_2(p-k) \rfloor$ and $Q_{k,t_{max}^k} = \{S_{k,t}, t \in \{0, .., t_{max}^k\}\}$ be a collection of linear subspace of $V_k^{\perp}$, where:

$\forall\ t \in \{0, .., t_{max}^k\} = \mathcal{T}_k$,

$$S_{k,t} = \text{span}(X_{k+1}, .., X_{k+2^t}) \cap V_k^{\perp}. \tag{2}$$

With this collection, $D_{k,t} = 2^t$ and $N_{k,t} = n - (k + 2^t)$.

As mentioned before, our procedure consists in doing successively the tests $(H_k)_{k\geq 0}$ until the null hypothesis is accepted; with this choice of the collection of linear subspaces $(V_k)_{0\leq k\leq p}$ and $(Q_{k,t_{max}})_{0\leq k<p}$, an estimation of k_0 with our procedure is $\hat{k} = \inf\{k \geq 0, H_k \text{ is accepted}\}$. The estimated set of relevant variables is then $\hat{J} = \{1, .., \hat{k}\}$.

A procedure to test the null hypothesis H_k is introduced in the following. Set: $\forall \alpha \in]0, 1[, \forall k < p$,

$$T_{k,\alpha} = \sup_{t\in\mathcal{T}_k} \left\{ \frac{N_{k,t}\|\Pi_{S_{k,t}} Y\|_n^2}{D_{k,t}\|Y - \Pi_{V_{k,t}} Y\|_n^2} - \bar{F}^{-1}_{D_{k,t},N_{k,t}}(\alpha_t) \right\}, \tag{3}$$

where $\{\alpha_t,\ t \in \mathcal{T}_k\}$ is a collection of number in]0,1[such that:

$$\forall \mu \in V_k, \qquad \mathbb{P}_\mu(T_{k,\alpha} > 0) \leq \alpha \tag{4}$$

The null hypothesis H_k is rejected when $T_{k,\alpha}$ is positive.
We choose the collection $\{\alpha_t, t \in \mathcal{T}_k\}$ in accordance with one of the two following procedures:

P1. For all $t \in \mathcal{T}_k, \alpha_t = \alpha_n$ where α_n is the α-quantile of the random variable

$$\inf_{t \in \mathcal{T}_k} \bar{F}_{D_{k,t},N_{k,t}} \left\{ \frac{N_{k,t}\|\Pi_{S_{k,t}}\epsilon\|_n^2}{D_{k,t}\|\epsilon - \Pi_{V_{k,t}}\epsilon\|_n^2} \right\},$$

P2. The α_t's satisfy the inequality

$$\sum_{t \in \mathcal{T}} \alpha_t \leq \alpha.$$

Procedure P1 gives a test H_k of size α whereas procedure P2 gives a test H_k of level α. The final multiple testing procedure, which consists in calculating successively $T_{k,\alpha}$ from $k = 0$ until $T_{k,\alpha}$ is negative, is proved to be powerful; an upper bound of the probability to wrongly estimate k_0 is given in the following theorem. For $k = 0, ..., p-1$, for $\gamma \in]0,1[$ and for all $t \in \mathcal{T}_k$, let $L_t = \log(1/\alpha_t), L = \log(2/\gamma), m_t = 2\exp(4L_t/N_{k,t})$, and for $u > 0$ let

$$K_t(u) = 1 + 2\sqrt{\frac{u}{N_{k,t}}} + 2m_t\frac{u}{N_{k,t}},$$

$$C_1(k,t) = 2.5(1 + K_t(L_t) \vee m_t)\frac{D_{k,t} + L_t}{N_{k,t}},$$

$$C_2(k,t) = 2.5\sqrt{1 + K_t^2(L)}\left(1 + \sqrt{\frac{D_{k,t}}{N_{k,t}}}\right),$$

$$C_3(k,t) = 2.5\left[\left(\frac{m_t K_t(L)}{2}\right) \vee 5\right]\left(1 + 2\frac{D_{k,t}}{N_{k,t}}\right),$$

Theorem 2.1. *Let Y obey to model* (1)*. We assume that $p < n$ and that $(X_i)_{1 \leq i \leq p}$ is a linearly independent family. We denote by J the set $\{j, \beta_j \neq 0\} = \{1, ..., k_0\}$. Let γ and α be fixed in $]0,1[$. The testing procedure estimates k_0 by $\hat{k} = inf\{k \geq 0, T_{k,\alpha} \leq 0\}$, where $T_{k,\alpha}$ is defined by* (3)*. Let $\{\alpha_t, t \in \mathcal{T}_k\}$ be calculated according to the procedure P1 or P2.*
If $\forall k \leq k_0 - 1$ the condition (R_k) holds

$$(R_k) : \exists\, t \in \mathcal{T}_k / \ \left\|\Pi_{S_{k,t}}(\mu)\right\|_n^2 \geq C_1(k,t)\left\|\Pi_{V_{k,t}^\perp}(\mu)\right\|_n^2 + \frac{\sigma^2}{n}\left[C_2(k,t)\sqrt{2^t\, log\left(\frac{2k_0}{\alpha_t\gamma}\right)} + C_3(k,t) log\left(\frac{2k_0}{\alpha_t\gamma}\right)\right]$$

then

$$\mathbb{P}_\mu(\hat{k} \neq k_0) \leq \gamma + \alpha \tag{5}$$

Remark 2.2. For k fixed, $C_1(k,t), C_2(k,t), C_3(k,t)$ behave like constants if the following conditions are verified:

for all $t \in \mathcal{T}_k, \alpha_t \geq \exp(-N_{k,t}/10)$, $\gamma \geq 2\exp(-N_{k,t}/21)$ and the ratio $\dfrac{D_{k,t}+L_{k,t}}{N_{k,t}}$ remains bounded.

Under these conditions, the following inequalities hold:

$$C_1(k,t) \leq 10\frac{D_{k,t}+\log(1/\alpha_t)}{N_{k,t}}, \qquad C_2(k,t) \leq 5\left(1+\sqrt{\frac{D_{k,t}}{N_{k,t}}}\right), \qquad C_3(k,t) \leq 12.5\left(1+2\frac{D_{k,t}}{N_{k,t}}\right).$$

It is important to note that Theorem 2.1 is non asymptotic, hence the strength of our procedure. Baraud et al. [2] proposed an adaptive procedure to test in model (1) that $\mu = X\beta$ belongs to some linear subspace V of $\mathbb{R}^n$. Our variable selection procedure as well as the results of Theorem 2.1 are inspired by this paper. An asymptotic property can be deduced to proved the consistency of our estimator $\hat{k}$ of k_0:

Corollary 2.3. *Assume that $p \leq An$ with $A < 1$. Then, using the same notation as in Theorem 2.1, we obtain*

$$\mathbb{P}_\mu(\hat{k} \neq k_0) \underset{n\to\infty}{\longrightarrow} 0. \tag{6}$$

Remark 2.4. We say that μ satisfies condition (R) if $\forall k \leq k_0 - 1$, (R_k) holds. According to Theorem 2.1, our procedure is powerful under the condition (R). A condition on the coefficients β_J underlies in (R) since the projection of Y onto a space spanned by a subset of the family $(X_i)_{1\leq i\leq p}$ depends both on β and on the matrix X. These conditions on β_J appear explicitly when $(X_i)_{1\leq i\leq p}$ is an orthonormal family. Assume in the following that $(X_i)_{1\leq i\leq p}$ is an orthonormal family. Thus (1) becomes:

$$Y = \underbrace{X_1\beta_1 + .. + X_k\beta_k}_{V_k} + \underbrace{X_{k+1}\beta_{k+1} + ... + X_p\beta_p}_{\subset V_k^\perp} + \epsilon. \tag{7}$$

With the new decomposition (7), the projection of Y on any subspace $S_{k,t}$ only depends on $(\beta_j)_{j\geq k+1}$. Thus the condition (R_k) can be written in a different form, making explicit use of the β's:

$$\exists\, t \in \mathcal{T}_k / \ \beta_{k+1}^2 + .. + \beta_{k+2^t}^2 \geq C_1(k,t)\sum_{j=k+2^t+1}^{p}\beta_j^2 + \frac{\sigma^2}{n}\left[C_2(k,t)\sqrt{2^t log\left(\frac{2k_0}{\alpha_t\gamma}\right)} + C_3(k,t) log\left(\frac{2k_0}{\alpha_t\gamma}\right)\right].$$

2.2. The case $p \geq n$

After pointing out the properties of the procedure in the common case $p < n$, let us discuss a really important framework: the high-dimensional case, i.e. $p \geq n$. The family $(X_i)_{1\leq i\leq p}$ can no longer be a linearly independent family. We have to make the assumption that the decomposition of μ is unique, i.e $\exists! J \subset \{1, ..., p\} / \mu = \sum_{j\in J} X_j\beta_j$.
Let recall that in this section $(X_i)_{1\leq i\leq p}$ is supposed to be ordered regardless of Y and that the relevant variables are supposed to be in the first places, i.e. we still have $J = \{1, ..., k_0\}$ and $|J| = k_0$.

Let define $a = \dim(\text{span}(X_1, ..., X_p))$, note that $a \leq n$. In this part, $\forall\ k < a-1$, $t^k_{max} = \lfloor log_2(a-k-1) \rfloor$, this is in order to always have $N_{k,t} \neq 0$ and so to be able to calculate $T_{k,\alpha}$ defined by (3) for all $k < a-1$. Denote $V_k = \text{span}(X_1, ..., X_{s_k})$ where s_k is defined by $s_k = \inf\{s / \dim(\text{span}(X_1, ..., X_s)) = k\}$ and $S_{k,t} = \text{span}(X_{s_k+1}, ..., X_{s_k+q_{k,t}}) \cap V_k^{\perp}$ where $q_{k,t}$ is defined by $q_{k,t} = \inf\{q / \dim(\text{span}(X_{s_k+1}, ..., X_{s_k+q})) = 2^t\}$.

The condition (R_k) in Theorem 2.1 gives no restriction on the growth of p. Thus Theorem 2.1 applies with the new notations for any $p \geq n$, but for $k_0 < n$. This is no strong restriction since $k_0 > n$ means that we do not have sufficient observations to estimate k_0, whatever the method.

Results from a simulation study in Section 4 will show the power of our procedure; either when $p < n$ or when $p \geq n$.

3. Non-ordered variable selection

In Section 2 we defined a procedure based on multiple hypotheses testing in order to estimate J, the set of relevant variables of a sparse linear model (1). As the family $(X_i)_{1 \leq i \leq p}$ was given an order independent of Y, the estimation of $J = \{1, .., k_0\}$ was reduced to the estimation of k_0. This present section is dedicated to a more common case: $(X_i)_{1 \leq i \leq p}$ is not assumed to be given an order anymore, so J is not necessarily equal to $\{1, .., k_0\}$. We define here a general two-step procedure to estimate J; the first step orders the variables and the second estimates $|J|$. After the first step of the general procedure, the ordered variables will be denoted as $X_{(1)}, ..., X_{(p)}$.

The first step of our procedure consists in ordering the variables. In this paper two ways to order $(X_i)_{1 \leq i \leq p}$ taking into account the observations Y are proposed:
– Variables ordered by increasing p-values: when $p < n$, a p-value is calculated for each variable using the least squares estimate and then the variables are sorted by increasing p-value. When $p \geq n$, a p-value is calculated for each variable using the decomposition of Y onto that variable.
– Variables ordered with the Bolasso technique, introduced by Bach [1]. It is a bootstrapped version of the Lasso which improves its stability: several independent bootstrap samples are generated and the Lasso is performed on each of them. This approach is proved to make the irrelevant variables asymptotically disappear. A modification is applied to the Bolasso to adapt it to non asymptotic analysis. An appearance frequency is calculated for each variable X_i by counting the number of times the variable X_i is selected over the bootstrap samples. A high frequency denotes a good prediction ability of the variable X_i, at a given penalty. To avoid the use of a penalty, we set the first ordered variable to be the first one to reach a frequency of 1 from a decreasing penalty; and so on for the other variables. We proceed by dichotomy to order the variables.

The first method -ordered p-values- is the one demanding less computational time, but as shown in Section 4, the Bolasso technique gives a better order and thus better results. Indeed, as we will see throughout this section, the crucial step is the first one; the correct ordering of the variables. Indeed, the ability to estimate J with this procedure depends on the ability to get the relevant variables in the first places.

