Multiple Testing and Variable Selection along the path of the Least Angle Regression

JEAN-MARC AZAÏS
Institut de Mathématiques de Toulouse
Université Paul Sabatier, 118 route de Narbonne, F-31062 Toulouse, France
jean-marc.azais@univ-toulouse.fr

AND

YOHANN DE CASTRO*,
Institut Camille Jordan UMR 5208, École Centrale Lyon
36 Avenue Guy de Collongue, F-69134 Écully, France
*Corresponding author: yohann.de-castro@ec-lyon.fr

We investigate multiple testing and variable selection using the Least Angle Regression (LARS) algorithm in high dimensions under the assumption of Gaussian noise. LARS is known to produce a piecewise affine solution path with change points referred to as the *knots of the LARS path*. The key to our results is an expression in closed form of the exact joint law of a K-tuple of knots conditional on the variables selected by LARS, the so-called *post-selection* joint law of the LARS knots. Numerical experiments demonstrate the perfect fit of our findings.

This paper makes three main contributions. First, we build testing procedures on variables entering the model along the LARS path in the general design case when the noise level can be unknown. These testing procedures are referred to as the Generalized t-Spacing tests (GtSt) and we prove that they have an exact non-asymptotic level (i.e., the Type I error is exactly controlled). This extends work of Tibshirani et al. [2016] where the spacing test works for consecutive knots and known variance. Second, we introduce a new exact multiple testing procedure after model selection in the general design case when the noise level may be unknown. We prove that this testing procedure has exact non-asymptotic level for general design and unknown noise level. Third, we prove exact control of the false discovery rate under orthogonal design assumption. Monte Carlo simulations and a real data experiment are provided to illustrate our results in this case. Of independent interest, we introduce an equivalent formulation of the LARS algorithm based on a recursive function.

Keywords: Multiple Testing; False Discovery Rate; High-Dimension; Selective Inference.

2000 Math Subject Classification: Primary 62E15, 62F03, 60G15, 62H10, 62H15; secondary 60E05; 60G10; 62J05; 94A08

1. Introduction

In the past decades, statistical problems have become increasingly high-dimensional, *i.e.,* they require estimation of more parameters than the number of available samples/observations. Some examples range from signal processing [Candès et al., 2006, Chen et al., 1998] to genomics [Barber et al., 2015, Rhee et al., 2006]. Some successful techniques of estimation have been developed and a popular approach is based on optimizing a suitable regularized likelihood function. Most models of statistical parameters are well approximated by sparse vectors; and sparsity promoting regularizations, such as the ℓ_1-norm, are now well recognized to tackle high-dimensional problems. Recent advances have focused on a deeper understanding of the law of the estimates of ℓ_1-regularization procedures in high-dimension. One goal is to quantify the uncertainty of some linear statistic of the outcomes of sparse regression estimation. Such estimators are non-linear and non-explicit. They are defined as the minimum of some optimization program, or as the outcomes of some greedy method. Most of them estimate some set of relevant parameters, *i.e.,* a small number of parameters that may explain the observation. This non-linear framework makes it impossible to characterize the distribution of the estimator. One possibility is to look at some conditional distribution of the estimator and this is the scope of the so-called selective inference, which produces an uncertainty quantification conditional on the set of indices of nonzero estimated parameters, referred to as the selection event. Selective inference aims at building some confidence intervals and some testing procedures on the estimates (see [van de Geer, 2016, Chapter 6] and references therein), or controlling the false discovery rate, *e.g.*, Barber et al. [2015] for instance.

One of the most popular regularized estimation procedure in high-dimensions is LASSO [Chen et al., 1998] and its asymptotic de-biased version referred to as the debiased LASSO. Controlling the FDR (resp., confidence intervals (CI)) built upon the debiased LASSO procedure has been studied in Javanmard et al. [2019] (resp., Javanmard and Montanari [2014]) which provides an FDR with asymptotic control (resp., the CI with asymptotic control of the confidence level) for designs with some independent sub-Gaussian rows. The LASSO is based on ℓ_1-norm regularization and one of its offsprings is the sorted-ℓ_1 regularization, referred to as the SLOPE, which achieves minimax rate of prediction and estimation. Controlling the FDR for SLOPE with the Benjamini-Hochberg (BH) selection procedure has been achieved in Bogdan et al. [2015] for orthogonal designs.

Inference after model selection has been studied in several papers, such as Fithian et al. [2014], Taylor and Tibshirani [2015] (resp., Tian et al. [2018]) for selective inference (resp., for a joint estimate of the noise level). These works give the non-asymptotic law of any linear statistics, *i.e.,* any linear combination of the estimates of the parameters, conditional on the selection event. For the first time, this paper provides the non-asymptotic joint law of **several linear statistics** conditional on the selection event. These linear statistics are given by the knots of the LARS procedure. One may note that, conditional on the selection event, the law of three consecutive knots has been studied by Lockhart et al. [2014] who refer to it as the spacing test (ST) [Tibshirani et al., 2015]. The article Azaïs et al. [2018] proved that the spacing test is unbiased and introduce a studentized version of this test. In the same direction, inference after model selection has been studied in several papers, such as Fithian et al. [2014], Taylor and Tibshirani [2015] and respectively Tian et al. [2018] for *selective inference* and respectively a joint estimate of the noise level.

In the present paper, our test is based on the conditional joint law of three, **not necessarily consecutive**, knots. In this way, we extend the work from Tibshirani et al. [2016] where the spacing test works for consecutive knots. We refer to these new tests as the generalized spacing tests (GSt). Furthermore, the exact formulation of the spacing test of the pioneering work of Tibshirani et al. [2016] requires extra computations of the term denoted by M^+ in [Tibshirani et al., 2016, Lemma 5]. They proved that the spacing test is asymptotically equivalent to the conservative spacing test. We remove this restriction and we prove that it suffices to check wether the so-called *Irrepresentable Check* Condition holds to get a non-asymptotic equivalence between the Spacing test and the conservative Spacing test. Finally, we theoretically prove that working with non-consecutive knots can render the testing procedure more powerful.

1.1 *Joint law of LARS knots in Post-Selection Inference*

In this paper, we consider linear models in high-dimensions where the number of observations n may be less than the number of predictors p. We denote by $Y \in \mathbb{R}^n$ the response variable and we assume that

$$Y = X\beta^0 + \eta \sim \mathscr{N}_n(X\beta^0, \sigma^2 \mathrm{Id}_n), \tag{1.1}$$

where $\eta \sim \mathscr{N}_n(0, \sigma^2 \mathrm{Id}_n)$ is a Gaussian noise, the noise level $\sigma > 0$ may be known or may have to be estimated (depending on the context), and $X \in \mathbb{R}^{n\times p}$ has rank $r > 0$. We consider the LARS and denote by $(\lambda_k)_{k\geqslant 1}$ the sequence of knots and by $(\bar{\imath}_k, \varepsilon_k)_{k\geqslant 1}$ the sequence of variables $\bar{\imath}_k \in [p]$ and signs $\varepsilon_k \in \{\pm 1\}$ that enter the model along the LARS path. We encode by

$$\widehat{\imath}_k := \bar{\imath}_k + p\left(\frac{1-\varepsilon_k}{2}\right) \in [2p],$$

both the variables $\bar{\imath}_k \in [p]$ and the signs $\varepsilon_k \in \{\pm 1\}$, calling them the 'signed variables'. Section 5.1 recalls LARS (Algorithm 2) and present equivalent formulations in Algorithm 3 (using orthogonal projections) and Algorithm 4 (using a recursion). In particular, Algorithm 4 consists in three lines, applying the same function recursively, see Section 5.2. As far as we know, **Algorithm 4 is new**.

For a short moment, consider the simplest linear model, where one observes the target vector $\beta^0 = (\beta^0_1, \ldots, \beta^0_p) \in \mathbb{R}^p$, namely there is no noise and the design $X = \mathrm{Id}_p$ is the identity. In this case, LASSO and LARS give the same knots $\lambda_1, \lambda_2, \ldots$ and the estimate of the LASSO is the outcome of the proximal operator of the ℓ_1-norm at point $\beta^0 \in \mathbb{R}^p$, see for instance [Tibshirani et al., 2015, Chapter 2]. In this simple case, we deduce that the knots are

$$\lambda_k = \beta^0_{(k)}, \tag{1.2}$$

where we have considered the reordering $\beta^0_{(1)} \geqslant \beta^0_{(2)} \geqslant \ldots$ of the entries of the target. Obviously, this is no longer true for general designs in high-dimensions with noise, but one may ask:

[Q1] *What is the* **joint law** *of the LARS knots $\lambda_1, \lambda_2, \ldots$ and how do they relate to the target β^0?*

We will answer [**Q1**] in high-dimensions under the assumption of Gaussian noise in Section 3.1.3 and Section 3.2. Working with the so-called 'Irrepresentable Check' Condition[1] ($\mathscr{A}_{\mathrm{Irr.}}$), which can be efficiently checked in practice, we are able to provide the joint law of the LAR's knots conditional on the so-called 'selection event' defined by

$$\mathscr{E} := \left\{\widehat{\imath}_1 = \imath_1, \ldots, \widehat{\imath}_K = \imath_K, \lambda_{K+1}\right\}.$$

This selection event states that the signed variable $\imath_k$ has been selected by the LARS algorithm at its k^{th} step for $k = 1, \ldots, K$. This is the cornerstone of the paper, showing that the conditional joint distribution of the LARS knots is a mixture of Gaussian order statistics, as presented in the next theorem.

[1] See Section 2.3 for a definition and detailed comments on this assumption.

THEOREM 1.1 (Conditional Joint Law of the LARS knots) *Let $(\lambda_1,\dots,\lambda_K,\lambda_{K+1})$ be the first knots of the LARS and let $(\widehat{\imath}_1,\dots,\widehat{\imath}_K)$ be the first variables entering along the LARS path. If $(\widehat{\imath}_1,\dots,\widehat{\imath}_K)$ satisfies $(\mathscr{A}_{\mathrm{Irr.}})$, then, conditional on the selection event $\{\widehat{\imath}_1,\dots,\widehat{\imath}_K,\lambda_{K+1}\}$, the vector $(\lambda_1,\dots,\lambda_K)$ obeys a law with the following density (w.r.t. Lebesgue measure)*

$$\mathrm{Q}^{-1}_{(\widehat{\imath}_1,\dots,\widehat{\imath}_K,\lambda_{K+1})}\Big(\prod_{k=1}^{K}\varphi_{m_k,v_k^2}(\ell_k)\Big)\mathbb{1}_{\{\ell_1\geqslant\ell_2\geqslant\dots\geqslant\ell_K\geqslant\lambda_{K+1}\}}\ \textit{at point}\ (\ell_1,\ell_2,\dots,\ell_K),$$

where $\mathrm{Q}_{(\widehat{\imath}_1,\dots,\widehat{\imath}_K,\lambda_{K+1})}$ is a normalizing constant, φ_{m_k,v_k^2} is the standard Gaussian density with mean m_k and variance $v_k^2 := \sigma^2\rho_k^2$, are explicitly given by (3.12) *and* (3.13).

The proof of this theorem is given in Section 3.2.1. Now, let us describe the dependency between (m_k,ρ_k^2) and $\overline{\mu}^0 := X^\top X\beta^0$. For a design matrix X with columns $(X_j)_{j=1}^p$, we denote by[2]

$$\{0\} =: H_0 \subset H_1 \subset \cdots \subset H_k := \mathrm{Span}(X_{\bar{\imath}_1},\dots,X_{\bar{\imath}_k}) \subset \cdots \subset H_K.$$

By (3.12) and (3.13), one has

$$\forall k\in[K],\quad m_k = c_k\,\varepsilon_k\,\langle X_{\bar{\imath}_k}, P_{k-1}^{\perp}(X\beta^0)\rangle \quad\text{and}\quad \rho_k^2 = d_k\sin\measuredangle(X_{\bar{\imath}_k},H_{k-1}),$$

where $c_k, d_k > 0$ are constants that depends only on $X_{\bar{\imath}_1},\dots,X_{\bar{\imath}_{k-1}}$, $P_{k-1}^{\perp}$ denotes the orthogonal projection onto the orthogonal of H_{k-1}, ε_k is the sign of the k^{th} variable entering the LARS path, and $\measuredangle(X_{\bar{\imath}_k},H_{k-1})$ is the angle between $X_{\bar{\imath}_k}$ and H_{k-1}.

1.2 *The Generalized t-Spacing test (GtSt)*

This paper introduces a class of exact tests built from ℓ_1-minimization regression in high-dimensions. More precisely, we design a testing procedure for a null hypothesis of the form

$$\mathbb{H}_0\ :\ `X\beta^0 \in H_{a_0}\text{'},$$

where $H_{a_0} := \mathrm{Span}(X_{\bar{\imath}_1},\dots,X_{\bar{\imath}_{a_0}})$. Note that the null $\mathbb{H}_0$ is equivalent to the hypothesis that all the true positives (i.e., the support of β^0) are among the first a_0 variables selected by LARS, namely $\{\bar{\imath}_1,\dots,\bar{\imath}_{a_0}\}$. Following the original idea of Lockhart et al. [2014], we study testing procedures of $\mathbb{H}_0$ based on the knots of the LARS path. Note that, conditional on the selection event, the law of three consecutive knots has been studied by Lockhart et al. [2014], where it was referred to as the *spacing test* (ST) [Tibshirani et al., 2015]. The article Azaïs et al. [2018] proved that the spacing test is unbiased, and introduced a Studentized version of this test. In the same direction, inference after model selection has been studied in several papers, such as Fithian et al. [2014], Taylor and Tibshirani [2015] and respectively Tian et al. [2018] for selective inference and respectively a joint estimate of the noise level. This raises the following questions.

[Q2] *Can we provide exact testing procedures based on knots that are not consecutive?*

[Q3] *What is the most powerful test among these spacing tests?*

[Q4] *Can we provide exact testing procedures when the noise level is not known?*

[2] Recall that the selected variables $\widehat{\imath}_k \in [2p]$ are decomposed into $\widehat{\imath}_k := \bar{\imath}_k + p\,(\frac{1-\varepsilon_k}{2})$.

• **Contribution** (**i**) : First, our test is based on the conditional joint law of three, not necessarily consecutive, knots $a_0 \leqslant a < b < c \leqslant K+1$. In this way, we extend the work of Tibshirani et al. [2016], where the spacing test works for consecutive knots. We present this framework in Section 3.2 and we refer to these new tests are Generalized Spacing tests (GSt). At the first reading of the next theorem, one can set $\widehat{m} = a_0$ for a fixed value $0 \leqslant a_0 \leqslant a$. The selection procedure, defining $\widehat{m}$, will be presented in Section 2.5 with the notion of an 'admissible procedure' ($\mathscr{A}_{\text{Stop}}$).

THEOREM 1.2 *Let* $a, b,$ *and* c *be such that* $0 \leqslant a < b < c \leqslant K+1$. *Let* $(\lambda_1, \dots, \lambda_K, \lambda_{K+1})$ *be the first knots and let* $(\widehat{\imath}_1, \dots, \widehat{\imath}_K)$ *be the first variables entering along the LARS path. If* $(\widehat{\imath}_1, \dots, \widehat{\imath}_K)$ *satisfies* ($\mathscr{A}_{\text{Irr.}}$) *and* $\widehat{m}$ *is chosen according to a procedure satisfying* ($\mathscr{A}_{\text{Stop}}$)*, then under the null hypothesis*

$$\mathbb{H}_0 : \text{“} X\beta^0 \in H_a \text{”},$$

and conditional on the selection event $\{\widehat{m} \leqslant a\}$*, it follows that*

$$\widehat{\alpha}_{abc} = \widehat{\alpha}_{abc}(\lambda_a, \lambda_b, \lambda_c, \widehat{\imath}_1, \dots, \widehat{\imath}_{c-1}) := 1 - \frac{\mathbb{F}_{abc}(\lambda_b)}{\mathbb{F}_{abc}(\lambda_a)} \sim \mathscr{U}(0,1), \tag{1.3}$$

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

The proof of Theorem 1.2 is presented in Section 3.2.3. The construction of the p-values $\widehat{\alpha}_{abc}$ and of $\mathbb{F}_{abc}$ is given in (3.19) and Section 3.2.2 respectively. We consider the following Generalized Spacing test procedures (GSt):

$$\mathscr{S}_{abc} := \mathbb{1}_{\{\widehat{\alpha}_{abc} \leqslant \alpha\}}, \tag{1.4}$$

that rejects if the p-value $\widehat{\alpha}_{abc}$ is less than the level α of the test. One may remark that

the p-value $\widehat{\alpha}_{abc}$ detects abnormally large values of λ_b relatively to the interval (λ_a, λ_c).

