Multiple Testing and Variable Selection along Least Angle Regression’s path

Jean-Marc Azaïs* and Yohann De Castro*

*Institut de Mathématiques de Toulouse
Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse, France

Abstract: In this article we investigate the outcomes of the standard Least Angle Regression (LAR) algorithm in high dimensions under the Gaussian noise assumption. We give the exact law of the sequence of knots conditional on the sequence of variables entering the model, i.e., the “post-selection” law of the knots of the LAR. Based on this result, we prove an exact of the False Discovery Rate (FDR) in the orthogonal design case and an exact control of the existence of false negatives in the general design case.

First, we build a sequence of testing procedures on the variables entering the model and we give an exact control of the FDR in the orthogonal design case when the noise level can be unknown. Second, we introduce a new exact testing procedure on the existence of false negatives when the noise level can be unknown. This testing procedure can be deployed after any support selection procedure that will produce an estimation of the support (i.e., the indexes of nonzero coefficients) for any designs. The type I error of the test can be exactly controlled as long as the selection procedure follows some elementary hypotheses, referred to as “admissible selection procedures”. These support selection procedures are such that the estimation of the support is given by the first variables entering the model where the random variable is a stopping time. Monte-Carlo simulations and a real data experiment are provided to illustrate our results.

MSC 2010 subject classifications: Primary 62E15, 62F03, 60G15, 62H10, 62H15; secondary 60E05, 60G10, 62J05, 94A08.

Keywords and phrases: Multiple Testing, False Discovery Rate, High-Dimensional Statistics, Post-Selection Inference.

Preprint version of June 27, 2019

1. Introduction

Parsimonious models have become an ubiquitous tool to tackle high-dimensional representations with a small budget of observations. Successful applications is signal processing (see for instance the pioneering works [13, 12] and references therein), biology (see for instance [3] or [11, Chapter 1.4] and references therein)... have shown that the existence of an (almost) sparse representations in some well chosen basis is reasonable assumption in practice. These important successes have put focus on High-Dimensional Statistics in the past decades and they may be due to the deployment of tractable algorithms with strong theoretical guarantees. Among the large panoply of methods, we have seen emerged \( \ell_1 \)-regularization which may have found a fine balance between tractability and performances. Nowadays, sparse regression techniques based on \( \ell_1 \)-regularization are a common and powerful tool in high-dimensional settings. Popular estimators, among which one may point LASSO [27] or SLOPE [10], are known to achieve minimax rate of prediction and to satisfy sharp oracle inequalities under conditions on the design such as Restricted Eigenvalue [7, 4] or Compatibility [11, 30]. Recent avances have focused on on a deeper understanding of these techniques looking at confidence intervals and testing procedures (see [30, Chapter 6] and references therein) or false discovery rate control (e.g., [3]) for instance. These new results aim at describing the (asymptotic or non asymptotic) law of the outcomes of \( \ell_1 \)-minimization regression. This line of works adresses important issues encountered in practice. Assessing the uncertainty of popular estimators give strong guarantees on the estimation produced, e.g., the false discovery rate is controlled or a confidence interval on linear statistics of the estimator can be given.
1.1. Least Angle Regression algorithm, Support Selection and FDR

Least Angle Regression (LAR) algorithm has been introduced in the seminal article [14]. This forward procedure produces a sequence of knots \( \lambda_1, \lambda_2, \ldots \) based on a control of the residuals in \( \ell_\infty \)-norm. This sequence of knots is closely related to the sequence of knots of LASSO [27], as they differ by only one rule: “Only in the LASSO case, if a nonzero coefficient crosses zero before the next variable enters, drop it from the active set and recompute the current joint least-squares direction”, as mentioned in [28, Page 120] for instance. This paper focuses on the LAR algorithm and presents three useful equivalent formulations of the LAR algorithm, see Section A.1. As far as we known, the last formulation given by Algorithm 1, based on a recursive formulation, is new.

One specific task is to estimate the support of the target sparse vector, namely identify the true positives. In particular, one may take the support of a LASSO (or SLOPE) solution as an estimate of the support solution. This strategy has been intensively studied in the literature, one may consider [32, 10, 30, 4] and references therein. Support selection has been studied under the so-called “Irrepresentable Condition” (IC), as presented for instance in the books [30, Page 53] and [11, Sec. 7.5.1] and also referred to as the “Mutual Incoherence Condition” [32]. Under the so-called “Beta-Min Condition”, one may prove [11, 30] that the LASSO asymptotically returns the true support. In this article, we investigate the existence of false non-negatives and we present exact non-asymptotic testing procedures to handle this issue, see Section 2.4. Our procedure does not require IC but a much weaker condition, referred to as the “Empirical Irrepresentable Check” (See Section 2.1.3), that can be checked in polynomial time.

Another recent issue is to control the False Discovery Rate (FDR) in high-dimensional setting, as for instance in [3] and references therein; or the Joint family-wise Error Rate as in [8] and references therein. In this paper, we investigate the consecutive spacings of knots of the LAR as testing statistics and we prove an exact FDR control using a Benjamini–Hochberg procedure [5] in the orthogonal design case, see Section 2.3. Our proof (see Appendix C.7) is based on the Weak Positive Regression Dependency (WPRDS), the reader may consult [9] or the survey [23], and Knothe-Rosemblatt transport, see for instance [24, Sec.2.3, P.67] or [31, P.20], which is based on conditional quantile transforms.

1.2. Post-Selection Inference in High-Dimensions

This paper presents a class of new tests issued from \( \ell_1 \)-minimization regression in high-dimensions. Following the original idea of [20], we study tests based on the knots of the LAR’s path. Note that, conditional on the sequence of indexes selected by LAR, the law of two consecutive knots has been studied for the first time by [20] referred to as the Spacing test [29]. Later, the article [1] proved that this test is unbiased and introduce a studentized version of this test. On the same note, inference after model selection has been studied in several papers, as for instance [15, 25] or [26] for selective inference and a joint estimation of the noise level. This line of works studies a single test on a linear statistics while one may ask for a simultaneous control of several tests as in multiple testing frame. To the best of our knowledge, this paper is the first to study the joint law of multiple spacing tests of LAR’s knots in a non-asymptotic frame, see Sections 2.2 and 2.3.

One may point others approaches for building confidence intervals and/or testing procedures in high-dimensional settings as follows. Simultaneous controls of confidence intervals independently of the selection procedure have been studied under the concept of post-selection constants as introduced in [6] and studied for instance in [2]. Asymptotic confidence intervals can be built using the de-sparsified LASSO, the reader may refer to [30, Chapter 5] and references therein. We also point a recent study [18] of the problem of FDR control as the sample size tends to infinity using debiased LASSO.
Algorithm 1: LAR algorithm ("recursive" formulation)

Data: Correlations vector $Z$ and variance-covariance matrix $R$.
Result: Sequence $((\lambda_k, \tau_k, \varepsilon_k))_{k \geq 1}$ where $\lambda_1 \geq \lambda_2 \geq \ldots > 0$ are the knots, and $\tau_1, \tau_2, \ldots$ are the variables that enter the model with signs $\varepsilon_1, \varepsilon_2, \ldots (\varepsilon_k = \pm 1)$.

/* Define the recursive function Rec() that would be applied repeatedly. The inputs of Rec() are $Z$ a vector, $R$ a SDP matrix and $T$ a vector. */

Function Rec($R, Z, T$):

Compute

$$\lambda = \max_{(j, T_j < 1)} \left\{ \frac{Z_j}{1 - T_j} \right\} \quad \text{and} \quad i = \arg \max_{(j, T_j < 1)} \left\{ \frac{Z_j}{1 - T_j} \right\}.$$  

/* The following recursions are given by (24), (25) and (26). */

Update

$$x = R_i / R_{ii}$$
$$R \leftarrow R - xR_i^T$$
$$Z \leftarrow Z - xZ_i$$
$$T \leftarrow T + x(1 - T_i)$$

return $(R, Z, T, \lambda, i)$

1. Set $k = 0$, $T = 0$, $Z = (\bar{Z}, -\bar{Z})$ and $R$ as in (10).

/* Use the following recursion function to compute the LAR path. */

2. Update $k \leftarrow k + 1$ and compute

$$(R, Z, T, \lambda_k, \tau_k) = \text{Rec}(R, Z, T)$$

Set $\tau_k = \tilde{\tau}_k \mod p$ and $\varepsilon_k = 1 - 2(\tilde{\tau}_k - \tau_k)/p \in \pm 1$.

1.3. Outline of the paper

In Section 2, we introduce our method along with the framework and the notion of “Empirical Irrepresentable Check” (EIC), a condition that can be checked in polynomial time, see Section 2.1.3. We divided our contributions in three part: Section 2.2 presents the joint law of the knots of LAR under EIC; Section 2.3 gives an exact control of FDR in the orthogonal design case; and Section 2.4 presents a general framework to exactly detect false negatives for general designs and some selection procedures referred to as “admissible”.

Detailed mathematical statements are given in Section 3 including a method to “studentize” all the tests by an independent estimation of variance.

Section 4 presents practical applications including an illustration on real data.

The Appendix gives details on the different algorithms to compute LAR, proofs of the statements and a list of notation.

2. Exact Controls using Least Angle Regression, an introductory presentation

2.1. Notation, LAR formulations and Assumptions

2.1.1. Least Angle Regression (LAR)

Consider the linear model in high-dimensions where the number of observations $n$ may be less than the number of predictors $p$. Consider the Least Angle Regression (LAR) algorithm where we denote by $(\lambda_k)_{k \geq 1}$ the sequence of knots and by $(\tau_k, \varepsilon_k)_{k \geq 1}$ the sequence of variables $\tau_k \in [p]$ and signs $\varepsilon_k \in \{\pm 1\}$ that enter the model along the LAR path, see for instance [28, Chapter 5.6] or [14] for standard description of this algorithm. We recall this algorithm in Algorithm 2 and we present equivalent formulations in Algorithm 3 (using orthogonal projections) and Algorithm 1 (using a recursion). The interested reader may find their analysis in Appendices A and B.

In particular, we present here Algorithm 1 that consists in three lines, applying the same function recursively. We introduce some notation that we will be useful throughout this paper.
We denote by \((\hat{r}_1, \ldots, \hat{r}_n) \in [2p]^k\) the “signed” variables that enter the model along the LAR path with the convention that

\[
\hat{r}_\ell := r_\ell + p \left( \frac{1 - \varepsilon_\ell}{2} \right),
\]

so that \(\hat{r}_\ell \in [2p]\) is a useful way of encoding both the variable \(r_\ell \in [p]\) and its sign \(\varepsilon_\ell = \pm 1\) as used in Algorithm 1. We denote by \(Z\) the vector such that \(Z_k\) is the scalar product between the \(k\)-th predictor and the response variable, and we denote by \(R\) its correlation matrix, see (9) and (10) for further details.

2.1.2. Nested Models

Our analysis is based conditionally to \((\hat{r}_1, \ldots, \hat{r}_n)\) and in this spirit it can be referred to as a “Post-Section” procedure. The selected model \(\hat{S}\) would be chosen among the family of nested models

\[
\{\hat{r}_1\} \subset \{\hat{r}_1, \hat{r}_2\} \subset \{\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_k\} \subset \cdots \subset \{\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_n\}.
\]

(2)

Respectively, denote

\[
E_1 \subset \cdots \subset E_k := \text{Span}(X_{r_1}, \ldots, X_{r_k}) \subset \cdots \subset E_n = \mathbb{R}^n,
\]

the corresponding family of nested subspaces of \(\mathbb{R}^n\). In the sequel, denote by \(P_k(Y)\) (resp. \(P_k^\perp(Y)\)) the orthogonal projection of the observation \(Y\) onto \(E_k\) (resp. the orthogonal of \(E_k\)) for all \(k \geq 1\).

2.1.3. Empirical Irrepresentable Check

In the sequel we will build our testing statistics on the joint law of the first \(K\) knots of the LAR. A detailed discussion on the choice of \(K\) is given in Section 2.5.1, in particular \(K\) will be such that the so-called “Empirical Irrepresentable Check of order \(K\)” holds.

This property will be defined and studied in this section. Note that the “Irrepresentable Condition” is a standard condition, as presented for instance in the books [30, Page 53] and [11, Sec. 7.5.1] and also referred to as the “Mutual Incoherence Condition” [32].

**Definition 1** (Irrepresentable Condition of order \(K\)). The design matrix \(X\) satisfies the Irrepresentable Condition of order \(K\) if and only if

\[
\forall S \subset [p] \text{ s.t. } \#S \leq K, \quad \max_{j \in [p] \setminus S} \max_{\|v\| \leq 1} \|X_j^T X_S (X_S^T X_S)^{-1} v\| < 1, \quad (A_{\text{Irr}})
\]

where \(X_j\) denotes the \(j\)-th column of \(X\) and \(X_S\) the submatrix of \(X\) obtained by keeping the columns indexed by \(S\).

**Remark 1.** This condition has been intensively studied in the literature and it is now well established that some random matrix models satisfy it with high probability. For instance, one may refer to the article [32] where it is shown that a design matrix \(X \in \mathbb{R}^{n \times p}\) whose rows are drawn independently with respect to a centered Gaussian distribution with variance-covariance matrix satisfying \((A_{\text{Irr}})\) (for instance the Identity matrix) satisfies \((A_{\text{Irr}})\) with high probability when \(n \gtrsim K \log(p-K)\), where \(\gtrsim\) denotes an inequality up to some multiplicative constant.

In practice, the Irrepresentable Condition \((A_{\text{Irr}})\) is a strong requirement on the design \(X\) and, additionally, this condition cannot be checked in polynomial time. One important feature of our result, see Theorem 3, is that we do not require Irrepresentable Condition but a much weaker requirement referred to as the “Empirical Irrepresentable Check” of order \(K\).
**Definition 2** (Empirical Irrepresentable Check of order $K$ w.r.t. $(\hat{\imath}_1,\ldots,\hat{\imath}_K)$). The design matrix $X$ satisfies the Empirical Irrepresentable Check of order $K$ with respect to $(\hat{\imath}_1,\ldots,\hat{\imath}_K)$ if and only if for all the supports $(S^k)_{k \in [K]}$ and all the signs $(\varepsilon^k)_{k \in [K]}$ which enter the model at steps $k \in [K]$

$$\forall k \in [K], \forall j \notin S^k := \{\hat{\imath}_1,\ldots,\hat{\imath}_k\}, \quad X_j^\top X_{S^k} (X_{S^k}^\top X_{S^k})^{-1} \varepsilon^k < 1 \quad (\hat{\mathcal{A}}_{\text{irr}}),$$

where $S^k$ is the selected support at step $k$ (see (2)) and $\varepsilon^k := \{\varepsilon_1,\ldots,\varepsilon_k\}$ are the corresponding signs of the selected variables.

**Remark 2.** Observe that Irrepresentable Condition ($\hat{\mathcal{A}}_{\text{irr}}$) of order $K$ implies Empirical Irrepresentable Check ($\hat{\mathcal{A}}_{\text{irr}}$) of order $K$ with respect to any possible $(\hat{\imath}_1,\ldots,\hat{\imath}_K)$.

**Remark 3.** Note that this condition can be checked in $O(n \times p \times K + n \times K^{cX,\text{inv}})$ time where $O(n \times K^{cX,\text{inv}})$ accounts for the computational time of the $K$ operations $X_S (X_S^\top X_S)^{-1} e$ where $e \in \mathbb{R}^K$ is a fixed vector and $S \subseteq [p]$ has size $k$ for $k = 1,\ldots,K$. A standard bound for this last term would be $c_{X,\text{inv}} \leq 6.3728639$, see for instance [19].

**Remark 4.** When computing the LAR path, one has to compute the values $X_j^\top X_{S^k} (X_{S^k}^\top X_{S^k})^{-1} \varepsilon^k$, see for instance Algorithm 2 or Algorithm 3 where these values are given by $\theta$ as shown by Proposition 2. It means that, in practice, one has for free the maximal order $K$ for which Empirical Irrepresentable Check holds. Furthermore, this latter remark shows that the maximal order $K$ for which Empirical Irrepresentable Check holds is a statistic of the variables $\hat{\imath}_k \in [p]$ and signs $\varepsilon_k \in \{\pm 1\}$ that enter the model along the LAR path.

In Figure 1, in the particular case where the design matrix $X \in \mathbb{R}^{n \times p}$ is constructed from independent column vectors uniformly distributed on the sphere and an independent $y \in \mathbb{R}^n$ with i.i.d. standard Gaussian entries, and we computed the maximal order $K$ for which the Empirical Irrepresentable Check holds. Due to symmetry of the law of the design and of the independent observation $y$, note that the probability to observe a given path $P\{\hat{\imath}_1 = i_1,\ldots,\hat{\imath}_n = i_n\}$ is almost the same for all possible paths, namely $\simeq 1/(n!)$. Then we can view the plot in Figure 1 as the target value $K$ for which ($\hat{\mathcal{A}}_{\text{irr}}$) holds when the support $\{\hat{\imath}_1 = i_1,\ldots,\hat{\imath}_K = i_K\}$ is uniformly distributed.

In Figure 1, we chose $p = 250, 500, 1000, 2000$, different values of $n$ and we plot a shaded blue line representing 95% of the values $\rho$ (among 1,000 repetitions). Interestingly, these values are...
concentrated around the solid blue line referred to as the “empirical phase transition” for the Gaussian model.

**Example 5.** For example, we found that for \( p = 1,000 \) and \( n = 300 \) (\( \delta = n/p = 0.3 \)), Empirical Irrepresentable Check (\( \hat{\text{Irr}} \)) of order \( K \) holds when \( K \) is about \( K \approx p \times 300 = 0.13 \times 300 = 39 \).

For every couple (\( \delta = n/p, \rho = K/n \)) below this curve we observe that we have (\( \hat{\text{Irr}} \)) of order \( K \) with overwhelming probability as \( p \) is sufficiently large. We did not pursue this issue in the present paper but we would like to emphasize that (\( \hat{\text{Irr}} \)) holds empirically for values \( K \) of the order of 10\% of \( n \) at least. In addition, on real data of Section 4.2, with about \( p = 200 \) variables and \( n = 700 \) observations with found \( K \) of the order of 30.