From now on, we assume that the decomposition of μ is unique, i.e $\exists! J \subset \{1, ..., p\} / \mu = \sum_{j \in J} X_j \beta_j$. We have $|J| = k_0$. We introduce here an event that will be useful in the following of

this section:

$$A_k = \{\text{the } k \text{ first ordered variables are relevant}\} = \{\{(1), .., (k)\} \subset J\}. \tag{8}$$

On the event A_{k_0}, the k_0 first ordered variables are relevant, so $\{(1), .., (k_0)\} = J$; an estimation of J is then obtained from an estimation of $|J| = k_0$.
The second step of the general procedure consists in testing successively the null hypothesis:

$$\hat{H}_k : \quad \{|J| = k\} \text{ against the alternative that } \{|J| > k\}. \tag{9}$$

The procedure stops when the null hypothesis is accepted; $\mathring{k} = \inf\{k \geq 0; \hat{H}_k \text{ is accepted}\}$ is an estimation of k_0 with our procedure and therefore $\hat{J} = \{(1), .., (\mathring{k})\}$ is an estimation of the set of relevant variables.

Two cases are distinguished to test the null hypothesis $\hat{H}_k$, either σ is known or not. The first step remains the same for both procedures. A procedure 'A' is proved powerful under some conditions on the data if σ is known; if σ is unknown, we provide another two-step procedure 'B' that is also proved to be powerful under some conditions on the data.

3.1. The case $p < n$ and σ is known

In this section, we define a procedure called Procedure 'A' under the assumption that the variance σ^2 is known. Assume that the family $(X_i)_{1 \leq i \leq p}$ is a linearly independent family and that the first step of Procedure 'A' has already been done; variables have been ordered. The second step is a testing procedure that will be described in the following.

Let us adapt the notation of Section 2.1 to this section: we first recall that $\forall\, k \leq p-1, t^k_{max} = \lfloor log_2(p-k) \rfloor$, $\mathcal{T}_k = \{0, ..., t^k_{max}\}$, we define $V_{(k)} = \text{span}(X_{(1)}, ..., X_{(k)})$, let $Q_{(k),t^k_{max}} = \{S_{(k),(t)}, t \in \mathcal{T}_k\}$ be a collection of linear subspaces of $V^{\perp}_{(k)}$, where $\forall\, t \in \mathcal{T}_k$, $S_{(k),(t)} = \text{span}(X_{(k+1)}, .., X_{(k+2^t)}) \cap V^{\perp}_{(k)}$. With the definition of $S_{(k),(t)}$, we have $\dim(S_{(k),(t)}) = D_{k,t} = 2^t$. Let us denote $V_{(k),(t)} = V_{(k)} \oplus S_{(k),(t)}$. Some notations of the previous section will also be used.

3.1.1. The general case

We introduce new statistics; for all $0 \leq k < p$ and for all $t \in \mathcal{T}_k$, let

$$U_{k,t} = \frac{||\Pi_{S_{(k),(t)}} Y||^2_n}{\sigma^2}.$$

The second step of the procedure 'A' presented here consists in doing successively several tests of the null hypothesis $\hat{H}_k$ defined by (9) at level α_t, where $\{\alpha_t, t \in \mathcal{T}_k\}$ is a suitable collection of number in $]0, 1[$, using the test statistics $U_{k,t}$. The final multiple testing procedure consists in rejecting $\hat{H}_k$ if one of those tests rejects this hypothesis. We want the final test to be of level α.

As the distribution of the statistic $U_{k,t}$ is unknown, an upper bound has to be found in order to build the same kind of testing procedure as in Section 2, indeed the space $S_{(k),(t)}$ is random and depends on Y.

Let $k \in \{0, ..., p-1\}$, $\epsilon' \sim \mathcal{N}_n(0, \sigma^2 I_n)$. The family $(X_i)_{1 \leq i \leq p}$ is ordered by a permutation σ_1 defined by:
$\forall\, j \in \{1, ..., k\}, \sigma_1(j) = (j)$ and $\forall\, j \in \{k+1, ..., p\}, X_{\sigma_1(j)}$ is the variable that maximizes:
$\{||\Pi_{X_i \cap < X_{\sigma_1(1)}, ..., X_{\sigma_1(j-1)} >^{\perp}} \epsilon'||^2_n, \forall\, i \in \{1, ..., p\} \backslash \{\sigma_1(1), ..., \sigma_1(j-1)\}\}$.

We can then calculate the statistics :

$$U^1_{k,t} = \frac{\|\Pi_{S_{(k),\sigma_1(t)}}\epsilon'\|^2_n}{\sigma^2}, \text{ where } S_{(k),\sigma_1(t)} = \text{span}(X_{\sigma_1(k+1)}, .., X_{\sigma_1(k+2^t)}) \cap V^{\perp}_{(k)}.$$

Lemma 3.1. *We have a stochastic upper bound of $U_{k,t}$ for all $0 \leq k < p$ and for all $t \in \mathcal{T}_k$: under $\hat{H}_k$ given by* (9) *and on the event A_k defined by* (8) *:*

$$U_{k,t} \leq U^1_{k,t}. \tag{10}$$

Let $\overline{U^1_{k,t}}(u)$ denote the probability for the statistic $U^1_{k,t}$ to be larger than u.
Set $\forall \alpha \in]0,1[, \forall 0 \leq k < p$,

$$M^1_{k,\alpha} = \sup_{t \in \mathcal{T}_k} \left\{ U_{k,t} - \overline{U^1_{k,t}}^{-1}(\alpha_t) \right\} \tag{11}$$

where $\{\alpha_t,\ t \in \mathcal{T}_k\}$ is a collection of number in]0,1[chosen in accordance to the following procedure:

P3. For all $t \in \mathcal{T}_k, \alpha_t = \alpha_n$ where α_n is the α-quantile of the random variable

$$\inf_{t \in \mathcal{T}_k} \overline{U^1_{k,t}} \left\{ U^1_{k,t} \right\}.$$

The null hypothesis $\hat{H}_k$ is rejected when $M^1_{k,\alpha}$ is positive. In fact, the second step of the procedure 'A' is to calculate $M^1_{k,\alpha}$ from $k = 0$ until $M^1_{k,\alpha}$ is negative. The calculation of the collection $\{\alpha_t,\ t \in \mathcal{T}_k\}$ with the procedure P3 gives a test $\hat{H}_k$ of level α.

In summary, the two-step procedure 'A' when σ is known consists in ordering the p variables and then estimating J by $\hat{J} = \{(1), .., (\mathring{k}_A)\}$ where $\mathring{k}_A = \inf\{k \geq 0; M^1_{k,\alpha} \leq 0\}$. The testing procedure 'A' is proved to be powerful and we give an upper bound of the probability to wrongly estimate J in the next theorem.

Theorem 3.2. *Let Y obey to model* (1)*. We assume that $p < n$ and that $(X_i)_{1 \leq i \leq p}$ is a linearly independent family. We denote by J the set $\{j, \beta_j \neq 0\}$. Let α and γ be fixed in*]0, 1[*.*
The procedure estimates J by $\hat{J} = \{(1), .., (\mathring{k}_A)\}$ where $\mathring{k}_A = inf\{k \geq 0, M^1_{k,\alpha} \leq 0\}$, where $M^1_{k,\alpha}$ is defined by (11) *and $\{\alpha_t, t \in \mathcal{T}_k\}$ is calculated according to the procedure P3.*

We consider the condition $(R_{2,k})$ stated as $(R_{2,k})$: $\exists t \leq log_2(k_0 - k)$ such that

$$\frac{1}{2\sigma^2} inf\left\{ \|\Pi_S \mu\|^2_n, S \in B_{2^t} \right\} \geq \frac{2^t}{n} \left[10 + 4log\left(\frac{(p-k)k_0}{2^{2t}} \right) \right] + \frac{2}{n} \left[\sqrt{2^{t+1} log\left(\frac{k_0 |\mathcal{T}_k|}{\gamma\alpha} \right)} + log\left(\frac{k_0 |\mathcal{T}_k|}{\gamma\alpha} \right) \right]$$

where $\forall d \leq k_0, B_d = \{span(X_I), I \subset J, |I| = d\}$ and $|\mathcal{T}_k| = log_2(p-k) + 1$.

If $\forall k \leq k_0 - 1$ the condition $(R_{2,k})$ holds, then

$$\mathbb{P}_\mu(\hat{J} \neq J) \leq \gamma + \alpha + \delta \tag{12}$$

where $\delta = \mathbb{P}_\mu(A^c_{k_0}) = P_\mu(\exists\, (j) \leq k_0 / \beta_{(j)} = 0)$.

This theorem is non asymptotic and shows that a crucial step is to correctly order the variables. Indeed, δ stands for the weight of the chosen order, if the k_0 relevant variables are not the first ones in the first step of the procedure, then we will not have $J = \hat{J}$.

3.1.2. The particular case where $(X_i)_i$ is an orthonormal family

When the family $(X_i)_{1\le i\le p}$ is orthonormal, the upper bound of the statistics $U_{k,t}$ in Lemma 3.1 can be expressed differently.
Let $D > 0$ and $W_1, ..., W_D$ be D i.i.d. standard Gaussian variables ordered as $|W_{(1)}| \ge ... \ge |W_{(D)}|$. We define: $\forall d = 1, ..., D$,

$$Z_{d,D} = \sum_{j=1}^{d} W_{(j)}^2 \tag{13}$$

Let $\overline{Z}_{d,D}(u)$ denote the probability for the statistic $Z_{d,D}$ to be larger than u.

Lemma 3.3. *We have a stochastic upper bound of $U_{k,t}$ for all $0 \le k < p$ and for all $t \in \mathcal{T}_k$: under $\hat{H}_k$ and on the event A_k:*

$$U_{k,t} \le Z_{D_{k,t},p-k}/n.$$

Set $\forall \alpha \in]0,1[$, $\forall 0 \le k < p$,

$$M_{k,\alpha} = \sup_{t\in\mathcal{T}_k} \left\{ U_{k,t} - \overline{Z}^{-1}_{D_{k,t},p-k}(\alpha_t)/n \right\} \tag{14}$$

where $\{\alpha_t,\ t \in \mathcal{T}_k\}$ is a collection of number in]0,1[chosen in accordance to the following procedure:

P4. For all $t \in \mathcal{T}_k$, $\alpha_t = \alpha_n$ where α_n is the α-quantile of the random variable

$$\inf_{t\in\mathcal{T}_k} \overline{Z}_{D_{k,t},p-k}\left\{ Z_{D_{k,t},p-k} \right\}.$$

The null hypothesis $\hat{H}_k$ is rejected when $M_{k,\alpha}$ is positive. The procedure P4 gives a test $\hat{H}_k$ of level α. The major benefit of Procedure 'A' when the family $(X_i)_{1\le i\le p}$ is orthonormal is that the upper bound of the statistics $U_{k,t}$ in Lemma 3.3 does not depend on the family $(X_i)_{1\le i\le p}$ nor on the order on that family. Thus the calculation of P4 only depends on k and t, with p and n fixed.

We have the next corollary in the particular case where $(X_i)_{1\le i\le p}$ is an orthonormal family, making explicit use of the β's.

Corollary 3.4. *Let Y obey to model* (1)*. We assume that $p < n$ and that $(X_i)_{1\le i\le p}$ is an orthonormal family. We denote by J the set $\{j, \beta_j \ne 0\}$. Let α and γ be fixed in* $]0,1[$.
The procedure estimates J by $\hat{J} = \{(1), .., (\mathring{k}_{Abis})\}$ where $\mathring{k}_{Abis} = inf\{k \ge 0, M_{k,\alpha} \le 0\}$, where $M_{k,\alpha}$ is defined by (14) *and $\{\alpha_t, t \in \mathcal{T}_k\}$ is calculated according to the procedure P4.*

We consider the condition $(R_{2bis,k})$ stated as $(R_{2bis,k}) : \exists\, t \le log_2(k_0 - k)$ such that

$$\frac{1}{2\sigma^2}\sum_{j=1}^{2^t}\beta^2_{\sigma_2(j)} \ge \frac{2^t}{n}\left[10 + 4log\left(\frac{(p-k)k_0}{2^{2t}}\right)\right] + \frac{2}{n}\left[\sqrt{2^{t+1}log\left(\frac{k_0|\mathcal{T}_k|}{\gamma\alpha}\right)} + log\left(\frac{k_0|\mathcal{T}_k|}{\gamma\alpha}\right)\right]$$

where σ_2 is defined by $|\beta_{\sigma_2(1)}| \le ... \le |\beta_{\sigma_2(k_0)}|$ and $|\mathcal{T}_k| = log_2(p-k) + 1$.

If $\forall k \le k_0 - 1$ the condition $(R_{2bis,k})$ holds, then

$$\mathbb{P}_\mu(\hat{J} \ne J) \le \gamma + \alpha + \delta \tag{15}$$

where $\delta = \mathbb{P}_\mu(A^c_{k_0}) = P_\mu(\exists\, (j) \le k_0/\beta_{(j)} = 0)$.