When the noise variance is unknown, we introduce the Generalized t-Spacing tests (GtSt) whose theoretical guarantees are given in the next theorem. The estimator of the variance $\widehat{\sigma}^2$ is given in Section 2.4.

THEOREM 1.3 *Let* $a, b,$ *and* c *be such that* $0 \leqslant a < b < c \leqslant K+1$. *Let* $(\lambda_1, \dots, \lambda_{K+1})$ *be the first knots and let* $(\widehat{\imath}_1, \dots, \widehat{\imath}_K)$ *be the first variables entering along the LARS path. If* $(\widehat{\imath}_1, \dots, \widehat{\imath}_K)$ *satisfies* ($\mathscr{A}_{\text{Irr.}}$) *and* $\widehat{m}$ *is chosen according to a procedure satisfying* ($\mathscr{A}_{\text{Stop}}$) *then under the null hypothesis*

$$\mathbb{H}_0 : \text{‘} X\beta^0 \in H_a \text{’},$$

and conditional on the selection event $\{\widehat{m} \leqslant a\}$*, it follows that*

$$\widehat{\beta}_{abc} = \widehat{\beta}_{abc}(\Lambda_a, \Lambda_b, \Lambda_c, \widehat{\imath}_1, \dots, \widehat{\imath}_K) := 1 - \frac{\widetilde{\mathbb{F}}_{abc}(\Lambda_b)}{\widetilde{\mathbb{F}}_{abc}(\Lambda_a)} \sim \mathscr{U}(0,1),$$

where $\Lambda_k := \lambda_k / \widehat{\sigma}$.

The proof of Theorem 1.3 is presented in Section 3.4.1. The construction of the p-values $\widehat{\beta}_{abc}$, of $\widetilde{\mathbb{F}}_{abc}$ and of the estimation of the noise $\widehat{\sigma}$ is given in (3.24), Section 3.23 and Section 2.4 respectively. One may remark that

the p-value $\widehat{\beta}_{abc}$ detects abnormally large values of Λ_b relatively to the interval (Λ_a, Λ_c).

• **Contribution** (**ii**) : Working with three consecutive knots, we recover the spacing test of Tibshirani et al. [2016] and even in this framework, the present paper improves the current state of knowledge. We prove that:

○ *Under ($\mathscr{A}_{\mathrm{Irr.}}$), the Spacing test procedure defined in [Tibshirani et al., 2016, Theorem 1] is exact, and is equal to the so-called 'conservative' spacing test defined in [Tibshirani et al., 2016, Theorem 2].*

The exact formulation of the spacing test of the pioneering work of Tibshirani et al. [2016] requires extra computations of the term denoted by M^+ in [Tibshirani et al., 2016, Lemma 5]. They proved that the spacing test is asymptotically equivalent to the conservative spacing test. We remove this restriction and we prove that it suffices to check wether the Irrepresentable Check Condition ($\mathscr{A}_{\mathrm{Irr.}}$) holds to get a non-asymptotic equivalence between the spacing test and the conservative spacing test.

• **Contribution** (**iii**) : We theoretically prove that working with non-consecutive knots can allow obtaining higher power for the testing procedure.

THEOREM 1.4 *Assume that the design X is orthogonal, namely $X^\top X = \mathrm{Id}_p$. Let a_0 be an integer such that $0 \leqslant a_0 \leqslant K-1$. If $\widehat{m}$ is chosen according to a procedure satisfying ($\mathscr{A}_{\mathrm{Stop}}$), then under the null hypothesis*

$$\mathbb{H}_0 : \text{`}X\beta^0 \in H_{a_0}\text{'},$$

and conditional on the selection event $\{\widehat{m} = a_0\}$, it follows that the test $\mathscr{S}_{a_0,a_0+1,K+1}$ is uniformly more powerful than any of the tests $\mathscr{S}_{a,b,c}$ for $a_0 \leqslant a < b < c \leqslant K+1$.

The proof of this result is given in Appendix 7.5. This shows that the most powerful test among the set of tests $(\mathscr{S}_{a,b,c})_{a_0 \leqslant a<b<c \leqslant K+1}$ is given by

the GSt test $\mathscr{S}_{a_0,a_0+1,K+1}$ with the smallest a and the largest c.

More precisely, in the proof of Theorem 1.4, it is shown that

$$\widehat{\alpha}_{ab(c+1)} \preccurlyeq \widehat{\alpha}_{abc} \text{ and } \widehat{\alpha}_{a(b-1)c} \preccurlyeq \widehat{\alpha}_{abc} \text{ and } \widehat{\alpha}_{(a-1)bc} \preccurlyeq \widehat{\alpha}_{abc},$$

for orthogonal designs, where $\preccurlyeq$ denotes stochastic ordering.

1.3 *A new exact testing procedure on false negatives (FN) after support selection*

One specific task is to estimate the support of the target sparse vector, namely identify the true positives in the context of a multiple testing procedure. In particular, one may take the support of the LASSO (or SLOPE) solution as an estimate of the support of the solution. This strategy has been intensively studied in the literature, one may consider Bellec et al. [2018], Bogdan et al. [2015], van de Geer [2016], Wainwright [2009] and references therein. Support selection has been studied under the so-called 'Irrepresentable Condition' (IC), as presented for instance in [van de Geer, 2016, Page 53] and [Bühlmann and van de Geer, 2011, Sec. 7.5.1] and also referred to as the 'Mutual Incoherence Condition' [Wainwright, 2009]. Under the so-called 'Beta-Min Condition', one may prove [Bühlmann and van de Geer, 2011, van de Geer, 2016] that the LASSO asymptotically returns the true support. Following this line of thought, one may ask:

[Q5] *Can we provide a false negative testing procedure with a controlled Type I error?*

In this article, we build an exact non-asymptotic multiple test for false non-negatives, see Sections 3.3 and 3.4. The control of the false negatives after model selection in the case of an unknown

```
ALGORITHM 1
Exact false negative testing after model selection
Data: $K$ satisfying (2.9), selection procedure $\widehat{m}$ satisfying ($\mathscr{A}_{\text{Stop}}$), couple $(X,Y)$ giving design and
    response.
Result: $p$-value $\widehat{\alpha}$ on the existence of false negative.
/* $\mathbb{1}_{\{\widehat{\alpha}\leqslant\alpha\}}$ is a testing procedure with level exactly $\alpha$ */
Compute the LARS path from $(X,Y)$.
Check that $(\widehat{\imath}_1,\ldots,\widehat{\imath}_K)$ satisfies ($\mathscr{A}_{\text{Irr.}}$). If not Stop.
Compute $\widehat{m}$, the size of the selected model.
Return $\widehat{\alpha}=\widehat{\alpha}_{\widehat{m}(\widehat{m}+1)(K+1)}$, see (1.3).
 /* When variance is unknown, $\widehat{\alpha}=\widehat{\beta}_{\widehat{m}(\widehat{m}+1)(K+1)}$, see (3.24). */
```

noise level is given in Section 3.4 and the procedure is introduced in Algorithm 1. We assume ($\mathscr{A}_{\text{Stop}}$), which assumes that the model has been selected using an 'admissible' procedure, which basically means that the decision to select a model of size a only depends on the orthogonal projection of the observation Y onto $H_a=\text{Span}(X_{\widehat{\imath}_1},\ldots,X_{\widehat{\imath}_a})$. Assuming further that ($\mathscr{A}_{\text{Irr.}}$) holds, we provide an exact testing method for false negatives. In order to reach high power, the test statistic is the p-value of three non-consecutive knots of the LARS path. To compute this p-value, one needs to marginalize the joint law of the knots, leading to a numerical integration whose complexity grows exponentially with the space between the indices of the knots. We propose to use QMC techniques to compute the statistic, see Appendix 8.

1.4 *False Discovery Rate control for LARS*

Simultaneous controls of confidence intervals independently of the selection procedure have been studied under the concept of *post-selection constants* as introduced in Berk et al. [2013] and studied for instance in Bachoc et al. [2018]. Asymptotic confidence intervals can be build using the *de-sparsified LASSO*, the reader may refer to [van de Geer, 2016, Chapter 5] and references therein. We also point a recent study [Javanmard et al., 2019] of the FDR control as the sample size tends to infinity using *de-biased LASSO*, which has been implemented in Section 4.2. Asymptotic FDR control has been studied in Barber et al. [2015] and references therein, which has been implemented in Section 4.2. Let us point recent control of the *Joint family-wise Error Rate* as in Blanchard et al. [2017] and references therein. Following these lines of work, one may ask:

[Q6] *Can we provide multiple Spacing Tests with a controlled False Discovery Rate (FDR)?*

To the best of our knowledge, this paper is the first to study the joint law and an exact control of multiple spacing tests of LARS knots in a non-asymptotic frame, see Sections 3.2 and 3.6. We investigate the consecutive spacings of the knots of the LARS as test statistics and we prove an exact FDR control using a Benjamini–Hochberg procedure [Benjamini and Hochberg, 1995] in the orthogonal design case, see Theorem 3.8 and Section 3.6. Our proof (see Appendix 7.8) is based on the *Weak Positive Regression Dependency* (WPRDS), the reader may consult Blanchard et al. [2008] or the survey Roquain [2011], and *Knothe-Rosemblatt transport*, see for instance [Santambrogio, 2015, Sec.2.3, Page 67] or [Villani, 2008, Page 20], which is based on conditional quantile transforms.

1.5 *Additional related works on high-dimensional statistics*

Parsimonious models have become ubiquitous tools to tackle high-dimensional representations with a small budget of observations. Successful applications may be found in signal processing (see for instance the pioneering works of Candès et al. [2006], Chen et al. [1998] and references therein) and biology (see for instance Barber et al. [2015] or [Bühlmann and van de Geer, 2011, Chapter 1.4] and references therein). These applications have shown that there are interesting *almost sparse representations* in some well chosen basis. Nowadays, in many practical situations, this sparsity assumption is recognized as reasonable.

These important successes have put a focus on High-Dimensional Statistics and Compressed Sensing in the past decades, which may be due to the deployment of tractable algorithms with strong theoretical guarantees. Among the large panoply of methods, one may consider ℓ_1-regularization, which benefits from a remarkable tractability, empirical performance, and theoretical guarantees. Nowadays, sparse regression techniques based on ℓ_1-regularization are a common and powerful tool in high-dimensional settings. Popular estimators, among which one may point to the LASSO [Tibshirani, 1996] and SLOPE [Bogdan et al., 2015], are known to achieve a minimax rate of prediction and to satisfy the sharp oracle inequalities under conditions on the design, such as Restricted Eigenvalue [Bellec et al., 2018, Bickel et al., 2009] or Compatibility [Bühlmann and van de Geer, 2011, van de Geer, 2016]. The sharp oracle inequalities show that the estimation errors, in ℓ_1 and ℓ_2 norm, of these estimators are optimal, see for instance [van de Geer, 2016, Chapter 2.7].

Variable selection has also been investigated, and it has been proven, see for instance [Bühlmann and van de Geer, 2011, Theorem 7.5], that the LASSO selects the true variables (*i.e.,* there are no false negatives) under the Compatibility condition and the so-called beta-min condition (which assumes that the true parameters are large enough with respect to some threshold that scales linearly with the regularization parameter λ of LASSO). Under a stronger assumption, referred to as the 'irrepresentable condition', one can prove, see for instance [Bühlmann and van de Geer, 2011, Theorem 7.1], that the ℓ_∞-estimation error scales linearly with the regularization parameter λ of LASSO. As the regularization parameter λ tends to zero, when the number of observation goes to infinity and under some assumption on the noise, these results show that LASSO produces a consistent selection of the variables (it asymptotically finds the true support with no errors).

1.6 *Outline of the paper*

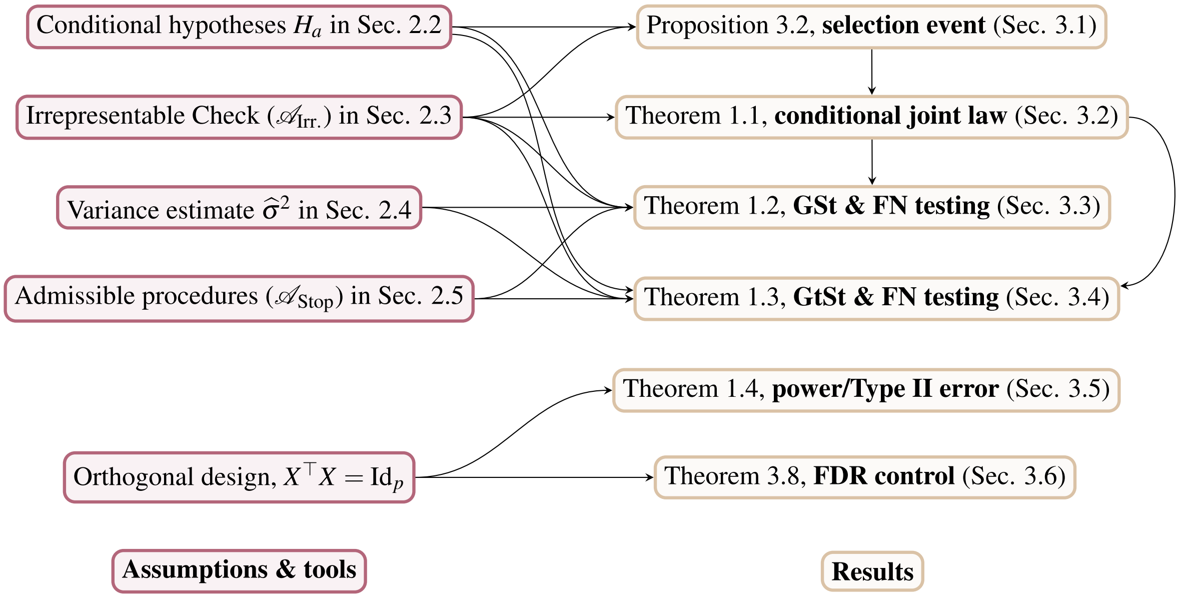
1.6.1 *Detailed outline.* Section 2 introduces the notation (see also Section 1.6.3 for a summary), assumptions ($\mathscr{A}_{\text{Irr.}}$) and ($\mathscr{A}_{\text{Stop}}$), and variance estimate $\widehat{\sigma}^2$. The variance estimate is a key step in our testing procedures: we introduce new variance estimate with properties useful for deriving exact and non-asymptotic post-selection laws, see Section 2.4.

The main assumption is based on the Irrepresentable Check condition ($\mathscr{A}_{\text{Irr.}}$), which can be checked in practice, see Section 2.3. Under ($\mathscr{A}_{\text{Irr.}}$), we obtain a new characterization of the selection event in Proposition 3.2 of Section 3.1.

Section 3 gives the main results: Section 3.2 describes the joint distribution of the LARS knots as a mixture of Gaussian order statistics and the GST and GtST tests. The power in the orthogonal case is considered in Section 3.5. The control of the false negatives in a post selection inference with estimation of the variance is presented in Section 3.4 (when the variance is known, this procedure is studied in Section 3.3). A procedure to control the FDR in the orthogonal case is presented in Section 3.6.

Illustrations of our method, both on simulated data and on real data, are presented in Section 4. A Zenodo repository of the code used in all our experiments can be found at De Castro [2021].