2.2. Result one: Mixture of Gaussian Order Statistics in LAR

Recall that this article concerns the linear model in high-dimensions where the number of predictors \( p \) can be much greater than the number of samples \( n \). We denote by \( Y \in \mathbb{R}^n \) the response variable and we assume that

\[
Y = X^0 \beta^0 + \eta \sim \mathcal{N}_n(X^0 \beta^0, \sigma^2 \Sigma),
\]

where \( \eta \sim \mathcal{N}_n(0, \sigma^2 \Sigma) \) is some Gaussian noise, the matrix \( \Sigma > 0 \) is some known positive definite matrix, the noise level \( \sigma > 0 \) may be known or that has to be estimated depending on the context, and \( X^0 \in \mathbb{R}^{n \times p} \). In this paper, we are interested in selecting the true support \( S^0 \) of \( \beta^0 \), where the support is defined by

\[
S^0 := \{ k \in [p] : \beta^0_k \neq 0 \}.
\]

Provided that \( \Sigma \) is known, one can always consider the homoscedastic-independent version of the responses, namely

\[
\Sigma^{-\frac{1}{2}} Y = X \beta^0 + \Sigma^{-\frac{1}{2}} \eta \sim \mathcal{N}_n(X \beta^0, \sigma^2 \mathbb{I}_n),
\]

where the design \( X \) is given by \( X := \Sigma^{-\frac{1}{2}} X^0 \). From the observation \( \Sigma^{-\frac{1}{2}} Y \) and the design \( X \) (or equivalently from (\( \mathbb{Z}, \mathbb{R} \) defined in (9)), one can compute the Least Angle Regression (LAR) knots \( (\lambda_k)_{k \geq 1} \) and the sequence \( (\tau_k, \varepsilon_k)_{k \geq 1} \) of variables \( \tau_k \in [p] \) and signs \( \varepsilon_k \in \{ \pm 1 \} \). In the sequel, the law (resp. the conditional law) of a random vector \( V \) (resp. conditionally to \( W \)) is denoted by \( \mathcal{L}(V) \) (resp. \( \mathcal{L}(V|W) \)). One of the main discovery of this paper (see Theorem 3) is the following:

**Under Empirical Irrepresentable Check (\( \hat{\text{Irr}} \)) of order \( K \), the LAR knots \( (\lambda_1, \ldots, \lambda_K) \) behave as a mixture of Gaussian order statistics with heterogeneous variances, namely it holds**

\[
\mathcal{L}(\lambda_1, \ldots, \lambda_K|\lambda_{K+1})
\]

\[
= \sum_{(\tau_1, \ldots, \tau_K) \in [p]^K} \pi_{(\tau_1, \ldots, \tau_K, \lambda_{K+1})} \mathcal{Z}_{(\tau_1, \ldots, \tau_K, \lambda_{K+1})}^{-1} \prod_{k=1}^{K} \gamma_{m_k, v_k^2} \mathbb{I}_{\{\lambda_1 \geq \cdots \geq \lambda_K \geq \lambda_{K+1}\}} \mathcal{L}(\lambda_1, \ldots, \lambda_K|\tau_1 = \cdots = \tau_K = i_k, \lambda_{K+1}),
\]

where \( \pi_{(\tau_1, \ldots, \tau_K, \lambda_{K+1})} = \mathbb{P}\{\tau_1 = i_1, \ldots, \tau_K = i_K | \lambda_{K+1}\} \) can be interpreted as mixing probabilities, \( \mathcal{Z}_{(\tau_1, \ldots, \tau_K, \lambda_{K+1})} \) is a normalizing constant, and \( \gamma_{m_k, v_k^2} \) is a short notation for the Gaussian law such that

- its mean \( m_k \) is given by (15) and depends only on the covariance \( \sigma^2 \Sigma \), the selected support \( (\tau_1, \ldots, \tau_k) \) at step \( k \), and the orthogonal projection of \( \tilde{\beta}^0 := X^\top X \beta^0 \) given by (5);
- and its variance \( v_k^2 := \sigma^2 \rho_k^2 \) is given by (13) and depends only on the covariance \( \sigma^2 \Sigma \) and the selected support \( (\tau_1, \ldots, \tau_k) \) at step \( k \).
As presented in the sequel, this result is the key step to prove an exact control of the FDR, see Section 2.3, and an exact control of the False Negatives of some selection procedures, see Section 2.4. These controls are obtained from \( p \)-values \( \hat{\alpha}_{a,b,c} \) described in Theorem 5 and (19). We will see that the \( p \)-value \( \hat{\alpha}_{a,b,c} \) detects abnormal large values of \( \lambda \) conditional on \( (\lambda_a, \lambda_c) \) and that, in the orthogonal design case, Theorem 6 shows that the test based on \( \hat{\alpha}_{a,a+1,K+1} \) is uniformly more powerful than tests based on \( \hat{\alpha}_{x,y,z} \) with \( a \leq x < y < z \leq K + 1 \).

### 2.3. Result two: Control of False Discovery Rate in the Orthogonal Design case

#### 2.3.1. Presentation in the general case

Assume that Empirical Irrepresentable Check \( \hat{A}_{\text{irr.}} \) of order \( K \) holds. The sequence of testing procedures under consideration is given by the sequence of \( p \)-values \( \hat{\alpha}_{k-1,k+1} \) that are referred to as “Spacing tests” [29] in the literature, they are recalled in Remark 10 below. These \( p \)-values account for “abnormally large” values of \( \lambda \) conditional on \( (\lambda_{k-1}, \lambda_{k+1}) \). Invoking (4), this conditional law of \( \lambda \) depends on two unobserved values: the mean \( \mu_k \), given by (15), and the variance \( \sigma^2 \) (in the case where the variance may be unknown). As discussed in Section 2.4.1, the variance \( \sigma^2 \) will be estimated on residuals as described by the heuristic of Remark 6. It entails that one may efficiently estimate the variance and, as we will see in Section 3.4, one can plug this independent estimator of the variance so as to get a studentized version of the testing procedures described here. For sake of readability, we will assume that \( \sigma \) is known and we refer to Section 3.4 for details on how to studentize our procedure. We understand that the law of testing statistics are parametrized by the hypotheses \((m_k)_{k \in [K]}\), where \( m_k \) is given by (15).

We recall that we denote \( \pi^0 = X^\top X \beta^0 \) and \( \pi_i^0 \) it \( i \)th coordinate. Assuming that predictors are normalised, in the general case, this quantity is the sum of \( \beta_i^0 \)’s whose predictors \( X_j \) are highly correlated with the predictor \( X_j \). Now, given \( \tau_1, \ldots, \tau_k \in [p] \) and signs \( \varepsilon_1, \ldots, \varepsilon_k \in \{\pm 1\}^k \), we denote by \( (P_{\tau_1, \ldots, \tau_{k-1}}(\pi^0))_{\tau_k} \) the orthogonal projection given by

\[
(P_{\tau_1, \ldots, \tau_{k-1}}(\pi^0))_{\tau_k} := \varepsilon_k X_{\tau_k}^\top \left[ I_{d_n} - X (X^\top X)^{-1} \text{diag}(\varepsilon_{\tau_k}^{-1})(X^\top X)^{-1} \text{diag}(\varepsilon_{\tau_k}^{-1}) X^\top X^{-1} \right] X \beta^0 .
\]

where \( \text{diag}(\varepsilon_{\tau_k}^{-1}) := \text{diag}(\varepsilon_1, \ldots, \varepsilon_{k-1}) \).

The tested hypotheses are conditional on the sequence of variables \((\tau_1, \ldots, \tau_K) \in [p]^K \) and signs \( \varepsilon_1, \ldots, \varepsilon_K \in \{\pm 1\}^K \) entering the model. The \( p \)-values under consideration here are given by

- \( \hat{\rho}_1 := \hat{\alpha}_{0,1,2} \) is the \( p \)-value testing \( H_{0,1} : \\text{“}m_1 = 0\text{”} \) namely \( \pi_{\tau_1}^0 = 0 \);
- \( \hat{\rho}_2 := \hat{\alpha}_{1,2,3} \) is the \( p \)-value testing \( H_{0,1} : \\text{“}m_2 = 0\text{”} \) namely \( (P^1_{\tau_1}(\pi^0))_{\tau_2} = 0 \);
- \( \hat{\rho}_3 := \hat{\alpha}_{2,3,4} \) is the \( p \)-value testing \( H_{0,1} : \\text{“}m_3 = 0\text{”} \) namely \( (P^2_{\tau_1}(\pi^0))_{\tau_3} = 0 \);
- \( \hat{\rho}_4 := \hat{\alpha}_{3,4,5} \) is the \( p \)-value testing \( H_{0,1} : \\text{“}m_4 = 0\text{”} \) namely \( (P^3_{\tau_1}(\pi^0))_{\tau_4} = 0 \);
- \( \hat{\rho}_5 := \hat{\alpha}_{4,5,6} \) is the \( p \)-value testing \( H_{0,1} : \\text{“}m_5 = 0\text{”} \) namely \( (P^4_{\tau_1}(\pi^0))_{\tau_5} = 0 \);
- \( \text{and so on...} \)

We write \( I_0 \) fo the set

\[
I_0 = \{ k \in [K] : H_{0,k} \text{ is true} \} .
\]

Given a subset \( \hat{R} \subseteq [K] \) of hypotheses that we consider as rejected, we call false positive (FP) and true positive (TP) the quantities \( \text{FP} = \text{card}(\hat{R} \cap I_0) \) and \( \text{TP} = \text{card}(\hat{R} \setminus I_0) \).

Denote by \( \hat{\rho}_1 \leq \ldots \leq \hat{\rho}_K \) the \( p \)-values ranked in a nondecreasing order. Let \( \alpha \in (0,1) \) and consider the Benjamini-Hochberg procedure, see for instance [5], defined by a rejection set \( \hat{R} \subseteq [K] \) such that \( \hat{R} = \emptyset \) when \( \{ k \in [K] : \hat{\rho}(k) \leq \alpha K / K \} = \emptyset \) and

\[
\hat{R} = \{ k \in [K] : \hat{\rho}(k) \leq \alpha K / K \} \quad \text{where} \quad \hat{k} = \max \{ k \in [K] : \hat{\rho}(k) \leq \alpha K / K \} .
\]
Recall the definition of FDR as the mean of False Discovery Proportion (FDP), namely

\[
\text{FDR} := E \left[ \frac{\text{FP}}{\text{FP} + \text{TP}} \mathbb{I}_{\text{FP} + \text{TP} \geq 1} \right],
\]

where the expectation is unconditional to the sequence of variables entering the model, while the hypotheses that are being tested are conditional on the sequence of variables entering the model. This FDR can be understood invoking the following decomposition

\[
\text{FDR} = \sum_{(i_1, \ldots, i_K) \in [p]^K} \pi_{(i_1, \ldots, i_K)} E[\text{FDP}|\bar{I}_1 = i_1, \ldots, \bar{I}_K = i_K],
\]

where \( \pi_{(i_1, \ldots, i_K)} = P\{\bar{I}_1 = i_1, \ldots, \bar{I}_K = i_K\} \).

2.3.2. FDR control of Benjamini–Hochberg procedure in the orthogonal design case

We now consider the orthogonal design case where \( X^\top X = \text{Id}_p \) and the set of \( p \)-values given by (6). Note that \( I_0 \) is simply the set of null coordinates of \( \beta \). Remark also that, Irrepresentable Condition (\( A_{\text{Irrep}} \)) of order \( p \) holds and so does Empirical Irrepresentable Check (\( \hat{A}_{\text{Irrep}} \)), see Proposition 2. Note also that \( I_0 \) is simply the set of null coordinates of \( \beta \). It implies that on can consider any value \( K \in [p] \) in the following result.

**Theorem 1.** Assume that the design is orthogonal, namely it holds \( X^\top X = \text{Id}_p \), and let \( K \in [p] \). Let \( (\bar{I}_1, \ldots, \bar{I}_K) \) be the first variables entering along the LAR’s path. Consider the \( p \)-values given by (6) and the set \( \tilde{R} \) given by (7). Then

\[
E[\text{FDP}|\bar{I}_1 = i_1, \ldots, \bar{I}_K = i_K] \leq \alpha,
\]

and so FDR is upper bounded by \( \alpha \).

The proof of this result is given in Appendix C.7.

One interpretation of post-selection type may be given as follows: if one looks at all the experiments giving the same sequence of variables entering the model \( \{ \bar{I}_1 = i_1, \ldots, \bar{I}_K = i_K \} \) and if one considers the Benjamini–Hochberg procedure for the hypotheses described in Section 2.3.1, then the FDR is exactly controlled by \( \alpha \).

2.4. Result three: Exact Testing Procedure on False Negatives

From the result of Section 2.2, one can present a method to select a model and propose an exact test on false negatives in the general design case. More precisely, we assume here for simplicity that the design \( X \) has full row rank \( n \). In theory we are able to compute the LAR knots up to \( \lambda_n \). Of course this is often out of reach due to numerical issues and one has to stop earlier to avoid bad conditioning, but we forget this possibility to avoid heavy notation. One can adapt easily the following text to take into account this possibility. Let us compute the LAR knots \( (\lambda_1, \ldots, \lambda_n) \) and the “signed” variables \( (\bar{\imath}_1, \ldots, \bar{\imath}_n) \in [2p]^n \) entering the model as in (1).

2.4.1. Admissible Support Selection Procedures

In order to select \( \hat{S} \), one may be interested in an estimation \( \hat{\sigma}_{\text{select}} \) of the noise level since, in practice, one usually does not know it. Our procedure is valid as long as the following property (\( P_1 \)) is satisfied

**Noise Level Estimation for Selection (\( P_1 \)):** The estimated noise level \( \hat{\sigma}_{\text{select}} \) is a measurable function of \( P_{n,\text{select}}^\perp(Y) \) where \( n_{\text{select}} \) may depend only on \( (\bar{I}_1, \ldots, \bar{\imath}_n) \).
In practice, one can chose a fixed \( n_{\text{select}} = n - B \) giving a fixed budget of \( B \geq 1 \) independent observations to estimate the variance\(^1\). Obviously, if the noise level is known, one can simply take \( n_{\text{select}} = n \) and \( \sigma_{\text{select}} = \sigma \).

**Remark 6.** Observe that if the true support \( S^0 \) is included in \( \{r_1, r_2, \ldots, r_k\} \) then \( P_k^\perp(Y) \) is centered Gaussian vector with known covariance (up to \( \sigma^2 \)) and it is standard to compute an estimator of the noise level \( \sigma \). In particular, the information contained in \( P_k^\perp(Y) \) does not depend on the target \( \beta^0 \), since \( P_k^\perp(X^0 \beta^0) = 0 \), and is purely due to the “noise” part \( \eta \), as defined in (3).

Note that, for large enough \( n_{\text{select}} \), the heuristic described in Remark 6 may be true and it entails that one may efficiently estimate the variance on \( P_{n_{\text{select}}}^\perp(Y) \).

Now, note that choosing a model \( \hat{S} \) is equivalent to choosing a model size \( \hat{m} \in [n] \) so that

\[
\hat{S} = \{\hat{r}_1, \hat{r}_2, \ldots, \hat{r}_{\hat{m}}\}.
\]

Our procedure is flexible on this point and it allows any choice of \( \hat{m} \) as long as the following property (\( P_2 \)) is satisfied

**Stopping Rule (\( P_2 \))**: The estimated model size \( \hat{m} \) is a “stopping time”, i.e. \( 1_{\{\hat{m} \leq k\}} \) is a measurable function of \( (k(Y), \hat{\sigma}_{\text{select}}) \) for all \( k \in [n_{\text{select}}] \).

In other words, the decision to select a model of size \( \hat{m} = k \) depends only on the part of the observation \( Y \) explained by the predictors \( X_{\hat{r}_1}, \ldots, X_{\hat{r}_k} \). In the sequel, we will present some examples of such “stopping time” procedures, see Section 3.5. We would like to emphasize that our analysis reveal the following conditional independence

\[
\mathcal{L}(P_{n_{\text{select}}}(Y), \hat{\sigma}_{\text{select}})(\hat{r}_1, \ldots, \hat{r}_n) = \mathcal{L}(P_{n_{\text{select}}}(Y)|\hat{r}_1, \ldots, \hat{r}_n) \otimes \mathcal{L}(\hat{\sigma}_{\text{select}}|\hat{r}_1, \ldots, \hat{r}_n)
\]

which can be advantageously invoked to build Student-type testing statistics to define \( \hat{m} \).

### 2.4.2. Exact Testing Procedure on False Negatives

Once one has selected a model of size \( \hat{m} \), one may be willing to test if \( \hat{S} \) contains the true support \( S^0 \) by considering the null hypothesis \( H_0 : “S^0 \subseteq \hat{S}” \), namely there is no false negatives. Equivalently, one aim at testing the null hypothesis

\[
H_0 : “X\beta^0 \in E_{\hat{m}}”
\]

at an exact significance level \( \alpha \in (0, 1) \).

In this article, we introduce a new exact testing procedure that can be deployed when (\( P_1 \)) and (\( P_2 \)) hold, namely an “admissible” selection procedure is used to build \( \hat{S} \). The variance estimator \( \hat{\sigma}_{\text{select}} \) may have been used in the definition of \( \hat{m} \) and our method forbid to use it again to “studentized” our testing procedure. We need to estimate the variance again to preserve some “independence” in the spirit of (8). As in the previous variance estimation, we advocate a “budget” to this task, namely we use the “strata” \( \{n_{\text{test}}, n_{\text{select}}\} \) (as in the vocabulary of analysis of variance) to build \( \hat{\sigma}_{\text{test}}^2 \) an estimation of the variance. More precisely, we use the orthogonal projection \( P_{\{n_{\text{test}}, n_{\text{select}}\}}(Y) \) which is the the orthogonal projector onto the space \( F \) defined by

\[
E_{n_{\text{select}}} = E_{n_{\text{test}}} \perp \perp F.
\]

Note that this space is generated by \( X_{n_{\text{test}}+1}, \ldots, X_{n_{\text{select}}} \).

The resulting test is based on \( (\lambda_{\hat{m}}/\hat{\sigma}_{\text{test}}, \ldots, \lambda_{n_{\text{test}}}/\hat{\sigma}_{\text{test}}) \). More precisely, we know the law of \( \lambda_{n_{\text{select}}+1}/\hat{\sigma}_{\text{test}} \) conditional on \( \hat{m}, \lambda_{\hat{m}}, \lambda_{n_{\text{test}}} \) under \( H_0 \).