3.2. The case $p < n$ and σ is unknown

In this section, we define a procedure 'B' under the assumption that the variance σ^2 is unknown. Assume that the family $(X_i)_{1\leq i\leq p}$ is a linearly independent family and that the first step of this procedure 'B' has already been done; variables have been ordered. In this section, some notations of Section 3.1 are used: $\forall\ k < p, t^k_{max} = \lfloor log_2(p-k)\rfloor$, $\mathcal{T}_k = \{0, ..., t^k_{max}\}$, we define $V_{(k)} = \text{span}(X_{(1)}, ..., X_{(k)})$, let $Q_{(k),t^k_{max}} = \{S_{(k),(t)}, t \in \mathcal{T}_k\}$ be a collection of linear subspaces of $V^{\perp}_{(k)}$, where $\forall\ t \in \mathcal{T}_k$, $S_{(k),(t)} = \text{span}(X_{(k+1)}, .., X_{(k+2^t)}) \cap V^{\perp}_{(k)}$.
We denote $V_{(k),(t)} = V_{(k)} \oplus S_{(k),(t)}$. With the definition of $S_{(k),(t)}$, we have $\dim(S_{(k),(t)}) = D_{k,t} = 2^t$ and $\dim(V^{\perp}_{(k),(t)}) = N_{k,t} = n - (k + 2^t)$.

3.2.1. The general case

Set the following statistics: for all $0 \leq k < p$ and for all $t \in \mathcal{T}_k$,

$$\tilde{U}_{D_{k,t},N_{k,t}} = \frac{N_{k,t}\|\Pi_{S_{(k),(t)}}Y\|^2_n}{D_{k,t}\|Y - \Pi_{V_{(k),(t)}}Y\|^2_n}.$$

The second step of the procedure 'B' presented here consists in doing successively several tests of the null hypothesis $\hat{H}_k$ defined by (9) at level α_t, suitable collection of number in $]0, 1[$, using the test statistics $\tilde{U}_{D_{k,t},N_{k,t}}$. As in the previous Section 3.1.1, an upper bound of $\tilde{U}_{D_{k,t},N_{k,t}}$ needs to be found.

Let $k \in \{0, ..., p-1\}$, $\epsilon' \sim \mathcal{N}_n(0, \sigma^2 I_n)$. The family $(X_i)_{1\leq i\leq p}$ is ordered by a permutation σ_1 defined by $\forall\ j \in \{1, ..., k\}, \sigma_1(j) = (j)$ and $\forall\ j \in \{k+1, ..., p\}, X_{\sigma_1(j)}$ is the variable that maximizes:
$\{\|\Pi_{X_i \cap < X_{\sigma_1(1)},...,X_{\sigma_1(j-1)} >^{\perp}}\epsilon'\|^2_n, \forall\ i \in \{1, ..., p\}\backslash\{\sigma_1(1), ..., \sigma_1(j-1)\}\}$.
We can then calculate the statistics $\Upsilon_{k,t} = \dfrac{N_{k,t}\|\Pi_{S_{(k),\sigma_1(t)}}\epsilon'\|^2_n}{D_{k,t}\|\epsilon' - \Pi_{V_{(k),\sigma_1(t)}}\epsilon'\|^2_n}$
where $S_{(k),\sigma_1(t)} = \text{span}(X_{\sigma_1(k+1)}, .., X_{\sigma_1(k+2^t)}) \cap V^{\perp}_{(k)}$ and $V_{(k),\sigma_1(t)} = S_{(k),\sigma_1(t)} \oplus V_{(k)}$.

Lemma 3.5. *We have a stochastic upper bound of $\tilde{U}_{D_{k,t},N_{k,t}}$ for all $0 \leq k < p$ and for all $t \in \mathcal{T}_k$: under $\hat{H}_k$ given by* (9) *and on the event A_k defined by* (8) *:*

$$\tilde{U}_{D_{k,t},N_{k,t}} \leq \Upsilon_{k,t}. \tag{16}$$

Let $\overline{\Upsilon}_{k,t}(u)$ denote the probability for the statistic $\Upsilon_{k,t}$ to be larger than u.
Set $\forall\alpha \in]0, 1[, \forall 0 \leq k < p$,

$$\hat{M}_{k,\alpha} = \sup_{t\in\mathcal{T}_k}\left\{\tilde{U}_{D_{k,t},N_{k,t}} - \overline{\Upsilon}^{-1}_{k,t}(\alpha_t)\right\} \tag{17}$$

where $\{\alpha_t,\ t \in \mathcal{T}_k\}$ is a collection of number in]0,1[chosen in accordance to the following procedure:

P5. For all $t \in \mathcal{T}_k, \alpha_t = \alpha_n$ where α_n is the α-quantile of the random variable

$$\inf_{t\in\mathcal{T}_k}\overline{\Upsilon}_{k,t}\{\Upsilon_{k,t}\},$$

The null hypothesis $\hat{H}_k$ is rejected when $\hat{M}_{k,\alpha}$ is positive. In fact, the second step of the procedure 'B' is to calculate $\hat{M}_{k,\alpha}$ from $k = 0$ until $\hat{M}_{k,\alpha}$ is negative. The calculation of the collection $\{\alpha_t,\ t \in \mathcal{T}_k\}$ with procedure P5 gives a final test $\hat{H}_k$ of level α.

In summary, this two-step procedure 'B' when σ is unknown consists in ordering the p variables and then estimating $|J|$ by $\hat{J} = \{(1), .., (\mathring{k}_B)\}$ where $\mathring{k}_B = \inf\{k \geq 0; \hat{M}_{k,\alpha} \leq 0\}$. The procedure is proved to be powerful in the next theorem; we give an upper bound of the probability to wrongly estimate J under some conditions on the data. Let us introduce some notations that will be used in the following theorem:

$L_t = \log(|\mathcal{T}_k|/\alpha)$, $m_t = \exp(4L_t/N_{k,t})$, $m_p = \exp\left(\frac{4D_{k,t}}{N_{k,t}} log\left(\frac{e(p-k)}{D_{k,t}}\right)\right)$, $M = 2m_t m_p$. Denote $\Lambda_1(k,t) = \sqrt{1 + \frac{D_{k,t}}{N_{k,t}}}$, $\Lambda_2(k,t) = \left(1 + 2\frac{D_{k,t}}{N_{k,t}}\right) M$ and $\Lambda_3(k,t) = 2\Lambda_1(k,t) + \Lambda_2(k,t)$.

Theorem 3.6. *Let Y obey to model* (1)*. We assume that $p < n$ and that $(X_i)_{1 \leq i \leq p}$ is a linearly independent family. We define by J the set $\{j, \beta_j \neq 0\}$. Let α and γ be fixed in* $]0, 1[$.
The procedure estimates J by $\hat{J} = \{(1), .., (\mathring{k}_B)\}$ where $\mathring{k}_B = inf\{k \geq 0, \hat{M}_{k,\alpha} \leq 0\}$, where $\hat{M}_{k,\alpha}$ is defined by (17) *and $\{\alpha_t, t \in \mathcal{T}_k\}$ is calculated according to the procedure P5.*

We consider the condition $(R_{3,k})$ stated as $(R_{3,k})$: $\exists t \leq log_2(k_0 - k)$ such that

$$\frac{1}{2} inf\left\{\|\Pi_S \mu\|_n^2, S \in B_{2^t}\right\} \geq$$

$$\frac{A(k,t)}{N_{k,t}}\left[\|\mu\|_n^2 + \sigma^2\left(2 + \frac{3}{n} log\left(\frac{2k_0}{\gamma}\right)\right)\right] + \frac{\sigma^2}{n}\left[2^t\left(6 + 4log\left(\frac{k_0}{2^t}\right)\right) + 3log\left(\frac{2k_0}{\gamma}\right)\right], \tag{18}$$

where $A(k,t) = 2^t\left[2 + \frac{2^t}{N_{k,t}} + \Lambda_3(k,t) log\left(\frac{e(p-k)}{2^t}\right)\right] + (1 + \Lambda_2(k,t))\, log\left(\frac{log_2(p-k)+1}{\alpha}\right)$
and $\forall d \leq k_0$, $B_d = \{span(X_I), I \subset J, |I| = d\}$.

If $\forall k \leq k_0 - 1$ the condition $(R_{3,k})$ holds, then

$$\mathbb{P}_\mu(\hat{J} \neq J) \leq \gamma + \alpha + \delta \tag{19}$$

where $\delta = \mathbb{P}_\mu(A_{k_0}^c) = P_\mu(\exists\, (j) \leq k_0 / \beta_{(j)} = 0)$.

This theorem is non asymptotic and shows that under some conditions on the data, the testing procedure 'B' presented in this section is powerful. As for Theorem 3.2 of Section 3.1.1, the first step of the procedure -the ordering of the variables- has an important part in Theorem 3.6.

Remark 3.7. The condition $(R_{3,k})$ can be simplified under the assumption that $2^t \leq (n-k)/2$ and $log(p-k) > 1$. Indeed, in this case, the right hand (18) is upper bounded by

$$C(\|\mu\|_n, \gamma, \alpha, \sigma) 2^t\left[\frac{log(p-k)}{N_{k,t}} + \frac{log(k_0)}{n}\right], \tag{20}$$

where $C(\|\mu\|_n, \gamma, \alpha, \sigma)$ is a constant depending on $\|\mu\|_n$, γ, α and σ.

A simulation study in Section 4 will show that this testing procedure combined with a good way to order variables -in order to minimize δ- performs well.

3.2.2. The particular case where $(X_i)_{1\leq i\leq p}$ is an orthonormal family

When $(X_i)_{1\leq i\leq p}$ is an orthonormal family, the condition $(R_{3,k})$ of Theorem 3.6 can be expressed differently, making explicit use of the β's. The new condition obtained in the case of an orthonormal family is also easier to satisfy.

Corollary 3.8. *Let Y obey to model* (1)*. We assume that $p < n$ and that $(X_i)_{1\leq i\leq p}$ is an orthonormal family. We define J by the set $\{j, \beta_j \neq 0\}$. Let α and γ be fixed in $]0,1[$.*
The procedure estimates J by $\hat{J} = \{(1), .., (\mathring{k}_B)\}$ where $\mathring{k}_B = inf\{k \geq 0, \hat{M}_{k,\alpha} \leq 0\}$, where $\hat{M}_{k,\alpha}$ is defined by (17) *and $\{\alpha_t, t \in \mathcal{T}_k\}$ is calculated according to the procedure P5.*

We consider the condition $(R_{3bis,k})$ stated as $(R_{3bis,k})$: $\exists t \leq log_2(k_0 - k)$ such that

$$\frac{1}{2\sigma^2}\sum_{j=1}^{2^t}\beta^2_{\sigma_2(j)} \geq$$

$$\frac{A(k,t)}{N_{k,t}}\left[\sum_{j=k+2^t}^{j=k_0}\beta^2_{\sigma_2(j)} + \sigma^2\left(2 + \frac{3}{n}log\left(\frac{2k_0}{\gamma}\right)\right)\right] + \frac{\sigma^2}{n}\left[2^t\left(6 + 4log\left(\frac{k_0}{2^t}\right)\right) + 3log\left(\frac{2k_0}{\gamma}\right)\right],$$

where $A(k,t) = 2^t\left[2 + \frac{2^t}{N_{k,t}} + \Lambda_3(k,t)log\left(\frac{e(p-k)}{2^t}\right)\right] + (1 + \Lambda_2(k,t))\, log\left(\frac{log_2(p-k)+1}{\alpha}\right)$
and σ_2 is defined such that $|\beta_{\sigma_2(1)}| \leq ... \leq |\beta_{\sigma_2(k_0)}|$.

If $\forall k \leq k_0 - 1$ the condition $(R_{3bis,k})$ holds, then

$$\mathbb{P}_\mu(\hat{J} \neq J) \leq \gamma + \alpha + \delta \tag{21}$$

where $\delta = \mathbb{P}_\mu(A^c_{k_0}) = P_\mu(\exists\, (j) \leq k_0 / \beta_{(j)} = 0)$.

Remark 3.7 is also verified in the particular case where $(X_i)_{1\leq i\leq p}$ is an orthonormal family.

3.3. The case $p \geq n$

We will now discuss the high-dimensional case with non-ordered variables, $p \geq n$. This section fits the two-step procedures previously introduced to high-dimensional analysis. Verzelen [15] shows that when $k_0 log(ep/k_0)/[nlog(n)] \geq 4$, called the ultra-high dimensional case, it is almost impossible to estimate the support of β. We will then consider that we are not in the ultra-high dimensional case.