1.6.2 *Dependency diagram.* The outline can be depicted by the following *dependency diagram*:


Conditional hypotheses H_a in Sec. 2.2
Irrepresentable Check ($\mathscr{A}_{\text{Irr.}}$) in Sec. 2.3
Variance estimate $\widehat{\sigma}^2$ in Sec. 2.4
Admissible procedures ($\mathscr{A}_{\text{Stop}}$) in Sec. 2.5
Orthogonal design, $X^\top X = \mathrm{Id}_p$
Assumptions & tools
Proposition 3.2, selection event (Sec. 3.1)
Theorem 1.1, conditional joint law (Sec. 3.2)
Theorem 1.2, GSt & FN testing (Sec. 3.3)
Theorem 1.3, GtSt & FN testing (Sec. 3.4)
Theorem 1.4, power/Type II error (Sec. 3.5)
Theorem 3.8, FDR control (Sec. 3.6)
Results

1.6.3 *Notation and commands.*

General notation

$[a]$	the set of integers $\{1,...,a\}$
$Y = X\beta^0 + \eta$	Linear Model (1.1), X is $n \times p$ design matrix with rank r
σ^2	the variance of the errors η
K	the number of knots $\lambda_1, \dots, \lambda_K$ that are considered, see (2.9)
n_1,n_2	number of d.o.f. used for constructing $\widehat{\sigma}$
φ_{m_k, v_k^2}	standard Gaussian density with mean m_k and variance $v_k^2 := \sigma^2 \rho_k^2$
$\widetilde{\varphi}$	multivariate t-distribution with $\nu = n_2$ degrees of freedom, mean $m = (m_1, \dots, m_K)$ and variance-covariance matrix $\mathrm{Diag}(\rho_1, \dots, \rho_K)$
m_k, v_k^2	conditional mean, see (3.12), and conditional variance $v_k^2 = \sigma^2 \rho_k^2$, see (3.13)
$\widehat{\alpha}_{abc}$	the p-value of the generalized spacing test (GSt), see (1.3)
$\mathscr{S}_{abc}$	$\mathbb{1}_{\{\widehat{\alpha}_{abc} \leqslant \alpha\}}$, the generalized spacing test (GSt) see (1.4)
Λ_k	t-knots defined by (3.22)
$\widehat{\beta}_{abc}$	the p-value of the generalized t-spacing test (GtSt), see (3.24)
$\mathscr{T}_{abc}$	$\mathbb{1}_{\{\widehat{\beta}_{abc} \leqslant \alpha\}}$, the *generalized t-spacing test* (GtSt), see (3.26)

Technical notation

$\widehat{\imath}_k$	a way of coding both indices and signs, see (2.1)
$\bar{\imath}_k; \varepsilon_k$	the indices and the signs of the variables that enter in the LARS path
$j_1, \ldots, j_k; s_1, \ldots, s_k$	a generic value of the sequences above
$i_1, \ldots, i_k$	a generic value of the sequence $\widehat{\imath}_k$
Z	the vector of correlations, obtained by symmetry from $\overline{Z}$ defined by (2.2)
R	the variance-covariance matrix of Z, see (2.4)
$M_{i_1,\ldots,i_\ell}$	sub-matrix of R indexed by $\{i_1, \ldots, i_\ell\}$, see (2.11)
$\overline{S}^k$	$\{\bar{\imath}_1, \ldots, \bar{\imath}_k\}$, a possible selected support (2.6)
S_0	the true support
$\widehat{S}$	the chosen set of variables : $\overline{S}^{\widehat{m}}$, see (2.7)
$\widehat{m}$	the chosen size
$(\mathscr{A}_{\mathrm{Stop}})$	stopping rule, see Section 2.5
H_k	$\mathrm{Span}(X_{\bar{\imath}_1}, \ldots, X_{\bar{\imath}_k})$
$P_k(P_k^\perp)$	Orthogonal projection on (the orthogonal of) H_k
$(\mathscr{A}_{\mathrm{Irr.}})$	Irrepresentable Check, see $(\mathscr{A}_{\mathrm{Irr.}})$
$\theta_j(i_1, \ldots, i_k)$	expectation of Z_j conditional on $Z_{i_1} = \cdots = Z_{i_k} = 1$, see (2.10), and $\theta^\ell := \theta(\widehat{\imath}_1, \ldots, \widehat{\imath}_\ell)$
$Z_j^{(i_1,\ldots,i_k)}$	frozen residual, see (3.1)
$\Pi_{i_1,\ldots,i_k}(Z_j)$	regression of Z_j on $(Z_{i_1}, \ldots, Z_{i_k})$, see (3.2)
$\lambda_k^f := Z_{i_k}^{i_1,\ldots i_{k-1}}$	the k^{th} frozen knot, see (3.5)
$m_k^f, \sigma\rho_k^f$	mean (3.7) and standard deviation (3.8) of λ_k^f
$\mathbb{F}_{abc}(t)$	up to some numerical constant, the CDF of $\lambda_b \mid \lambda_a, \lambda_c$, see (3.19)
$F_i; \mathscr{P}_{ij}$	$F_i := \Phi_i(\lambda_i) := \Phi(\lambda_i/(\sigma\rho_i))$ and $\mathscr{P}_{ij}$ is given by (3.16)
$\widetilde{\mathbb{F}}_{abc}(t)$	up to some numerical constant, the CDF of $\Lambda_b \mid \Lambda_a, \Lambda_c$, see (3.23)
$\boldsymbol{T}_k$	up to some numerical constant, the CDF of centered t-Student distribution, see (3.27)

2. Assumptions, Variance Estimation and Admissible Procedures

2.1 *Signed variables of LARS*

We give some notation that will be useful. We denote by $(\widehat{\imath}_1, \ldots, \widehat{\imath}_k) \in [2p]^k$ the 'signed' variables that enter the model along the LARS path with the convention that

$$\widehat{\imath}_k := \bar{\imath}_k + p\left(\frac{1-\varepsilon_k}{2}\right), \tag{2.1}$$

so that $\widehat{\imath}_k \in [2p]$ is a useful way of encoding both the variable $\bar{\imath}_k \in [p]$ and its sign $\varepsilon_k = \pm 1$ as used in Algorithm 4. We denote by $\overline{Z} := X^\top Y$ the *correlation* vector such that $\overline{Z}_k$ is the scalar product between the k^{th} predictor and the response variable, and we denote by $\sigma^2 \overline{R}$ its variance-covariance matrix. For the sake of presentation, we may consider the $2p$-vector

$$Z := (\overline{Z}, -\overline{Z}) = (X^\top Y, -X^\top Y), \tag{2.2}$$

whose mean is given by

$$\mu^0 := (\overline{R}\beta^0, -\overline{R}\beta^0) = (X^\top X\beta^0, -X^\top X\beta^0) = (\overline{\mu}^0, -\overline{\mu}^0), \tag{2.3}$$

and whose variance-covariance matrix is $\sigma^2 R$ with

$$R = \begin{bmatrix} \overline{R} & -\overline{R} \\ -\overline{R} & \overline{R} \end{bmatrix} = \begin{bmatrix} X^\top X & -X^\top X \\ -X^\top X & X^\top X \end{bmatrix}. \tag{2.4}$$

We also denote by

- $\widehat{\imath}_1, \ldots, \widehat{\imath}_k$, the first k signed variables entering the LARS,
- $i_1, \ldots, i_k$, a generic value of the sequence above,
- $\bar{\imath}_1, \ldots, \bar{\imath}_k$, the first k variables entering the LARS,
- $j_1, \ldots, j_k$, a generic value of the sequence above,
- $\varepsilon_1, \ldots, \varepsilon_k$, the first k signs of the coefficients of the variables entering in the LARS,
- $s_1, \ldots, s_k$, a generic value of the sequence above.

The quantities above are related by (2.1) and

$$i_k := j_k + p\left(\frac{1-s_k}{2}\right). \tag{2.5}$$

2.2 *Models, conditional hypotheses, and the notation K*

We are interested in selecting the true support S^0 of β^0, where the support is defined by

$$S^0 := \{k \in [p] \ : \ \beta_k^0 \neq 0\}.$$

To estimate this support, we will consider the models that appear along the LARS path: the selected model $\widehat{S}$ would be chosen from the family of nested models

$$\underbrace{\{\bar{\imath}_1\}}_{\overline{S}^1} \subset \underbrace{\{\bar{\imath}_1, \bar{\imath}_2\}}_{\overline{S}^2} \subset \cdots \subset \underbrace{\{\bar{\imath}_1, \bar{\imath}_2, \ldots, \bar{\imath}_a\}}_{\overline{S}^a} \subset \cdots \subset \underbrace{\{\bar{\imath}_1, \bar{\imath}_2, \ldots, \bar{\imath}_K\}}_{\overline{S}^K}, \tag{2.6}$$

where K denotes the maximal model size. We denote by $\widehat{m}$ the size of the selected model $\widehat{S}$, and then

$$\widehat{S} = \overline{S}^{\widehat{m}}. \tag{2.7}$$

Respectively, denote

$$\{0\} =: H_0 \subset H_1 \subset \cdots \subset H_a := \mathrm{Span}(X_{\bar{\imath}_1}, \ldots, X_{\bar{\imath}_a}) \subset \cdots \subset H_K, \tag{2.8}$$

the corresponding family of nested subspaces of $\mathbb{R}^n$. Once the model has been selected, we will construct tests based on the $K+1$ first knots of the LARS.

REMARK 2.1 *The testing procedures under consideration are not standard since the $(H_a)_{a=1}^K$ are random subspaces. We are interested in the framework of selective testing, namely, testing procedures conditional on the selection event $\{\widehat{m} = a, \widehat{\imath}_1 = i_1, \ldots, \widehat{\imath}_K = i_K\}$, for some fixed $a \in [K-1]$. Conditional on the event, note that H_a is fixed. By convention, we may consider the case $a = 0$, that is, testing the global null hypothesis.*

Throughout this paper, we assume that

$$K \text{ is fixed and such that } 1 \leqslant K < \min(n,r) \text{ where } r = \operatorname{rank}(X). \tag{2.9}$$

In practice, K can be considerably much smaller than n. Our analysis is conditional on $(\widehat{\imath}_1,\ldots,\widehat{\imath}_K)$ and in this spirit it can be referred to as a 'Post-Section' procedure, see *e.g.* Taylor and Tibshirani [2015], Tibshirani et al. [2015; 2016].

2.3 *Irrepresentable Check on the Active sets*

We define the set of *Active Sets* $\mathscr{A}_K$ as all the sequences $i_1,\ldots,i_K$ of signed variables such that $j_1,\ldots,j_K$ are pairwise different, where the j's are defined by (2.5), namely

$$\mathscr{A}_K := \big\{(i_1,\ldots,i_K) \in [2p]^K \ :\ j_1,\ldots,j_K \text{ are pairwise different}\big\}.$$

Sometimes it would be useful to consider $\mathscr{A}_{K+1}$, the set of active sets of size $K+1$. We introduce the notion of 'Irrepresentable Check', which is the only assumption on the design and the selected active set in most of our results.

DEFINITION 2.1 (Irrepresentable Check) An active set $(i_1,\ldots,i_K) \in \mathscr{A}_K$ is said to satisfy the *Irrepresentable Check* condition if

$$\forall k \in [K],\ \forall j \notin T^k := \{j_1,\ldots,j_k\},\quad X_j^\top X_{T^k}\big(X_{T^k}^\top X_{T^k}\big)^{-1} s_k < 1, \tag{$\mathscr{A}_{\text{Irr.}}$}$$

where j_k and s_k are defined from i_k using (2.5). By a slight abuse of notation we will denote by ($\mathscr{A}_{\text{Irr.}}$) the set of sequences $(i_1,\ldots,i_K)$ that satisfy this property.

In our procedures and theoretical results, we will limit our attention to sequences $\widehat{\imath}_1,\ldots,\widehat{\imath}_K$ chosen by LARS that satisfy ($\mathscr{A}_{\text{Irr.}}$). A particular case is when the property is true for all possible active sets. This is equivalent to the Irrepresentable Condition that we will now recall.

DEFINITION 2.2 (Irrepresentable Condition of order K) The design matrix X satisfies the Irrepresentable Condition of order K if and only if

$$\forall S \subset [p] \text{ s.t. } \#S \leqslant K,\quad \max_{j\in[p]\setminus S}\ \max_{\|v\|_\infty\leqslant 1} X_j^\top X_S\big(X_S^\top X_S\big)^{-1} v < 1, \tag{Irrep.}$$

where X_j denotes the j^{th} column of X and X_S the sub-matrix of X obtained by keeping the columns indexed by S.

REMARK 2.2 *Note that the Irrepresentable Condition is a standard condition, as presented for instance, in [van de Geer, 2016, Page 53] and [Bühlmann and van de Geer, 2011, Sec. 7.5.1], and is also referred to as the Mutual Incoherence Condition [Wainwright, 2009].*

REMARK 2.3 *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 Wainwright [2009], 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* (Irrep.) *(for instance the identity matrix) satisfies* (Irrep.) *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 (Irrep.) is a strong requirement on the design X and, in addition, this condition cannot be checked in polynomial time. One important feature of our results is that we do not require the Irrepresentable Condition (Irrep.) but only the weaker requirement of Irrepresentable Check ($\mathscr{A}_{\mathrm{Irr.}}$) on the selected active set. Namely, we will assume that

$$\text{For } K \text{ defined by (2.9), } (\widehat{\imath}_1,\ldots,\widehat{\imath}_K) \text{ satifies } (\mathscr{A}_{\mathrm{Irr.}}). \qquad \text{(Assumption)}$$

Given $(\widehat{\imath}_1,\ldots,\widehat{\imath}_K)$, note that this condition can be checked in polynomial time.

EXAMPLE 2.3 Taking the (signed) variables entering the model with LARS in an iid Gaussian design and as response variable a centered Gaussian vector with iid entries from $10,000$ Monte Carlo repetitions, Figure 1 illustrates the law of the maximal order $K_{\max}$ for which the Irrepresentable Check condition holds. For example, we found that for $p = 1,000$ and $n = 100$ (with ratio $n/p = 0.1$) resp. $n = 500$ (with ratio $n/p = 0.5$), the Irrepresentable Check condition ($\mathscr{A}_{\mathrm{Irr.}}$) of order $K_{\max}$ holds when $K_{\max}$ is about $K_{\max} \simeq 0.16 \times n = 16$ respectively $K_{\max} \simeq 0.12 \times n = 60$, see Figure 1.

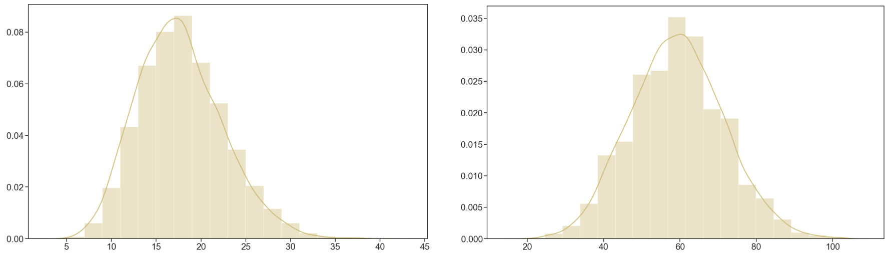


Figure 1: The law of the maximal order K for which Irrepresentable Check holds, taking the (signed) variables entering the model with LARS from an iid Gaussian design and with response variable a centered Gaussian vector with iid entries using $10,000$ Monte Carlo repetitions. There are $p = 1,000$ predictors and $n = 100$ (left) $n = 500$ (right) observations, and we observe that $K_{\max} \in [10,27]$ (left) and $K_{\max} \in [39,81]$ (right) for 95% of the values.

2.3.1 *Irrepresentable Check: An equivalent formulation.*

Now, we can define

$$\forall (i_1,\ldots,i_k) \in [2p]^k, \quad \theta_j(i_1,\ldots,i_k) := \big(R_{j,i_1} \cdots R_{j,i_k}\big) M^{-1}_{i_1,\ldots,i_k}(1,\ldots,1), \tag{2.10}$$

where $(1,\ldots,1)$ is the column vector of size k whose entries are equal to one; $\sigma^2 M_{i_1,\ldots,i_k}$ is the variance-covariance matrix of the vector $(Z_{i_1},\cdots,Z_{i_k})$ 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 $\{i_1,\ldots,i_k\}$, namely

$$M_{i_1,\ldots,i_k} := (R_{i,j})_{i,j=i_1,\ldots,i_k}. \tag{2.11}$$

Remark that

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

when $\mathbb{E}Z = 0$. Then Proposition 2.4 shows that the Irrepresentable Condition (Irrep.) of order K is equivalently given by

$$\forall k \leqslant K,\ \forall (i_1,\ldots,i_k) \in [2p]^k,\ \forall j \notin \{i_1,\ldots,i_k\}, \quad \theta_j(i_1,\ldots,i_k) < 1, \tag{2.12}$$

where $\theta_j(i_1,\ldots,i_k)$ is given by (2.10).