---

\(^1\) One can also look at the conditioning of the variance-covariance matrix of \( P_k^\perp(Y) \) for \( k = n, n - 1, \ldots \) to choose \( n_{\text{select}} \). A more detailed discussion on this issue can be found in Section 2.5.
1: Compute the largest order of Empirical Irrepresentability Check

2: Use a stopping rule to select the support (the variance is estimated on the tail a first time)

3: Estimate the variance a second time for the test statistic

4: Compute the test statistics

\[
K = \min(K_{\text{irr}}, \hat{m} + \Delta_{\text{num}})
\]

Figure 2. The testing procedure computes the significance of knot \( \lambda_{1 + \hat{m}} \) w.r.t. knots \( \lambda_{\hat{m}} \) and \( \lambda_K \) (see Step 4 above), where \( K \) is a compromise between the numerical limitations \( \hat{m} + \Delta_{\text{num}} \) of the cost of integrating a function on \([0,1]^{\Delta_{\text{num}}^{-2}}\) and the statistical limitations of the design as depicted by the Empirical Irrepresentable Check.

2.5. Practical Framework

We start by presenting the meta-algorithm in which our testing procedure can be deployed. First, we compute the sequence of LAR knots \( (\lambda_1, \ldots, \lambda_n) \) and “signed” variables \( (\hat{\iota}_1, \ldots, \hat{\iota}_n) \in [2p]^n \) entering the model as in (1). Conditionally to the sequence \( \hat{\iota}_1, \ldots, \hat{\iota}_n \) the quantities displayed in Figure 2 are introduced in the following order.

- Compute \( K_{\text{irr}} \), which is the maximal value such that \( (\hat{A}_{\text{irr}}) \) holds true.
- Fixing a budget \( B \) for variance estimation \( \hat{\sigma}_{\text{select}} \) we fix \( n_{\text{select}} = n - B \). In most of the applications, \( n \) is large and consequently \( n_{\text{select}} \) is also large.
- From the estimation \( \hat{\sigma}_{\text{select}} \) we deduce \( \hat{m}_{\text{select}} \). Here, the practitioner is free to use any selection procedures as soon as it satisfies \((P_1)\) and \((P_2)\).
- Multivariate integration programs have some dimension limitation, this implies that we can only handle simultaneously \( \Delta_{\text{num}} \) knots. As a consequence \( K \) must be smaller or equal than \( \hat{m}_{\text{select}} + \Delta_{\text{num}} \).
- In most cases the quantity

\[
K = \min(K_{\text{irr}}, \hat{m}_{\text{select}} + \Delta_{\text{num}})
\]

is convenient because \( K < n_{\text{select}} \) and there is enough budget in the strata \( (K, n_{\text{select}}) \) to build the second estimator of variance, namely \( \hat{\sigma}_{\text{test}} \).
- Otherwise we have to make a “trade-off”: diminishing \( K \) to obtain a sufficient budget of degree of freedom in the strata \( (K, n_{\text{select}}) \) to estimate the variance while maintaining \( K \) large to gain power.

An example of values of these parameters is

\[
n = 200, \ p = 300, \ K_{\text{irr}} = 35 \text{ (for a Gaussian design)}, \ \text{and} \ \Delta_{\text{num}} = 10.
\]

This framework has been deployed on real data in Section 4. Examples of admissible procedures satisfying \((P_1)\) and \((P_2)\) are presented in Section 3.5.
3. Main Results

Consider the correlations vector $\overline{Z}$ of the homoscedastic-independent observation $\Sigma^{-\frac{1}{2}}Y$ with the design matrix $X$, so that

$$\overline{Z} := X^0\Sigma^{-1}Y = \overline{R}\beta^0 + X^0\Sigma^{-1}\eta \sim N_p(\overline{R}\beta^0, \sigma^2\overline{R})$$

and

$$\overline{R} := X^T X = X^0\Sigma^{-1}X^0.$$  \hfill (9)

From $\overline{Z}$ and $\overline{R}$ we can compute the Least Angle Regression (LAR) knots $(\lambda_k)_{k\geq 1}$ and the sequence $(\tau_k, \varepsilon_k)_{k\geq 1}$ of variables $\tau_k$ and signs $\varepsilon_k$ that enter the model along the LAR path. We will use these statistics as testing statistics. For sake of presentation, we may consider the $2p$-vector $\overline{Z} := (Z, -Z)$ whose mean is given by $\mu^0 := (\overline{R}\beta^0, -\overline{R}\beta^0) = (X^T\beta^0, -X^T\beta^0) = (\mu^0, -\mu^0)$ and its variance-covariance matrix is $\sigma^2\overline{R}$ with

$$R = \begin{bmatrix} \overline{R} & -\overline{R} \\ -\overline{R} & \overline{R} \end{bmatrix} = \begin{bmatrix} X^TX & -X^TX \\ -X^TX & X^TX \end{bmatrix}. \quad \hfill (10)$$

3.1. Problem reformulation and Assumptions

Remark that $Z = (Z_i)_i$ is a Gaussian vector of size $2p$ such that $Z_i = -Z_{i+p}$ where the indices are considered modulo $2p$. Note also that we know the covariance of $Z$ up to some multiplicative constant $\sigma^2$, namely the variance-covariance matrix of $Z$ is given by $\sigma^2R$ where $R \succeq 0$ is known and $\sigma^2$ is some parameter that may be known or not.

Now, we can define

$$\forall (i_1, \ldots, i_k) \in [2p]^k, \quad \theta_j(i_1, \ldots, i_k) := (R_{j,i_1} \cdots R_{j,i_k})M_{i_1,\ldots,i_k}^{-1}(1, \ldots, 1),$$

where $(1, \ldots, 1)$ is the column vector of size $k$ whose entries are equal to one and $\sigma^2M_{i_1,\ldots,i_k}$ is the variance-covariance matrix of the vector $(Z_{i_1}, \ldots, Z_{i_k})$, and $(R_{j,i_1} \cdots R_{j,i_k})$ is a row vector of size $k$. Note that $M_{i_1,\ldots,i_k}$ is the submatrix of $R$ obtained by keeping the columns and the rows indexed by $\{i_1, \ldots, i_k\}$. Remark that

$$\theta_j(i_1, \ldots, i_k) = \mathbb{E}(Z_j \mid Z_{i_1} = 1, \ldots, Z_{i_k} = 1),$$

when $\mathbb{E}Z = 0$. In our context, the Irrepresentable Condition of order $K$ is given by

$$\forall k \leq K, \forall (i_1, \ldots, i_k) \in [2p]^k, \forall j \notin \{i_1, \ldots, i_k\}, \quad \theta_j(i_1, \ldots, i_k) < 1,$$

where $\theta_j(i_1, \ldots, i_k)$ is given by (11). In Proposition 2 we show that (12) is equivalent to the Irrepresentable Condition ($A_{irr.}$) on the design matrix $X$. A proof can be found in Appendix C.2.

**Proposition 2.** Assume $X$ and $R$ satisfy (10) then the following assumptions are equivalent:

- the design matrix $X$ satisfies ($A_{irr.}$) of order $K$,
- the variance-covariance matrix $R$ satisfies (12) of order $K$.

Furthermore, if one of these two assumptions hold then “Empirical Irrepresentable Check” of order $K$ holds, namely

$$\max_{j \neq i_1} \left[ \max_{j \neq i_1, \ldots, i_k} \theta_j(i_1, \ldots, i_k) \right] < 1$$

which is an equivalent formulation of ($A_{irr.}$) of Definition 2.

**Remark 7.** One may require that the design is “normalized” so that $R_{i,i} = 1$, namely its columns have unit Euclidean norm. Under this normalization, one can check that $R$ satisfies ($A_{irr.}$) of order $K = 1$. Hence, up to some normalization, one can always assume ($A_{irr.}$) of order $K = 1$. 
3.2. Conditional Joint Law of the Knots

In this section, we are interested in the joint law of the knots \((\lambda_1, \ldots, \lambda_K)\) of the LAR conditional on \(\lambda_{K+1}\). Let \((\hat{t}_1, \ldots, \hat{t}_K)\) be the first variables entering along the LAR path. Define the first correlation by \(\rho_1 := \sqrt{R_{\hat{t}_1, \hat{t}_1}}\) and the others by

\[
\rho_\ell := \frac{\sqrt{R_{\hat{t}_1, \hat{t}_\ell}} - (R_{\hat{t}_{\ell-1}, \hat{t}_1} \cdots R_{\hat{t}_{\ell-1}, \hat{t}_\ell}) M_{\hat{t}_1, \ldots, \hat{t}_{\ell-1}}^{-1} \left( R_{\hat{t}_1, \hat{t}_1}, \ldots, R_{\hat{t}_{\ell-1}, \hat{t}_{\ell-1}} \right)}{1 - \theta_{\hat{t}_1}^{-1}}, \quad \text{for } \ell \geq 2, \tag{13}
\]

where

\[
\theta_{\hat{t}_1}^{-1} := \theta(\hat{t}_1, \ldots, \hat{t}_{\ell-1}), \quad \text{for } \ell \geq 2,
\]

is defined by (11) and \(M_{\hat{t}_1, \ldots, \hat{t}_{\ell-1}}\) is the submatrix of \(R\) obtained by keeping the columns and the rows indexed by \(\{\hat{t}_1, \ldots, \hat{t}_{\ell-1}\}\). Furthermore, denote

\[
F_i := \Phi_i(\lambda_i) \quad \text{and} \quad P_{i,j} := \Phi_i \circ \Phi_j^{-1}, \quad \text{for } i, j \in [K + 1], \tag{14}
\]

where \(\Phi_k(\cdot) := \Phi(\cdot/(\sigma_k))\) is the CDF of the centered Gaussian law with variance \(\sigma_k^2\) for \(k \geq 1\), \(\lambda_0 = \infty\) and \(F_0 = 1\) by convention. The main discovery of this paper is the following theorem.

**Theorem 3** (Conditional Joint Law of the LAR Knots). Let \((\lambda_1, \ldots, \lambda_K, \lambda_{K+1})\) be the first knots and let \((\hat{t}_1, \ldots, \hat{t}_K)\) be the first variables entering along the LAR path. If (\(A_{\text{irr.}}\)) holds then, conditional on \(\{\hat{t}_1, \ldots, \hat{t}_K, \lambda_{K+1}\}\), the vector \((\lambda_1, \ldots, \lambda_K)\) has law with density (w.r.t. the Lebesgue measure)

\[
Z_{\hat{t}_1, \ldots, \hat{t}_K, \lambda_{K+1}}^{-1} \left( \prod_{k=1}^{K} \varphi_{m_k, v_k^2}(\ell_k) \right) 1(\ell_1 \geq \ell_2 \geq \cdots \geq \ell_K \geq \lambda_{K+1}) \quad \text{at point } (\ell_1, \ell_2, \ldots, \ell_K),
\]

where \(\varphi_{m_k, v_k^2}\) is the standard Gaussian density with mean

\[
m_k := \mu_k^0 - (R_{\hat{t}_{k-1}, \cdots, \hat{t}_1} \cdots R_{\hat{t}_{\ell-1}, \hat{t}_1}) M_{\hat{t}_1, \cdots, \hat{t}_{\ell-1}}^{-1} (\mu_{\hat{t}_1}^0, \ldots, \mu_{\hat{t}_{\ell-1}}^0) \quad \text{with} \quad \mu^0 := (\mu^0, -\mu^0), \tag{15}
\]

and variance \(v_k^2 := \sigma_k^2\).

A useful corollary of this theorem is the following. It shows that one can explicitly describe the joint law of the LAR’s knots after having selected a support of size \(\hat{m}\) with any procedure satisfying \((P_1)\) and \((P_2)\).

**Corollary 4.** Under the conditions of Theorem 3, let \(\hat{m}\) be chosen according to a procedure satisfying \((P_1)\) and \((P_2)\) with \(n_{\text{select}} \geq K\). Then under the null hypothesis, namely

\[
\mathbb{H}_0 : \quad "X\beta^0 \in E_{\hat{m}}", \nonumber
\]

and conditional on the selection event \(\{\hat{m} = a, F_0, F_{a+1}, \hat{t}_1, \ldots, \hat{t}_K, \hat{t}_{K+1}\}\) with \(0 \leq a \leq K - 1\), the vector \((F_{a+1}, \ldots, F_K)\) is uniformly distributed on

\[
\mathcal{D}_{a+1, K} := \left\{ (f_{a+1}, \ldots, f_K) \in \mathbb{R}^{K-a} : \right. \nonumber
\]

\[
\left. P_{a+1, a}(F_a) \geq f_{a+1} \geq P_{a+1, a+2}(f_{a+2}) \geq \cdots \geq P_{a+1, K}(f_K) \geq P_{a+1, K+1}(F_{K+1}) \right\},
\]

where \(P_{i,j}\) are described in (14).

**Remark 8.** The previous statement is consistent with the case \(a = 0\) corresponding to the global null hypothesis \(\mathbb{H}_0 : \quad "X\beta^0 = 0"\) (or equivalently \(EZ = 0\)). Therefore, if \(Z\) is centered then, conditional on \(F_{K+1}, (F_1, \ldots, F_K)\) is uniformly distributed on

\[
\mathcal{D}_{1, K} := \left\{ (f_1, \ldots, f_K) \in \mathbb{R}^K : \right. \nonumber
\]

\[
\left. 1 \geq f_1 \geq P_{1, 2}(f_2) \geq \cdots \geq P_{1, K}(f_K) \geq P_{1, K+1}(F_{K+1}) \right\}. \nonumber
\]
Remark 9. In the orthogonal case where $\mathbf{R} = \mathbf{Id}$, note that $\theta_j(i_1, \ldots, i_\ell) = 0$ for all $\ell \geq 1$ and all $i_1, \ldots, i_\ell \neq j$, $\rho_j = 1$ and $P_{i,j}(f) = f$. We recover that $D_{i,K}$ is the set of order statistics

$$1 \geq f_1 \geq f_2 \geq \ldots \geq f_K \geq \Phi(\lambda_{K+1}).$$

In this case, knots $\lambda_i$ are Gaussian order statistics $\lambda_1 = Z_{\hat{1}} \geq \lambda_2 = Z_{\hat{2}} \geq \ldots \lambda_K = Z_{\hat{K}} \geq \lambda_{K+1}$ of the vector $Z$.

3.3. Testing Procedures

From Theorem 3, we deduce several testing statistics. To this end, we introduce some notation. First, define

$$I_{ab}(s,t) := \int \frac{df_a}{P_{(a+1),a}(s)} \int \frac{df_{a+1}}{P_{(a+2),a+1}(f_{a+1})} \int \frac{df_{a+2}}{P_{(a+3),a+2}(f_{a+2})} \ldots \int \frac{df_{b-1}}{P_{(b-1),a+b-2}(f_{b-2})}$$

for $0 \leq a < b$ and $s,t \in \mathbb{R}$, with the convention that $I_{ab} = 1$ when $b = a + 1$,

and also

$$F_{abc}(t) := \mathbb{I}_{(\lambda_a \leq t \leq \lambda_a)} \int_{\phi_b(t)} I_{ab}(F_a, f_b) I_{bc}(f_b, F_c) df_b$$

(16)

for $0 \leq a < b < c \leq K + 1$, $t \in \mathbb{R}$ where $F_a = \Phi_a(\lambda_a)$ and $F_c = \Phi_c(\lambda_c)$.

Then, a simple integration shows that Corollary 4 implies that

$$P[\lambda_b \leq t \mid \lambda_a, \lambda_c, \hat{1}, \hat{2}, \ldots, \hat{c-1}] = \frac{F_{abc}(t)}{F_{abc}(\lambda_a)},$$

(17)

whenever $EZ = 0$. Actually, this result can be refined in the case when $EZ \neq 0$ by the next result.

Theorem 5. Let $(\lambda_1, \ldots, \lambda_K, \lambda_{K+1})$ be the first knots and let $(\hat{1}, \ldots, \hat{K})$ be the first variables entering along the LAR path. If $(A_{Irr})$ holds and $\hat{m}$ is chosen according to a procedure satisfying $(P_1)$ and $(P_2)$ with $n_{select} \geq K$, then under the null hypothesis, namely

$$H_0 : "X\beta^0 \in E_{\hat{m}}",$$

and conditional on the selection event $\{\hat{m} = a\}$ with $a \leq K - 1$, and for any integers $b, c$ such that $a < b < c \leq K + 1$, it holds that

$$\tilde{\alpha}_{abc} := 1 - \frac{F_{abc}(\lambda_b)}{F_{abc}(\lambda_a)} \sim \mathcal{U}(0, 1),$$

(18)

namely, it is uniformly distributed over $(0, 1)$.

This testing statistic generalizes previous testing statistics that appeared in “Spacing Tests”, as presented in [29, Chapter 5] for instance, and will be referred to as the Generalized Spacing test.

Remark 10. If one considers $a = 0$, $b = 1$ and $c = 2$ then one gets

$$\tilde{\alpha}_{012} = 1 - \frac{\Phi_1(\lambda_1) - \Phi_1(\lambda_2)}{\Phi_1(\lambda_0) - \Phi_1(\lambda_2)} = \frac{1 - \Phi_1(\lambda_1)}{1 - \Phi_1(\lambda_2)}.$$  

Similarly, taking $b = a + 1$ and $c = a + 2$ one gets

$$\tilde{\alpha}_{a(a+1)(a+2)} = \frac{\Phi_{a+1}(\lambda_{a+1}) - \Phi_{a+1}(\lambda_{a+2})}{\Phi_{a+1}(\lambda_a) - \Phi_{a+1}(\lambda_{a+2})},$$

which is the spacing test as presented in [29, Chapter 5].
Given $\alpha \in (0, 1)$, one can consider the following testing procedures

$$S_{abc} := I_{\{\hat{\alpha}_{abc} \leq \alpha\}},$$

(19)

that rejects if the $p$-value $\hat{\alpha}_{abc}$ is less than the level $\alpha$ of the test. Now, recall (17) to witness that the smaller $\hat{\alpha}_{abc}$, the bigger $\lambda_b$ conditionally to $(\lambda_a, \lambda_c)$. So the $p$-value $\hat{\alpha}_{abc}$ detects abnormal large values of $\lambda_b$ conditional on $(\lambda_a, \lambda_c)$.

One may investigate the power of these tests to detect false negatives, namely to detect the alternative given by: there exists $k \in S^0$ such that $k \notin \{\hat{\tau}_1, \ldots, \hat{\tau}_n\}$. In particular, what is the most powerful test among these latter testing procedures? A comprehensive study of the orthogonal case is given in the following theorem.