The family $(X_i)_{1\leq i\leq p}$ is now a dependent family. As said at the beginning of Section 3, we assume that the decomposition of μ is unique, i.e. $\exists! J \subset \{1, ..., p\} / \mu = \sum_{j\in J} X_j\beta_j$. We still have $|J| = k_0$. The general procedure defined at the beginning of Section 3 remains the same; the first step orders the variables and the second step estimates k_0.

Procedure 'A' defined in Section 3.1 when the variance is known and procedure 'B' defined in Section 3.2 when the variance is unknown are still applicable, but with some minor modifications. The first modification we have to make concerns the definition of the subspaces $V_{(k)}$ and $S_{(k),(t)}$. Indeed, we have to take into account that the family $(X_i)_{1\leq i\leq p}$ is a dependent family. Let define $a = \dim(\text{span}(X_1, ..., X_p))$, note that $a \leq n$, hence we set:
$\forall k < a - 1, t_{max} = \lfloor log_2(a - k - 1)\rfloor$ and note $V_{(k)} = \text{span}(X_{(1)}, ..., X_{(s_k)})$ where s_k is defined by $s_k = \inf\{s / \dim(\text{span}(X_{(1)}, ..., X_{(s)})) = k\}$ and $S_{(k),(t)} = \text{span}(X_{(s_k+1)}, ..., X_{(s_k+q_{k,t})}) \cap V^{\perp}_{(k)}$ where $q_{k,t}$ is defined by $q_{k,t} = \inf\{q / \dim(\text{span}(X_{(s_k+1)}, ..., X_{(s_k+q)})) = 2^t\}$. With these definitions, we have

$\dim(V_k) = k$ and $\dim(S_{k,t}) = 2^t$.
The second modification we have to make concerns the construction of the respective upper bound $U^1_{k,t}$ and $\Upsilon_{k,t}$ in Lemma 3.1 and Lemma 3.5. Indeed, as $p \geq n$, these upper bounds are constructed as following:

Let $k \in \{0, ..., a-2\}$, $\epsilon' \sim \mathcal{N}_n(0, \sigma^2 I_n)$. The family $(X_i)_{1\leq i\leq p}$ is ordered by σ_1 defined by: $\forall\ j \in \{1, ..., s_k\}, \sigma_1(j) = (j)$ and $\forall\ j \in \{s_k + 1, ..., p\}, X_{\sigma_1(j)}$ is the variable that maximizes: $\{||\Pi_{X_i \cap < X_{\sigma_1(1)}, ..., X_{\sigma_1(j-1)} >^{\perp}} \epsilon'||^2_n, \forall\ i \in \{1, ..., p\} \backslash \{\sigma_1(1), ..., \sigma_1(j-1)\}\}$.
Once we have j such that $\dim(< X_{\sigma_1(1)}, ..., X_{\sigma_1(j)} >) = a$, the next ordered variables $X_{\sigma_1(j+1)}$ can be any remaining unordered variables in $\{1, ..., p\} \backslash \{\sigma_1(1), ..., \sigma_1(j)\}$. We could complete σ_1 by an arbitrary order on the remaining variables, but since we achieve to construct a family $(X_{\sigma_1(1)}, ..., X_{\sigma_1(j)})$ that describes $\mathbb{R}^a$, we do not care about the remaining unordered variables.

We can then calculate the statistics:

$$U^1_{k,t} = \frac{||\Pi_{S_{(k),\sigma_1(t)}} \epsilon'||^2_n}{\sigma^2}, \text{ and } \Upsilon_{k,t} = \frac{N_{k,t} ||\Pi_{S_{(k),\sigma_1(t)}} \epsilon'||^2_n}{D_{k,t} ||\epsilon' - \Pi_{V_{(k),\sigma_1(t)}} \epsilon'||^2_n}$$

where $S_{(k),\sigma_1(t)} = \text{span}(X_{\sigma_1(s_k+1)}, ..., X_{\sigma_1(s_k+r_{k,t})}) \cap V^{\perp}_{(k)}$ where $r_{k,t}$ is defined by
$r_{k,t} = \inf\{q / \dim(\text{span}(X_{\sigma_1(s_k+1)}, ..., X_{\sigma_1(s_k+q)})) = 2^t\}$ and $V_{(k),\sigma_1(t)} = S_{(k),\sigma_1(t)} \oplus V_{(k)}$.

With those two modifications, Theorem 3.2 of Section 3.1.1 and Theorem 3.6 of Section 3.2.1 apply assuming $k_0 < a$. A simulation study is given in the next section showing that our procedure 'B' performs well.

4. Simulation study

In this section, we comment the results of the simulation study which are presented in the Appendix A. Our aim was to test the performances of our selection methods. Six methods were compared; the procedure described in Section 2 with ordered variables, denoted "pre-ordered" in the tables of Appendix A, the two-step procedure 'B' described in Section 3 with non-ordered variables, either with ordered p-values denoted "procpval" or with the Bolasso order denoted "procbol", the FDR procedure described in Bunea et al. [6], the Lasso method and the Bolasso technique. For the purpose of comparison, we considered the design of the simulations of Bunea et al. [6]. The comparison of the first method and the others is unfair and was not performed because of prior information being available on the relative importance of the variables. The two kinds of method have to be compared separately.

The simulation was performed in several frameworks: in the common case when $(X_i)_{1\leq i\leq p}$ is a linearly independent family, in a more precise case (the orthonormal case), and in a more general case (the high-dimensional case). For the latter, the FDR procedure of Bunea et al. [6] cannot be computed as p-values can not be obtained with the least squares estimate with all p variables. In this case we compared an adjusted FDR (denoted FDR2); a p-value was calculated for each variable X_i from the regression of Y onto the variable concerned. As mentioned in the introduction, this is a natural extension of the FDR procedure in high-dimensional analysis and extended FDR is currently widely used in biology for QTL research and transcriptome analysis.

Concerning the design of our simulations, we simulated p independent vectors $X^*_j \sim \mathcal{N}_n(0, I_n)$, and set the predictors $X_j = X^*_j / ||X^*_j||$, for $j = 1, ..., p$. The response variable Y was computed via $Y = \beta_{i_1} X_{i_1} + ... + \beta_{i_{k_0}} X_{i_{k_0}} + \epsilon$, where ϵ is a vector of independent standard Gaussian variables,

$\{i_1, .., i_{k_0}\} = J \subset \{1, .., p\}$ and $\beta_J \in \{\sqrt{n}, 6\}$. We considered two instances of k_0 (5 or 10). In each instance, samples of $n = 100$ and 500 in the low-dimensional case, and $n = 100$ in the high-dimensional case have been simulated. We let p to vary with the sample size n. In the orthonormal case, we set the predictors X_j, $j = 1, ..., p$ as an orthonormal basis of span$(X_1^*, .., X_p^*)$ (principal component for example). When $p > n$, the number of non-zero coefficients, $|J|$, was checked. Indeed, we assume that the decomposition of μ was unique, so we had to check the possibility that, even though k_0 non-zero coefficients were simulated, several other variables might be included in span$(X_{i_1}, ..., X_{i_{k_0}})$ because of collinearity. All variables included in span$(X_{i_1}, ..., X_{i_{k_0}})$ were looked for. In all simulations described above and reported in the tables A.1-A.3 in Appendix A, there were no other variables in J than the one used to simulate Y. Thus the aim of all simulations remained the same, i.e the estimation of $J = \{i_1, .., i_{k_0}\}$.

When $(X_i)_{1 \le i \le p}$ is not an orthonormal family, the calculation of $T_{k,\alpha}$ with (3) demands a lot of computational time, as a calculation of $V_k^{\perp}$ and $Q_{k,t_{max}}$ is needed for each k. Since a variable selection method is not only judged on its results but also on its fastness, useless calculations in our procedure had to be avoided. The Gram-Schmidt process was used to get an orthonormal family out of $(X_i)_{1 \le i \le p}$. Thus the calculation of $(V_k^{\perp})_{k \ge 0}$ was done once and for all.
Decompose $\forall\, l > 0$:

$$X_{k+l} = \Pi_{V_k}(X_{k+l}) + \Pi_{V_k^{\perp}}(X_{k+l})$$

Note $(e_j)_{j=1..k}$ an orthonormal basis of V_k, then:

$$\Pi_{V_k}(X_{k+l}) = \sum_{j=1}^{k} < X_{k+l}, e_j > e_j \quad \text{and} \quad \Pi_{V_k^{\perp}}(X_{k+l}) = X_{k+l} - \sum_{j=1}^{k} < X_{k+l}, e_j > e_j$$

The family $(X_1, .., X_p)$ was modified into $\left(X_1, \frac{\Pi_{V_1^{\perp}}(X_2)}{||\Pi_{V_1^{\perp}}(X_2)||}, \frac{\Pi_{V_2^{\perp}}(X_3)}{||\Pi_{V_2^{\perp}}(X_3)||}, ..., \frac{\Pi_{V_{p-1}^{\perp}}(X_p)}{||\Pi_{V_{p-1}^{\perp}}(X_p)||} \right)$
We called that orthonormal family $(\tilde{X}_1, .., \tilde{X}_p)$. Y had been decomposed as:

$$Y = \underbrace{\tilde{X}_1\tilde{\beta}_1 + .. + \tilde{X}_k\tilde{\beta}_k}_{V_k} + \underbrace{\tilde{X}_{k+1}\tilde{\beta}_{k+1} + ... + \tilde{X}_p\tilde{\beta}_p}_{\subset V_k^{\perp}} + \epsilon \tag{22}$$

Then $S_{k,t} = \text{span}(\tilde{X}_{k+1}, .., \tilde{X}_{k+2^t})$ and so $||\Pi_{S_{k,t}} Y||_n^2 = \tilde{\beta}_{k+1}^2 + ... + \tilde{\beta}_{k+2^t}^2$. This technique avoided a lot of useless and redundant calculations.
The decomposition of Gram-Schmidt has also been used in the non-ordered variables case with the two-step procedure 'A' and 'B' once the variables have been ordered.

When $(X_i)_{1 \le i \le p}$ is an orthonormal family, we used another upper bound of the statistics $\tilde{U}_{D_{k,t},N_{k,t}}$ in our simulations than the one in Lemma 3.5. Indeed, we can obtain an upper bound which does not depend on the family $(X_i)_{1 \le i \le p}$ nor on the order on that family.
Let $I_1, ..., I_p$ be p i.i.d. standard Gaussian variables, and let $|I_{(1)}| \ge ... \ge |I_{(p)}|$.
We define: $\forall k = 0, ..., p-1,\ \forall D = 0, ..., p-k-1$, $L_{k,D} = \sum_{j=k+D+1}^{p} I_{(j)}^2$
We have a stochastic upper bound of $\tilde{U}_{D_{k,t},N_{k,t}}$ for all $0 \le k < p$ and for all $t \in \mathcal{T}_k$: under $\hat{H}_k$ and on the event A_k:

$$\tilde{U}_{D_{k,t},N_{k,t}} \le \frac{N_{k,t}}{D_{k,t}} \frac{Z_{D_{k,t},p-k}}{L_{k,D_{k,t}} + K_{n-p}}$$

where K_{n-p} is a chi-square variable with $n-p$ degrees of freedom and $Z_{d,D}$ is defined by (13). The simulations were performed with this new upper bound.

The results of the simulation study are presented in Table A.1-A.3. The method displayed as FDR in the simulation tables corresponds to the procedure described in Bunea et al. [6], by choosing q (user level) as 0.1 and 0.05.

The l^1 penalty of the Lasso was tuned via 10-cross validation. Concerning the Bolasso technique, we choose $it = 100$ bootstrap iterations; the frequency threshold and the penalty were also tuned via 10-cross validation.

Concerning the Bolasso for ordering, we chose to stop the dichotomy algorithm (see Section 3) as soon as $\min(p, 60)$ variables were ordered. The objective was to spare calculation because it is uneasy to distinguish the remaining variables after the $\min(p, 60)$th position. The dichotomy algorithm assumes that when a variable reaches a frequency of 1, the frequency stays at 1 when the penalty decreases. In practice this assumption might be wrong, the algorithm is then restarted.

Concerning the three procedures presented in this paper, the results are displayed for a level $\alpha \in \{0.1, 0.05\}$. For the ordered case, $(X_i)_{1\leq i\leq p}$ became (X_J, X_{-J}) and the collection $\{\alpha_t, t \in \mathcal{T}_k\}$ was chosen in accordance to the procedure P1, which demanded more computational time than P2, but which was far much powerful. For the non-ordered case, the collection $\{\alpha_t, t \in \mathcal{T}_k\}$ was chosen with the procedure P5, when $(X_i)_{1\leq i\leq p}$ was not an orthonormal family, as the variance was considered unknown in the simulation.