PROPOSITION 2.4 Let X and R be defined by (2.4). Then, the following assumptions are equivalent:

- the design matrix X satisfies (Irrep.) of order K,
- the variance-covariance matrix R satisfies (2.12) of order K.

Furthermore, they imply that for all $(i_1,\ldots,i_K)\in\mathscr{A}_K$ one has

$$\max\Big[\max_{j\neq i_1}\theta_j(\imath_1),\,\ldots,\,\max_{j\neq i_1,\ldots,i_K}\theta_j(i_1,\ldots,i_K)\Big]<1$$

which is an equivalent formulation of $(i_1,\ldots,i_K)$ satisfying $(\mathscr{A}_{\mathrm{Irr.}})$.

Proof. Let $S=\{j_1,\ldots,j_k\}\subset[p]$ and $j\in[2p]\setminus S$. Let $\overline{v}=(\overline{v}_1,\ldots,\overline{v}_k)\in\{-1,1\}^k$ and define $i_\ell=j_\ell+p(1-\overline{v}_\ell)/2$ for $\ell\in[k]$. Note that

$$\begin{aligned}\theta_j(i_1,\ldots,i_k)&=\big(R_{j,i_1}\cdots R_{j,i_k}\big)M^{-1}_{i_1,\ldots,i_k}(1,\ldots,1)\\&=\Big[X_j^\top X_S\mathrm{Diag}(\overline{v})\Big]M^{-1}_{i_1,\ldots,i_k}(1,\ldots,1)\\&=\Big[X_j^\top X_S\mathrm{Diag}(\overline{v})\Big]M^{-1}_{i_1,\ldots,i_k}\Big[\mathrm{Diag}(\overline{v})\overline{v}\Big]\\&=X_j^\top X_S\Big[\mathrm{Diag}(\overline{v})M^{-1}_{i_1,\ldots,i_k}\mathrm{Diag}(\overline{v})\Big]\overline{v}\\&=X_j^\top X_S\big(X_S^\top X_S\big)^{-1}\overline{v}.\end{aligned}$$

Now, observe that

$$\max_{\|v\|_\infty\leqslant 1}v^\top\big(X_S^\top X_S\big)^{-1}X_S^\top X_j=\max_{\overline{v}\in\{-1,1\}^k}X_j^\top X_S\big(X_S^\top X_S\big)^{-1}\overline{v},$$

showing the equivalence between the two assumptions. □

REMARK 2.4 *One may require that the design be 'normalized' so that $R_{i,i}=1$, namely its columns have unit Euclidean norm. Under this normalization, one can check that R satisfies* (Irrep.) *of order $K=1$. Hence, up to some normalization, one can always assume* (Irrep.) *of order $K=1$.*

REMARK 2.5 *When computing the LARS path, one has to compute the values*

$$X_j^\top X_{\overline{S}^k}\big(X_{\overline{S}^k}^\top X_{\overline{S}^k}\big)^{-1}\varepsilon^k,$$

see for instance Algorithm 2 or Algorithm 3, where these values are given by θ, as shown by Proposition 2.4. This implies that, in practice, along the LARS path, one witnesses the maximal order K for which Irrepresentable Check $(\mathscr{A}_{\mathrm{Irr.}})$ holds.

2.4 *The estimator of the variance*

In our analysis, we introduce an estimate of the variance $\widehat{\sigma}^2$ to perform post-selection inference when the noise level σ is unknown. The degree of freedom to the estimation of the variance is $n-K$. Let us fix, for the moment, $j_1,\ldots,j_K$, the indices that are the putative indices for the selected variables. Let $P_K^\perp$ be the orthogonal projection on the orthogonal to $H_K^{j_1,\ldots,j_K}:=\mathrm{Span}(X_{j_1},\ldots,X_{j_K})$. We define

$$\widehat{\sigma}^{j_1,\ldots,j_K}:=\frac{\|P_K^\perp Y\|_2}{\sqrt{n-K}}.\tag{2.13}$$

By a slight abuse of notation, we can index the estimator above by the signed indexes $i_1,\ldots,i_K$. Eventually, we set

$$\widehat{\sigma} := \widehat{\sigma}^{\bar{\imath}_1,\ldots,\bar{\imath}_K},$$

the estimates of the standard deviation σ.

2.5 *Admissible Selection Procedures*

Note that choosing a model $\widehat{S}$ is equivalent to choosing a model size $\widehat{m}$ so that

$$\widehat{S} = \big\{\bar{\imath}_1,\bar{\imath}_2,\ldots,\bar{\imath}_{\widehat{m}}\big\}. \tag{2.14}$$

Our procedure is flexible on this point and allows any choice of $\widehat{m}$ as long as the following property $(\mathscr{A}_{\text{Stop}})$ is satisfied:

> **Stopping Rule:** *The estimated model size $\widehat{m}$ is a 'stopping time': $\widehat{m} \in [K-1]$ and, for all $a \in [K-1]$,*
>
> $$\mathbb{1}_{\{\widehat{m}\leqslant a\}} \text{ is a measurable function of } \big(\lambda_1,\ldots,\lambda_a,\widehat{\imath}_1,\ldots,\widehat{\imath}_K\big)\,. \tag{$\mathscr{A}_{\text{Stop}}$}$$

In other words, the decision to select a model of size $\{\widehat{m}=a\}$ depends only on the first a variables entering the LARS.

REMARK 2.6 *We now give an example to show that $(\mathscr{A}_{\text{Stop}})$ implies some restriction. Suppose, for example, that we want to decide wether the target β^0 is two sparse or one sparse. A natural decision rule is to look at large values of the second knot λ_2, if "$\lambda_2 >$ (some threshold)" choose $m=2$ otherwise choose $m=1$. This rule does not satisfy $(\mathscr{A}_{\text{Stop}})$, since looking at λ_2 we can choose only sizes m greater than or equal to 2.*

Denote by $P_k(Y)$ (resp. $P_k^{\perp}(Y)$) the orthogonal projection of the observation Y onto H_k (resp. the orthogonal of H_k) for all $k \geqslant 1$ where H_k are defined by (2.8). Given h any measurable function,

$$\mathbb{1}_{\{\widehat{m}\leqslant a\}} = h(P_a(Y)),$$

determines a class of selection procedures satisfying $(\mathscr{A}_{\text{Stop}})$. These procedures decide whether to stop at $\{\widehat{m}=a\}$ based on the information given by $P_a(Y)$. Once one has selected a model of size $\widehat{m}$, one may be willing to test if $\widehat{S}$ contains the true support S^0 by considering the null hypothesis

$$\mathbb{H}_0\,:\,\text{‘}S^0 \subseteq \widehat{S}\text{’},$$

namely there are no false negatives. Equivalently, one aims at testing the null hypothesis

$$\mathbb{H}_0\,:\,\text{‘}X\beta^0 \in H_{\widehat{m}}\text{’}, \tag{2.15}$$

at an exact significance level $\alpha \in (0,1)$, where $(H_a)_{a=0}^{K-1}$ is defined by (2.8).

3. Exact Controls using Least Angle Regression: Main Results

3.1 *Key notion: the 'frozen' knots, their means and variances*

3.1.1 *Frozen knots.*

Given K as defined in (2.9) and fixed $i_1,\ldots,i_{K+1} \in [2p]$, one may define

$$\forall j \text{ s.t. } \theta_j(i_1,\ldots,i_k) \neq 1, \quad Z_j^{(i_1,\ldots,i_k)} := \frac{Z_j - \Pi_{i_1,\ldots,i_k}(Z_j)}{1-\theta_j(i_1,\ldots,i_k)}, \tag{3.1}$$

where

$$\Pi_{i_1,\dots,i_k}(Z_j) := \big(R_{j,i_1} \cdots R_{j,i_k}\big) M^{-1}_{i_1,\dots,i_k}(Z_{i_1},\dots,Z_{i_k}) \tag{3.2}$$

and $\theta_j(i_1,\dots,i_k)$ is given by (2.10). When $\mathbb{E}Z = 0$, one may remark that $\Pi_{i_1,\dots,i_k}(Z_j)$ is the regression of Z_j on the vector $(Z_{i_1},\cdots,Z_{i_k})$ whose variance-covariance matrix is $\sigma^2 M_{i_1,\dots,i_k}$, namely

$$\text{When } \mathbb{E}Z = 0, \quad \Pi_{i_1,\dots,i_k}(Z_j) = \big(R_{j,i_1} \cdots R_{j,i_k}\big) M^{-1}_{i_1,\dots,i_k}(Z_{i_1},\dots,Z_{i_k}) = \mathbb{E}\big[Z_j | Z_{i_1},\cdots,Z_{i_k}\big].$$

From this point on, we introduce

$$\forall k \geqslant 0, \quad \lambda_{k+1}^{(i_1,\dots,i_k)} := \max_{j:\theta_j(i_1,\dots,i_k)<1} Z_j^{(i_1,\dots,i_k)}, \tag{3.3}$$

and we emphasize that

$$\forall k \geqslant 0, \quad \lambda_{k+1}\,\mathbb{1}_{\{\widehat{\imath}_1=i_1,\dots,\widehat{\imath}_k=i_k\}} = \lambda_{k+1}^{(i_1,\dots,i_k)}\,\mathbb{1}_{\{\widehat{\imath}_1=i_1,\dots,\widehat{\imath}_k=i_k\}}, \tag{3.4}$$

as proven in Appendix 5.4 (Eq. (5.3)) and Proposition 3.2. We are now able to define the "*frozen*" values of the knots:

$$\lambda_1^f := Z_{i_1}, \dots, \lambda_{K+1}^f := Z_{i_{K+1}}^{i_1,\dots i_K}. \tag{3.5}$$

They are the Gaussian random variables that coincide with $\lambda_1,\lambda_2,\cdots,\lambda_{K+1}$ when the random variables defined by the signed indices $\widehat{\imath}_1,\widehat{\imath}_2,\dots,\widehat{\imath}_{K+1}$ take the particular values $i_1,i_2,\dots,i_{K+1}$.

REMARK 3.1 *An interesting feature of the LARS knots is that they have a simple expression in terms of the partition given by the identity*

$$\sum_{(i_1,\dots,i_K)\in\mathscr{A}_K} \mathbb{1}_{\{\widehat{\imath}_1=i_1,\dots,\widehat{\imath}_K=i_K\}} = 1 \quad \textit{almost surely.}$$

As we have seen in (3.5)*,*

$$\forall k \in [K], \quad \lambda_k = \sum_{(i_1,\dots,i_k)\in\mathscr{A}_k} \mathbb{1}_{\{\widehat{\imath}_1=i_1,\dots,\widehat{\imath}_k=i_k\}} \underbrace{Z_{i_k}^{i_1,\dots i_{k-1}}}_{=:\lambda_k^f},$$

giving the definition of the frozen knots λ_k^f above.

3.1.2 *Mean and centering of the frozen knots.* Now, write

$$\forall y \in \mathbb{R}^n, \quad P^{(i_1,\dots,i_k)}(y) = \big(X_{j_1} \cdots X_{j_k}\big) M^{-1}_{j_1,\dots,j_k}(X_{j_1}^\top,\dots,X_{j_k}^\top)\, y \tag{3.6}$$

for the orthogonal projection of y onto $\mathrm{Span}(X_{j_1},\dots,X_{j_k})$. Recall that P_k is the orthogonal projection onto H_k, which is a random subspace. Recall also that, conditional on the event $\{\widehat{\imath}_1 = i_1,\dots,\widehat{\imath}_k = i_k\}$ the subspace H_k is fixed. Note that

$$P_k\,\mathbb{1}_{\{\widehat{\imath}_1=i_1,\dots,\widehat{\imath}_k=i_k\}} = P^{(i_1,\dots,i_k)}\,\mathbb{1}_{\{\widehat{\imath}_1=i_1,\dots,\widehat{\imath}_k=i_k\}},$$

$$P_k^\perp\,\mathbb{1}_{\{\widehat{\imath}_1=i_1,\dots,\widehat{\imath}_k=i_k\}} = \big(\mathrm{Id}_n - P^{(i_1,\dots,i_k)}\big)\,\mathbb{1}_{\{\widehat{\imath}_1=i_1,\dots,\widehat{\imath}_k=i_k\}},$$

$$\text{and for all } i \in [2p], \quad \Pi_{i_1,\dots,i_k}(Z_i) = s\langle X_j, P^{(i_1,\dots,i_k)}(Y)\rangle,$$

where $i = j + p(1-s)/2$. The mean m_k^f and standard deviation $\sigma\rho_k^f$ of λ_k^f are important values defined for all $k \in [K]$:

$$m_k \mathbb{1}_{\{\widehat{\imath}_1=i_1,\ldots,\widehat{\imath}_k=i_k\}} = m_k^f \mathbb{1}_{\{\widehat{\imath}_1=i_1,\ldots,\widehat{\imath}_k=i_k\}},$$
$$\rho_k \mathbb{1}_{\{\widehat{\imath}_1=i_1,\ldots,\widehat{\imath}_k=i_k\}} = \rho_k^f \mathbb{1}_{\{\widehat{\imath}_1=i_1,\ldots,\widehat{\imath}_k=i_k\}},$$

with

$$m_k^f = \frac{s_k\langle X_{j_k}, \big(\mathrm{Id}_n - P^{(i_1,\ldots,i_{k-1})}\big)X\beta^0\rangle}{1-\theta_{i_k}(i_1,\ldots,i_{k-1})}, \tag{3.7}$$

$$\rho_k^f = \frac{\sqrt{\langle X_{j_k}, \big(\mathrm{Id}_n - P^{(i_1,\ldots,i_{k-1})}\big)X_{j_k}\rangle}}{1-\theta_{i_k}(i_1,\ldots,i_{k-1})}, \tag{3.8}$$

and this definition is equivalent to (3.12) and (3.13), see Section 3.2. Recall that $\overline{\mu}^0 = \overline{R}\beta^0$ as defined in (2.3) and note that

$$m_k^f = 0 \quad\Leftrightarrow\quad \overline{\mu}^0_{i_k} - \Pi_{i_1,\ldots,i_{k-1}}(\overline{\mu}^0_{i_k}) = 0 \quad\Leftrightarrow\quad \langle X_{j_k}, \big(\mathrm{Id}_n - P^{(i_1,\ldots,i_{k-1})}\big)X\beta^0\rangle = 0, \tag{3.9}$$

which is true when the true support S^0 of β^0 is included in $\overline{S}^k$, defined by (2.6). This proves the next proposition.

PROPOSITION 3.1 For fixed $0 \leqslant a \leqslant K-1$, conditional on the selection event $\{\widehat{\imath}_1 = i_1,\ldots,\widehat{\imath}_K = i_K\}$, the hypothesis

$$\mathbb{H}_0 : \text{“}X\beta^0 \in H_a\text{”}$$

implies that $m_k^f = 0$ for all $a < k \leqslant K$, namely $(Z_{i_{a+1}}^{(i_1,\ldots,i_a)},\ldots,Z_{i_K}^{(i_1,\ldots,i_{K-1})})$ is centered.

This proposition is important for defining the hypothesis under consideration, see also Remark 2.1.

3.1.3 *A key result: The characterization of the selection event.*

Regarding the joint law of the frozen knots, one has the following important proposition whose proof can be found in Section 7.1.

PROPOSITION 3.2 Let $(i_1,\ldots,i_K,i_{K+1}) \in \mathscr{A}_{K+1}$, that is, a fixed active set of size $K+1$.