**Theorem 6.** Assume that the design is orthogonal, namely $\bar{R} = Id_p$. Let $\hat{m}$ be chosen according to a procedure satisfying $(P_1)$ and $(P_2)$ with $n_{select} \geq K$. Then under the null hypothesis, namely

$$H_0 : \ "X\beta^0 \in E_{\hat{m}}",$$

and conditional on the selection event $\{\hat{\tau}_1 = \hat{\tau}_1, \ldots, \hat{\tau}_n = \hat{\tau}_n\}$ with $1 \leq a_0 \leq K - 1$, it holds the test $S_{a_0, a_0 + 1, K + 1}$ is uniformly most powerful than any of the tests $S_{a, b, c}$ for $a_0 \leq a < b < c \leq K + 1$.

The proof of this result is given in Appendix C.6. It shows that the best choice among the tests $S_{a, b, c}$ is the test $S_{a_0, a_0 + 1, K + 1}$ with the smallest $a$ and the largest $c$.

### 3.4. “Studentization”: Adapting to Unknown Noise Level

The previous results can be further refined to the case where $\sigma^2$ is unknown. Let $K$ chosen following the considerations of Section 2.5, the key property is to build an estimation of the variance $\hat{\sigma}^2_{test}$ on the strata $(n_{test}, n_{select})$ which is independent from

$$(\hat{\lambda}_{\hat{m}}, \ldots, \hat{\lambda}_{K + 1}, \hat{\sigma}_{test}),$$

conditional on the selection event $\{\hat{\tau}_1 = \hat{\tau}_1, \ldots, \hat{\tau}_n = \hat{\tau}_n\}$ and under $H_0 : \ "X\beta^0 \in E_{\hat{m}}"$.

Because of our hypotheses, $\text{rank}(P_{(n_{test}, n_{select})})$ is maximal equal to $m := n_{select} - n_{test}$. Our variance estimator

$$\hat{\sigma}^2 := \hat{\sigma}^2_{test} = \frac{\|P_{(n_{test}, n_{select})}(Z)\|^2_2}{m},$$

follows a $\sigma^2 \chi^2(m)/m =: \sigma^2 W^2$. 

Figure 3. CDF of $p$-values $\hat{\alpha}_{abc}$ over 3,000 Monte-Carlo iterations with $n = 200$, $p = 300$ and a random design matrix $X$ given by 300 independent column vectors uniformly distributed on the Euclidean sphere $S^{299}$. Central panel represents an alternative composed by a 2 sparse vector, the right panel an alternative composed by a 2 sparse vector 5 times larger while the left panel corresponds to the null.
We set \( \hat{\lambda}_i := \frac{\lambda_i}{\sigma} = \frac{\lambda_i}{\sigma_{\text{select}}} \). Let us consider \( a = m < b < c \leq K + 1 \) and consider, under \( \mathbb{H}_0 \), the distribution conditional to \( \{ \hat{m} = a, \lambda_a, \lambda_{K+1}, \hat{r}_1, \ldots, \hat{r}_{K+1} \} \) then the variables \( (\lambda_a, \ldots, \lambda_c) \) and \( W \) are independent. More over the distribution of \( \lambda_b \) is given by (17). We have
\[
\mathbb{P}\{\hat{\lambda}_b \leq t | \hat{\lambda}_a \hat{\lambda}_c, W = w, \hat{r}_1, \ldots, \hat{r}_{K+1}\} = \mathbb{P}\left\{ \frac{\lambda_b}{\sigma} \leq w | \hat{\lambda}_a \hat{\lambda}_c, W = w, \hat{r}_1, \ldots, \hat{r}_{K+1}\right\}
\]
Then we can use Theorem 5 in the case \( \sigma = 1 \) to get that the expression above is equal to
\[
\frac{F_{abc}(tw)}{F_{abc}(\lambda_a w)}.
\]
Where \( F_{abc} \) is the function given by (17) in the case \( \sigma = 1 \). De-conditioning in \( w \) we get
\[
\mathbb{P}\{\hat{\lambda}_b \leq t | \hat{\lambda}_a \hat{\lambda}_c, i_1, \ldots, i_{K+1}\} = \int_{\mathbb{R}} \frac{F_{abc}(tw)}{F_{abc}(\lambda_a w)} dP_W(w).
\]
As a consequence the \( p \)-value of the test in now given by
\[
\hat{\beta}_{abc} = 1 - \int_{\mathbb{R}} \frac{F_{abc}(\hat{\lambda}_b w)}{F_{abc}(\lambda_a w)} dP_W(w).
\]

3.5. Examples of Admissible procedures

We are now able to present here an admissible procedure to build an estimate \( \hat{S} \) of the support that satisfies the properties \((P_1)\) and \((P_2)\). We chose a level \( \alpha' \) and we define a light modification of the Student test
\[
T_{abc} = 1_{\hat{\beta}_{abc} \leq \alpha'},
\]
by replacing \( \hat{\sigma}^2_{\text{test}} \) by \( \hat{\sigma}^2_{\text{select}} \). The number of degrees of freedom of the \( \chi^2 \) distribution is now \( m := n - n_{\text{select}} \). By a small abuse of notation, we still set
\[
\hat{\lambda}_i := \frac{\lambda_i}{\sigma_{\text{select}}}.
\]
We limit our attention to consecutive \( a, b, c \). In such a case it is easy to see that
\[
\hat{\beta}_{a(a+1)(a+2)} = \frac{T\left(\frac{\hat{\lambda}_{a+1}}{\sigma_{\rho_{a+1}}}\right) - T\left(\frac{\hat{\lambda}_a}{\sigma_{\rho_{a+1}}}\right)}{T\left(\frac{\hat{\lambda}_{a+2}}{\sigma_{\rho_{a+2}}}\right) - T\left(\frac{\hat{\lambda}_a}{\sigma_{\rho_{a+1}}}\right)},
\]
where \( T \) is the cumulative distribution of the Student \( T(m) \) law, so the quantity above is easy to compute. We are now in condition to present our algorithm.

- Begin with \( a = 0 \),
- at each step, perform the test \( T_{a,a+1,a+2} \) at the level \( \alpha' \),
- if the test is significative, we set \( a = a + 1 \) and keep on going,
- if it is non-significative, we stop and set \( \hat{m} = a + 2 \).

Recall that possible selected supports along the LAR’s path are nested models of the form (2).

Denote \( k^0 \geq 1 \) the smallest integer \( k \) such that the true support \( S^0 \) is contained in \( \overline{S}^k \) and denote
\[
S^0 \subseteq \overline{S}^0 \text{ and } S^0 \not\subseteq \overline{S}^{(k^0 - 1)}.
\]

We understand that admissible procedures depend on the sequence of false positives appearing along the LAR’s path. If the experimenter believes that there is no more than \( \gamma_{FP} \) consecutive false positives in \( S^0 \) a possible admissible procedure would be the following.
• Begin with $a = 0$,
• at each step, perform the test $T_{a,a+1,a+2}$ at the level $\alpha'$,
• if the test is significative, we set $a = a + 1$ and keep on going,
• if the $\gamma_{FP}$ consecutive tests $T_{a,a+1,a+2}, T_{a+1,a+2,a+3}, \ldots, T_{a+\gamma_{FP}-1,a+\gamma_{FP},a+\gamma_{FP}+1}$ are all non-significatives, we stop and set $\hat{m} = a + \gamma_{FP} + 1$.

This method has been deployed on real data in Section 4.2 with $\gamma_{FP} = 3$.

4. Numerical Experiments

4.1. Monte-Carlo experiment

To study the relative power in a non-orthogonal design case we have build a Monte-Carlo experiment with 3,000 repetitions. We have considered a model with $n = 200$, $p = 300$ and a random design matrix $X$ given by $300$ independent column vectors uniformly distributed on the Euclidean sphere $\mathbb{S}^{299}$. The computation of the function $F_{abc}$ given by (16) is an important issue that demands multivariate integration tools, see Appendix E for a solution using cubature of integral by lattice rule. This has lead to some limitations namely $\Delta_{num} \leq 4$ that implies $c \leq 5$ when $a = 1$ in our experimental framework.

A python notebook and codes are given at https://github.com/ydecastro/lar_testing. The base function is observed_significance_CBC(lars, sigma, start, end, middle) in the file multiple_spacing_tests.py. It gives the $p$-value of $T_{(start)(middle)(end)}$ of knots and indexes given by lars and an estimation of (or the true) standard deviation given by sigma. We have run 3,000 repetitions of this function to get the laws displayed in Figure 3. It presents the CDF of the $p$-value $\hat{\alpha}_{abc}$ under the null and under two $2$-sparse alternatives, one with low signal and one with 5 times more signal. Results show, in our particular case, that all the tests are exact and the test $S_{125}$ is the most powerful. More precisely, it holds, as in the orthogonal case, that $S_{123} \leq S_{124} \leq S_{125}$; $S_{234} \leq S_{245}$.

4.2. Real data

A detailed presentation in a Python notebook is available at https://github.com/ydecastro/lar_testing/blob/master/multiple_spacing_tests.ipynb.

We consider a data set about HIV drug resistance extracted from [3] and [22]. The experiment consists in identifying mutations on the genes of the HIV-virus that are involved with drug resistance. The data set contains about $p = 200$ and $n = 700$ observations. Since some protocol is used to remove some gene or some individuals, the exact numbers depend on the considered drug.

We used a procedure referred to as “spacing-BH” procedure which is a Benjamini–Hochberg procedure based on the sequence of spacing tests

$$\hat{\beta}_{012}, \hat{\beta}_{123}, \ldots, \hat{\beta}_{a(a+1)(a+2)}, \ldots$$

as described in Section 3.4 with $\alpha = 0.2$. The results for Knockoff of [3] and of Benjamini–Hochberg procedure on the coefficients of linear regression (BHq) are for the R-vignette knockoff of the dedicated web page of Stanford. All results are evaluated using the TSM data base that gives, in some sense, the list of true positives. A comparison of our results with those of [3] is displayed in Figure 4. Our procedure is a bit more conservative but, in most of the case, gives a better control of the FDP.

In addition we have performed on the same dataset a false negative detection as in Section 2.4, and Section 3.5 with $\gamma_{FP} = 3$. We refer to the aforementioned Python notebook for further details.
Figure 4. Comparison of numbers true and false positives for three procedures: Spacing-BH, Knockoff [3] and Benjamini–Hochberg procedure on the coefficient of linear regression BHq. For each drug we indicate the number of true positives in blue and false positives in orange. In the three procedures the aimed FDR is $\alpha = 20\%$. 
References


Appendix A: Representing the LAR Knots

A.1. The equivalent formulations of the LAR algorithm

We present here three equivalent formulations of the LAR that are a consequence of the analysis provided in Appendices A and B. One formulation is given by Algorithm 1.

### Algorithm 2: LAR algorithm (standard formulation)

**Data:** Correlations vector $Z$ and variance-covariance matrix $R$

**Result:** Sequence $((\lambda_k, t_k, \varepsilon_k))_{k \geq 1}$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ are the knots, and $t_1, t_2, \ldots$ are the variables that enter the model with signs $\varepsilon_1, \varepsilon_2, \ldots$ ($\varepsilon_k = \pm 1$).

1. Set $k = 1$, $\lambda_1 := \max|Z|$, $t_1 := \arg\max|Z|$ and $\varepsilon_1 = Z_{t_1}/\lambda_1 \in \pm 1$, and $\overline{N}^{(1)} := Z$.

2. Set $k \leftarrow k + 1$ and compute the least-squares fit

$$\tilde{\theta}_j := (\overline{R}_{j, t_1} \cdots \overline{R}_{j, t_{k-1}})M_{t_1, \ldots, t_{k-1}}^{-1}(\varepsilon_1, \ldots, \varepsilon_{k-1}), \quad j = 1, \ldots, p,$$

where $M_{t_1, \ldots, t_{k-1}}$ is the submatrix of $\overline{R}$ keeping the columns and the rows indexed by $\{t_1, \ldots, t_{k-1}\}$.

3. For $0 < \lambda \leq \lambda_{k-1}$ compute the "residuals" $\overline{N}^{(k)}(\lambda) = (\overline{N}^{(k)}_1(\lambda), \ldots, \overline{N}^{(k)}_p(\lambda))$ given by

$$\overline{N}^{(k)}_j(\lambda) := \overline{N}^{(k-1)}_j - (\lambda_{k-1} - \lambda)\overline{\theta}_j, \quad j = 1, \ldots, p,$$

and pick

$$\lambda_k := \max\{\beta > 0 : \exists j \notin \{t_1, \ldots, t_{k-1}\}, \text{s.t. } |\overline{N}^{(k)}_j(\beta)| = \beta\} \quad \text{and} \quad t_k := \arg\max_{j \notin \{t_1, \ldots, t_{k-1}\}} |\overline{N}^{(k)}_j(\lambda_k)|,$$

$$\varepsilon_k := \overline{N}^{(k)}(\lambda_k)/\lambda_k \in \pm 1 \quad \text{and} \quad \overline{N}^{(k)} := \overline{N}^{(k)}(\lambda_k).$$

Then, iterate from 2.

### Algorithm 3: LAR algorithm ("projected" formulation)

**Data:** Correlations vector $Z$ and variance-covariance matrix $R$

**Result:** Sequence $((\lambda_k, t_k, \varepsilon_k))_{k \geq 1}$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ are the knots, and $t_1, t_2, \ldots$ are the variables that enter the model with signs $\varepsilon_1, \varepsilon_2, \ldots$ ($\varepsilon_k = \pm 1$).

1. Define $Z = (\overline{Z}, \overline{\varepsilon})$ and $R$ as in (10), and set $k = 1$, $\lambda_1 := \max Z$, $\tilde{t}_1 := \arg\max Z$, $\tilde{t}_1 = \tilde{t}_1 \mod p$ and $\varepsilon_1 = 1 - 2(\tilde{t}_1 - \tilde{t}_1)/p \in \pm 1$.

2. Set $k \leftarrow k + 1$ and compute

$$\lambda_k = \max_{(j : \theta_j(\tilde{t}_1, \ldots, \tilde{t}_{k-1}) \leq 1)} \left\{ \frac{Z_j - P_{1, \ldots, \tilde{t}_{k-1}}(Z_j)}{1 - \theta_j(\tilde{t}_1, \ldots, \tilde{t}_{k-1})} \right\} \quad \text{and} \quad \hat{t}_k = \arg\max_{(j : \theta_j(\tilde{t}_1, \ldots, \tilde{t}_{k-1}) \leq 1)} \left\{ \frac{Z_j - P_{1, \ldots, \tilde{t}_{k-1}}(Z_j)}{1 - \theta_j(\tilde{t}_1, \ldots, \tilde{t}_{k-1})} \right\},$$

where

$$P_{1, \ldots, \tilde{t}_{k-1}}(Z_j) := (R_{j, \tilde{t}_1} \cdots R_{j, \tilde{t}_{k-1}})M_{\tilde{t}_1, \ldots, \tilde{t}_{k-1}}^{-1}(Z_{\tilde{t}_1}, \ldots, Z_{\tilde{t}_{k-1}})$$

and

$$\theta_j(\tilde{t}_1, \ldots, \tilde{t}_{k-1}) := (R_{j, \tilde{t}_1} \cdots R_{j, \tilde{t}_{k-1}})M_{\tilde{t}_1, \ldots, \tilde{t}_{k-1}}^{-1}(1, \ldots, 1)$$

and set $\tilde{t}_k = \hat{t}_k \mod p$ and $\varepsilon_k = 1 - 2(\hat{t}_k - \tilde{t}_k)/p \in \pm 1$. Then, iterate from 2.
A.2. Initialization: First Knot

The first step of the LAR algorithm (Step 1 in Algorithm 2) seeks the most correlated predictor with the observation. In our formulation, introduce the first residual $N^{(1)} := Z$ and observe that $N^{(1)} := (N^{(1)}_1, -N^{(1)}_1)$. We define the first knot $\lambda_1 > 0$ as

$$\lambda_1 = \max Z \quad \text{and} \quad \hat{t}_1 = \arg \max Z.$$ 

One may see that this definition is consistent with $\lambda_1$ in Algorithm 2 and note that $\hat{t}_1$ and $(\hat{t}_1, \varepsilon_1)$ are related as in (1).

The LAR algorithm is a forward algorithm that selects a new variable and maintains a residual at each step. We also define

$$N^{(2)}(\lambda) = N^{(1)} - (\lambda_1 - \lambda)\theta(\hat{t}_1), \quad 0 < \lambda \leq \lambda_1,$$ 

and one can check that $N^{(2)}(\lambda) = (N^{(2)}_1, -N^{(2)}_1$) where $N^{(1)}$ is defined in Algorithm 2. It is clear that the coordinate $\hat{t}_1$ of $N^{(2)}(\lambda)$ is equal to $\lambda$. On the other hand $N^{(1)} = Z$ attains its maximum at the single point $\hat{t}_1$. By continuity this last property is kept for $\lambda$ in a left neighborhood of $\lambda_1$. We search for the first value of $\lambda$ such that this property is not met, i.e. the largest value of $\lambda$ such that

$$\exists j \neq \hat{t}_1 \text{ such that } N^{(2)}(\lambda) = \lambda,$$

as in Step 3 of Algorithm 2. We call this value $\lambda_2$ and one may check that this definition is consistent with $\lambda_2$ in Algorithm 2.

Now, we can be more explicit on the expression of $\lambda_2$. Indeed, we may make the following discussion depending on the values of $\theta_j(\hat{t}_1)$.

- If $\theta_j(\hat{t}_1) \geq 1$, since $N^{(1)}_j < N^{(1)}_{\hat{t}_1}$ for $j \neq \hat{t}_1$ there is no hope to achieve the equality between $N^{(2)}_j(\lambda)$ and $N^{(2)}_{\hat{t}_1}(\lambda) = \lambda$ for $0 < \lambda \leq \lambda_1$ in view of (20).
- Thus we limit our attention to the $j$’s such that $\theta_j(\hat{t}_1) < 1$. We have equality $N^{(2)}_j(\lambda) = \lambda$ when

$$\lambda = \frac{N^{(1)}_j - \lambda_1 \theta_j(\hat{t}_1)}{1 - \theta_j(\hat{t}_1)}.$$ 

So we can also define the second knot $\lambda_2$ of the LAR as

$$\lambda_2 = \max_{j: \theta_j(\hat{t}_1) < 1} \left\{ \frac{Z_j - P_{\hat{t}_1}(Z_j)}{1 - \theta_j(\hat{t}_1)} \right\},$$

where $P_{\hat{t}_1}(Z_j) := Z_{\hat{t}_1}\theta_j(i_1)$. Remark that $P_{\hat{t}_1}(Z_j) = E(Z_j | Z_{\hat{t}_1})$ is the regression of $Z_j$ on $Z_{\hat{t}_1}$ when $E Z = 0$.