In all tables, the first row gives an estimation of $\delta = P_\mu(A^c_{k_0})$. This estimation is not mentioned for the first procedure as the variables have already been ordered, so $\delta = 0$. In low dimension, the parameter m reflect the well-conditioned of the matrix X; $m = \max_{1\leq j\leq p} m_{jj}$ where $(m_{ij})_{1\leq i,j\leq p} = (X^T X)^{-1}/n$, a low m means that the matrix X is well-conditioned. The second row "Truth" records the percentage of times the true model is selected; i.e. the pourcentage of time we actually found $\hat{J} = J$. The third row, labelled "Inclusions", records the number of variables selected, average over 500 replications. "Correct inclusions" records the number of relevant variables that are included in the selected model, average too. The MSE (Mean Squared Error) is calculated by average over all simulations: $MSE = \sum_{i=1}^{n}(\hat{Y}_i - (X\beta_J)_i)/n$, where $\hat{Y} = X\hat{\beta}$, with $\hat{\beta}$ an estimation of β with non zero values only on $\hat{J}$.

First, concerning the ordered case procedure, Tables A.1-A.3 all show that this procedure gave excellent results, even in the high-dimensional case with a higher number of variables than the number of observations (Table A.3). These results are not surprising because our choices of β verified condition (R) of Theorem 2.1 with a very small γ, so the probability of wrongly estimating k_0 was almost reduced to α.

Concerning all the other methods tested, Table A.1 shows that the FDR procedure performed slightly better in the orthonormal case when β_J was small. Table A.2 shows results in the common case when $(X_i)_{1\leq i\leq p}$ was not an orthonormal family. Our procedure with the Bolasso order gave the best results, especially compared to the FDR procedure which gave weak results.

Table A.3 focuses on the main aim of this paper, the high-dimensional case. We chose two alternatives for the number of variables, $p = 300$ and $p = 600$. The table shows that the FDR2 was far from satisfactory. Indeed, nearly no true model were recovered in the 500 simulations. In fact, Table A.3 shows that our "procbol" procedure outperformed the others when $p >> n$. However, a combination of a small β_J and a high number of variables induced a high $\hat{\delta}$ and consequently decreased the power of our "procbol" method. Moreover, the results of the "procbol" method become less satisfactory with an increase on the value of k_0 because of the overestimation of the statistics in Lemma 3.5.

5. Conclusion

This paper tackled the problem of recovering the set of relevant variables J in a sparse linear model, especially when the number of variables p was higher than the sample size n. We proposed three new methods based on hypotheses testing to estimate J: one when the variables were ordered and two when they were not; one if the variance is known and the other when the variance is unknown. The three procedures are proved to be powerful under some conditions on the data. The simulations showed that these new procedures outperformed all the other methods tested in a common case but also in the high-dimensional case, which was the aim of this study. For instance, a method commonly used in applied sciences gave inaccurate results in simulation. Finally, a crucial point in these new methods remains the way to order variables. To improve the two-step procedure presented in this paper, a better way to order variables than the Bolasso technique needs to be found.

Appendix A. Simulation results

Results	pre-ordered		procpval		procbol		FDR		Lasso	Bolasso
	α=0.1	α=0.05	α=0.1	α=0.05	α=0.1	α=0.05	q=0.1	q=0.05		
$k_0 = 10, n = 100, p = 80, \beta = \sqrt{n}, m = 0.01$										
			$\hat{\delta} = 0.00$		$\hat{\delta} = 0.00$		$\hat{\delta} = 0.00$			
Truth	0.89	0.95	0.98	0.99	0.98	0.99	0.88	0.95	0.80	0.75
Inclusions	10.58	10.22	10.06	10.01	10.05	10.02	10.16	10.07	10.93	10.80
Correct incl.	10.00	10.00	10.00	9.99	10.00	10.00	10.00	10.00	10.00	10.00
MSE	0.11	0.11	0.11	0.10	0.10	0.10	0.11	0.11	0.15	0.15
$k_0 = 5, n = 100, p = 80, \beta = 6, m = 0.01$										
			$\hat{\delta} = 0.00$		$\hat{\delta} = 0.01$		$\hat{\delta} = 0.01$			
Truth	0.89	0.96	0.75	0.70	0.80	0.76	0.81	0.78	0.73	0.72
Inclusions	5.6	5.19	4.82	4.68	4.89	4.78	4.97	4.80	5.91	5.82
Correct incl.	5.00	5.00	4.77	4.67	4.82	4.74	4.85	4.74	4.96	4.99
MSE	0.06	0.06	0.13	0.16	0.11	0.14	0.11	0.15	0.12	0.10

Table A.1: The orthonormal case

Results	pre-ordered		procpval		procbol		FDR		Lasso	Bolasso
	α=0.1	α=0.05	α=0.1	α=0.05	α=0.1	α=0.05	q=0.1	q=0.05		
$k_0 = 10, n = 100, p = 80, \beta = \sqrt{n}, m = 0.102$										
			$\hat{\delta} = 0.46$		$\hat{\delta} = 0.00$		$\hat{\delta} = 0.45$			
Truth	0.92	0.96	0.54	0.54	0.94	0.96	0.13	0.10	0.29	0.67
Inclusions	10.33	10.15	12.06	11.62	10.08	10.05	7.55	6.60	12.18	10.70
Correct incl.	10.00	10.00	9.92	9.90	10.00	10.00	7.34	6.53	10.00	9.99
MSE	0.12	0.11	0.20	0.22	0.11	0.11	2.97	3.72	0.18	0.14
$k_0 = 5, n = 100, p = 80, \beta = 6, m = 0.103$										
			$\hat{\delta} = 0.88$		$\hat{\delta} = 0.07$		$\hat{\delta} = 0.82$			
Truth	0.91	0.95	0.11	0.11	0.86	0.84	0.00	0.00	0.27	0.47
Inclusions	5.37	5.13	6.30	5.54	5.00	4.94	0.98	0.66	7.22	6.14
Correct incl.	5.00	5.00	4.05	3.90	4.91	4.87	0.86	0.62	4.94	4.94
MSE	0.06	0.06	0.40	0.44	0.08	0.09	1.42	1.45	0.16	0.13
$k_0 = 10, n = 500, p = 450, \beta = \sqrt{n}, m = 0.040$										
			$\hat{\delta} = 0.02$		$\hat{\delta} = 0.00$		$\hat{\delta} = 0.01$			
Truth	0.91	0.95	0.98	0.98	0.94	0.96	0.84	0.85	0.88	0.99
Inclusions	11.09	10.32	10.05	10.05	10.07	10.04	10.12	9.99	10.26	10.01
Correct incl.	10.00	10.00	10.00	10.00	10.00	10.00	9.94	9.90	10.00	10.00
MSE	0.02	0.02	0.02	0.02	0.02	0.02	0.30	0.31	0.02	0.02
$k_0 = 5, n = 500, p = 450, \beta = 6, m = 0.044$										
			$\hat{\delta} = 1.00$		$\hat{\delta} = 0.07$		$\hat{\delta} = 1.00$			
Truth	0.89	0.95	0.00	0.00	0.86	0.84	0.00	0.00	0.68	0.27
Inclusions	7.35	6.06	2.19	1.68	4.95	4.88	0.09	0.05	5.37	3.78
Correct incl.	5.00	5.00	1.22	1.16	4.90	4.85	0.07	0.05	4.91	3.78
MSE	0.02	0.01	0.27	0.28	0.02	0.02	0.36	0.36	0.03	0.09

Table A.2: The non orthonormal case, $p < n$

Results	pre-ordered		procpval		procbol		FDR2		Lasso	Bolasso
	α=0.1	α=0.05	α=0.1	α=0.05	α=0.1	α=0.05	q=0.1	q=0.05		
$k_0 = 10, n = 100, p = 300, \beta = \sqrt{n}$										
			$\hat{\delta} = 1.00$		$\hat{\delta} = 0.00$		$\hat{\delta} = 1.00$			
Truth	0.91	0.96	0.00	0.00	0.99	0.99	0.00	0.00	0.60	0.78
Inclusions	10.53	10.14	8.92	8.68	10.01	10.01	4.17	3.38	11.05	10.46
Correct incl.	10.00	10.00	8.35	8.36	10.00	10.00	4.17	3.38	10.00	10.00
MSE	0.11	0.10	1.56	1.63	0.10	0.10	5.21	6.04	0.15	0.13
$k_0 = 5, n = 100, p = 300, \beta = 6$										
			$\hat{\delta} = 0.65$		$\hat{\delta} = 0.09$		$\hat{\delta} = 0.60$			
Truth	0.93	0.96	0.33	0.33	0.79	0.74	0.03	0.01	0.38	0.56
Inclusions	5.438	5.16	4.62	4.50	4.88	4.78	3.22	2.74	7.57	6.32
Correct incl.	5.00	5.00	4.32	4.24	4.82	4.74	3.15	2.71	4.92	4.90
MSE	0.06	0.05	0.27	0.29	0.11	0.14	0.66	0.79	0.18	0.15
$k_0 = 10, n = 100, p = 600, \beta = \sqrt{n}$										
			$\hat{\delta} = 1.00$		$\hat{\delta} = 0.17$		$\hat{\delta} = 1.00$			
Truth	0.89	0.95	0.00	0.00	0.83	0.83	0.00	0.00	0.00	0.25
Inclusions	10.66	10.21	4.88	4.36	10.30	10.20	2.33	2.02	16.97	12.24
Correct incl.	10.00	10.00	4.68	4.23	9.99	9.99	2.33	2.02	9.99	9.99
MSE	0.12	0.11	4.11	4.56	0.11	0.11	6.34	6.69	0.31	0.20
$k_0 = 5, n = 100, p = 600, \beta = 6$										
			$\hat{\delta} = 0.95$		$\hat{\delta} = 0.30$		$\hat{\delta} = 0.92$			
Truth	0.912	0.96	0.05	0.05	0.62	0.56	0.00	0.00	0.11	0.26
Inclusions	5.43	5.12	3.36	3.22	4.62	4.48	1.48	1.18	10.52	7.49
Correct incl.	5.00	5.00	3.14	3.04	4.50	4.39	1.46	1.17	4.59	4.65
MSE	0.06	0.06	0.59	0.62	0.22	0.25	1.10	1.22	0.37	0.30

Table A.3: The high-dimensional case, $p \geq n$

Appendix B. Proofs

Proof of Theorem 2.1. Let $k \leq k_0 - 1$ and assume that (R_k) holds. According to Baraud et al. [2], the power of the test H_k, $\mathbb{P}_\mu(T_{k,\alpha} > 0)$, is greater than $1 - \gamma/k_0$. This is equivalent to $\mathbb{P}_\mu(H_k \text{ is accepted}) \leq \gamma/k_0$.
Moreover, for all $k \geq k_0$, $\mathbb{P}_\mu(T_{k,\alpha} > 0) \leq \alpha$, since α is the level of the test H_k.
Then we have:

$$\mathbb{P}_\mu(\hat{k} > k_0) \leq \mathbb{P}_\mu(H_{k_0} \text{ is rejected}) = \mathbb{P}_\mu(T_{k_0,\alpha} > 0) \leq \alpha$$

and

$$\begin{aligned} \mathbb{P}_\mu(\hat{k} < k_0) &\leq \sum_{j=0}^{k_0-1} \mathbb{P}_\mu(H_j \text{ is accepted}) \\ &\leq k_0\gamma/k_0. \end{aligned}$$

Hence we obtain

$$\mathbb{P}_\mu(\hat{k} \neq k_0) \leq \mathbb{P}_\mu(\hat{k} < k_0) + \mathbb{P}_\mu(\hat{k} > k_0) \leq \gamma + \alpha$$

which concludes the proof of (5).

□

Proof of Corollary 2.3. Let $p < n$ and $k_0 \leq p$ such that $p = An$, where $A < 1$. We set $\alpha_n = \gamma_n = 1/n$. We set $\forall k, \forall t,\ \alpha_t = \dfrac{\alpha_n}{\log_2(n)}$, thus $\sum_{t\in\mathcal{T}} \alpha_t \leq \alpha_n$. With these choices, we have that the conditions of Remark 2.2 are verified. Indeed $\alpha_t = \dfrac{\log_2(n)}{n} \geq \exp(-N_{k,t}/10)$ because $k + 2^t \leq p = An$. Moreover $\gamma_n = 1/n \geq 2\exp(-N_{k,t}/21)$ and the ratio $\dfrac{D_{k,t} + L_{k,t}}{N_{k,t}} = \dfrac{2^t + log(1/\alpha_t)}{n - k - 2^t} \leq \dfrac{An + log(1/\alpha_t)}{(1-A)n}$ remains bounded.