- If $(i_1,\ldots,i_K)$ satisfies $(\mathscr{A}_{\mathrm{Irr.}})$ then

$$\begin{aligned}
&\big\{\widehat{\imath}_1 = i_1,\ldots,\widehat{\imath}_{k+1} = i_{k+1}\big\}\\
&= \big\{\lambda_{k+1}^{(i_1,\ldots,i_k)} = Z_{i_{k+1}}^{(i_1,\cdots,i_k)} \leqslant Z_{i_k}^{(i_1,\cdots,i_{k-1})} \leqslant \cdots \leqslant Z_{i_2}^{(i_1)} \leqslant Z_{i_1}\big\}\\
&= \big\{\lambda_{k+1}^{(i_1,\ldots,i_k)} = Z_{i_{k+1}}^{(i_1,\cdots,i_k)} \leqslant Z_{i_k}^{(i_1,\cdots,i_{k-1})} \leqslant \cdots \leqslant Z_{i_{a+1}}^{(i_1,\cdots,i_a)} \leqslant Z_{i_a}^{(i_1,\cdots,i_{a-1})} = \lambda_a^{(i_1,\ldots,i_{a-1})}\big\}\\
&\quad \bigcap \big\{\lambda_{k+1} = Z_{i_{k+1}}^{(i_1,\cdots,i_k)},\ldots,\lambda_a = Z_{i_a}^{(i_1,\cdots,i_{a-1})}\big\} \bigcap \big\{\widehat{\imath}_1 = i_1,\ldots,\widehat{\imath}_a = i_a\big\},
\end{aligned}$$

  for any $0 \leqslant a < k \leqslant K$ with the convention $\lambda_0 = \infty$.

- It holds that

$$(Z_j^{(i_1,\ldots,i_k)})_{j\neq i_1,\ldots,i_k} \perp\!\!\!\perp Z_{i_k}^{(i_1,\ldots,i_{k-1})} \perp\!\!\!\perp Z_{i_{k-1}}^{(i_1,\ldots,i_{k-2})} \perp\!\!\!\perp \cdots \perp\!\!\!\perp Z_{i_2}^{(i_1)} \perp\!\!\!\perp Z_{i_1}$$

  are mutually independent, for any $k \in [K]$.

Furthermore, if $X\beta^0 \in H_K$ then

$$\widehat{\sigma}^{i_1,\ldots,i_K} \perp\!\!\!\perp \Big(\frac{Z_j^{(i_1,\ldots,i_K)}}{\widehat{\sigma}^{i_1,\ldots,i_K}}\Big)_{j\neq i_1,\ldots,i_K} \perp\!\!\!\perp Z_{i_K}^{(i_1,\ldots,i_{K-1})} \perp\!\!\!\perp \cdots \perp\!\!\!\perp Z_{i_2}^{(i_1)} \perp\!\!\!\perp Z_{i_1}. \tag{3.10}$$

- If $(i_1,\ldots,i_K)$ satisfies $(\mathscr{A}_{\mathrm{Irr.}})$ then

$$\begin{aligned}&\big\{\widehat{\imath}_1 = i_1,\ldots,\widehat{\imath}_K = i_K\big\}\\ &= \Big\{\lambda_{K+1}^{(i_1,\ldots,i_K)} \leqslant Z_{i_K}^{(i_1,\cdots,i_{K-1})} \leqslant \cdots \leqslant Z_{i_2}^{(i_1)} \leqslant Z_{i_1}\Big\}\\ &= \left\{\Lambda_{K+1}^{(i_1,\ldots,i_K)} := \frac{\lambda_{K+1}^{(i_1,\ldots,i_K)}}{\widehat{\sigma}^{i_1,\ldots,i_K}} \leqslant \frac{Z_{i_K}^{(i_1,\ldots,i_{K-1})}}{\widehat{\sigma}^{i_1,\ldots,i_K}} \leqslant \cdots \leqslant \frac{Z_{i_2}^{(i_1)}}{\widehat{\sigma}^{i_1,\ldots,i_K}} \leqslant \frac{Z_{i_1}}{\widehat{\sigma}^{i_1,\ldots,i_K}}\right\}.\end{aligned}$$

REMARK 3.2 **Is Proposition 3.2 a new polyhedral lemma?** *The characterization of the selection event for the inference of a single testing statistic has been known as the 'polyhedral lemma' in the literature, see for instance [Tibshirani et al., 2015, Figure 6.9] and references therein. This result is the cornerstone of selective inference with sparse models. It is based on two ingredients: First, the selection event can be expressed as a polyhedra; Second, conditional on the selection event, any linear statistics is distributed according to a truncated Gaussian with independent truncation bounds.*

A first remark is that the polyhedral lemma is shown for one linear statistic and, as far as we known, there is no polyhedral lemma for multiple linear statistics. The interesting point is that our result (Proposition 3.2) can be seen as a polyhedral lemma for multiple linear statistics. Under $(\mathscr{A}_{\mathrm{Irr.}})$, the selection event $\{\widehat{\imath}_1,\ldots,\widehat{\imath}_K\}$ corresponds to a polyhedra described by the $Z_{i_k}^{(i_1,\cdots,i_{k-1})}$ variables in the third point of Proposition 3.2. Our main result shows that the joint law of these multiple linear statistics are the Gaussian distribution restricted to the polyhedra $\{\ell_1 \geqslant \ldots \geqslant \ell_K \geqslant \lambda_{K+1}\}$, see Theorem 1.1.

Note that the selection event has to include λ_{K+1}. As discussed above, our polyhedral lemma (Theorem 1.1 and Proposition 3.2) shows that, conditional on the selection event, $\lambda_1,\ldots,\lambda_K$ are distributed with respect to a Gaussian distribution restricted to the polyhedra $\{\ell_1 \geqslant \ldots \geqslant \ell_K \geqslant \lambda_{K+1}\}$. If one does not include λ_{K+1} in the selection event, then one has to integrate this latter conditional law with respect to the distribution of λ_{K+1} which is not known.

PROPOSITION 3.3 Assume that the design X is such that the Irrepresentable Condition (Irrep.) of order K holds. Almost surely, one has

- Among all possible sets $(i_1,\ldots,i_K) \in \mathscr{A}_K$, there is one and only one such that

$$\max_{i_{K+1}\neq i_1,\ldots,i_K} Z_{i_{K+1}}^{(i_1,\cdots,i_K)} \leqslant Z_{i_K}^{(i_1,\cdots,i_{K-1})} \leqslant \cdots \leqslant Z_{i_2}^{(i_1)} \leqslant Z_{i_1}. \tag{3.11}$$

- This set is the set selected by LARS, namely $\widehat{\imath}_1 = i_1,\ldots,\widehat{\imath}_K = i_K$,
- and, for all $(i_1,\ldots,i_K) \in \mathscr{A}_K$,

$$\mathbb{P}\big(\widehat{\imath}_1 = i_1,\ldots,\widehat{\imath}_K = i_K\big) = \mathbb{P}\Big(\max_{i_{K+1}\neq i_1,\ldots,i_K} Z_{i_{K+1}}^{(i_1,\cdots,i_K)} \leqslant Z_{i_K}^{(i_1,\cdots,i_{K-1})} \leqslant \cdots \leqslant Z_{i_2}^{(i_1)} \leqslant Z_{i_1}\Big).$$

Proof. Note that (Irrep.) implies $(\mathscr{A}_{\mathrm{Irr.}})$ by Proposition 2.4. Then apply the first point of Proposition 3.2 to conclude. □

Finding the set $\{\widehat{\imath}_1 = i_1, \ldots, \widehat{\imath}_K = i_K\}$ selected by LARS may be related to a combinatorial search testing (3.11) all possible candidates $(i_1, \ldots, i_K) \in \mathscr{A}_K$. Under the Irrepresentable Condition, the support selected by LARS is given by (3.11), which can be seen as the extension of (1.2) introducing **[Q1]** in Section 1.1.

3.2 *Main results: Joint law and construction of post-selection tests*

We assume that K is defined as in (2.9). Except in Section 3.4, σ^2 is assumed to be known. Let $(\widehat{\imath}_1, \ldots, \widehat{\imath}_K)$ be the first signed variables entering along the LARS path. In this section, we are interested in the joint law of the LARS knots $(\lambda_1, \ldots, \lambda_K)$ conditional on λ_{K+1} and $(\widehat{\imath}_1, \ldots, \widehat{\imath}_K)$. To determine this joint law, we need to make precise the centering parameters m_k, by (see also (3.7))

$$m_k := \frac{\mu^0_{\widehat{\imath}_k} - \big(R_{\widehat{\imath}_k,\widehat{\imath}_1} \cdots R_{\widehat{\imath}_k,\widehat{\imath}_{k-1}}\big) M^{-1}_{\widehat{\imath}_1,\ldots,\widehat{\imath}_{k-1}} \big(\mu^0_{\widehat{\imath}_1}, \cdots, \mu^0_{\widehat{\imath}_{k-1}}\big)}{1 - \theta^{k-1}_{\widehat{\imath}_k}} \tag{3.12}$$

the first standard deviation $\sigma\rho_1$ with $\rho_1 := \sqrt{R_{\widehat{\imath}_1,\widehat{\imath}_1}}$, and the others $\sigma\rho_k$ by (see also (3.8))

$$\rho_\ell := \frac{\sqrt{R_{\widehat{\imath}_\ell,\widehat{\imath}_\ell} - \big(R_{\widehat{\imath}_\ell,\widehat{\imath}_1} \cdots R_{\widehat{\imath}_\ell,\widehat{\imath}_{\ell-1}}\big) M^{-1}_{\widehat{\imath}_1,\ldots,\widehat{\imath}_{\ell-1}} \big(R_{\widehat{\imath}_\ell,\widehat{\imath}_1}, \cdots, R_{\widehat{\imath}_\ell,\widehat{\imath}_{\ell-1}}\big)}}{1 - \theta^{\ell-1}_{\widehat{\imath}_\ell}} \quad \text{for } 2 \leqslant \ell \leqslant K+1, \tag{3.13}$$

where

$$\theta^{\ell-1} := \theta(\widehat{\imath}_1, \ldots, \widehat{\imath}_{\ell-1}), \quad \text{for } 2 \leqslant \ell \leqslant K+1,$$

is defined by (2.10) and $M_{\widehat{\imath}_1,\ldots,\widehat{\imath}_{\ell-1}}$ is defined by (2.11).

3.2.1 *Proof of Theorem 1.1.*

From the definition of the Gaussian random variable $Z^{(i_1,\ldots,i_{k-1})}_{i_k}$ in (3.1) one can deduce that its mean m_k is given by (3.12) and its standard deviation v_k by (3.13), considering putative indices for the selected variables. By the second point of Proposition 3.2, we know that these variables are independent. We deduce that their joint density $(Z_{i_1}, Z^{(i_1)}_{i_2}, \ldots, Z^{(i_1,\ldots,i_{K-1})}_{i_K})$ is

$$\prod_{k=1}^{K} \varphi_{m_k, v_k^2}(\ell_k),$$

with respect to Lebesgue measure. For now on, we condition on $\mathscr{E} := \{\widehat{\imath}_1 = i_1, \ldots, \widehat{\imath}_K = i_K, \lambda_{K+1}\}$ and we assume that $(i_1, \ldots, i_K)$ satisfies $(\mathscr{A}_{\mathrm{Irr.}})$. By the first equality of the third point of Proposition 3.2 we known that $\mathscr{E} = \big\{\lambda_{K+1} \leqslant Z^{(i_1,\ldots,i_{K-1})}_{i_K} \leqslant \cdots \leqslant Z_{i_1}\big\}$, and on the event $\mathscr{E}$,

$$(Z_{i_1}, Z^{(i_1)}_{i_2}, \ldots, Z^{(i_1,\ldots,i_{K-1})}_{i_K}) = (\lambda_1, \lambda_2, \ldots, \lambda_K). \tag{3.14}$$

Conditional on $\mathscr{E}$, the joint density of $(Z_{i_1}, Z^{(i_1)}_{i_2}, \ldots, Z^{(i_1,\ldots,i_{K-1})}_{i_K})$ is proportional to

$$\Big(\prod_{k=1}^{K} \varphi_{m_k, v_k^2}(\ell_k)\Big) \mathbb{1}_{\{\ell_1 \geqslant \ell_2 \geqslant \cdots \geqslant \ell_K \geqslant \lambda_{K+1}\}}, \tag{3.15}$$

with respect to Lebesgue measure, and by (3.14) it is the conditional density of the knots.

3.2.2 *Construction of the Generalized Spacing test.* A useful consequence of Theorem 1.1 is that one can explicitly describe the joint law of the LARS knots after having selected a support $\widehat{S}$ of size $\widehat{m}$ with any procedure satisfying ($\mathscr{A}_{\text{Stop}}$). In the sequel, we write

$$F_i := \Phi_i(\lambda_i) := \Phi\Big(\frac{\lambda_i}{\sigma\rho_i}\Big) \quad \text{and} \quad \mathscr{P}_{i,j} := \Phi_i \circ \Phi_j^{-1}, \quad \text{for } i,j \in [K+1], \tag{3.16}$$

where $\lambda_0 = \infty$ and $F_0 = 1$ by convention.

PROPOSITION 3.4 Let $a \in \mathbb{N}$ be such that $0 \leqslant a \leqslant K-1$. Let $\widehat{m}$ be a selection procedure satisfying ($\mathscr{A}_{\text{Stop}}$). Under the conditions of Theorem 1.1, under the null hypothesis

$$\mathbb{H}_0 : \text{“}X\beta^0 \in H_a\text{”}, \tag{3.17}$$

and conditional on the selection event $\big\{\widehat{m} = a, F_a, F_{K+1}, \widehat{\imath}_1, \dots, \widehat{\imath}_K\big\}$, we have that $(F_{a+1}, \dots, F_K)$ is uniformly distributed on

$$\begin{aligned}\mathscr{D}_{a+1,K} :=\big\{&(f_{a+1},\dots,f_K) \in \mathbb{R}^{K-a} : \\ &\mathscr{P}_{a+1,a}(F_a) \geqslant f_{a+1} \geqslant \mathscr{P}_{a+1,a+2}(f_{a+2}) \geqslant \cdots \geqslant \mathscr{P}_{a+1,K}(f_K) \geqslant \mathscr{P}_{a+1,K+1}(F_{K+1})\big\},\end{aligned}$$

where the $\mathscr{P}_{i,j}$ are described in (3.16).

A proof of this proposition can be found in Appendix 7.2.

REMARK 3.3 *The previous statement is consistent with the case $a = 0$ corresponding to the* global null *hypothesis $\mathbb{H}_0$: ‘$X\beta^0 = 0$’ (or equivalently $\mathbb{E}Z = 0$). Therefore, if Z is centered, then, conditional on F_{K+1}, one has that $(F_1,\dots,F_K)$ is uniformly distributed on*

$$\mathscr{D}_{1,K} := \big\{(f_1,\dots,f_K) \in \mathbb{R}^K \ : \ 1 \geqslant f_1 \geqslant \mathscr{P}_{1,2}(f_2) \geqslant \cdots \geqslant \mathscr{P}_{1,K}(f_K) \geqslant \mathscr{P}_{1,K+1}(F_{K+1}))\big\}.$$

REMARK 3.4 *In the orthogonal case, where $\overline{R} = \mathrm{Id}$, note that $\theta_j(i_1,\dots,i_\ell) = 0$ for all $\ell \geqslant 1$ and all $i_1,\dots,i_\ell \neq j$, $\rho_j = 1$ and $\mathscr{P}_{i,j}(f) = f$. We recover that $\mathscr{D}_{1,K}$ is the set of order statistics*

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

In this case, the knots λ_i are Gaussian order statistics $\lambda_1 = Z_{\widehat{\imath}_1} \geqslant \lambda_2 = Z_{\widehat{\imath}_2} \geqslant \ldots \geqslant \lambda_K = Z_{\widehat{\imath}_K} \geqslant \lambda_{K+1}$ for the vector Z.