A.3. Recursion: Next Knots

The loop (2 ⇔ 3) in Algorithm 2 builds iteratively the knots $\lambda_1, \lambda_2 \ldots$ of the LAR algorithm and some “residuals" $N^{(1)}_1, N^{(2)}_1, \ldots$ defined in Step 3. We will present here an equivalent formulation of these knots.

Assume that $k \geq 2$ and we have build $\lambda_1, \ldots, \lambda_{k-1}$ and selected the “signed” variables $\hat{t}_1, \ldots, \hat{t}_{k-1}$. Introduce $N^{(k-1)} := (N^{(k-1)}_1, -N^{(k-1)}_1)$ and define

$$N^{(k)}(\lambda) = N^{(k-1)} - (\lambda_{k-1} - \lambda)\theta(\hat{t}_1, \ldots, \hat{t}_{k-1}), \quad 0 < \lambda \leq \lambda_{k-1}.$$ 

Check that $\theta_j(\hat{t}_1, \ldots, \hat{t}_{k-1}) = (\bar{\theta}_j, -\bar{\theta}_j)$ where we recall that we define

$$\bar{\theta}_j := (\bar{\mathcal{R}}_{j,1}, \ldots, \bar{\mathcal{R}}_{j,k-1})M_{1,\ldots,k-1}^{-1}(\varepsilon_1, \ldots, \varepsilon_{k-1}), \quad j = 1, \ldots, p,$$
at Step 2 and it holds that \( \hat{\tau}_i \) and \((\tau_i, \varepsilon_i)\) are related as in (1). From this equality, we deduce that it holds \( N^{(k)}(\lambda) = (N^{(k)}(\lambda), -N^{(k)}(\lambda)) \). One may also check that the coordinates \( \hat{\tau}_1, \ldots, \hat{\tau}_{k-1} \) of \( N^{(k)}(\lambda) \) are equal to \( \lambda \).

Again if we want to solve \( N_j^{(k)}(\lambda) = \lambda \) for some \( j \), we have to limit our attention to \( j \)'s such that \( \theta_j(\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}) < 1 \). Solving this latter equality yields to

\[
\lambda_k = \max_{j: \theta_j(\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}) < 1} \left\{ \frac{N_j^{(k-1)} - \lambda_{k-1} \theta_j(\hat{\tau}_1, \ldots, \hat{\tau}_{k-1})}{1 - \theta_j(\hat{\tau}_1, \ldots, \hat{\tau}_{k-1})} \right\}.
\]

This expression is consistent with \( \lambda_k \) in Algorithm 2.

Now, we can give an other expression of \( \lambda_k \) that will be useful in the proofs of our main theorems. Note that the residuals satisfy the relation

\[
N_j^{(k)} = N_j^{(k-1)} - (\lambda_{k-1} - \lambda_k) \theta(\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}),
\]

and that \( N_j^{(k-1)} = \lambda_{k-1} \) for \( j = \hat{\tau}_1, \ldots, \hat{\tau}_{k-1} \). The following lemma permits a drastic simplification of the expression of the knots. Its proof is given in Appendix C.3.

**Lemma 7.** It holds

\[
N^{(k-1)} - \lambda_{k-1} \theta(\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}) = Z - P_{\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}}(Z)
\]

where we denote \( P_{\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}}(Z) = (P_{\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}}(Z_1), \ldots, P_{\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}}(Z_{2p})) \) and for all \( j \in \{2p\} \) one has \( P_{\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}}(Z_j) = (R_j, i_1 \cdots R_j, i_{k-1}) M_{\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}}^{-1}(Z_{i_1}, \ldots, Z_{i_{k-1}}) \).

Using Lemma 7 we deduce that \( \lambda_k \) in Algorithm 2 is consistent with

\[
\lambda_k = \max_{j: \theta_j(\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}) < 1} \left\{ \frac{Z_j - P_{\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}}(Z_j)}{1 - \theta_j(\hat{\tau}_1, \ldots, \hat{\tau}_{k-1})} \right\},
\]

where \( P_{\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}}(Z_j) = (R_j, i_1 \cdots R_j, i_{k-1}) M_{\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}}^{-1}(Z_{i_1}, \ldots, Z_{i_{k-1}}) \). When \( E\in = 0 \), one may remark that \( P_{\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}}(Z_j) \) is the regression of \( Z_j \) on the vector \((Z_{i_1}, \ldots, Z_{i_{k-1}})\) whose variance-covariance matrix is \( M_{\hat{\tau}_1, \ldots, \hat{\tau}_{k-1}} \). This analysis leads to an equivalent formulation of the LAR algorithm (Algorithm 2). We present this formulation in Algorithm 3.

**Remark 11.** Note that Algorithm 2 implies that \( \hat{\tau}_1, \ldots, \hat{\tau}_k \) are pairwise different, but also that they differ modulo \( p \). In the rest of the paper we will limit our attention to such sequences.

**Appendix B: First Steps to Derive the Joint Law of the LAR Knots**

Given fixed \( i_1, \ldots, i_k \in \{2p\} \), one may define \( Z_j^{(i_1, \ldots, i_k)} := Z_j - P_{i_1, \ldots, i_k}(Z_j) \) for indices \( j \) such that \( \theta_j(i_1, \ldots, i_k) = 1 \) and

\[
\forall j \text{ s.t. } \theta_j(i_1, \ldots, i_k) \neq 1, \quad Z_j^{(i_1, \ldots, i_k)} := \frac{Z_j - P_{i_1, \ldots, i_k}(Z_j)}{1 - \theta_j(i_1, \ldots, i_k)},
\]

where we recall that \( P_{i_1, \ldots, i_k}(Z_j) = (R_j, i_1 \cdots R_j, i_k) M_{i_1, \ldots, i_k}^{-1}(Z_{i_1}, \ldots, Z_{i_k}) \). From this point, we can introduce

\[
\lambda_{k+1}^{(i_1, \ldots, i_k)} := \max_{j: \theta_j(i_1, \ldots, i_k) < 1} Z_j^{(i_1, \ldots, i_k)},
\]

and remark that \( \lambda_{k+1} = \lambda_{k+1}^{(\hat{\tau}_1, \ldots, \hat{\tau}_k)} \).
B.1. Law of the First Knot

One has the following lemma governing the law of $\lambda_1$.

**Lemma 8.** It holds that

- $Z_{i_1}$ is independent of $(Z_j^{(i_1)})_{j \neq i_1}$,
- If $\theta_j(i_1) < 1$ for all $j \neq i_1$ then
  \[ \{\tilde{t}_1 = i_1\} = \{\lambda_2^{(i_1)} \leq Z_{i_1}\}, \]
- If $\theta_j(i_1) < 1$ for all $j \neq i_1$ then, conditionally to $\{\tilde{t}_1 = i_1\}$ and $\lambda_2$, $\lambda_1$ is a truncated Gaussian random variable with mean $E(Z_{i_1})$ and variance $\rho_1^2 := R_{i_1,\tilde{t}_1}$ subject to be greater than $\lambda_2$.

**Proof.** The first point is clear and standard. Now, observe that

\[ \{\lambda_2^{(i_1)} \leq Z_{i_1}\} \iff \{\forall j \neq i_1, \frac{Z_j - Z_{i_1}\theta_j(i_1)}{1 - \theta_j(i_1)} \leq Z_{i_1}\} \]
\[ \iff \{\forall j \neq i_1, Z_j - Z_{i_1}\theta_j(i_1) \leq Z_{i_1} - Z_{i_1}\theta_j(i_1)\} \]
\[ \iff \{\forall j \neq i_1, Z_j \leq Z_{i_1}\} \]
\[ \iff \{\tilde{t}_1 = i_1\}, \]

as claimed. The last statement is a consequence of the two previous points. \qed

B.2. Recursive Formulation of the LAR

One has the following proposition whose proof can be found in Section C.4. As we will see in this section, this intermediate result as a deep consequence, the LAR algorithm can be stated in a recursive way applying the same function repeatedly, as presented in Algorithm 1.

**Proposition 9.** Set

\[ \tau_{j,ik} := \frac{R_{j,ik} - (R_{j,i_1} \cdots R_{j,i_{k-1}})M_{i_1,\ldots,i_{k-1}}^{-1} (R_{i_1,i_1}, \ldots, R_{i_k,i_{k-1}})}{(1 - \theta_j(i_1, \ldots, i_{k-1}))(1 - \theta_k(i_1, \ldots, i_{k-1}))}, \]

and observe that $\tau_{j,ik}$ is the covariance between $Z_j^{(i_1,\ldots,i_{k-1})}$ and $Z_k^{(i_1,\ldots,i_{k-1})}$. Furthermore, it holds

\[ \frac{\tau_{j,ik}}{\tau_{i_k,ik}} = 1 - \frac{1 - \theta_j(i_1, \ldots, i_k)}{1 - \theta_j(i_1, \ldots, i_{k-1})} \quad (22) \]

and

\[ \forall j \neq i_1, \ldots, i_k, \quad Z_j^{(i_1,\ldots,i_k)} = \frac{Z_j^{(i_1,\ldots,i_{k-1})} - Z_k^{(i_1,\ldots,i_{k-1})} \tau_{j,ik} / \tau_{i_k,ik}}{1 - \tau_{j,ik} / \tau_{i_k,ik}}. \quad (23) \]

Now, we present Algorithm 1. Define $R(0) := R$, $Z(0) = Z$ and $T(0) = 0$. For $k \geq 1$ and fixed $i_1, \ldots, i_k \in [2p]$, introduce

\[ R(k) := (R_{j,\ell} - (R_{j,i_1} \cdots R_{j,i_k}) M_{i_1,\ldots,i_{k-1}}^{-1} (R_{\ell,i_1}, \ldots, R_{\ell,i_k}))_{j,\ell}, \]
\[ Z(k) := Z - P_{i_1,\ldots,i_k}(Z) \]
\[ T(k) := (\theta_j(i_1, \ldots, i_k))_j, \]

and note that $R(k)$ is the variance-covariance matrix of the Gaussian vector $Z(k)$. The key property is following. Let $v_1, \ldots, v_k$, be $k$ linearly independent vectors of an Euclidean space and let $u$ be any vector of the space. Set

\[ v := P_{(v_1,\ldots,v_{k-1})}^\perp v_k, \]
the projection of $v_k$ orthogonally to $v_1, \ldots, v_{k-1}$. Then

$$P_{v_1, \ldots, v_k}^\perp u = P_v^\perp P_{v_1, \ldots, v_{k-1}} u.$$ Using this result we deduce that

$$Z(k) = P_{i_1, \ldots, i_k}^\perp (Z) = P_{i_k}^\perp (P_{i_1, \ldots, i_{k-1}}^\perp (Z)) = P_{i_k}^\perp (Z(k-1)) = Z(k-1) - x(k) Z(k-1),$$

(24)

where $x(k) = R_{i_k} (k-1)/R_{i_k,i_k} (k-1)$. It yields that

$$R(k) = R(k-1) - x(k) R_{i_k} (k-1).$$

(25)

Using (22) (or (31)), remark that

$$T(k) = T(k-1) - x(k) (1 - T_{i_k} (k-1)).$$

(26)

These relations give a recursive formulation of the LAR as presented in Algorithm 1.

**B.3. Joint Law**

Regarding the joint law of knots, one has the following proposition whose proof can be found in Section C.5.

**Proposition 10.** One has the following for any fixed $i_1, \ldots, i_{k+1} \in [2p]$.

- It holds that
  $$\exists j \neq i_1, \ldots, i_{k+1} \perp Z^{(i_1, \ldots, i_{k+1})}_{i_k} \perp Z^{(i_1, \ldots, i_{k-1})}_{i_k} \perp \cdots \perp Z^{(i_1)}_{i_1} \perp Z_{i_1},$$

  are mutually independent.

- If $(\hat{\alpha}_{\text{irr.}})$ of order $k$ holds then

  $$\{\hat{i}_1 = i_1, \ldots, \hat{i}_{k+1} = i_{k+1}\} = \{\lambda^{(i_1, \ldots, i_{k+1})} = Z^{(i_1, \ldots, i_{k+1})}_{i_{k+1}} \leq Z^{(i_1, \ldots, i_{k-1})}_{i_k} \leq \cdots \leq Z^{(i_1)}_{i_1} \leq Z_{i_1}\},$$

  $$\{\lambda^{(i_1, \ldots, i_{k+1})} = Z^{(i_1, \ldots, i_{k+1})}_{i_{k+1}} \leq Z^{(i_1, \ldots, i_{k-1})}_{i_k} \leq \cdots \leq Z^{(i_1 \ldots, i_a)}_{i_{a+1}} \leq Z^{(i_1 \ldots, i_{a-1})}_{i_a} = \lambda^{(i_1, \ldots, i_{a-1})}_{a}\}$$

  $$\bigcap \{\hat{i}_1 = i_1, \ldots, \hat{i}_a = i_a, \hat{i}_{k+1} = i_{k+1}\} \bigcap \{\lambda^{(i_1, \ldots, i_{k+1})}_{K+1} = Z^{(i_1, \ldots, i_{k+1})}_{i_{k+1}}, \ldots, \lambda^{(i_1, \ldots, i_{a-1})}_{a} = Z^{(i_1, \ldots, i_{a-1})}_{i_a}\},$$

for any $0 \leq a \leq k-1$ with the convention $\lambda_0 = \infty$. 24
Appendix C: Proofs

C.1. Proof of Theorem 3

Step 1

We prove by induction on \( k \) the following equality

\[
\{ \hat{\ell}_1 = i_1, \ldots, \hat{\ell}_{K+1} = i_{K+1} \} = \left\{ \lambda_{k+1}^{(i_1, \ldots, i_{k+1})} \leq \frac{e_{i_1, \ldots, i_{k+1}}}{Z_{i_{k+1}}^{(i_1, \ldots, i_{k+1})}} \leq \cdots \leq \left[ \frac{\lambda_1}{Z_{i_1}} \right] \right\},
\]

where the bracket means that, at any position you can choose indifferently the upper- or the lower-case to get an equivalent event.

Using Lemma 8 we have the following equivalences

\[
\{ \hat{\ell}_1 = i_1 \} = \{ \lambda_1 = Z_{i_1} \} = \{ \lambda_2^{(i_1)} \leq Z_{i_1} \} = \{ \lambda_2^{(i_1)} \leq \lambda_1 \}
\]

give the property for \( k = 1 \).

Using the induction hypothesis and Proposition 9 we can pass from \( k \) to \( k + 1 \) to get that

\[
\{ \hat{\ell}_1 = i_1, \ldots, \hat{\ell}_{k+1} = i_{k+1} \} = \left\{ \lambda_{k+2}^{(i_1, \ldots, i_{k+1})} \leq \lambda_{k+1}^{(i_1, \ldots, i_{k+1})} \leq \frac{e_{i_1, \ldots, i_{k+1}}}{Z_{i_{k+1}}^{(i_1, \ldots, i_{k+1})}} \leq \cdots \leq \left[ \frac{\lambda_1}{Z_{i_1}} \right] \right\}.
\]

The \( 2^k \) equivalent events above imply obviously the following \( 2^k \) events

\[
\left\{ \lambda_{k+2}^{(i_1, \ldots, i_{k+1})} \leq \frac{e_{i_1, \ldots, i_{k+1}}}{Z_{i_{k+1}}^{(i_1, \ldots, i_{k+1})}} \leq \lambda_{k+1}^{(i_1, \ldots, i_{k+1})} \leq \cdots \leq \left[ \frac{\lambda_1}{Z_{i_1}} \right] \right\}.
\]

To get the implication in the other direction we remark that

\[
\begin{align*}
\lambda_{k+2}^{(i_1, \ldots, i_{k+1})} &\leq \frac{e_{i_1, \ldots, i_{k+1}}}{Z_{i_{k+1}}^{(i_1, \ldots, i_{k+1})}} \\
\iff & \forall j \neq i_1, \ldots, i_{k+1}, \frac{Z_{i_1, \ldots, i_{k+1}}^{(i_1, \ldots, i_{k+1})}}{Z_{i_1, \ldots, i_{k+1}}^{(i_1, \ldots, i_{k+1})}} \leq \frac{e_{i_1, \ldots, i_{k+1}}}{Z_{i_{k+1}}^{(i_1, \ldots, i_{k+1})}} \\
\iff & \forall j \neq i_1, \ldots, i_{k+1}, \frac{Z_{i_1, \ldots, i_{k+1}}^{(i_1, \ldots, i_{k+1})}}{Z_{i_1, \ldots, i_{k+1}}^{(i_1, \ldots, i_{k+1})}} \leq \frac{e_{i_1, \ldots, i_{k+1}}}{Z_{i_{k+1}}^{(i_1, \ldots, i_{k+1})}} \\
\iff & \forall j \neq i_1, \ldots, i_{k+1}, \frac{Z_{i_1, \ldots, i_{k+1}}^{(i_1, \ldots, i_{k+1})}}{Z_{i_{k+1}}^{(i_1, \ldots, i_{k+1})}} \leq \frac{e_{i_1, \ldots, i_{k+1}}}{Z_{i_{k+1}}^{(i_1, \ldots, i_{k+1})}} \\
\iff & \lambda_{k+1}^{(i_1, \ldots, i_{k+1})} = Z_{i_{k+1}}^{(i_1, \ldots, i_{k+1})}.
\end{align*}
\]

using that (23) and that \( 1 - \tau_{j, i_{k+1}} / \tau_{i_{k+1}, i_{k+1}} > 0 \) (which is a consequence of (22) and \( (\hat{A}_{\text{irr}}) \)) in (27).

Step 2

Using the version 3 of the LAR algorithm we deduce directly the expressions of the expectation and variance of

\[
Z_{i_{k+1}}^{(i_1, \ldots, i_{k-1})}
\]
given by (13) (15) (Changing \( k \) into \( \ell \)). From this same version of this algorithm we get the independence of the \( Z_{i_{k+1}}^{(i_1, \ldots, i_{k-1})} \) when \( \ell \) varies.

Now suppose that we are conditional to \((a, \lambda, \lambda_{k+1}, \hat{\ell}_1, \ldots, \hat{\ell}_{K+1})\), then standard properties of order statistics imply that the density of \((\lambda_1, \ldots, \lambda_k)\), taken at \((\ell_{a+1}, \ldots, \ell_k)\) is proportional to

\[
\left( \prod_{k=a+1}^{K} \varphi_{m_k, \nu_k^2}(\ell_k) \right) \mathbb{I}_{\{\ell_{a+1} \geq \ell_{a+2} \geq \cdots \geq \ell_K \geq \lambda_K \}}.
\]
C.3. Proof of Lemma 7

Suppose now that we are conditional to \( \{ \tilde{m} = a, \lambda_a, \lambda_{K+1}, \tilde{t}_1, \ldots, \tilde{t}_K, \tilde{t}_{K+1} \} \) with \( 0 \leq a \leq K - 1 \). Because of \( \tilde{P}_2 \), \( \lambda_{a+1}, \ldots, \lambda_K \) are independent of the condition \( m = a \), so their conditional distribution is equal to the unconditional distribution given by (28).