With these conditions $C_2(k,t)$ and $C_3(k,t)$ behave like constants, and thus for $t = \min(\lfloor log_2(p-k) \rfloor, \inf\{t, 2^t \geq k_0\})$ the condition (R_k) is verified for all $k \leq k_0 - 1$ and n large enough:

$\left\|\Pi_{V_{k,t}^{\perp}}(\mu)\right\|_n^2 = 0$ and $\dfrac{\sigma^2}{n}\left[C_2(k,t)\sqrt{2^t\ log\left(\dfrac{2k_0}{\alpha_t\gamma}\right)} + C_3(k,t) log\left(\dfrac{2k_0}{\alpha_t\gamma}\right)\right] \underset{n\to\infty}{\longrightarrow} 0$, thus Theorem 2.1 can be applied and $\mathbb{P}_\mu(\hat{k} \neq k_0) \leq \gamma_n + \alpha_n$. In particular, $\mathbb{P}_\mu(\hat{k} \neq k_0) \underset{n\to\infty}{\longrightarrow} 0$.

□

Proof of Theorem 3.2. Let $k < k_0$.

We use the identity $\forall (a,b) \in \mathbb{R}^2, (a+b)^2 \geq \dfrac{1}{2}a^2 - b^2$. On the event A_{k_0}: $\forall t \in I = \{0, ..., log_2(k_0 - k)\}$:

$$\begin{aligned} \|\Pi_{S_{(k),(t)}} Y\|_n^2 &= \|\Pi_{S_{(k),(t)}}(\mu + \epsilon)\|_n^2 \\ &\geq \frac{1}{2}\|\Pi_{S_{(k),(t)}}\mu\|_n^2 - \|\Pi_{S_{(k),(t)}}\epsilon\|_n^2 \\ &\geq \frac{1}{2}\inf\left\{\|\Pi_S\mu\|_n^2, S \in B_{2^t}\right\} - \|\Pi_{S_{(k),(t)}}\epsilon\|_n^2 \end{aligned}$$

where $B_{2^t} = \{\text{span}(X_I), I \subset J, |I| = 2^t\}$. Hence:

$$\begin{aligned} &\mathbb{P}\left(\forall t \in I, \frac{1}{\sigma^2}\|\Pi_{S_{(k),(t)}} Y\|_n^2 \leq \overline{U_{k,t}^1}^{-1}(\alpha_t) \cap A_{k_0}\right) \\ &= \mathbb{P}\left(\forall t \in I, \frac{1}{\sigma^2}\|\Pi_{S_{(k),(t)}}(\mu+\epsilon)\|_n^2 \leq \overline{U_{k,t}^1}^{-1}(\alpha_t) \cap A_{k_0}\right) \\ &\leq \mathbb{P}\left(\forall t \in I, \frac{1}{2\sigma^2}\inf\left\{\|\Pi_S\mu\|_n^2, S \in B_{2^t}\right\} - \frac{1}{\sigma^2}\|\Pi_{S_{(k),(t)}}\epsilon\|_n^2 \leq \overline{U_{k,t}^1}^{-1}(\alpha_t) \cap A_{k_0}\right) \end{aligned}$$

We have on the event A_{k_0} and for $k + 2^t \leq k_0$ that $\|\Pi_{S_{(k),(t)}}\epsilon\|_n^2 \leq \sup\left\{\|\Pi_S\epsilon\|_n^2, S \in B_{2^t}\right\}$. Moreover, for $S \in B_{2^t}$, $\|\Pi_S\epsilon\|_n^2 \sim \dfrac{\sigma^2}{n}\chi^2_{2^t}$. Note that $|B_{2^t}| = \binom{k_0}{2^t}$. Let denote $Z_t = \dfrac{\|\Pi_{S_{(k),(t)}}\epsilon\|_n^2}{\sigma^2}$ and $\bar{Z}_t(u)$ the probability for the statistic Z_t to be larger than u. We denote $\bar{\chi}_d(u)$ the probability for a chi-square with d degrees of freedom to be larger than u. We have an upper bound of the $(1-u)$-quantile of the statistic Z_t: $\bar{Z}_t^{-1}(u) \leq \bar{\chi}^{-1}_{2^t}(u/|B_{2^t}|)/n$. Indeed:

$$\begin{aligned}\mathbb{P}\left(Z_t > \frac{\overline{\chi^2_{2^t}}^{-1}(u/|B_{2^t}|)}{n}\right) &\leq \mathbb{P}\left(\sup\left\{\frac{||\Pi_S \epsilon||_n^2}{\sigma^2}, S \in B_{2^t}\right\} > \frac{\overline{\chi^2_{2^t}}^{-1}(u/|B_{2^t}|)}{n}\right)\\ &\leq \sum_{S\in B_{2^t}} \mathbb{P}\left(||\Pi_S \epsilon||_n^2 > \frac{\sigma^2}{n}\overline{\chi^2_{2^t}}^{-1}(u/|B_{2^t}|)\right)\\ &\leq |B_{2^t}|\frac{u}{|B_{2^t}|} \leq u.\end{aligned}$$

Therefore, the following condition

$$(cond_k): \qquad \exists t \in I, \frac{1}{2\sigma^2}\inf\left\{||\Pi_S \mu||_n^2, S \in B_{2^t}\right\} \geq \frac{1}{n}\overline{\chi^2_{2^t}}^{-1}\left(\frac{\gamma/k_0}{|B_{2^t}|}\right) + \overline{U^1_{k,t}}^{-1}(\alpha_t)$$

implies that:

$$\mathbb{P}\left(\forall t \in I, \frac{1}{2\sigma^2}\inf\left\{||\Pi_S \mu||_n^2, S \in B_{2^t}\right\} - \frac{1}{\sigma^2}||\Pi_{S_{(k),(t)}}\epsilon||_n^2 \leq \overline{U^1_{k,t}}^{-1}(\alpha_t) \cap A_{k_0}\right) \leq \gamma/k_0. \tag{B.1}$$

Let us denote $\forall 0 < d$,

$$G_{k,d} = \{\text{span}(X_I), I \subset \{1,..,p\}\backslash\{(1),...,(k)\}, |I| = d\}. \tag{B.2}$$

Note that $|G_{k,d}| = \binom{p-k}{d}$. Then $U^1_{k,t} \leq \sup\left\{||\Pi_S \epsilon||_n^2, S \in G_{k,2^t}\right\}$. This inequality leads us to an upper bound of the (1-u)-quantile of $U^1_{k,t}$: $\overline{U^1_{k,t}}^{-1}(u) \leq \overline{\chi^2_{2^t}}^{-1}(u/|G_{k,2^t}|)/n$.

Using $\overline{U^1_{k,t}}^{-1}(u) \leq \overline{\chi^2_{2^t}}^{-1}(u/|G_{k,2^t}|)/n$ in the condition $(cond_k)$, we obtain the condition $(cond_{2,k})$ which still implies (B.1):

$$(cond_{2,k}) \qquad \exists t \in I, \frac{1}{2\sigma^2}\inf\left\{||\Pi_S \mu||_n^2, S \in B_{2^t}\right\} \geq \frac{1}{n}\left[\overline{\chi^2_{2^t}}^{-1}\left(\frac{\gamma/k_0}{|B_{2^t}|}\right) + \overline{\chi^2_{2^t}}^{-1}\left(\frac{\alpha_t}{|G_{k,2^t}|}\right)\right].$$

Moreover, Laurent and Massart [10] showed that for $K \sim \chi^2_d$:

$$\mathbb{P}\left(K \geq d + 2\sqrt{dx} + 2x\right) \leq e^{-x}. \tag{B.3}$$

Then for $d = 2^t$ and $x_u = log\left(\frac{|B_{2^t}|}{\gamma/k_0}\right)$ we have $\overline{\chi^2_{2^t}}^{-1}\left(\frac{\gamma/k_0}{|B_{2^t}|}\right) \leq 2^t + 2\sqrt{2^t x_u} + 2x_u$. Since $\binom{D}{d} \leq \left(\frac{eD}{d}\right)^d$, $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$ for all $u > 0, v > 0$ and since $\forall 1 < u$, $\sqrt{u} \leq u$, we obtain:
$x_u = 2^t log\left(\frac{ek_0}{2^t}\right) + log\left(\frac{k_0}{\gamma}\right)$, thus

$$\begin{aligned}\overline{\chi^2_{2^t}}^{-1}\left(\frac{\gamma/k_0}{|B_{2^t}|}\right) &\leq 2^t\left[1 + 2\sqrt{log\left(\frac{ek_0}{2^t}\right)} + 2log\left(\frac{ek_0}{2^t}\right)\right] + 2\left[\sqrt{2^t log(k_0/\gamma)} + log(k_0/\gamma)\right]\\ &\leq 2^t\left[5 + 4log\left(\frac{k_0}{2^t}\right)\right] + 2\left[\sqrt{2^t log(k_0/\gamma)} + log(k_0/\gamma)\right].\end{aligned}$$

For $d = 2^t$ and $x_u = log(|G_{k,2^t}|/\alpha_t)$, we obtain:

$$\begin{aligned}\overline{\chi}_{2^t}^{-1}(\alpha_t/|G_{k,2^t}|) &\leq 2^t\left[1 + 2\sqrt{log\left(\frac{e(p-k)}{2^t}\right)} + 2log\left(\frac{e(p-k)}{2^t}\right)\right] + 2\left[\sqrt{2^t log(1/\alpha_t)} + log(1/\alpha_t)\right] \\ &\leq 2^t\left[5 + 4log\left(\frac{p-k}{2^t}\right)\right] + 2\left[\sqrt{2^t log(1/\alpha_t)} + log(1/\alpha_t)\right].\end{aligned}$$

We also have an upper bound of $1/\alpha_t, \forall t \in \mathcal{T}_k$. Indeed , the construction of $\{\alpha_t, t \in \mathcal{T}_k\}$ with the procedure P3 gives that $\mathbb{P}\left(\exists t \in \mathcal{T}_k, U_{k,t}^1 > \overline{U_{k,t}^1}^{-1}(\alpha_t)\right) = \alpha$. Thus $\forall t \in \mathcal{T}_k, \alpha_t \geq \alpha/|\mathcal{T}_k|$, since $\mathbb{P}\left(\exists t \in \mathcal{T}_k, U_{k,t}^1 > \overline{U_{k,t}^1}^{-1}(\alpha/|\mathcal{T}_k|)\right) \leq \alpha$.
Hence we obtain:

$$\overline{\chi}_{2^t}^{-1}(\alpha_t/|G_{k,2^t}|) \leq 2^t\left[5 + 4log\left(\frac{p-k}{2^t}\right)\right] + 2\left[\sqrt{2^t log\left(\frac{|\mathcal{T}_k|}{\alpha}\right)} + log\left(\frac{|\mathcal{T}_k|}{\alpha}\right)\right].$$

Using the inequality $a\sqrt{u} + b\sqrt{v} \leq \sqrt{a^2+b^2}\sqrt{u+v}$ which holds for any positive numbers a, b, u, v, we finally get the condition $(R_{2,k})$ which implies (B.1):
$(R_{2,k})$: $\exists t \in I$ such that

$$\frac{1}{2\sigma^2}\inf\left\{||\Pi_S\mu||_n^2, S \in B_{2^t}\right\} \geq \frac{2^t}{n}\left[10 + 4log\left(\frac{(p-k)k_0}{2^{2t}}\right)\right] + \frac{2}{n}\left[\sqrt{2^{t+1}log\left(\frac{k_0|\mathcal{T}_k|}{\gamma\alpha}\right)} + log\left(\frac{k_0|\mathcal{T}_k|}{\gamma\alpha}\right)\right].$$

This leads to

$$\mathbb{P}\left(\forall t \in I, \frac{1}{\sigma^2}||\Pi_{S_{(k),(t)}}Y||_n^2 \leq \overline{U_{k,t}^1}^{-1}(\alpha_t) \cap A_{k_0}\right) \leq \gamma/k_0.$$

Hence

$$\mathbb{P}\left(\forall t \in I, U_{k,t} \leq \overline{U_{k,t}^1}^{-1}(\alpha_t) \cap A_{k_0}\right) \leq \gamma/k_0.$$

Then, $\forall k < k_0, \mathbb{P}\left(\mathring{k}_{Abis} = k \cap A_{k_0}\right) \leq \gamma/k_0$, where $\mathring{k}_{Abis} = |\hat{J}|$.