From Theorem 1.1, we deduce several test statistics. To this end, we introduce some notation defining

$$\mathscr{I}_{ab}(s,t) := \int_{\mathscr{P}_{(a+1),b}(t)}^{\mathscr{P}_{(a+1),a}(s)} \mathrm{d}f_{a+1} \int_{\mathscr{P}_{(a+2),b}(t)}^{\mathscr{P}_{(a+2),(a+1)}(f_{a+1})} \mathrm{d}f_{a+2} \int_{\mathscr{P}_{(a+3),b}(t)}^{\mathscr{P}_{(a+3),(a+2)}(f_{a+2})} \mathrm{d}f_{a+3} \cdots \int_{\mathscr{P}_{(b-1),b}(t)}^{\mathscr{P}_{(b-1),(b-2)}(f_{b-2})} \mathrm{d}f_{b-1} \tag{3.18}$$

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

and also

$$\mathbb{F}_{abc}(t) := \mathbb{1}_{\{\lambda_c \leqslant t \leqslant \lambda_a\}} \int_{\Phi_b(\lambda_c)}^{\Phi_b(t)} \mathscr{I}_{ab}(F_a, f_b)\, \mathscr{I}_{bc}(f_b, F_c)\, \mathrm{d}f_b \tag{3.19}$$

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

REMARK 3.5 *On the numerical side, note that this quantity can be computed using Quasi Monte Carlo (QMC) methods as in [Genz and Bretz, 2009, Chapter 5.1] or Appendix 8. The function* $\mathbb{F}_{abc}$ *gives the CDF of* λ_b *conditional on* λ_a, λ_c *and on some selection event, as shown in the next proposition.*

PROPOSITION 3.5 Let a, b, and c be such that $0 \leqslant a < b < c \leqslant K+1$. Let $(\lambda_1, \ldots, \lambda_K, \lambda_{K+1})$ be the first knots and let $(\widehat{\imath}_1, \ldots, \widehat{\imath}_K)$ be the first variables entering along the LARS path. If $(\widehat{\imath}_1, \ldots, \widehat{\imath}_K)$ satisfies $(\mathscr{A}_{\mathrm{Irr.}})$ and $\widehat{m}$ is chosen according to a procedure satisfying $(\mathscr{A}_{\mathrm{Stop}})$, then under the null hypothesis

$$\mathbb{H}_0 : \text{“} X\beta^0 \in H_a \text{”},$$

it holds that

$$\mathbb{P}\left[\lambda_b \leqslant t \mid \widehat{m} = a, \lambda_a, \lambda_c, \widehat{\imath}_1, \ldots, \widehat{\imath}_{c-1}\right] = \frac{\mathbb{F}_{abc}(t)}{\mathbb{F}_{abc}(\lambda_a)}. \tag{3.20}$$

A proof of this proposition can be found in Appendix 7.3.

REMARK 3.6 *Note that the deterministic choice* $\widehat{m} = a$*, for a fixed* $a \in [K-1]$*, is a procedure satisfying* $(\mathscr{A}_{\mathrm{Stop}})$ *and Proposition 3.5 holds. This shows that if* $(\widehat{\imath}_1, \ldots, \widehat{\imath}_K)$ *satisfies* $(\mathscr{A}_{\mathrm{Irr.}})$*, then, under the null hypothesis* $\mathbb{H}_0 :$ *“*$X\beta^0 \in H_a$*”,*

$$\mathbb{P}\left[\lambda_b \leqslant t \mid \lambda_a, \lambda_c, \widehat{\imath}_1, \ldots, \widehat{\imath}_{c-1}\right] = \frac{\mathbb{F}_{abc}(t)}{\mathbb{F}_{abc}(\lambda_a)}, \tag{3.21}$$

for any $0 \leqslant a < b < c \leqslant K+1$.

3.2.3 *Proof of Theorem 1.2.* From Proposition 3.5 we know that under $\mathbb{H}_0$ and conditional on the selection event $\{\widehat{m} = a\}$, Eq. (3.20) gives the conditional CDF of λ_b. As a consequence,

$$\frac{\mathbb{F}_{abc}(\lambda_b)}{\mathbb{F}_{abc}(\lambda_a)} \sim \mathscr{U}(0,1).$$

Finally, considerations of the distribution under the alternative show that to obtain a p-value, we must consider the complement to 1 of the quantity above.

3.2.4 *Monte Carlo simulations, Spacing tests, and Generalized Spacing tests.* Theorem 1.2 is illustrated numerically in Figure 2. Note that we have a perfect fit with the uniform law: the conditional law of the LARS knots obtained theoretically is numerically validated[3]. This test statistic generalizes previous test statistics that appeared in '*Spacing Tests*', as presented in [Tibshirani et al., 2015, Chapter 5] for instance, and will be referred to as the *Generalized Spacing test*.

REMARK 3.7 *If one takes* $a = 0$*,* $b = 1$*, and* $c = 2$ *then*

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

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

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

which is the conservative spacing test, see [Tibshirani et al., 2016, Theorem 2].

[3] A reproducible experiment given in a Python notebook is available at `https://github.com/ydecastro/lar_testing/blob/master/Law_LAR.ipynb`

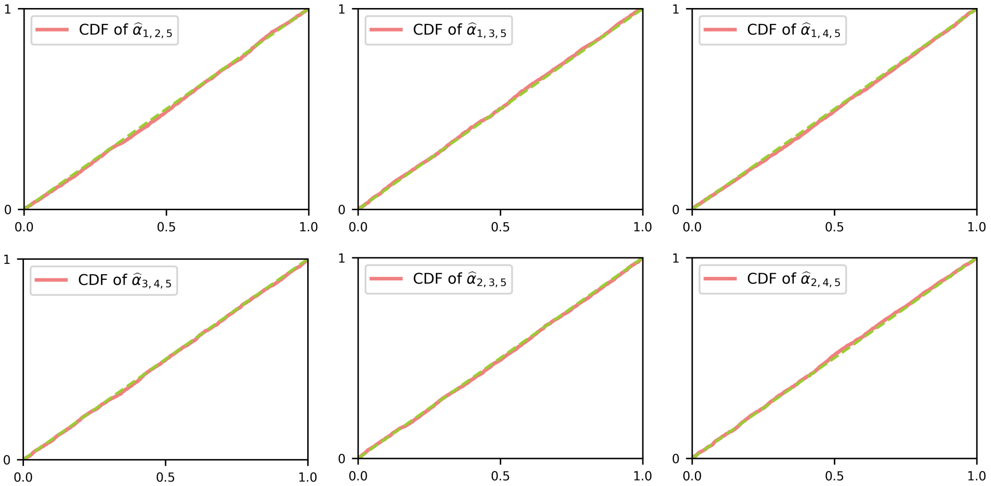


Figure 2: Observed empirical law of $\widehat{\alpha}_{abc}$ over $5,000$ Monte Carlo repetitions with $n = 200$ and $p = 300$. We considered a design $X \in \mathbb{R}^{n\times p}$ with independent column vectors uniformly distributed on the sphere and an independent $y \in \mathbb{R}^n$ with i.i.d. standard Gaussian entries, and we computed the indices $(\widehat{\imath}_1, \ldots, \widehat{\imath}_n)$ and the knots $(\lambda_1, \ldots, \lambda_n)$ entering the model with LARS. The empirical CDF of the $\widehat{\alpha}_{abc}$ are displayed. We observe a perfect fit with the uniform distribution: the conditional law of the LARS knots obtained theoretically is numerically validated.

3.3 *Exact false negative testing after model selection*

We return to the case of a general design. Given $\alpha \in (0,1)$ and using Theorem 1.2, one can consider the following exact testing procedure at level α on false negatives, see the pseudo-code in Algorithm 1. The theoretical guarantee of this algorithm is given by the next proposition. It shows that conditional on the event that 'there are no false negatives', namely '$X\beta^0 \in H_{\widehat{m}}$', the observed significance $\widehat{\alpha}$ obeys the uniform law and hence $\mathbb{1}_{\{\widehat{\alpha}\leqslant\alpha\}}$ is a testing procedure with level exactly α.

COROLLARY 3.1 Let $(\lambda_1, \ldots, \lambda_K, \lambda_{K+1})$ be the first knots and let $(\widehat{\imath}_1, \ldots, \widehat{\imath}_K)$ be the first variables entering along the LARS path. If $(\widehat{\imath}_1, \ldots, \widehat{\imath}_K)$ satisfies $(\mathscr{A}_{\mathrm{Irr.}})$ and $\widehat{m}$ is chosen according to a procedure satisfying $(\mathscr{A}_{\mathrm{Stop}})$, then, conditional on the null hypothesis

$$\mathbb{H}_0 : \text{“}X\beta^0 \in H_{\widehat{m}}\text{”},$$

it holds that

$$\widehat{\alpha}_{\widehat{m}(\widehat{m}+1)(K+1)} := 1 - \frac{\mathbb{F}_{\widehat{m}(\widehat{m}+1)(K+1)}(\lambda_{\widehat{m}+1})}{\mathbb{F}_{\widehat{m}(\widehat{m}+1)(K+1)}(\lambda_{\widehat{m}})} \sim \mathscr{U}(0,1),$$

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

Proof. By Theorem 1.2, the conditional law of $\widehat{\alpha}_{a(a+1)(K+1)}$ with respect to $\{\widehat{m} \leqslant a\}$ is the uniform distribution. Note that the conditional law (1.3) does not depend on $a, b = a+1, c = K+1$, hence this law is unconditional on $\widehat{m}$. □

When the variance σ^2 is unknown, one can 'Studentize' this test, as presented in the next section. The reader may consult Section 3.4 for a definition and check that the quantities $\widehat{\beta}_{abc}, \widetilde{\mathbb{F}}, \Lambda_k$ do not require σ to be computed.

COROLLARY 3.2 Let $(\lambda_1, \dots, \lambda_K, \lambda_{K+1})$ be the first knots and let $(\widehat{\imath}_1, \dots, \widehat{\imath}_K)$ be the first variables entering along the LARS path. If $(\widehat{\imath}_1, \dots, \widehat{\imath}_K)$ satisfies $(\mathscr{A}_{\mathrm{Irr.}})$ and $\widehat{m}$ is chosen according to a procedure satisfying $(\mathscr{A}_{\mathrm{Stop}})$, then, conditional on the null hypothesis

$$\mathbb{H}_0 : \text{“}X\beta^0 \in H_{\widehat{m}}\text{”},$$

it holds that

$$\widehat{\beta}_{\widehat{m}(\widehat{m}+1)(K+1)} := 1 - \frac{\widetilde{\mathbb{F}}_{\widehat{m}(\widehat{m}+1)(K+1)}(\Lambda_{\widehat{m}+1})}{\widetilde{\mathbb{F}}_{\widehat{m}(\widehat{m}+1)(K+1)}(\Lambda_{\widehat{m}})} \sim \mathscr{U}(0,1),$$

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

Proof. By Theorem 1.3, the conditional law of $\widehat{\beta}_{a(a+1)(K+1)}$ with respect to $\{\widehat{m} \leqslant a\}$ is the uniform distribution. Note that the conditional law (3.24) does not depend on $a, b = a+1, c = K+1$, hence this law is unconditional on $\widehat{m}$. □

3.4 *Exact Testing Procedure for False Negatives with Variance Estimation*

From the results of Section 3.2, one can present a method to select a model and propose an exact test of false negatives in the case of a general design, when the variance is unknown. We introduce a new exact testing procedure that can be deployed when $(\mathscr{A}_{\mathrm{Stop}})$ holds, namely an 'admissible' selection procedure is used to build $\widehat{S}$. We start by a preliminary result whose proof is in Appendix 7.4.

PROPOSITION 3.6 Let $(\lambda_1, \dots, \lambda_K, \lambda_{K+1})$ be the first knots of LARS and let $(\widehat{\imath}_1, \dots, \widehat{\imath}_K)$ be the first variables entering along the LARS path. If $(\widehat{\imath}_1, \dots, \widehat{\imath}_K)$ satisfies $(\mathscr{A}_{\mathrm{Irr.}})$, then

- conditional on $\{\widehat{\imath}_1, \dots, \widehat{\imath}_K, \lambda_{K+1}\}$, the random variables $(\lambda_1, \dots, \lambda_K)$ and $\widehat{\sigma}$ are independent;
- conditional on $\{\widehat{\imath}_1, \dots, \widehat{\imath}_K\}$ and under the null hypothesis $\mathbb{H}_0$: '$X\beta^0 \in H_K$', the random variables $(\lambda_{K+1}/\widehat{\sigma})$ and $\widehat{\sigma}$ are independent;
- conditional on $\{\widehat{\imath}_1, \dots, \widehat{\imath}_K, \lambda_{K+1}\}$ and under the null hypothesis $\mathbb{H}_0$: '$X\beta^0 \in H_K$', the distribution of $(\lambda_1, \dots, \lambda_K)$ is given by Theorem 1.1, while the distribution of $\widehat{\sigma}/\sigma$ is the same as the random variable

  $$(n-K)^{-\frac{1}{2}} \Big(\sum_{\ell=1}^{n-K} w_\ell^2 \Big)^{\frac{1}{2}},$$

  where $W := (w_1, \dots, w_{n-K})$ is a 'truncated' standard Gaussian vector with the truncation given by

  $$\|\mathrm{Diag}(\mathbb{1}_p - \theta^K)^{-1} \times X^\top U W\|_\infty = \lambda_{K+1}/\sigma,$$

  where $U \in \mathbb{R}^{n \times (n-K)}$ is any matrix such that $UU^\top = \mathrm{Id}_n - P^{(\widehat{\imath}_1, \dots, \widehat{\imath}_K)}$, $\theta^K := (\theta_j(\widehat{\imath}_1, \dots, \widehat{\imath}_K))_{j \in [p]}$, and with the convention $0/0 = 0$.

REMARK 3.8 *Under the null hypothesis* $\mathbb{H}_0$: *'*$X\beta^0 \in H_K$*', the Gaussian vectors* $P_K^\perp(Y)$ *(see* (2.13)*), defining the variance estimate* $\widehat{\sigma}^2$*, is centered. This null hypothesis means that the true support is included in the set of the K first indices chosen by LARS. One may choose K large enough to guarantee this null hypothesis.*

Recall that, up to some positive numerical constant, the probability density function of the multivariate t-distribution with $(n-K)$ degrees of freedom, mean $m=(m_1,\dots,m_K)$ and variance-covariance matrix $\mathrm{Diag}(\rho_1^2,\dots,\rho_K^2)$ is given by

$$\widetilde{\varphi}(t_1,\dots,t_K):=\left[1+\frac{1}{n-K}\sum_{k=1}^{K}\Big(\frac{t_k-m_k}{\rho_k}\Big)^2\right]^{-\frac{n}{2}}.$$

We have an analogue to Theorem 1.1 giving the joint law of

$$\Lambda_k:=\frac{\lambda_k}{\widehat{\sigma}}\quad\text{for}\quad k=1,\dots,K+1, \tag{3.22}$$

where $\widehat{\sigma}$ is given by (2.13) has $n-K$ degrees of freedom, see Proposition 3.6.

THEOREM 3.7 (Conditional Joint Law of the Studentized LARS knots) *Let $(\lambda_1,\dots,\lambda_K,\lambda_{K+1})$ be the first knots and let $(\widehat{\imath}_1,\dots,\widehat{\imath}_K)$ be the first variables entering along the LARS path. If $(\widehat{\imath}_1,\dots,\widehat{\imath}_K)$ satisfies $(\mathscr{A}_{\mathrm{Irr.}})$ then, under the null hypothesis*

$$\mathbb{H}_0\,:\ \text{`}X\beta^0\in H_K\text{'}$$

and conditional on the selection event $\{\widehat{\imath}_1,\dots,\widehat{\imath}_K,\Lambda_{K+1}\}$, the vector $(\Lambda_1,\dots,\Lambda_K)$ obeys a law with the density (w.r.t. Lebesgue measure)

$$\mathrm{P}^{-1}_{(\widehat{\imath}_1,\dots,\widehat{\imath}_K,\Lambda_{K+1})}\widetilde{\varphi}(t_1,\dots,t_K)\mathbb{1}_{\{t_1\geqslant t_2\geqslant\cdots\geqslant t_K\geqslant\Lambda_{K+1}\}},$$

at point $(t_1,t_2,\dots,t_K)$, where $\mathrm{P}_{(\widehat{\imath}_1,\dots,\widehat{\imath}_K,\Lambda_{K+1})}$ is a normalizing constant, m_k and ρ_k are as in (3.12) *and* (3.13).