C.2. Proof of Proposition 2

Let \( S = \{ j_1, \ldots, j_s \} \subset [p] \) and \( j \in [2p] \setminus S \). Let \( \mathbf{v} = (v_1, \ldots, v_k) \in \{-1, 1\}^k \) and define \( i_\ell = j_\ell + p(1 - v_\ell)/2 \) for \( \ell \in [k] \). Note that

\[
\theta_j(i_1, \ldots, i_k) = (R_{j, i_1} \cdots R_{j, i_k}) M_{i_1, \ldots, i_k}^{-1}(1, \ldots, 1) = \left[ X_j^T X_S \text{Diag}(\mathbf{v}) \right] M_{i_1, \ldots, i_k}^{-1}(1, \ldots, 1) = \left[ X_j^T X_S \text{Diag}(\mathbf{v}) \right] M_{i_1, \ldots, i_k}^{-1} \text{Diag}(\mathbf{v}) = X_j^T X_S S^{-1} v.
\]

Now, observe that

\[
\max_{||v||_1 \leq 1} v^T (X_j^T X_S)^{-1} X_j = \max_{\mathbf{v} \in \{-1, 1\}^k} X_j^T X_S (X_S^T X_S)^{-1} \mathbf{v},
\]

showing the equivalence between the two assumptions.

C.3. Proof of Lemma 7

The proof works by induction. Let us check the relation for \( k = 2 \), namely

\[
N^{(1)} - \lambda_1 \theta(\tilde{t}_1) = Z - Z_{\tilde{t}_1} \theta(\tilde{t}_1) = Z - P_{\tilde{t}_1}(Z).
\]

Now, let \( k \geq 3 \). First, the three perpendicular theorem implies that for every \( j, i_1, \ldots, i_{k-1} \),

\[
\theta_j(i_1, \ldots, i_{k-2}) = (R_{j, i_1} \cdots R_{j, i_{k-1}}) M_{i_1, \ldots, i_{k-1}}^{-1}(\theta_{i_1}(i_1, \ldots, i_{k-2}), \ldots, \theta_{i_{k-1}}(i_1, \ldots, i_{k-2})) = P_{i_1, \ldots, i_{k-2}}(Z_j).
\]

and

\[
P_{i_1, \ldots, i_{k-2}}(Z_j) = (R_{j, i_1} \cdots R_{j, i_{k-1}}) M_{i_1, \ldots, i_{k-1}}^{-1}(P_{i_1, \ldots, i_{k-2}}(Z_{i_1}), \ldots, P_{i_1, \ldots, i_{k-2}}(Z_{i_{k-1}})).
\]

By induction, using (21), we get that

\[
N^{(k-1)} = N^{(k-2)} - (\lambda_{k-2} - \lambda_{k-1}) \theta(\tilde{t}_1, \ldots, \tilde{t}_{k-2}),
\]

\[
= (N^{(k-2)} - \lambda_{k-2} \theta(\tilde{t}_1, \ldots, \tilde{t}_{k-2})) + \lambda_{k-1} \theta(\tilde{t}_1, \ldots, \tilde{t}_{k-2}),
\]

\[
= Z - P_{i_1, \ldots, i_{k-2}}(Z) + \lambda_{k-1} \theta(\tilde{t}_1, \ldots, \tilde{t}_{k-2}).
\]

Then, recall that \( N^{(k-1)} \) for \( j = \tilde{t}_1, \ldots, \tilde{t}_{k-1} \) and remark that

\[
\lambda_{k-1} \theta_j(\tilde{t}_1, \ldots, \tilde{t}_{k-2}) = (R_{j, \tilde{t}_1} \cdots R_{j, \tilde{t}_{k-1}}) M_{\tilde{t}_1, \ldots, \tilde{t}_{k-1}}^{-1}(N^{(k-1)}_{\tilde{t}_1}, \ldots, N^{(k-1)}_{\tilde{t}_{k-1}}).
\]

Using (29) at indices \( j = \tilde{t}_1, \ldots, \tilde{t}_{k-1} \), we deduce that

\[
\lambda_{k-1} \theta_j(\tilde{t}_1, \ldots, \tilde{t}_{k-1})
\]

\[
= (R_{j, \tilde{t}_1} \cdots R_{j, \tilde{t}_{k-1}}) M_{\tilde{t}_1, \ldots, \tilde{t}_{k-1}}^{-1}(N^{(k-1)}_{\tilde{t}_1}, \ldots, N^{(k-1)}_{\tilde{t}_{k-1}})
\]

\[
= (R_{j, \tilde{t}_1} \cdots R_{j, \tilde{t}_{k-1}}) M_{\tilde{t}_1, \ldots, \tilde{t}_{k-1}}^{-1}(Z_{\tilde{t}_1}, \ldots, Z_{\tilde{t}_{k-1}})
\]

\[
= (R_{j, \tilde{t}_1} \cdots R_{j, \tilde{t}_{k-1}}) M_{\tilde{t}_1, \ldots, \tilde{t}_{k-1}}^{-1}(P_{\tilde{t}_1, \ldots, \tilde{t}_{k-2}}(Z_{\tilde{t}_1}), \ldots, P_{\tilde{t}_1, \ldots, \tilde{t}_{k-2}}(Z_{\tilde{t}_{k-1}}))
\]

\[
+ \lambda_{k-1} (R_{j, \tilde{t}_1} \cdots R_{j, \tilde{t}_{k-1}}) M_{\tilde{t}_1, \ldots, \tilde{t}_{k-1}}^{-1}(\theta_{\tilde{t}_1}(\tilde{t}_1, \ldots, \tilde{t}_{k-2}), \ldots, \theta_{\tilde{t}_{k-1}}(\tilde{t}_1, \ldots, \tilde{t}_{k-2}))
\]

\[
= P_{\tilde{t}_1, \ldots, \tilde{t}_{k-2}}(Z_j) - P_{\tilde{t}_1, \ldots, \tilde{t}_{k-2}}(Z_j) + \lambda_{k-1} \theta_j(\tilde{t}_1, \ldots, \tilde{t}_{k-2}).
\]
Namely
\[ P_{t_1, \ldots, \hat{t}_{k-1}}(Z) - \lambda_{k-1}\theta(t_1, \ldots, \hat{t}_{k-1}) = P_{t_1, \ldots, \hat{t}_{k-2}}(Z) - \lambda_{k-1}\theta(t_1, \ldots, \hat{t}_{k-2}). \]

Using again (29) we get that
\[ N^{(k-1)} = Z - P_{t_1, \ldots, \hat{t}_{k-2}}(Z) + \lambda_{k-1}\theta(t_1, \ldots, \hat{t}_{k-2}), \]
\[ = Z - P_{t_1, \ldots, \hat{t}_{k-1}}(Z) + \lambda_{k-1}\theta(t_1, \ldots, \hat{t}_{k-1}), \]
as claimed.

**C.4. Proof of Proposition 9**

We denote
\[
R_j := (R_{j,i_1}, \ldots, R_{j,i_{k-1}}),
\]
\[
R_{ik} := (R_{ik,i_1}, \ldots, R_{ik,i_{k-1}}),
\]
\[
M := M_{i_1, \ldots, i_{k-1}},
\]
\[
\overline{M} := M_{i_1, \ldots, i_k} \begin{bmatrix} M & R_k \\ R^T_k & R_{ik,i_k} \end{bmatrix},
\]
\[
\overline{R} := (R_{j,i_1}, \ldots, R_{j,i_k}),
\]
\[
x := \frac{1 - \theta_j(i_1, \ldots, i_{k-1})}{1 - \theta_k(i_1, \ldots, i_{k-1})} \tau_{j,i_k},
\]

and observe that
\[
x = \frac{R_{j,ik} - R^T_j M^{-1} R_{ik}}{R_{ik,ik} - R^T_{ik} M^{-1} R_{ik}},
\]
\[
\overline{M}^{-1} = \begin{bmatrix} \text{Id}_{k-1} & -M^{-1} R_{ik} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} M^{-1} & 0 \\ 0 & (R_{ik,ik} - R^T_{ik} M^{-1} R_{ik})^{-1} \end{bmatrix} \begin{bmatrix} \text{Id}_{k-1} & 0 \\ -R^T_{ik} M^{-1} & 1 \end{bmatrix},
\]
\[
\overline{M}^{-1} \overline{R} = \begin{bmatrix} M^{-1}(R_j - xR_{ik}) \\ x \end{bmatrix},
\]

using the Schur complement of the block \( M \) of the matrix \( \overline{M} \) and a LU decomposition. Note also that
\[
\frac{Z^{(i_1, \ldots, i_{k-1})}_{i_k} - Z^{(i_1, \ldots, i_{k-1})}}{1 - \tau_{j,i_k}/\tau_{ik,i_k}} = \frac{Z_j - P_{i_1, \ldots, i_{k-1}}(Z_j) - x(Z_{ik} - P_{i_1, \ldots, i_{k-1}}(Z_{ik}))}{1 - \theta_j(i_1, \ldots, i_{k-1}) - x(1 - \theta_k(i_1, \ldots, i_{k-1}))}.
\]

To prove (23), it suffices to show that the R.H.S term above is equal to the following R.H.S term
\[
\frac{Z_j^{(i_1, \ldots, i_k)}}{1 - \theta_j(i_1, \ldots, i_k)}.
\]

We will prove that the numerators are equal and that the denominators are equal. For the denominators, we use that
\[
1 - \theta_j(i_1, \ldots, i_{k-1}) - x(1 - \theta_k(i_1, \ldots, i_{k-1}))
\]
\[= 1 - \theta_j(i_1, \ldots, i_{k-1}) - x + x \theta_k(i_1, \ldots, i_{k-1})
\]
\[= 1 - (\underbrace{1 - \theta_j(i_1, \ldots, i_{k-1})}_k \text{ times}) \begin{bmatrix} M^{-1}(R_j - xR_{ik}) \end{bmatrix}_x
\]
\[= 1 - \theta_j(i_1, \ldots, i_k),
\]

as claimed.
By induction and using (32), it holds that
\[
Z_j - P_{i_1,\ldots,i_{k-1}}(Z_j) - x \left( Z_{i_k} - P_{i_1,\ldots,i_k}(Z_{i_k}) \right) \\
= Z_j - P_{i_1,\ldots,i_{k-1}}(Z_j) - xZ_{i_k} + xP_{i_1,\ldots,i_{k-1}}(Z_{i_k}) \\
= Z_j - (Z_{i_1} \cdots Z_{i_k}) \left[ M^{-1}(R_j - xR_{i_k}) \right] \\
= Z_j - P_{i_1,\ldots,i_k}(Z_j).
\]
using (30).

**C.5. Proof of Proposition 10**

The proof of the first point can be lead by induction. The initialization of the proof is given by the first point of Lemma 8. Now, observe that $Z_{i_1}^{(1,\ldots,i_{k-1})}, \ldots, Z_{i_1}^{(i_1)}, Z_{i_1}$ are measurable functions of $(Z_{i_1}, \ldots, Z_{i_k})$ and one may check that the vector $(Z_{i_1}, \ldots, Z_{i_k})$ is independent of the vector $(Z_{i_{k+1}}^{(i_{k+1})}, (Z_{i_j}^{(i_1,\ldots,i_{k+1})})_{j \neq i_1, \ldots, i_{k+1}})$. We deduce that $Z_{i_k}^{(i_1,\ldots,i_{k-1})}, \ldots, Z_{i_1}^{(i_1)}, Z_{i_1}$ are independent of $(Z_{i_{k+1}}^{(i_{k+1})}, (Z_{i_j}^{(i_1,\ldots,i_{k+1})})_{j \neq i_1, \ldots, i_{k+1}})$. One can also check that $Z_{i_{k+1}}^{(i_1,\ldots,i_{k+1})}$ is independent of $(Z_{i_{k+1}}^{(i_{k+1})})_{j \neq i_1, \ldots, i_{k+1}}$.

The second point also works by induction. The initialization of the proof is given by the second point of Lemma 8. We will use Proposition 9 to prove the second point.

Now, we have
\[
\lambda^{(i_1,\ldots,i_k)}_{k+1} \leq Z_{i_1}^{(i_1,\ldots,i_{k-1})} \\
\Leftrightarrow \forall j \neq i_1, \ldots, i_k, Z_{i_1}^{(i_1,\ldots,i_{k-1})} \leq Z_{i_1}^{(i_1,\ldots,i_{k-1})} \\
\Leftrightarrow \forall j \neq i_1, \ldots, i_k, Z_{i_1}^{(i_1,\ldots,i_{k-1})} = Z_{i_1}^{(i_1,\ldots,i_{k-1})} \\
\Leftrightarrow \forall j \neq i_1, \ldots, i_k, Z_{i_1}^{(i_1,\ldots,i_{k-1})} \leq Z_{i_1}^{(i_1,\ldots,i_{k-1})} \\
\Leftrightarrow \lambda^{(i_1,\ldots,i_{k-1})} \leq Z_{i_1}^{(i_1,\ldots,i_{k-1})}.
\]
using that (23) and that $1 - \tau_{j,i_k}/\tau_{i_k,i_k} > 0$ (which is a consequence of (22) and $(\tilde{A}_{irr.})$) in (32). By induction and using (32), it holds that
\[
\{ \lambda^{(i_1,\ldots,i_k)}_{k+1} \leq Z_{i_1}^{(i_1,\ldots,i_{k-1})} \} \subseteq \{ \lambda^{(i_1,\ldots,i_{k-2})}_{k+1} \leq \cdots \leq \lambda^{(i_1)}_{k+1} \} \\
\Leftrightarrow \{ \forall j \neq i_1, \ldots, i_k, Z_{i_1}^{(i_1,\ldots,i_{k-1})} \leq Z_{i_1}^{(i_1,\ldots,i_{k-1})} \} \subseteq \{ \lambda^{(i_1,\ldots,i_{k-2})}_{k+1} \leq \cdots \leq \lambda^{(i_1)}_{k+1} \} \\
\Leftrightarrow \{ \lambda^{(i_1,\ldots,i_{k-2})}_{k+1} \leq \cdots \leq \lambda^{(i_1)}_{k+1} \}.
\]
\[
\{ \hat{\iota}_1 = i_1, \ldots, \hat{\iota}_k = i_k \}.
\]
Now, observe that $\hat{\iota}_{k+1}$ is the (unique) arg max of $\lambda^{(i_1,\ldots,i_k)}_{k+1}$ on the event $\{ \hat{\iota}_1 = i_1, \ldots, \hat{\iota}_k = i_k \}$. It yields that
\[
\{ \hat{\iota}_1 = i_1, \ldots, \hat{\iota}_{k+1} = i_{k+1} \} = \{ \lambda^{(i_1,\ldots,i_k)}_{k+1} = Z_{i_1}^{(i_1,\ldots,i_k)} \leq Z_{i_1}^{(i_1,\ldots,i_{k-1})} \leq \cdots \leq Z_{i_1}^{(i_1)} \leq Z_{i_1} \},
\]
as claimed. Stopping at \( a \) as in \((s_a)\) gives the second part of the statement.

### C.6. Orthogonal Case: Proof of Theorem 6

Let \( \mathcal{I} \) the set of admissible indexes

\[
\mathcal{I} := \{a, b, c : a_0 \leq a < b < c \leq K + 1\}.
\]

**Step 1:** We prove that, when the considered indexes such that \(c + 1 \leq K + 1\) belong to \( \mathcal{I} \), \( S_{abc} \) is more powerful than \( S_{a,b,c} \). Our proof is conditional to \( F_a = f_a, F_{c+1} = f_{c+1} \). Note that \((F_{a+1}, \ldots, F_c)\) has for distribution the uniform distribution on the simplex

\[
S := \{f_a > F_{a+1} > \cdots > F_c > f_{c+1}\}.
\]

This implies by direct calculations that

\[
\mathcal{I}_{ab}(s, t) = \frac{(s - t)^{b-a-1}}{(b-a-1)!}
\]

and that

\[
\frac{F_{abc}(\lambda_b)}{F_{abc}(\lambda_a)} = F_{\beta((b-a),(c-b))} \left( \frac{F_b - F_c}{f_a - F_c} \right),
\]

where \( F_{\beta} \) is the cumulative distribution of the Beta distribution in reference. Using monotony of this function the \( S_{abc} \) test has for rejection region

\[
(F_b - F_c) \geq z_1(f_a - F_c) \Leftrightarrow F_b \geq z_1f_a + (1 - z_1)F_c,
\]

where \( z_1 \) is some threshold depending on \( \alpha \) that belongs to \((0, 1)\).

Similarly \( S_{ab(c+1)} \) has for rejection region

\[
F_b \geq z_2(f_a - f_{c+1}) + f_{c+1},
\]

where \( z_2 \) is some other threshold belonging to \((0, 1)\). We use the following lemmas.

**Lemma 11.** Let \( c \leq K \). The density \( h_{\mu} \) of \( f_1, \ldots, f_c \), conditional to \( F_{c+1} \) with respect of the Lebesgue measure under the alternative is coordinate-wise non-decreasing and given by \((36)\).

**Proof.** Observe that it suffices to prove the result when \( \sigma = 1 \). Note that

\[
\lambda_{c+1}^{i_1, \ldots, i_c} = \max_{j \in \{p \} \setminus \{i_1, \ldots, i_c\}} |Z_j|.
\]

Thus its density \( p_{\mu_{\lambda_1}^{i_1, \ldots, i_c}, \ldots, \mu_{\lambda_1}^{i_1, \ldots, i_c}} \) does not depend on \( \mu_{\lambda_1}^{0, \ldots, 0} \). As a consequence the following variables have the same distribution ; \( \lambda_{c+1}^{i_1+e_1p, \ldots, i_c+e_cp} \), where \( e_1, \ldots, e_c \) take the value 0 or 1 and indices are taken modulo \( p \).
Because of independence of the different variables, the joint density, under the alternative, of \( \lambda_1, \ldots, \lambda_c+1 \) taken at \( \ell_1, \ldots, \ell_{c+1} \), on the domain \( \{ \lambda_1 > \cdots > \lambda_{c+1} \} \) takes the value

\[
(\text{Const}) \sum' (\phi(\ell_1 - \mu^{0}_{j_1}) + \phi(\ell_1 + \mu^{0}_{j_1})), \ldots, (\phi(\ell_c - \mu^{0}_{j_c}) + \phi(\ell_c + \mu^{0}_{j_c})) p_{\mu^{0},j_1,\ldots,j_c}(\ell_{k+1}).
\]

Here the sum \( \sum' \) is taken over all different \( j_1, \ldots, j_c \) belonging to \([1,p]\).