We can calculate $\mathbb{P}_\mu(\hat{J} \neq J)$:

$$\begin{aligned}\mathbb{P}_\mu(\hat{J} \neq J) &\leq \mathbb{P}_\mu(\hat{J} \neq J \cap A_{k_0}) + \mathbb{P}(A_{k_0}^c) \\ &\leq \left(\sum_{j=0}^{k_0-1}\mathbb{P}_\mu(\mathring{k}_{Abis} = j \cap A_{k_0}) + \mathbb{P}_\mu(\mathring{k}_{Abis} > k_0 \cap A_{k_0})\right) + \mathbb{P}(A_{k_0}^c) \\ &\leq k_0\gamma/k_0 + \alpha + \delta.\end{aligned}$$

And then (12) is proved. □

Proof of Lemma 3.3. Under $\hat{H}_k$ and on the event A_k :
$U_{k,t} = ||\Pi_{S_{(k),(t)}}Y||_n^2/\sigma^2 = ||\Pi_{S_{(k),(t)}}(\mu+\epsilon)||_n^2/\sigma^2 = ||\Pi_{S_{(k),(t)}}\epsilon||_n^2/\sigma^2$
The family $(X_i)_i$ is orthonormal, thus: $U_{k,t} = \sum_{j=k+1}^{k+2^t} < \epsilon, X_{(j)} >_n^2 /\sigma^2$.
As $\epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$, we have $\forall 1 \leq j \leq p, < \epsilon, X_j > \sim \mathcal{N}(0, \sigma^2)$ and the variables $< \epsilon, X_j >, j = 1, ..., p$ are i.i.d.. Thus $\{< \epsilon, X_{(j)} >, j > k\} = \{< \epsilon, X_m >, m \notin J\} = \{\sigma W_1, ..., \sigma W_{p-k}\}$.

So $\sum_{j=k+1}^{k+2^t} < \epsilon, X_{(j)} >_n^2 /\sigma^2 \leq \sum_{j=1}^{2^t} W_{(j)}^2/n = Z_{k,D_{k,t}}/n$.

□

Proof of Corollary 3.4. Let $k < k_0$.
σ_2 is defined such that $|\beta_{\sigma_2(1)}| \leq ... \leq |\beta_{\sigma_2(k_0)}|$, note $\epsilon_{(j+1)} = ||\Pi_{S_{(j),0}}\epsilon||$, $\forall\, j \in \{k+1, ..., k+2^t\}$ with $k+2^t \leq k_0$.
Similarly as in the proof of Theorem 3.2, using the equality $\inf\left\{||\Pi_S\mu||_n^2, S \in B_{2^t}\right\} = \sum_{j=1}^{2^t}\beta_{\sigma_2(j)}^2$, we get that:

$$\mathbb{P}\left(\forall t \in I, \frac{1}{\sigma^2}||\Pi_{S_{(k),(t)}}Y||_n^2 \leq \frac{\bar{Z}^{-1}_{D_{k,t},p-k}(\alpha_t)}{n} \cap A_{k_0}\right)$$

$$\leq \mathbb{P}\left(\forall t \in I, \frac{1}{2\sigma^2}\sum_{j=1}^{2^t}\beta_{\sigma_2(j)}^2 - \frac{1}{n\sigma^2}\sum_{j=k}^{k+2^t-1}\epsilon_{(j+1)}^2 \leq \frac{\bar{Z}^{-1}_{D_{k,t},p-k}(\alpha_t)}{n} \cap A_{k_0}\right)$$

On the event A_{k_0}, $\{< \epsilon, X_{(j+1)} >, k \leq j \leq k + 2^t - 1\} \subset \{< \epsilon, X_j >, j \in J\}$, which implies that we have an stochastic upper bound: $\sum_{j=k}^{k+2^t-1}\epsilon_{(j+1)}^2 \leq \sigma^2 Z_{2^t,k_0}$.
Hence the following condition

$$(cond_{3,k}) : \exists\, t \leq log_2(k_0 - k)/\frac{1}{2\sigma^2}\sum_{j=1}^{2^t}\beta_{\sigma_2(j)}^2 \geq \frac{1}{n}\left[\bar{Z}^{-1}_{D_{k,t},p-k}(\alpha_t) + \bar{Z}^{-1}_{D_{k,t},k_0}(\gamma/k_0)\right]$$

implies that

$$\mathbb{P}\left(\forall t \in I, \frac{1}{2\sigma^2}\sum_{j=1}^{2^t}\beta_{\sigma_2(j)}^2 - \frac{1}{n\sigma^2}\sum_{j=k}^{k+2^t-1}\epsilon_{(j+1)}^2 \leq \frac{\bar{Z}^{-1}_{D_{k,t},p-k}(\alpha_t)}{n} \cap A_{k_0}\right) \leq \gamma/k_0.$$

This leads to

$$\mathbb{P}\left(\forall t \in I, \frac{1}{\sigma^2}||\Pi_{S_{(k),(t)}}Y||_n^2 \leq \bar{Z}^{-1}_{D_{k,t},p-k}(\alpha_t) \cap A_{k_0}\right) \leq \gamma/k_0. \tag{B.4}$$

Let $0 < u < 1$, $0 < D$ and $d < D$. In the following, we study the behavior of the $(1-u)$ quantile of the statistic $Z_{d,D}$ in order to obtain a more explicit condition than $(cond_{3,k})$.
Let define $V_{d,D} = \{I \subset \{1, ..., D\}/|I| = d\}$. Note that $|V_{d,D}| = \binom{D}{d}$. Let recall that $Z_{d,D}$ is defined by (13) as $Z_{d,D} = \sum_{j=1}^{d} W_{(j)}^2$ where $W_1, ..., W_D$ are D i.i.d. standard Gaussian variables ordered as $|W_{(1)}| \geq ... \geq |W_{(D)}|$. We have that: $Z_{d,D} \leq \sup\left\{\sum_{i\in I} W_i^2, I \in V_{d,D}\right\}$. Note that for $I \in V_{d,D}$, $\sum_{i\in I} W_i^2, \sim \chi_d^2$.
We obtain that the $(1-u)$-quantile of $Z_{d,D}$ is lower than $\bar{\chi_d}^{-1}(u/|V_{d,D}|)$:

$$\begin{aligned}
\mathbb{P}\left(Z_{d,D} > \bar{\chi_d}^{-1}(u/|V_{d,D}|)\right) &\leq \mathbb{P}\left(\sup\left\{\sum_{i\in I} W_i^2, \forall I \in V_{d,D}\right\} > \bar{\chi_d}^{-1}(u/|V_{d,D}|)\right) \\
&\leq \sum_{I\in V_{d,D}} \mathbb{P}\left(\sum_{i\in I} W_i^2 > \bar{\chi_d}^{-1}(u/|V_{d,D}|)\right) \\
&\leq |V_{d,D}|\frac{u}{|V_{d,D}|} \leq u.
\end{aligned}$$

Using the expression of the upper bound of $\bar{\chi_d}^{-1}(u)$ from the proof of Theorem 3.2, we get the condition $(R_{2bis,k})$ from an upper bound of the right part in the condition $(cond_{3,k})$. The end of the proof is the same as in the proof of Theorem 3.2. □

Proof of Theorem 3.6. Let $k < k_0$ and $0 < \gamma < 1$. Denote $I = \{0, ..., \lfloor log_2(k_0 - k) \rfloor\}$.
From the proof of Theorem 3.2 (more precisely the condition $(cond_k)$), we have that if the following condition is verified:

$$\exists t \in I / \frac{1}{2}\inf\left\{||\Pi_S \mu||_n^2, S \in B_{2^t}\right\} \geq \bar{\Upsilon}_{k,t}^{-1}(\alpha_t) Q_{1-\gamma/2k_0} \frac{D_{k,t}}{N_{k,t}} + \frac{\sigma^2}{n} \bar{\chi}_{2^t}^{-1}\left(\frac{\gamma/2k_0}{|B_{2^t}|}\right) \tag{B.5}$$

where Q_{1-u} denote the $(1-u)$-quantile of the statistics $||Y - \Pi_{V_{(k),(t)}} Y||_n^2$ under the event A_{k_0}, then we have:

$$\mathbb{P}\left(\forall t \in I, ||\Pi_{S_{(k),(t)}} Y||_n^2 \leq \bar{\Upsilon}_{k,t}^{-1}(\alpha_t) Q_{1-\gamma/2k_0} \frac{D_{k,t}}{N_{k,t}} \cap A_{k_0}\right) \leq \gamma/2k_0.$$

Since $\mathbb{P}\left(\forall t \in I, \tilde{U}_{D_{k,t},N_{k,t}} < \bar{\Upsilon}_{k,t}^{-1}(\alpha_t) \cap A_{k_0}\right) \leq \inf_{t \in I}\left\{\mathbb{P}\left(\tilde{U}_{D_{k,t},N_{k,t}} < \bar{\Upsilon}_{k,t}^{-1}(\alpha_t) \cap A_{k_0}\right)\right\}$ and since

$$\begin{aligned}
\mathbb{P}\left(\tilde{U}_{D_{k,t},N_{k,t}} < \bar{\Upsilon}_{k,t}^{-1}(\alpha_t) \cap A_{k_0}\right) &\leq \underbrace{\mathbb{P}\left(||Y - \Pi_{V_{(k),(t)}} Y||_n^2 > Q_{1-\gamma/2k_0} \cap A_{k_0}\right)}_{\leq \gamma/2k_0} \\
&\quad + \mathbb{P}\left(\frac{||\Pi_{S_{(k),(t)}} Y||_n^2}{D_{k,t}} \leq \bar{\Upsilon}_{k,t}^{-1}(\alpha_t) \frac{Q_{1-\gamma/2k_0}}{N_{k,t}} \cap A_{k_0}\right) \\
&\leq \gamma/k_0.
\end{aligned}$$

we have that the condition (B.5) implies that

$$\mathbb{P}\left(\forall t \in I, \tilde{U}_{D_{k,t},N_{k,t}} < \bar{\Upsilon}_{k,t}^{-1}(\alpha_t) \cap A_{k_0}\right) \leq \gamma/k_0. \tag{B.6}$$

In the following, we give an upper bound of the right part in (B.5). For this doing, we have to give an upper bound of $\bar{\Upsilon}_{k,t}^{-1}(\alpha_t)$ and $Q_{1-\gamma/2k_0}$.

Assume we are on the event A_k and under $\hat{H}_k$, then

$$\Upsilon_{k,t} = \frac{N_{k,t}||\Pi_{S_{(k),\sigma_1(t)}} Y||_n^2}{D_{k,t}||Y - \Pi_{V_{(k),\sigma_1(t)}} Y||_n^2} = \frac{N_{k,t}||\Pi_{S_{(k),\sigma_1(t)}} \epsilon||_n^2}{D_{k,t}||Y - \Pi_{V_{(k)}} Y - \Pi_{S_{(k),\sigma_1(t)}} \epsilon||_n^2}.$$

As we are on the event A_k and under $\hat{H}_k$, the space $V_{(k)}$ is not a random space. Thus for any subspaces S of dimension $D_{k,t} = 2^t$, we have that $||\Pi_S Y||_n^2 = ||\Pi_S \epsilon||_n^2 \sim \sigma^2 \chi_{2^t}^2/n$ and we have that $||Y - \Pi_{V_{(k)}} Y - \Pi_S Y||_n^2 = ||\Pi_{(S \oplus V_{(k)})^\perp} \epsilon||_n^2 \sim \sigma^2 \chi_{n-(2^t+k)}^2/n$. Hence $\frac{N_{k,t}||\Pi_S Y||_n^2}{D_{k,t}||Y - \Pi_{V_{(k)}} Y - \Pi_S Y||_n^2} \sim F_{D_{k,t},N_{k,t}}$.
Thus on the event A_k and under $\hat{H}_k$, $\Upsilon_{k,t} \leq \sup\left\{\frac{N_{k,t}||\Pi_S \epsilon||_n^2}{D_{k,t}||\epsilon - \Pi_{V_{(k)}+S} \epsilon||_n^2}, S \in G_{k,2^t}\right\}$, where $G_{k,2^t}$ is defined by (B.2).
We deduce that the $(1-u)$-quantile of $\Upsilon_{k,t}$ is lower that $\bar{F}_{D_{k,t},N_{k,t}}^{-1}(u/|G_{k,2^t}|)$. Indeed:

$$\begin{aligned}
\mathbb{P}\left(\Upsilon_{k,t} > \bar{F}_{D_{k,t},N_{k,t}}^{-1}(u/|G_{k,2^t}|)\right) &\leq \mathbb{P}\left(\sup\left\{\frac{N_{k,t}||\Pi_S \epsilon||_n^2}{D_{k,t}||\epsilon - \Pi_{V_{(k)}+S} \epsilon||_n^2}, S \in G_{k,2^t}\right\} > \bar{F}_{D_{k,t},N_{k,t}}^{-1}(u/|G_{k,2^t}|)\right) \\
&\leq \sum_{S \in G_{k,2^t}} \mathbb{P}\left(\frac{N_{k,t}||\Pi_S \epsilon||_n^2}{D_{k,t}||\epsilon - \Pi_{V_{(k)}+S} \epsilon||_n^2} > \bar{F}_{D_{k,t},N_{k,t}}^{-1}(u/|G_{k,2^t}|)\right) \\
&\leq |G_{k,2^t}| \frac{u}{|G_{k,2^t}|} \leq u.
\end{aligned}$$