Proof of Theorem 3.7. Let us fix some values $i_1,\dots,i_K$. From the definition of the Gaussian random variable $Z_{i_k}^{(i_1,\dots,i_{k-1})}$ in (3.1), one can deduce that its mean m_k is given by (3.12) and its standard deviation $v_k:=\sigma\rho_k$ by (3.13), considering putative indices for the selected variables. By the proof of Proposition 3.6, we know that these variables are independent of $\widehat{\sigma}^{i_1,\dots,i_K}$. We deduce that the vector $(Z_{i_1}/\widehat{\sigma}^{i_1,\dots,i_K},\dots,Z_{i_K}^{(i_1,\dots,i_{K-1})}/\widehat{\sigma}^{i_1,\dots,i_K})$ has density a multivariate t-distribution with $n-K$ degrees of freedom, mean $m=(m_1,\dots,m_K)$ and variance-covariance matrix $\mathrm{Diag}(\rho_1,\dots,\rho_K)$. Furthermore, by (3.10) of Proposition 3.6, we know that this vector is independent of $(Z_j^{(i_1,\dots,i_K)}/\widehat{\sigma}^{i_1,\dots,i_K})_{j\neq i_1,\dots,i_K}$, and, in particular, independent of $\Lambda_{K+1}^{(i_1,\dots,i_K)}:=\max_j\{Z_j^{(i_1,\dots,i_K)}/\widehat{\sigma}^{i_1,\dots,i_K}\}$. Recall that, conditional on

$$\mathscr{E}:=\{\widehat{\imath}_1=i_1,\dots,\widehat{\imath}_K=i_K,\lambda_{K+1}\},$$

and assuming that $(i_1,\dots,i_K)$ satisfies $(\mathscr{A}_{\mathrm{Irr.}})$, Proposition 3.2 implies that

$$\mathscr{E}=\left\{\Lambda_{K+1}^{(i_1,\dots,i_K)}\leqslant\frac{Z_{i_K}^{(i_1,\dots,i_{K-1})}}{\widehat{\sigma}^{i_1,\dots,i_K}}\leqslant\cdots\leqslant\frac{Z_{i_1}}{\widehat{\sigma}^{i_1,\dots,i_K}}\right\}.$$

Furthermore, on the event $\mathscr{E}$ we have

$$(Z_{i_1}/\widehat{\sigma}^{i_1,\dots,i_K},\dots,Z_{i_K}^{(i_1,\dots,i_{K-1})}/\widehat{\sigma}^{i_1,\dots,i_K},\Lambda_{K+1}^{(i_1,\dots,i_K)})=(\Lambda_1,\dots,\Lambda_K,\Lambda_{K+1}).$$

Because of the independence above, this implies that the conditional distribution is the one claimed. □

For $0 \leqslant a < b \leqslant K+1$, we introduce

$$\widetilde{\varphi}_{ab}(t_{a+1},\dots,t_{b-1}) := \left[1 + \frac{1}{n-K}\sum_{k=a+1}^{b-1}\Big(\frac{t_k}{\rho_k}\Big)^2\right]^{-\frac{n-K+b-a}{2}}$$

$$\widetilde{\mathscr{I}}_{ab}(s,t) := \int_{\{s \geqslant t_{a+1} \geqslant \dots \geqslant t_{b-1} \geqslant t\}} \widetilde{\varphi}_{ab}(t_{a+1},\dots,t_{b-1})\mathrm{d}t_{a+1}\cdots\mathrm{d}t_{b-1},$$

with the convention $\widetilde{\mathscr{I}}_{ab}(s,t) = 1$ when $b = a+1$: and also

$$\widetilde{\mathbb{F}}_{abc}(t) := \mathbb{1}_{\{\Lambda_c \leqslant t \leqslant \Lambda_a\}} \int_{\Lambda_c}^{t} \widetilde{\mathscr{I}}_{ab}(\Lambda_a,\ell_b)\,\widetilde{\mathscr{I}}_{bc}(\ell_b,\Lambda_c)\left[1 + \frac{1}{n-K}\Big(\frac{\ell_b}{\rho_b}\Big)^2\right]^{-\frac{n-K+1}{2}} \mathrm{d}\ell_b \tag{3.23}$$

$$\text{for } 0 \leqslant a < b < c \leqslant K+1,\ t \in \mathbb{R}.$$

When $m_{a+1} = \cdots = m_{c-1} = 0$, the function $\widetilde{\mathbb{F}}_{abc}$ gives the CDF of Λ_b conditional on Λ_a, Λ_c and on some selection event, as shown below in Theorem 1.3 and (3.25). For $0 \leqslant a < b < c \leqslant K+1$, we introduce the p-value

$$\widehat{\beta}_{abc} = \widehat{\beta}_{abc}(\Lambda_a,\Lambda_b,\Lambda_c,\widehat{\imath}_1,\dots,\widehat{\imath}_K) = 1 - \frac{\widetilde{\mathbb{F}}_{abc}(\Lambda_b)}{\widetilde{\mathbb{F}}_{abc}(\Lambda_a)} \tag{3.24}$$

On the numerical side, note that this quantity can be computed using *Quasi Monte Carlo* (QMC) methods as in [Genz and Bretz, 2009, Chapter 5.1].

3.4.1 *Proof of Theorem 1.3.* Fix a such that $0 \leqslant a \leqslant K-1$ and consider any selection procedure $\widehat{m}$ satisfying ($\mathscr{A}_{\text{Stop}}$). From Proposition 3.1, conditional on

$$\mathscr{F} := \big\{\widehat{\imath}_1 = i_1,\dots,\widehat{\imath}_K = i_K, \Lambda_a, \Lambda_{K+1}\big\}$$

and under the null hypothesis $\mathbb{H}_0$: '$X\beta^0 \in H_a$', we know that $m_{a+1} = \ldots = m_K = 0$. From Theorem 3.7 we know that the density of $(\Lambda_{a+1},\Lambda_{a+2},\dots,\Lambda_K)$ conditional on $\mathscr{F}$ is given by

$$(const)\left[1 + \frac{1}{n-K}\sum_{k=a+1}^{K}\Big(\frac{t_k}{\rho_k}\Big)^2\right]^{-\frac{n-a}{2}} \mathbb{1}_{\Lambda_a \geqslant \ell_{a+1} \geqslant \cdots \geqslant \ell_K \geqslant \Lambda_{K+1}}.$$

From the definition of assumption ($\mathscr{A}_{\text{Stop}}$), and on the event $\mathscr{F}$, we know that the indicator $\mathbb{1}_{\{\widehat{m}=a\}}$ is a measurable function of $\lambda_1,\dots,\lambda_{a-1}$, which are respectively equal to $Z_{i_1},\dots,Z_{i_{a-1}}^{(i_1,\dots,i_{a-2})}$ on $\mathscr{F}$ by (3.5). By (3.10) of Proposition 3.2, we deduce that $\mathbb{1}_{\{\widehat{m}=a\}}$ is independent of $(\lambda_{a+1}/\widehat{\sigma}^{i_1,\dots,i_K},\dots,\lambda_K/\widehat{\sigma}^{i_1,\dots,i_K})$ and of $\Lambda_{K+1} := \lambda_{K+1}/\widehat{\sigma}^{i_1,\dots,i_K}$ conditional on $\mathscr{F}$. We deduce that the conditional density above is also the conditional density on the event

$$\mathscr{G} := \big\{\widehat{m} = a, \widehat{\imath}_1 = i_1,\dots,\widehat{\imath}_K = i_K, \Lambda_a, \Lambda_{K+1}\big\}.$$

Now, a simple integration shows that

$$\mathbb{P}\big[\Lambda_b \leqslant t \mid \widehat{m} = a, \Lambda_a, \Lambda_c, \widehat{\imath}_1,\dots,\widehat{\imath}_K\big] = \frac{\widetilde{\mathbb{F}}_{abc}(t)}{\widetilde{\mathbb{F}}_{abc}(\Lambda_a)}. \tag{3.25}$$

As a consequence and under the same conditioning, one has

$$\frac{\widetilde{\mathbb{F}}_{abc}(\Lambda_b)}{\widetilde{\mathbb{F}}_{abc}(\Lambda_a)} \sim \mathscr{U}(0,1).$$

Finally, considerations of the distribution under the alternative show that to obtain a p-value we must consider the complement to 1 of the quantity above.

3.4.2 *t-Spacing tests and Generalized t-Spacing tests.*

Consider the following testing procedures:

$$\mathscr{T}_{abc} := \mathbb{1}_{\{\widehat{\beta}_{abc} \leqslant \alpha\}}, \tag{3.26}$$

that rejects if the p-value $\widehat{\beta}_{abc}$ is less than the level α of the test. This test statistic generalizes previous test statistics that appeared in t-Spacing Tests, as presented in Azaïs et al. [2018] for instance, and will be referred to as the *Generalized t-Spacing test* (GtSt).

REMARK 3.9 *If one takes $a=0$, $b=1$ and $c=2$, then one gets*

$$\widehat{\beta}_{012} = 1 - \frac{\boldsymbol{T}_1(\Lambda_1) - \boldsymbol{T}_1(\Lambda_2)}{\boldsymbol{T}_1(\Lambda_0) - \boldsymbol{T}_1(\Lambda_2)} = \frac{1-\boldsymbol{T}_1(\Lambda_1)}{1-\boldsymbol{T}_1(\Lambda_2)}.$$

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

$$\widehat{\beta}_{a(a+1)(a+2)} = \frac{\boldsymbol{T}_{a+1}(\Lambda_{a+1}) - \boldsymbol{T}_{a+1}(\Lambda_a)}{\boldsymbol{T}_{a+1}(\Lambda_{a+2}) - \boldsymbol{T}_{a+1}(\Lambda_a)}.$$

which is the t-spacing test as presented in Azaïs et al. [2018], where

$$\boldsymbol{T}_k(\ell) := \int_{-\infty}^{\ell} \left[1 + \frac{1}{n-K}\left(\frac{\ell}{\rho_k}\right)^2\right]^{-\frac{n-K+1}{2}} \mathrm{d}\ell \tag{3.27}$$

is, up to some positive numerical constant, the CDF of a centered t-Student distribution with variance ρ_k^2 and $n-K$ degrees of freedom.

3.5 *Power studies*

3.5.1 *Power when the design is orthogonal.*

One may investigate the power of these tests at detecting false negatives, namely, the alternatives given by: there exists $k \in S^0$ such that $k \notin \{\bar{\imath}_1, \ldots, \bar{\imath}_a\}$. In particular, what is the most powerful test among these latter (1.4) testing procedures? A comprehensive study for the case of orthogonal designs is given by Theorem 1.4.

3.5.2 *Numerical studies on the power for the general design case.*

In the case of an orthogonal design, Theorem 1.4 shows that the test based on $\widehat{\alpha}_{a,a+1,K+1}$ is uniformly more powerful than tests based on $\widehat{\alpha}_{x,y,z}$ with $a \leqslant x < y < z \leqslant K+1$. Numerical experiments on the power of these tests are presented in Figure 3 and they witness the same phenomenon for Gaussian designs. It presents the CDF of the p-value $\widehat{\alpha}_{abc}$ under the null and under two 2-sparse alternatives, one with low signal and one with 5 times more signal. The numerical results show that all the tests are exact (leftmost panel) and the test $\mathscr{S}_{125}$ is the most powerful. A detailed presentation of this is given in Section 4.1.

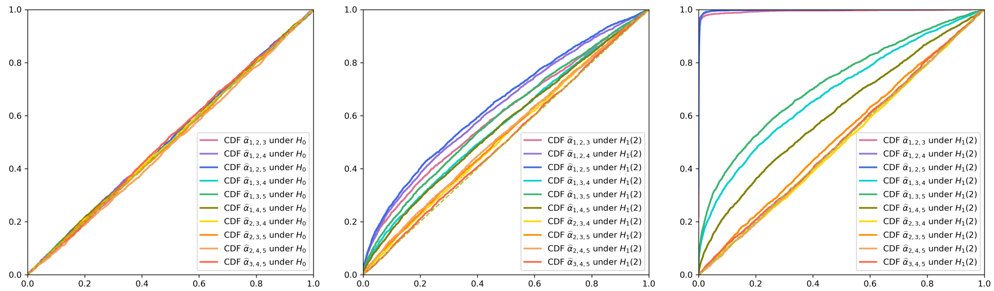


Figure 3: CDF of p-values $\widehat{\alpha}_{abc}$ over $3,000$ Monte Carlo iterations and a random design $X \in \mathbb{R}^{n\times p}$ given by $p = 300$ independent column vectors uniformly distributed on the Euclidean sphere $\mathbb{S}^{199}$ ($n = 200$). Central panel represents alternative composed by 2-sparse vector, right panel alternative composed by 2-sparse vector 5 times larger while left panel corresponds to the null.

More precisely, one has, as proved in the orthogonal case by Theorem 1.4 (and its proof), that

- $\widehat{\alpha}_{125} \preccurlyeq \widehat{\alpha}_{124} \preccurlyeq \widehat{\alpha}_{123}$;
- $\widehat{\alpha}_{125} \preccurlyeq \widehat{\alpha}_{135} \preccurlyeq \widehat{\alpha}_{235} \preccurlyeq \widehat{\alpha}_{234}$;
- $\widehat{\alpha}_{125} \preccurlyeq \widehat{\alpha}_{135} \preccurlyeq \widehat{\alpha}_{145} \preccurlyeq \widehat{\alpha}_{245}$.

where $\preccurlyeq$ denotes stochastic ordering. In the proof of Theorem 1.4, it was shown that

$$\widehat{\alpha}_{ab(c+1)} \preccurlyeq \widehat{\alpha}_{abc} \text{ and } \widehat{\alpha}_{a(b-1)c} \preccurlyeq \widehat{\alpha}_{abc} \text{ and } \widehat{\alpha}_{(a-1)bc} \preccurlyeq \widehat{\alpha}_{abc},$$

for orthogonal designs.

3.6 *Control of False Discovery Rate in the Orthogonal Design case*

3.6.1 *Presentation in the general case.* For the sake of readability, we will assume, for the moment, that σ is known. We understand that the law of test statistics are parametrized by the hypotheses $(m_k)_{k\in[K]}$, where m_k is given by (3.12).

We recall that we write $\overline{\mu}^0 = X^\top X\beta^0$ and $\overline{\mu}^0_i$ for its ith coordinate. Assuming that the predictors are normalised, in the general case, this quantity is the sum of β^0_i and a linear combination of the β^0_j's whose predictors X_j are highly correlated with the predictor X_i. Now, given the variables $\bar\imath_1,\dots,\bar\imath_k \in [p]$ and signs $\varepsilon_1,\dots,\varepsilon_k \in \{\pm 1\}^k$, we denote by $(\Pi^{\perp}_{\bar\imath_1,\dots,\bar\imath_{k-1}}(\overline{\mu}^0))_{\bar\imath_k}$ the orthogonal projection given by

$$(\Pi^{\perp}_{\bar\imath_1,\dots,\bar\imath_{k-1}}(\overline{\mu}^0))_{\bar\imath_k} := \varepsilon_k X_{\bar\imath_k}^\top \Big[\mathrm{Id}_n - X_{\overline{S}^{k-1}}\big(X_{\overline{S}^{k-1}}^\top X_{\overline{S}^{k-1}}\big)^{-1} X_{\overline{S}^{k-1}}^\top\Big] X\beta^0. \tag{3.28}$$

The tested null hypotheses are conditional on some sub-sequence of variables $(\bar\imath_1,\dots,\bar\imath_{K+1}) \in [p]^{K+1}$ and signs $\varepsilon_1,\dots,\varepsilon_{K+1} \in \{\pm 1\}^{K+1}$ entering the model. The p-values under consideration are

○ $\widehat{p}_1 := \widehat{\alpha}_{0,1,2}$ is the p-value testing $\mathbb{H}_{0,1}$: "$m_1 = 0$" namely $\overline{\mu}^0_{\bar\imath_1} = 0$;

○ $\widehat{p}_2 := \widehat{\alpha}_{1,2,3}$ is the p-value testing $\mathbb{H}_{0,2}$: "$m_2 = 0$" namely $(\Pi^{\perp}_1(\overline{\mu}^0))_{\bar\imath_2} = 0$;

○ $\widehat{p}_3 := \widehat{\alpha}_{2,3,4}$ is the p-value testing $\mathbb{H}_{0,3}$: "$m_3 = 0$" namely $(\Pi^{\perp}_2(\overline{\mu}^0))_{\bar\imath_3} = 0$; (3.29)

○ and so on...