Then the density, conditional to \( F_{c+1} = f_{c+1} \), of \( F_1, \ldots, F_c \) at \( f_1, \ldots, f_c \) takes the value

\[
(\text{const}) \sum' \cosh(\mu_{j_1} f_1) \ldots \cosh(\mu_{j_c} f_c) \mathbb{1}_{f_1 > \cdots > f_c > f_{c+1}},
\]

implying that this density is coordinate-wise non-decreasing.

**Lemma 12.** Let \( \nu_0 \) the image on the plane \((F_b,F_c)\) on the uniform probability on \( S \): it is the distribution under the null of \((F_b,F_c)\). The two rejection regions : \( R_1 \) associated to \((34)\) and \( R_2 \) associated to \((35)\) have of course the same probability \( \alpha \) under \( \nu_0 \). Let \( \eta_{\nu_0} \) the density w.r.t. \( \nu_0 \) of the distribution of \((F_b,F_c)\) under the alternative.

Then \( \eta_{\nu_0} \) is non decreasing coordinate-wise.

**Proof.** Integration yields that the density of \( \nu_0 \) w.r.t. the Lebesgue measure taken at point \((f_b,f_c)\) is

\[
\frac{(f_a-f_b)^{b-a-1}(f_b-f_c)^{c-b-1}}{(b-a-1)!(c-b-1)!}.
\]

The density of \( \nu_{\nu_0} \) w.r.t. Lebesgue measure is

\[
\int_{f_b}^{f_a} df_{a+1} \ldots \int_{f_b}^{f_{c-2}} df_{a+1} \int_{f_b}^{f_{c-2}} df_{b-1} \int_{f_b}^{f_{c-2}} df_{b+1} \ldots \int_{f_b}^{f_c} df_{c-1} h_{\nu_0}(f_a,\ldots,f_c).
\]

Thus \( \eta_{\nu_0} \) which is the quotient of these two quantities is just a mean value of \( h_{\nu_0} \) on the domain of integration \( D_{f_a,f_c} \) in \((37)\).

Suppose that \( f_b \) and \( f_c \) increase, then all the borns of the domain \( D_{f_a,f_c} \) increase also. By Lemma 11 the mean value increases.

**We finish now the proof of Step 1:** For a given level \( \alpha \) let us consider the two rejection regions \( R_{a,b,c} \) and \( R_{a,b,(c+1)} \) of the two considered tests in the plane \( F_b,F_c \) and set

\[
A := R_{a,b,c} \setminus R_{a,b,(c+1)} \quad \text{and} \quad B := R_{a,b,(c+1)} \setminus R_{a,b,c},
\]

see Figure 5. These two regions have the same \( \nu_0 \) measure. By elementary geometry there exist a point \( K = (K_b,K_c) \) in the plane such that

- For every point of \( A \), \( F_b \leq K_b, F_c \leq K_c \),
- For every point of \( B \), \( F_b \geq K_b, F_c \geq K_c \).

By transport of measure there exists a transport function \( T \) that preserve the measure \( \nu_0 \) and that is one-to one \( A \to B \). As a consequence the transport by \( T \) improve the probability under the alternative: the power of \( S_{a,b,c+1} \) is larger than that of \( S_{a,b,c} \).

- **Step 2:** We prove that, when the considered indexes belong to \( I \) such that \( a < b-1 \), \( S_{a,(b-1),c} \) is more powerful than \( S_{a,b,c} \). Our proof is conditional to \( F_a = f_a, F_b = f_b \) and is located in the plane \((F_{b-1},F_c)\).

The rejection region \( R_{a,b,c} \) takes the form \( F_c \leq \frac{1}{1-z_2} f_b - \frac{z_1}{1-z_2} f_a \) for some threshold \( z_1 \) belonging to \((0,1)\).

The rejection region \( R_{a,(b-1),c} \) takes the form \( F_c \leq \frac{1}{1-z_2} F_{b-1} - \frac{z_1}{1-z_2} f_a \) for some other threshold \( z_2 \) belonging to \((0,1)\).

These regions as well as the regions \( A \) and \( B \) and the point \( K \) are indicated in Figure 5.
Transport of measure and the convenient modification of Lemma 12 imply that the power of $S_{a,(b-1),c}$ is greater of equal that of $S_{a,b,c}$.

**Step 3:** We prove that, when the considered indexes belong to $I$ such that $a + 1 < b$, $S_{a,b,c}$ is more powerful than $S_{(a+1),b,c}$. Our proof is conditional to $F_a = f_a, F_c = f_c$ and is located in the plane $F_{a+1}, F_b$. The rejection region $R_{a,b,c}$ takes the form $F_b \geq z_1f_a + (1 - z_1)f_c$ for some threshold $z_1$ belonging to $(0, 1)$.

The rejection region $R_{a+1,b,c}$ takes the form $F_b \geq z_2F_{a+1} + (1 - z_2)f_c$ for some other threshold $z_2$ belonging to $(0, 1)$.

These regions as well as the regions $A$ and $B$ and the point $K$ are indicated in Figure 5.

Transport of measure and the convenient modification of Lemma 12 imply that the power of $S_{a,b,c}$ is greater of equal that of $S_{(a+1),b,c}$.

Considering the three cases above, we get the desired result.

### C.7. Proof of Theorem 1

We rely on the Weak Positive Regression Dependency (WPRDS) property to prove the result, one may consult [17, Page 173] for instance. We say that a function $g : [0, 1]^K \rightarrow \mathbb{R}^+$ is nondecreasing if for any $p, q \in [0, 1]^K$ such that $p_k \geq q_k$ for any $k = 1, \ldots, K$, we have $g(p) \geq g(q)$. We say that a Borel set $\Gamma \in [0, 1]^K$ is nondecreasing if $g = 1_\Gamma$ is nondecreasing. In other words if $y \in \gamma$ and if $z \geq 0$, then $y + z \in \gamma$. We say that the $p$-values $(\hat{p}_1 = \hat{\alpha}_{0,1,K+1}, \ldots, \hat{p}_K = \hat{\alpha}_{K-1,1,K,K+1})$ satisfy the WPRDS if for any nondecreasing set $\Gamma$ and for all $k^0 \in I_0$, the function

$$u \mapsto \mathbb{P}_{\mu^0}[(\hat{p}_1, \ldots, \hat{p}_K) \in \Gamma | \hat{p}_{k^0} \leq u]$$

is nondecreasing where $\mu^0 = \beta^0$ in our orthogonal design case, and we recall that

$$I_0 = \{k \in [K] : H_{0,k} \text{ is true}\}.$$

To prove Theorem 1, note that it is sufficient [17, Chapter 8] to prove that

$$u \mapsto \mathbb{E}_\mu[(\hat{p}_1, \ldots, \hat{p}_K) \in \Gamma | \hat{p}_{k^0} \leq u]$$

is nondecreasing (38)

where $\mathbb{E}, \mathbb{P}$ will denote that expectations and probabilities are conditional on $\{t_1, \ldots, t_K, \lambda_{K+1}\}$ and under the hypothesis that $\mu^0 = X^TX\beta^0$. Note that one can integrate in $\lambda_{K+1}$ to get the statement of Theorem 1.

**Step 1:** We start by giving the joint law of the LAR’s knots under the alternative in the orthogonal design case. Lemma 11 and (36) show that, conditional on $\{t_1, \ldots, t_K, \lambda_{K+1}\}$, $(\lambda_1, \ldots, \lambda_K)$ is distributed on the set $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K \geq \lambda_{K+1}$ and it has a coordinate-wise nondecreasing density. Now we can assume without loss of generality that $\sigma^2 = 1$, in addition because of orthogonality $\rho^2_k = 1$ implying that $F_k = \Phi(\lambda_k) \mathcal{P}_{i,j} = \Phi = \Phi^{-1} = \text{Id}$. We deduce that, conditional on $\{t_1, \ldots, t_K, F_{K+1}\}$, $(F_1, \ldots, F_K)$ is distributed on the set

$$\{(f_1, \ldots, f_K) \in \mathbb{R}^K : 1 \geq f_1 \geq f_2 \geq \cdots \geq f_K \geq F_{K+1}\},$$

it has an explicit density given by (36), and we denote it by $h_{\mu^0}$. By the change of variables $G_k := \frac{F_k - F_{K+1}}{F_K - F_{K+1}}$ one obtains that the distribution of $(G_1, \ldots, G_K)$ is supported on $[0, 1]^K$. More precisely, define

$$\psi(f_1, \ldots, f_K) := (g_1, \ldots, g_K) := \left(\frac{f_1 - F_{K+1}}{1 - F_{K+1}}, \ldots, \frac{f_K - F_{K+1}}{1 - F_{K+1}}\right)$$

$$\psi^{-1}(g_1, \ldots, g_K) := \left((1 - F_{K+1})g_1 + F_{K+1}, \ldots, (1 - F_{K+1})g_2 \cdots g_K + F_{K+1}\right),$$

31
We deduce that the density of \((G_1, \ldots, G_K)\) at point \(g\) with respect to Lebesgue measure is

\[
p(g) := (\text{const}) I_{g \in (0,1)^K} \prod_{k=1}^K g_k^{K-k} \cosh [\beta_{\Gamma_k}(1 - F_{K+1}) \prod_{\ell=1}^k (1 - g_\ell + F_{K+1})],
\]

where we have used (36). From (18) and (33), one has

\[
\hat{p}_k = 1 - F_{\beta(1,K-k+1)}(G_k) = 1 - F_{\beta(1,K-k+1)}(G_k),
\]

where \(F_{\beta}\) is the cumulative distribution of the Beta distribution in reference. We deduce that for any \(v \in (0,1)\) and for any \(\ell \in [K],

\[
\hat{p}_k = v \iff (G_1, \ldots, G_K) \in [0,1]^K \cap \{F_{\beta(1,K-k+1)}(1-v) = G_k\},
\]

so that

\[
\mathbb{P}[\hat{p}_1, \ldots, \hat{p}_K \in \Gamma | \hat{p}_{k^0} \leq u] = \mathbb{P}[(G_1, \ldots, G_K) \in \Gamma | G_{k^0} \geq F_{\beta(1,K-k+1)}(1-u)],
\]

where \(\Gamma\) can be proved to be a nonincreasing Borel set from (40).

- **Step 2:** Let \(0 < x < y < 1\) and denote by \(\mu_x\) the following conditional law

\[
\mu_x := \text{law}[(G_1, \ldots, G_K) | \{\tau_1, \ldots, \tau_K, F_{K+1}, G_{k^0} \geq x\}].
\]

Remark that if there exists a measurable \(T : [0,1]^K \rightarrow [0,1]^K\) such that

- \(T\) is nondecreasing, meaning that for any \(g \in [0,1]^K, T(g) \geq g\);
- \(T\) is such that push-forward of \(\mu_x\) by \(T\) gives \(\mu_y\), namely \(T^\# \mu_x = \mu_y\);

then it holds

- \(\mathbb{1}_{\{T(g) \in \Gamma\}} \leq \mathbb{1}_{\{g \in \Gamma\}};
- \text{law} [T(G) | \{\tau_1, \ldots, \tau_K, F_{K+1}, G_{k^0} \geq x\}] = \text{law} [G | \{\tau_1, \ldots, \tau_K, F_{K+1}, G_{k^0} \geq y\}]\) where \(G = (G_1, \ldots, G_K)\).

In this case, we deduce that

\[
\mathbb{P}[G \in \Gamma | G_{k^0} \geq x] \geq \mathbb{P}[T(G) \in \Gamma | G_{k^0} \geq x] = \mathbb{P}[G \in \Gamma | G_{k^0} \geq y].
\]

If one can prove that such function \(T\) exists for any \(0 < x < y < 1\), it proves that

\[
x \mapsto \mathbb{P}[G \in \Gamma | G_{k^0} \geq x] \text{ is nonincreasing ,}
\]

and, in view of (41), it proves (38). Proving that such function \(T\) exists is done in the next step.

- **Step 3:** Let \(0 < x < y < 1\). Consider the Knothe-Rosenblatt transport map \(T\) of \(\mu_x\) toward \(\mu_y\) following the order

\[
k^0 \rightarrow k^0 + 1 \rightarrow \cdots \rightarrow K \rightarrow k^0 - 1 \rightarrow k^0 - 2 \rightarrow \cdots \rightarrow 1.
\]
It is based on a sequence of conditional quantile transforms defined following the ordering above. Its construction is presented for instance in [24, Sec.2.3, P.67] or [31, P.20]. The transport \( T \) is defined as follows. Given \( z, z' \in [0, 1]^K \) such that \( z' = T(z) \) it holds

\[
\begin{align*}
z'_{k0} & = T^{(k^0)}(z_{k0}); \\
z'_{k0+1} & = T^{(k^0+1)}(z_{k0+1}, z'_k); \\
& \vdots \\
z'_K & = T^{(K)}(z_K, z'_{K-1}, \ldots, z'_k); \\
z'_{k0-1} & = T^{(k^0-1)}(z_{k0-1}, z'_K, \ldots, z'_k); \\
& \vdots \\
z'_1 & = T^{(1)}(z_1, z'_2, \ldots, z'_{k-1}, z'_K, \ldots, z'_{k0});
\end{align*}
\]

where \( T^{(k^0)}, T^{(k^0+1)}, \ldots, T^{(K)}, T^{(k^0-1)}, \ldots, T^{(1)} \) will be build in the sequel, in which we will drop their dependencies in the \( z''_k \)'s to ease notations. It remains to prove that

- \( T \) is nondecreasing, meaning that for any \( g \in [0, 1]^K \), \( T(g) \geq g \);
- \( T \) is such that push-forward of \( \mu_x \) by \( T \) gives \( \mu_y \), namely \( T_# \mu_x = \mu_y \);

to conclude. The last point is a property of the Knothe-Rosenblatt transport map. Proving the first point will be done in the rest of the proof.

{o Step 3.1:} We start by the first transport map \( T^{(k^0)} : [0, 1] \mapsto [0, 1] \). Denote \( \mu^{(k^0)}_x \) the following conditional law

\[
\mu^{(k^0)}_x := \text{law} \left[ G_{k^0} | (\bar{t}_1, \ldots, \bar{t}_K, F_{K+1}, G_{k^0} \geq x) \right],
\]

and \( F_x^{(k^0)} \) its cdf. Note that the Knothe-Rosenblatt construction gives \( T^{(k^0)} = (F_y^{(k^0)})^{-1} \circ F_x^{(k^0)} \).

We would like to prove that \( T^{(k^0)}(t) \geq t \) for all \( z \in (0, 1) \). This is equivalent to prove that it holds \( F_x^{(k^0)} \geq F_y^{(k^0)} \). For \( t \leq y \), \( F_y^{(k^0)}(t) = 0 \) and it implies that \( F_x^{(k^0)}(t) \geq F_y^{(k^0)}(t) \). Let \( t > y \), using the conditional density \( p \) defined in (39), note that

\[
F_x^{(k^0)}(t) \geq F_y^{(k^0)}(t) \iff \frac{\int_x^t p}{\int_x^1 p} \geq \frac{\int_y^t p}{\int_y^1 p}
\]

\[
\iff \int_x^t \int_y^1 p \otimes p \geq \int_y^t \int_x^1 p \otimes p,
\]

where, for example

\[
\int_x^t \text{ means the integral over the hyper rectangle } [x, t] := \left\{ (g_1, \ldots, g_K) \in [0, 1]^K : x \leq g_{k0} \leq t \right\}.
\]

A simple calculation (see also Figure 6) gives that

\[
\int_x^t \int_y^1 p \otimes p = \int_y^t \int_x^1 p \otimes p + \int_{[x,y] \times [t,1]} p \otimes p,
\]

and it proves that \( F_x^{(k^0)} \geq F_y^{(k^0)} \).

{o Step 3.2:} We continue with the second transport map in Knothe-Rosenblatt construction. Let \( z_{k0} \in (x, 1) \) and denote \( \mu^{(k^0+1)}_{z_{k0}} \) the following conditional law

\[
\mu^{(k^0+1)}_{z_{k0}} := \text{law} \left[ G_{k0+1} | (\bar{t}_1, \ldots, \bar{t}_K, F_{K+1}, G_{k0} = z_{k0}) \right],
\]

33
and $F_{x_{k_0}}^{(k_0+1)}$ its cdf. Let $z'_{k_0} := T^{(k_0)}(z_k)$ and denote $\mu_{z'_{k_0}}^{(k_0+1)}$ the following conditional law

$$\mu_{z'_{k_0}}^{(k_0+1)} := \text{law} \left[ G_{k_0+1} \{ 0, \ldots, 0, \xi_{k_0+1}, F_{k_0+1}, G_{k_0} = z'_{k_0} \} \right],$$

and $F_{x_{k_0}}^{(k_0+1)}$ its cdf. Note that $x < z_{k_0} \leq z'_{k_0} = T^{(k_0)}(z_k) \leq 1$. Again, we would like to prove that $F_{x_{k_0}}^{(k_0+1)} \geq F_{z'_{k_0}}^{(k_0+1)}$ which implies that the transport map $T^{(k_0+1)} := (\mu_{z'_{k_0}}^{(k_0+1)})^{-1} \circ F_{z'_{k_0}}^{(k_0+1)}$ satisfies $T^{(k_0+1)}(u) \geq u$ for all $u \in (0, 1)$.