Baraud et al. [2] gave an upper bound of $\overline{F}^{-1}_{D,N}(u)$, for $0 < D$,$0 < N$ and $0 < u$:

$$D\overline{F}^{-1}_{D,N}(u) \leq D + 2\sqrt{D\left(1+\frac{D}{N}\right)log\left(\frac{1}{u}\right)} + \left(1+2\frac{D}{N}\right)\frac{N}{2}\left[exp\left(\frac{4}{N}log\left(\frac{1}{u}\right)\right)-1\right].$$

Since $\exp(u) - 1 \leq u\exp(u)$ for any $u > 0$, $\sqrt{u+v} \leq \sqrt{u} + \sqrt{v}$ for all $u > 0, v > 0$ and since $\alpha_t \geq \alpha/|\mathcal{T}_k|$, we derive that:

$$2^t\overline{\Upsilon}^{-1}_{k,t}(\alpha_t) \leq 2^t\left[1+\Lambda_3(k,t)log\left(\frac{e(p-k)}{2^t}\right)\right] + 2\left[\sqrt{2^t\left(1+\frac{2^t}{N_{k,t}}\right)log\left(\frac{|\mathcal{T}_k|}{\alpha}\right)} + \frac{\Lambda_2(k,t)}{2}log\left(\frac{|\mathcal{T}_k|}{\alpha}\right)\right],$$

where $\Lambda_1(k,t) = \sqrt{1+\frac{D_{k,t}}{N_{k,t}}}$, $\Lambda_2(k,t) = \left(1+2\frac{D_{k,t}}{N_{k,t}}\right)M$ and $\Lambda_3(k,t) = 2\Lambda_1(k,t) + \Lambda_2(k,t)$ with $L_t = \log(|\mathcal{T}_k|/\alpha)$, $m_t = \exp(4L_t/N_{k,t})$, $m_p = \exp\left(\frac{4D_{k,t}}{N_{k,t}}log\left(\frac{e(p-k)}{2^t}\right)\right)$, $M = 2m_tm_p$.
Since $\sqrt{ab} + mb \leq a/2 + (m+1/2)b$ holds for any positive numbers a, b, m, we obtain that:

$$2^t\overline{\Upsilon}^{-1}_{k,t}(\alpha_t) \quad \leq \quad 2^t\left[1+\Lambda_1^2(k,t)+\Lambda_3(k,t)log\left(\frac{e(p-k)}{2^t}\right)\right] + (1+\Lambda_2(k,t))\,log\left(\frac{|\mathcal{T}_k|}{\alpha}\right) \quad \text{(B.7)}$$

We have now to find an upper bound of $Q_{1-\gamma/2k_0}$.
$Q_{1-\gamma/2k_0}$ is defined by $\mathbb{P}\left(||Y - \Pi_{V_{(k),(t)}}Y||_n^2 > Q_{1-\gamma/2k_0} \cap A_{k_0}\right) \leq \gamma/2k_0$.
We always have that: $||Y - \Pi_{V_{(k),(t)}}Y||_n^2 \leq ||\mu||_n^2 + ||\epsilon||_n^2$. Thus $\forall\ 0 < u < 1$, the $(1-u)$-quantile of $||Y - \Pi_{V_{(k),(t)}}Y||_n^2$ is lower than the $(1-u)$-quantile of $||\mu||_n^2 + ||\epsilon||_n^2$.
As $||\epsilon||_n^2 \sim \sigma^2\chi_n^2/n$, we can use the equation (B.3) for $x_u = log(2k_0/\gamma)$ and we obtain that $\overline{\chi}^{-1}_n(\gamma/2k_0) \leq n + 2\sqrt{nx_u} + 2x_u$.
Therefore

$$Q_{1-\gamma/2k_0} \leq ||\mu||_n^2 + \sigma^2\frac{n + 2\sqrt{nx_u} + 2x_u}{n} \quad \text{(B.8)}$$

and as $1 + 2\sqrt{u} + 2u \leq 2 + 3u$, we get

$$Q_{1-\gamma/2k_0} \leq ||\mu||_n^2 + \sigma^2\left(2 + \frac{3}{n}log\left(\frac{2k_0}{\gamma}\right)\right). \quad \text{(B.9)}$$

Combining (B.7) (B.9) in (B.5) and using that

$$\begin{aligned}\overline{\chi}^{-1}_{2^t}\left(\frac{\gamma/2k_0}{|B_{2^t}|}\right) \quad &\leq \quad 2^t\left[5+4log\left(\frac{k_0}{2^t}\right)\right] + 2\left[\sqrt{2^tlog(2k_0/\gamma)} + log(2k_0/\gamma)\right] \\ &\leq \quad 2^t\left[6+4log\left(\frac{k_0}{2^t}\right)\right] + 3log(2k_0/\gamma)\end{aligned}$$

we obtain the following condition:

$(R_{3,k})$: $\exists t \in I$ such that

$$\begin{aligned}
\frac{1}{2}\inf\left\{\|\Pi_S\mu\|_n^2, S \in B_{2^t}\right\} &\geq \frac{D_{k,t}\bar{F}^{-1}_{D_{k,t},N_{k,t}}(\alpha_t/|G_{2^t}|)}{N_{k,t}}\left[\|\mu\|_n^2 + \sigma^2\left(2 + \frac{3}{n}log\left(\frac{2k_0}{\gamma}\right)\right)\right] \\
&\quad + \frac{\sigma^2}{n}\left[2^t\left(6 + 4log\left(\frac{k_0}{2^t}\right)\right) + 3log\left(\frac{2k_0}{\gamma}\right)\right] \\
&\geq \frac{A(k,t)}{N_{k,t}}\left[\|\mu\|_n^2 + \sigma^2\left(2 + \frac{3}{n}log\left(\frac{2k_0}{\gamma}\right)\right)\right] \\
&\quad + \frac{\sigma^2}{n}\left[2^t\left(6 + 4log\left(\frac{k_0}{2^t}\right)\right) + 3log\left(\frac{2k_0}{\gamma}\right)\right],
\end{aligned}$$

where $A(k,t) = 2^t\left[2 + \frac{2^t}{N_{k,t}} + \Lambda_3(k,t)log\left(\frac{e(p-k)}{2^t}\right)\right] + (1 + \Lambda_2(k,t))\,log\left(\frac{|\mathcal{T}_k|}{\alpha}\right)$.

The condition $(R_{3,k})$ leads to (B.6) and thus

$$\begin{aligned}
\mathbb{P}_\mu(\hat{J} \neq J) &\leq \mathbb{P}_\mu(\hat{J} \neq J \cap A_{k_0}) + \mathbb{P}(A_{k_0}^c) \\
&\leq \left(\sum_{j=0}^{k_0-1}\mathbb{P}_\mu(\mathring{k}_B = j \cap A_{k_0}) + \mathbb{P}_\mu(\mathring{k}_B > k_0 \cap A_{k_0})\right) + \mathbb{P}(A_{k_0}^c) \\
&\leq k_0\gamma/k_0 + \alpha + \delta.
\end{aligned}$$

And then (19) is proved. □

Proof of Remark 3.7. In the following, $C(a,b)$ denote a constant depending on the parameters a and b. Under the assumption that $2^t \leq (n-k)/2$ and since $\forall\ x \geq 2$, $\frac{log(x)}{x} \leq 1$ we have that:

$$\frac{D_{k,t}}{N_{k,t}}log\left(\frac{p-k}{D_{k,t}}\right) \leq \frac{2^t}{n-k-2^t}log\left(\frac{n-k}{2^t}\right) \leq 2\frac{2^t}{n-k}log\left(\frac{n-k}{2^t}\right) \leq 2.$$

Moreover the ratio $D_{k,t}/N_{k,t}$ is bounded by 1, thus $log(m_p) \leq 4\frac{D_{k,t}}{N_{k,t}} + 4\frac{D_{k,t}}{N_{k,t}}log\left(\frac{p-k}{D_{k,t}}\right) \leq 12$.
As the ratio $4L_{k,t}/N_{k,t}$ is bounded by $C'(\alpha)$ and since $M \leq 2\exp(C'(\alpha))\exp(12)$, we have that M is bounded by $C''(\alpha)$. Thus $\Lambda_1(k,t) \leq \sqrt{2}$, $\Lambda_2(k,t) \leq 3C''(\alpha)$ and $\Lambda_3(k,t) \leq 2\sqrt{2} + 3C''(\alpha)$.
We obtain under the condition $log(p-k) > 1$ that $A(k,t) \leq 2^t C(\alpha)log(p-k)$.
We also have that $\left[\|\mu\|_n^2 + \sigma^2\left(2 + \frac{3}{n}log\left(\frac{2k_0}{\gamma}\right)\right)\right] \leq C(\|\mu\|_n, \gamma, \sigma)$ since $log(k_0)/n \leq 1$,
and that $2^t\left(6 + 4log\left(\frac{k_0}{2^t}\right)\right) + 3log\left(\frac{2k_0}{\gamma}\right) \leq 2^t\left[6 + 4log\left(\frac{k_0}{2^t}\right) + 3log\left(\frac{2k_0}{\gamma}\right)\right] \leq 2^t C(\gamma)log(k_0)$.

We finally obtain equation (20). □

Proof of Corollary 3.8. The differences between the two conditions $(R_{3,k})$ and $(R_{3bis,k})$ lie in the fact that $\inf\left\{\|\Pi_S\mu\|_n^2, S \in B_{2^t}\right\} = \sum_{j=1}^{2^t}\beta^2_{\sigma_2(j)}$ and that the upper bound of $Q_{1-\gamma/2k_0}$ is modified, where $Q_{1-\gamma/2k_0}$ is defined by $\mathbb{P}\left(\|Y - \Pi_{V_{(k),(t)}}Y\|_n^2 > Q_{1-\gamma/2k_0} \cap A_{k_0}\right) \leq \gamma/2k_0$.
Indeed, on the event A_{k_0} we have that $\|Y - \Pi_{V_{(k),(t)}}Y\|_n^2 \leq \sum_{j=k+2^t}^{j=k_0}\beta^2_{\sigma_2(j)} + \|\epsilon\|_n^2$, where σ_2 is defined such that $|\beta_{\sigma_2(1)}| \leq ... \leq |\beta_{\sigma_2(k_0)}|$. We get from there the condition $(R_{3bis,k})$. □

References

[1] Bach, F., 2009. Model-consistent sparse estimation through the bootstrap. Technical report, hal-00354771, version 1.

[2] Baraud, Y., Huet, S., Laurent, B., 2003. Adaptative test of linear hypotheses by model selection. The Annals of Statistics 31, 225–251.

[3] Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing. Journal of the Royal Statistical Society B 57, 289-300.

[4] Bickel, P.J., Ritov, Y., Tsybakov, A.B., 2009. Simultaneous analysis of lasso and dantzig selector. The Annals of Statistics 37, 1705–1732.

[5] Bunea, F., Tsybakov, A., Wegkamp, M., 2007. Sparsity oracle inequalities for the lasso. Electron. J. Statist. 1, 169–194.

[6] Bunea, F., Wegkamp, M., Auguste, A., 2006. Consistent variable selection in high dimensional regression via multiple testing. Journal of Statistical Planning and Inference 136, 4349–4363.

[7] Candes, E., Tao, T., 2007. The dantzig selector: Statistical estimation when p is much larger than n. The Annals of Statistics 35, 2313–2351.

[8] Chesneau, C., Hebiri, M., 2008. Some theoretical results on the grouped variables lasso. Mathematical Methods of Statistics 17, 317–326.

[9] Huang, J., Ma, S., Zhang, C.H., 2008. Adaptative lasso for sparse high-dimensional regression models. Stat. Sin. 18, 1603–1618.

[10] Laurent, B., Massart, P., 2000. Adaptive estimation of a quadratic functional by model selection. The Annals of Statistics 28, 1302–1338.

[11] Lê Cao, K.A., Rossouw, D., Robert-Granié, C., Besse, P., 2008. A sparse pls for variable selection when integrating omics data. Statistical applications in genetics and molecular biology 7, Article 35.

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

[13] Tenenhaus, M., 1998. La régression PLS: théorie et pratique. Editions Technip.

[14] Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society B 58, 267–288.

[15] Verzelen, N., 2010. Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons.

[16] Zhang, C.H., Huang, J., 2006. Model-selection consistency of the lasso in high-dimensional linear regression. Technical report, No. 2006-003.