Recall that the conditional density $\rho$ of $G \{ 0, \ldots, 0, \xi_{k_0+1}, F_{k_0+1}, G_{k_0} = z \}$ is

$$\rho(\xi_{k_0+1} | 0, \ldots, 0, F_{k_0+1}, G_{k_0} = z) = (\text{const}) \mathbb{1}_{g \in (0, 1)^{K}} \prod_{g_{k_0} = z} \prod_{k < k_0} g_{k}^{K-k} \cosh \left[ \mu_{k}^{(0)} \left( (1 - F_{k+1}) \prod_{\ell=1}^{k} g_{\ell} + F_{k+1} \right) \right] \prod_{k > k_0} g_{k}^{K-k} \cosh \left[ \mu_{k}^{(0)} \left( (1 - F_{k+1}) z \prod_{1 \leq \ell \neq k_0 \leq k} g_{\ell} + F_{k+1} \right) \right].$$

Set $\tau := z'_{k_0}/z_{k_0} \geq 1$ and $G'_{k_0+1} = \tau G_{k_0+1}$ so that

$$z_{k_0} G_{k_0+1} = z'_{k_0} G'_{k_0+1}.$$ 

Denote $G' := (G_1, \ldots, G_{k_0-1}, G'_{k_0+1}, G_{k_0+2}, \ldots, G_{K}) \in (0, 1)^{K-1} \times (0, \tau) \times (0, 1)^{K-k^0-1}$ and note that the conditional density of $G' \{ 0, \ldots, 0, F_{k_0+1}, G_{k_0} = z \}$ is

$$(\text{const}) \mathbb{1}_{g \in (0, 1)^{K-1} \times (0, \tau) \times (0, 1)^{K-k^0-1}} \prod_{k < k_0} g_{k}^{K-k} \cosh \left[ \mu_{k}^{(0)} \left( (1 - F_{k+1}) \prod_{\ell=1}^{k} g_{\ell} + F_{k+1} \right) \right] \prod_{k > k_0} g_{k}^{K-k} \cosh \left[ \mu_{k}^{(0)} \left( (1 - F_{k+1}) z \prod_{1 \leq \ell \neq k_0 \leq k} g_{\ell} + F_{k+1} \right) \right],$$

which, up to some normalising constant, is the same as (42) up to the following change of support

$$\mathbb{1}_{g \in (0, 1)^{K}} \leftrightarrow \mathbb{1}_{g' \in (0, 1)^{K-1} \times (0, \tau) \times (0, 1)^{K-k^0-1}}.$$ 

By an abuse of notation, we denote by $\rho$ this function, namely

$$\rho(\xi_{k_0+1} | 0, \ldots, 0, F_{k_0+1}, G_{k_0} = z) = (\text{const}) \prod_{k < k_0} g_{k}^{K-k} \cosh \left[ \mu_{k}^{(0)} \left( (1 - F_{k+1}) \prod_{\ell=1}^{k} g_{\ell} + F_{k+1} \right) \right] \prod_{k > k_0} g_{k}^{K-k} \cosh \left[ \mu_{k}^{(0)} \left( (1 - F_{k+1}) z \prod_{1 \leq \ell \neq k_0 \leq k} g_{\ell} + F_{k+1} \right) \right].$$
Lemma 13. Let $a,b \geq 0$. The function

$$z \mapsto \cosh(a \, z + b) \times \cosh(a / z + b)$$

is non-decreasing on the domain $[1, \infty)$. 

Now, let $(g_1, \ldots, g_{k_0-1}, g_{k_0+2}, \ldots, g_K) \in (0,1)^{K-1}$ be fixed in the integrals (43). We are the looking at the weights of the domains $(h_1, h_2) \in (0,t) \times (0,\tau)$ and $(h_3, h_4) \in (0,\tau t) \times (0,1)$ for the weight function $w$ given by

$$w(h_1, h_2) = C_1 h_1^{K-k_0-1} \cosh [\mu_{\tau_k}^0 ((1-F_{K+1}) z_{k_0} \prod_{1 \leq \ell \leq k_0} g_{\ell} \times h_1 + F_{K+1})] \prod_{k > k_0+1} \cosh [\mu_{\tau_{k_0+k}^0} ((1-F_{K+1}) z_{k_0} \prod_{1 \leq \ell \leq k_0} g_{\ell} \times h_2 + F_{K+1})] \times h_2^{K-k_0-1} \cosh [\mu_{\tau_{k_0+k}^0} ((1-F_{K+1}) z_{k_0} \prod_{1 \leq \ell \leq k_0} g_{\ell} \times h_2 + F_{K+1})],$$

where the constant $C_1$ depends on $(g_1, \ldots, g_{k_0-1}, g_{k_0+2}, \ldots, g_K) \in (0,1)^{K-1}$. By the change of variables $h_1' = h_3/t$ and $h_2' = th_4$, the right hand term of (43) is given by the integration on the
domain \((h'_1, h'_2) \in (0, t) \times (0, \tau)\) of the weight function \(w'\) given by

\[
\begin{align*}
w'(h'_1, h'_2) &= C_1 h^1_1 K^{-k^0 - 1} \cosh [\mu^{0}_{\tau_0} ((1 - F_{K+1}) z_{k^0} \prod_{1 \leq \ell < k^0} g_{\ell} \times t \times h'_1 + F_{K+1})] \\
& \quad \times \prod_{k > k^0 + 1} \cosh [\mu^{0}_{\tau_0} ((1 - F_{K+1}) z_{k^0} \prod_{1 \leq \ell < k^0, \ell \neq k^0, k^0 + 1 \leq k} g_{\ell} \times t \times h'_2 + F_{K+1})] \\
& \quad \times h^2_2 K^{-k^0 - 1} \cosh [\mu^{0}_{\tau_0} ((1 - F_{K+1}) z_{k^0} \prod_{1 \leq \ell < k^0} g_{\ell} \times h'_2 / t + F_{K+1})] \\
& \quad \times \prod_{k > k^0 + 1} \cosh [\mu^{0}_{\tau_0} ((1 - F_{K+1}) z_{k^0} \prod_{1 \leq \ell < k^0, \ell \neq k^0, k^0 + 1 \leq k} g_{\ell} \times h'_2 / t + F_{K+1})] .
\end{align*}
\]

Now, invoke Lemma 13 with

\[
\begin{align*}
a &= \mu^{0}_{\tau_0} ((1 - F_{K+1}) z_{k^0} \prod_{1 \leq \ell < k^0} g_{\ell} \times h \\
b &= \mu^{0}_{\tau_0} F_{K+1} \\
z &= t \geq 1,
\end{align*}
\]

where \(h = h_1\) or \(h_2\), to get that \(w' \geq w\) and so

\[
\int_{D(t) \times D(\tau)} p \otimes p \geq \int_{D(\tau) \times D(1)} p \otimes p ,
\]

which concludes this part of the proof.

\textbf{\textcircled{3.3}}: We continue by induction with the other transport maps in Knothe-Rosenblatt’s construction. Assume that we have built \(z' := (z'_k, \ldots, z'_k)\) and \(z := (z_k, \ldots, z_{k^0})\) for some \(k > k^0\).

Denote \(\mu^{(k+1)}_z\) the following conditional law

\[
\mu^{(k+1)}_z := \text{law} \left[ G_{k+1} \left\{ \tau_1, \ldots, \tau_K, F_{K+1}, G_k = z_k, \ldots, G_{k^0} = z_{k^0} \right\} \right] ,
\]

and \(F^{(k+1)}_z\) its cdf. Denote \(\mu^{(k+1)}_{z'}\) the following conditional law

\[
\mu^{(k+1)}_{z'} := \text{law} \left[ G_{k+1} \left\{ \tau_1, \ldots, \tau_K, F_{K+1}, G_k = z'_k, \ldots, G_{k^0} = z'_{k^0} \right\} \right] ,
\]

and \(F^{(k+1)}_{z'}\) its cdf. Note that \(z \leq z' = T^{(k)}(z) \leq 1\). Again, we would prove that \(F^{(k+1)}_z \geq F^{(k+1)}_{z'}\) which implies that the transport map \(T^{(k+1)} := (F^{(k+1)}_z)^{-1} \circ F^{(k+1)}_{z'}\) satisfies \(T^{(k+1)}(u) \geq u\) for all \(u \in (0, 1)\).

For \(z \in (0, 1)^{k-k^0} \times (x, 1)\), the conditional density of \(G \{ \tau_1, \ldots, \tau_K, F_{K+1}, G^{(k-k^0)} = z \}\) is

\[
\begin{align*}
\text{const} \, 1_{g \in (0, 1)^K} 1_{g^{(k,k^0)} = z} \prod_{m < k^0} g^{K-m}_m \cosh \left[ \mu^{0}_{\tau_0} ((1 - F_{K+1}) \prod_{\ell=1}^m g_{\ell} + F_{K+1}) \right] \\
& \quad \times \prod_{k^0 \leq m < k} z^{K-m}_m \cosh \left[ \mu^{0}_{\tau_0} ((1 - F_{K+1}) \prod_{1 \leq \ell < k^0, \ell \neq k^0, k^0 + 1 \leq k} g_{\ell} \times m \times z_m + F_{K+1}) \right] \\
& \quad \times \prod_{k < m} g^{K-m}_m \cosh \left[ \mu^{0}_{\tau_0} ((1 - F_{K+1}) \prod_{1 \leq \ell < k^0} g_{\ell} \times m \times z_m \prod_{k < \ell \leq m} g_{\ell} + F_{K+1}) \right].
\end{align*}
\]
Set \( \tau := \prod_{n=k^0}^{k} z'_n / \prod_{n=k^0}^{k} z_n \geq 1 \) and \( G'_k = \tau G_{k+1} \) so that
\[
\left[ \prod_{n=k^0}^{k} z'_n \right] G_{k+1} = \left[ \prod_{n=k^0}^{k} z_n \right] G'_{k+1}.
\]

Then the proof follows the same idea as in Step 3.2 and we will not detail it here.

\( \circ \) Step 3.4: This is the last step of the proof. Assume that we have built \( z' := (z'_K, \ldots, z'_k) \) and \( z := (z_K, \ldots, z_k) \). Denote \( \mu_{z'}^{(k^0-1)} \) the following conditional law
\[
\mu_{z'}^{(k^0-1)} := \text{law}[G_{k^0-1}|\{t_1, \ldots, t_K, F_{K+1}, G^{[K,k^0]} = z}],
\]
and \( F_{z'}^{(k^0-1)} \) its cdf. Denote \( \mu_{z'}^{(k^0-1)}(T_{z'}^{(k^0-1)}) \) the following conditional law
\[
\mu_{z'}^{(k^0-1)} := \text{law}[G_{k^0-1}|\{t_1, \ldots, t_K, F_{K+1}, G^{[K,k^0]} = z}],
\]
and \( F_{z'}^{(k^0-1)} \) its cdf. Note that \( z \leq z' = T^{(K)}(z) \leq 1 \). Again, we would prove that \( F_{z'}^{(k^0-1)} \geq F_{z'}^{(k^0-1)} \) which implies that the transport map \( T^{(k^0-1)} := (F_{z'}^{(k^0-1)})^{-1} \circ F_{z'}^{(k^0-1)} \) satisfies \( T^{(k^0-1)}(u) \geq u \) for all \( u \in (0,1) \).

For \( z \in (0,1)^{K-k^0} \times (x,1) \), the conditional density of \( G^{[1, \ldots, t_K, F_{K+1}, G^{[K,k^0]} = z]} \) is

\[
(const) \prod_{g \in (0,1)^{x \times (K,k^0)}} \prod_{m<k^0} g_m^{K-m} \cosh \left[ \mu_{m}^{(0)} \left( (1 - F_{K+1}) \prod_{\ell=1}^{m} g_{\ell} \times F_{K+1} \right) \right]
\]

\[
\times \prod_{k^0 \leq m \leq K} z_m^{K-m} \cosh \left[ \mu_{m}^{(0)} \left( (1 - F_{K+1}) \prod_{1 \leq \ell < k^0} g_{\ell} \prod_{n=k^0}^{k^0} z_n \times F_{K+1} \right) \right].
\]

Now, let \((g_1, \ldots, g_{k^0-2}) \in (0,1)^{k^0-2} \) be fixed and denote by

\[
\forall g \in (0,1), \quad w_z(g) := g^{K-k^0-1} \cosh \left[ \mu_{m}^{(0)} \left( (1 - F_{K+1}) \prod_{\ell=1}^{m} g_{\ell} \times g \times F_{K+1} \right) \right]
\]

\[
\times \prod_{k^0 \leq m \leq K} z_m^{K-m} \cosh \left[ \mu_{m}^{(0)} \left( (1 - F_{K+1}) \prod_{n=k^0}^{k^0} z_n \prod_{\ell=1}^{m} g_{\ell} \times g \times F_{K+1} \right) \right].
\]

and, substituting \( z \) by \( z' \), define \( w_{z'} \) as well. Let \( t \in (0,1) \). Following the idea of Step 3.2, one can check that it is sufficient to prove that
\[
\int_{0}^{t} \left( \int_{0}^{1} w_z(g)w_{z'}(g')dg' \right)dg \geq \int_{0}^{1} \left( \int_{0}^{t} w_z(g)w_{z'}(g')dg' \right)dg .
\]

Substituting
\[
\int_{0}^{t} \left( \int_{0}^{1} w_z(g)w_{z'}(g')dg' \right)dg
\]
on both parts, one is reduced to prove that
\[
\int_{0}^{t} \left( \int_{0}^{t} w_z(g)w_{z'}(g')dg' \right)dg \geq \int_{0}^{t} \left( \int_{0}^{1} w_{z'}(g)w_z(g')dg' \right)dg .
\]

Observe that \( g \leq g' \) in the last two integrals. Now, we have this lemma whose proof is omitted.
Lemma 14. Let $0 < a \leq a'$ and $b > 0$. The function

$$z \mapsto \frac{\cosh(az + b)}{\cosh(a'z + b)}$$

is non-increasing on the domain $(0, \infty)$.

Let $g \leq g'$. From Lemma 14, we deduce that $\cosh(ag + b) \cosh(a'g' + b) \geq \cosh(ag' + b) \cosh(a'g + b)$, proving that $w_z(g)w_z(g') \geq w_z(g)w_z(g')$. It proves that $T(k^{(k-1)}(u) \geq u$ for all $u \in (0, 1)$.

We then proceed by induction for $k^0 = k^0 - 1 \rightarrow k^0 - 2 \rightarrow \cdots \rightarrow 1$. The proof follows the same line as above, Step 3.4.

Proof of Theorem 5 From Corollary 4 we deduce that, under $H_0$ and conditionally on the selection event $\{\hat{m} = a\}$ with $a \leq K - 1$, (17) remains true, giving the conditional cumulative distribution function of $\lambda_0$. As a consequence and under the same conditioning,

$$\frac{F_{abc}(\lambda_0)}{F_{abc}(\lambda_a)} \sim U(0, 1).$$

Since the conditional distribution doesn’t depend on the condition, it is also the unconditional distribution. Finally considerations of distribution under the alternative show that to obtain a $p$-value we must consider the complement to 1 of the quantity above.

Appendix D: Main notation

- $[a]$ the set of integers $\{1, \ldots, a\}$
- $X$ a full rank $n \times p$ design matrix
- $\sigma^2$ the variance of the errors
- $r_k, \varepsilon_k$ the indexes and the signs of the variables that enter in the LAR path
- $\hat{s}_k$ a way of coding both indexes and signs, see (1)
- $S^k \{t_1, \ldots, t_k\}$, a possible selected support
- $S_0$ the true support
- $E_k$ Span($X_{t_1}, \ldots, X_{t_k}$)
- $P_k$ projector on (the orthogonal of) $E_k$
- $\gamma_{m_k, \nu^2_k}(\phi_{m_k, \nu^2_k})$ The normal distribution (density) with mean $m_k$ given (13) and variance $\nu^2_k = \sigma^2 \rho^2_k$ given by (15)
- $Z$ the vector of correlations, obtained by sym. from $\overline{Z}$ defined by (9)
- $R$ the variance-covariance matrix of $Z$, see (10)
- $K$ the number of knots that are considered, see Figure 2 and Section 2.5
- $K_{select}$ see Figure 2
- $\hat{m}$ the chosen size
- $\hat{S}$ the chosen set of variables : $\overline{S}^{\hat{m}}$
- $\rho^2_k$ $\nu^2_k/\sigma^2$, see (13)
- $\theta_j(i_1, \ldots, i_k)$ see (11)
- $\theta^t$ $\theta(i_1, \ldots, i_t)$
- $M_{i_1, \ldots, i_k}$ the submatrix of $R$ obtained by keeping the columns, and the rows indexed by $\{i_1, \ldots, i_k\}$
- $F_{ij} \colon P_{ij} := \Phi_j(\lambda_i) := \Phi(\lambda_i/\sigma \rho_i)$ and $P_{ij}$ is given by (14)
- $F_{abc}(t)$ see (16)
- $\hat{\alpha}_{abc}$ the $p$-value of the generalized spacing test, see (18)
- $S_{abc} \mathbf{1}(\hat{\alpha}_{abc} \leq \alpha)$, the generalized spacing test see (19)
Appendix E: Cubature of integral by lattice rule

Our goal is to compute the integral of some function $f$ on the hypercube of dimension $d$, namely

$$I := \int_{[0,1]^d} f(x) dx.$$

We want to approximate it by a finite sum over $n$ points

$$I_n := \frac{1}{n} \sum_{i=1}^{n} f(x^{(i)}).$$

A convenient way of constructing the sequence $x^{(i)}, i = 1, \ldots, n$ is the so-called *lattice rule*: from the first point $x^{(1)}$ we deduce the others $x^{(i)}$ by

$$x^{(i)} = \{ i.x^{(1)} \},$$

where the $\{ \}$ brackets mean that we take the fractional part coordinate by coordinate. In such a case the error given by

$$E(f, n, x^{(1)}) = I - I_n$$

is a function, in particular, of starting point $x^{(1)}$.

The Fast-rank algorithm [21] is a fast algorithm that finds, component by component and as a function of the prime $n$, the sequence of coordinates of $x^{(1)}$ that minimizes the maximal error when $f$ varies in a unit ball $\mathcal{E}$ of some *RKHS*, namely a tensorial product of *Koborov spaces*. In addition it gives an expression of its minimax error, namely

$$\max_{f \in \mathcal{E}}(f, n, x^{(1)}).$$

In practice, very few properties are known on the function $f$, so the result above is not directly applicable. Nevertheless for many functions $f$, it happens that the convergence of $I_n$ to $I$ is "fast": typically of the order $1/n$ while the Monte-Carlo method (choosing the $x^{(i)}$ at random) converges at rate $1/\sqrt{n}$.

A reliable estimation of the error is obtained by adding a *Monte-Carlo layer* as in [16] for instance. This can be done as follows. Let $U$ a *unique* uniform variable on $[0,1]^d$, we define

$$x_{U}^{(i)} := \{ i.x^{(1)} + U \}, \quad I_{n, U} := \frac{1}{n} \sum_{i=1}^{n} f(x_{U}^{(i)}).$$

Classical computations show that $I_{n, U}$ is now an unbiased estimator of $I$. In a final step, we perform $N$ (in practice 15-20) independent repetitions of the experiment above an we compute usual asymptotic confidence intervals for independent observations.