We write I_0 of the set $I_0 = \big\{k \in [K]\ :\ \mathbb{H}_{0,k} \text{ is true}\big\}$. Given a subset $\widehat{R} \subseteq [K]$ of hypotheses that we consider as rejected, we call *false positive* (FP) and *true positive* (TP) the quantities $\mathrm{FP} = \mathrm{card}(\widehat{R}\cap I_0)$ and $\mathrm{TP} = \mathrm{card}(\widehat{R}\setminus I_0)$. Denote by $\widehat{p}_{(1)} \leqslant \ldots \leqslant \widehat{p}_{(K)}$ the p-values ranked in a nondecreasing order. Let $\alpha \in (0,1)$ and consider the Benjamini–Hochberg procedure, see for instance Benjamini and Hochberg [1995], defined by a rejection set $\widehat{R} \subseteq [K]$ such that $\widehat{R} = \emptyset$ when $\{k \in [K]\ :\ \widehat{p}_{(k)} \leqslant \alpha k/K\} = \emptyset$ and

$$\widehat{R} = \{k \in [K]\ :\ \widehat{p}_k \leqslant \alpha\widehat{k}/K\} \quad \text{where} \quad \widehat{k} = \max\big\{k \in [K]\ :\ \widehat{p}_{(k)} \leqslant \alpha k/K\big\}. \tag{3.30}$$

Recall the definition of the FDR as the mean of the False Discovery Proportion (FDP), namely

$$\mathrm{FDR} := \mathbb{E}\Big[\underbrace{\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TP}}\mathbb{1}_{\mathrm{FP}+\mathrm{TP}\geqslant 1}}_{\mathrm{FDP}}\Big],$$

where the expectation is unconditional on 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 by invoking the following decomposition

$$\mathrm{FDR} = \sum_{(\imath_1,\ldots,\imath_K)\in[p]^K} \overline{\pi}_{(\imath_1,\ldots,\imath_K)}\, \mathbb{E}\big[\mathrm{FDP}|\bar{\imath}_1 = i_1,\ldots,\bar{\imath}_K = i_K\big],$$

where $\overline{\pi}_{(\imath_1,\ldots,\imath_K)} = \mathbb{P}\big\{\bar{\imath}_1 = \imath_1,\ldots,\bar{\imath}_K = \imath_K\big\}$.

3.6.2 *Control of the FDR by the Benjamini–Hochberg procedure in the orthogonal design case.* We now consider the case of an orthogonal design where $X^\top X = \mathrm{Id}_p$ and the set of p-values is given by (3.29). Note that I_0 is simply the set of null coordinates of β^0. Remark also that the Irrepresentable Condition (Irrep.) of order p holds and so does Empirical Irrepresentable Check ($\mathscr{A}_{\mathrm{Irr.}}$), see Proposition 2.4.

THEOREM 3.8 *Assume that the design is orthogonal, i.e. $X^\top X = \mathrm{Id}_p$, and let $K \in [p]$. Let $(\bar{\imath}_1,\ldots,\bar{\imath}_K)$ be the first variables entering along the LARS path. Consider the p-values given by* (3.29) *and the set $\widehat{R}$ given by* (3.30)*. Then*

$$\mathbb{E}\big[\mathrm{FDP}|\bar{\imath}_1 = i_1,\ldots,\bar{\imath}_K = i_K\big] \leqslant \alpha,$$

and so FDR *is bounded above by α.*

The proof of this result is given in Appendix 7.8. 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 $\{\bar{\imath}_1 = i_1,\ldots,\bar{\imath}_K = i_K\}$ and if one considers the Benjamini–Hochberg procedure for the hypotheses described in Section 3.6.1, then the FDR is exactly controlled by α.

4. Testing procedures: Numerical studies

4.1 *Power in the non-orthogonal case*

To study the power in the case of a *non-orthogonal* design, we built a Monte-Carlo experiment with:

- a model with $n = 200$ observations and $p = 300$ predictors,
- a random design matrix X given by 300 independent column vectors uniformly distributed on the Euclidean sphere $\mathbb{S}^{199}$,

- and we ran $3,000$ Monte Carlo experiments.

- The results are presented in Figure 3.

The computation of the function $\mathbb{F}_{abc}$ given by (3.19) requires multivariate integration tools. All our test statistics can be efficiently computed using Quasi Monte Carlo methods (QMC) for Multi-Variate Normal (MVN) and t (MVT) distributions, see the book Genz and Bretz [2009] for a comprehensive treatment of this topic or Appendix 8 for a short overview of the method we used. We compute spacings of length at most 4, which implies that $c \leqslant 5$ 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 $\widehat{\alpha}_{(\texttt{start})(\texttt{middle})(\texttt{end})}$ of the knots and indices given by `lars` and an estimate of (or the true) standard deviation given by `sigma`. We ran $3,000$ repetitions of this function to get the laws displayed in Figure 3. It presents the CDF of the p-value $\widehat{\alpha}_{abc}$ under the null and under two 2-sparse alternatives, one with low signal and one with 5 times more signal. The results show, in our particular case, that all the tests are exact and the test $\mathscr{S}_{125}$ is the most powerful, see Section 3.5.2 for further details.

4.2 *A comparison of FDR control and power on simulated data*

We take the experiments introduced in [Javanmard et al., 2019, Section 5]. As in this reference, we consider a linear model with design X with independent rows drawn with respect to $\mathscr{N}_p(0,\Sigma)$. The covariance $\Sigma \in \mathbb{R}^{p\times p}$ is such that $\Sigma_{ij} = r^{|i-j|}$, for some parameter $r \in (0,1)$. We then normalize the columns of X to have unit Euclidean norm. We draw a k-sparse vector $\beta^0 \in \mathbb{R}^p$ by choosing a support of size k at random with values $\{\pm A\}$ uniformly at random, where $A > 0$ denotes the absolute value of the amplitudes. The Gaussian noise term η is drawn from $\mathscr{N}_n(0, \mathrm{Id}_n)$.
We compare the performances of three procedures:

- **[Knockoff]** Knockoff filters for FDR control [Barber et al., 2015] and we use `knockoff+` as implemented on `https://web.stanford.edu/group/candes/knockoffs/`;

- **[FCD]** False Discovery Control via Debiasing [Javanmard et al., 2019, Section 5] and we use the implementation of debiased lasso presented on the webpage `https://web.stanford.edu/~montanar/sslasso/` with the theoretical value $\lambda = 2\sqrt{(2\log p)/n}$ for the regularizing parameter. When the sample size is larger than the number of predictors ($n \geqslant p$), the debiasing step in FCD is superfluous as the decorrelating matrix (M) can be the inverse of the sample covariance. So, in this case, we start with an unbiased estimator upfront (which is Ordinary Least Squares OLS). The FCD then becomes thresholding the test statistics $|T_i|$ obtained from OLS;

- **[GtSt-BH]** Generalized t-Spacing tests on successive entries of the LARS path combined with a Benjamini–Hochberg procedure Benjamini and Hochberg [1995] based on the sequence of spacings $\widehat{\beta}_{012}, \widehat{\beta}_{123}, \ldots, \widehat{\beta}_{a(a+1)(a+2)}, \ldots$ as described in Section 3.4 with nominal value $\alpha = 0.1$;

we numerically investigate the effects of the level of sparsity, the magnitude of the signal, the correlation between the features, and the empirical power. In all simulations, we set the target level FDR to $\alpha = 0.1$.

4.2.1 *The effect of the amplitude of the signal.* We chose $n = 200$ (sample size), $p = 100$ (predictors), $k = 20$ (sparsity), $\eta = 0.1$ (features correlation) and varied the amplitude within the set $A \in [25]$. We computed the FDR and power by averaging over $3,000$ realizations of the noise and generations of the coefficients of the vector β^0. The results are plotted in Figure 4. Recall that, in the case $n \geqslant p$, FCD is a thresholded OLS and it might be considered as the best test here. One may note that it presents the best features (low FDR and high power). GtSt controls the FDR below the nominal value $\alpha = 0.1$ with a slightly lower power than FCD. Knockoff+ has controlled FDR and matches the power of FCD.

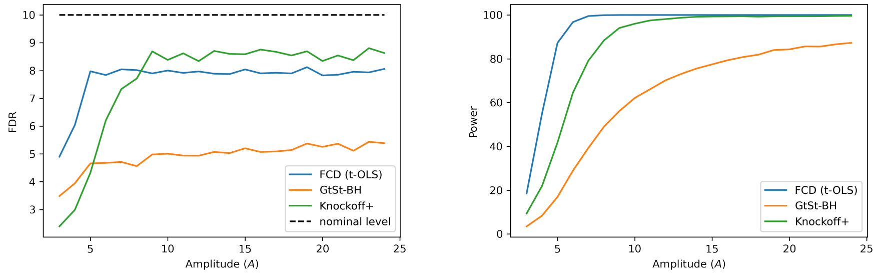


Figure 4: Comparison of the FDR control and power for GtSt, FCD and Knockoff+ when the amplitude of the signal varies from 1 (low signal) to 25 (strong signal) over $3,000$ trials.

4.2.2 *Effect of feature correlation.* We test the effect of correlations between the features with $n = 200$, $p = 100$, $k = 20$, and $A = 10$. Recall that the rows of the design matrix X are generated from an $\mathscr{N}_p(0,\Sigma)$ distribution, with $\Sigma_{ij} = \eta^{|i-j|}$, and then the columns of X are normalized to have unit norm. We vary the parameter η within the set $\{0.1, 0.15, 0.2, \ldots, 0.75, 0.8, 0.85\}$. For each value of η, we compute the FDR and power by averaging over $3,000$ realizations of the noise and design matrix X. The results are displayed in Figure 5. One may note that, in the case $n \geqslant p$, FCD is a thresholded OLS and might be considered as the best estimation here (with low FDR and high power). Knockoff+ has controlled FDR and matches the power of FCD for small feature correlations. GtSt controls the FDR below the nominal value $\alpha = 0.1$ with a lower power than Knockoff+.

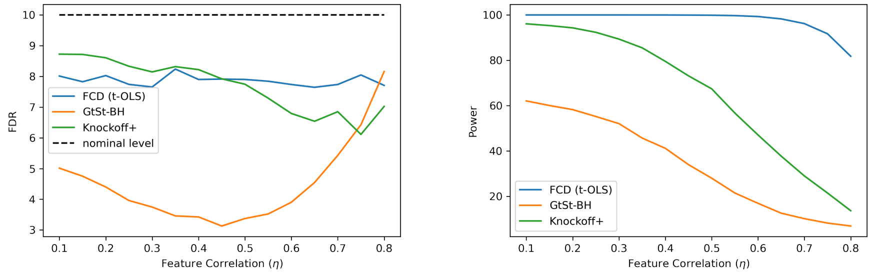


Figure 5: Comparison of the FDR control and power for GtSt, FCD and Knockoff when the correlation between features varies from 0.1 (low correlation) to 0.85 (strong correlation) over $3,000$ trials.

4.2.3 *Effect of sparsity.* We set $n = 200$, $p = 100$, $A = 10$, and $\eta = 0.1$, and varied the level of sparsity of the coefficients within the set $k \in (10,40)$. The power and the FDR are computed by averaging over $3,000$ trials of the noise and generations of the coefficients of the vector β^0. The results are displayed in Figure 6. Knockoff+ has controlled FDR and matches the power of FCD. GtSt controls the FDR below the nominal value $\alpha = 0.1$ with a lower power than Knockoff+.

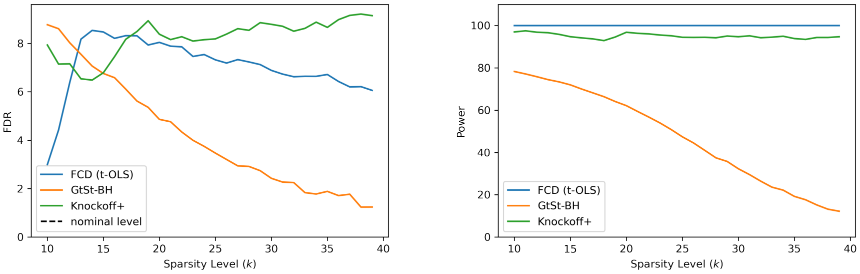


Figure 6: Comparison of the FDR control and power for GtSt, FCD and Knockoff when the feature correlation varies from 1 (very sparse) to 35 (slightly sparse) over $3,000$ trials.

4.3 *FDR on 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 Barber et al. [2015] and Rhee et al. [2006]. The experiment consists in identifying mutations of the genes of the HIV that are involved with drug resistance. The data set contains about $p = 200$ and $n = 700$ observations. Since some protocol was used to remove some genes and some individuals, the exact numbers depend on the considered drug.

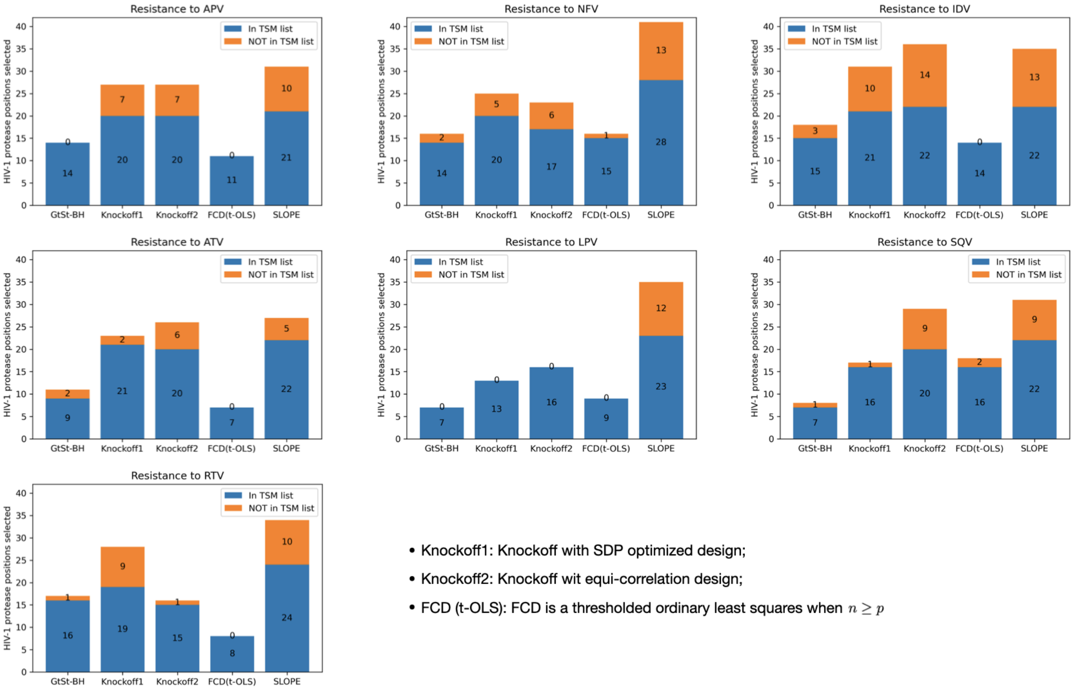
The methods considered are **[Knockoff]**, **[FCD]**, **[GtSt-BH]**, and:

- **[Slope]** Slope for FDR control, as presented in [Bogdan et al., 2015].

The comparison is displayed in Figure 7. It appears that GtSt-BH and FCD procedures are more conservative but they give a better control of the False Discovery Proportion (FDP). SLOPE and Knockoff are more powerful but their FDP is greater than the expected FDR $\alpha = 0.2$ (in 6 experiments out of 7 for SLOPE, in 3 out of 7 for Knockoff1, and in 5 out of 7 for Knockoff2).

Acknowledgements

The authors would like to thank anonymous referees for their time, comments and useful remarks on preliminary versions of this paper. The authors are in debt to Quentin Duchemin for his valuable comments and remarks on this work.



- Knockoff1: Knockoff with SDP optimized design;
- Knockoff2: Knockoff wit equi-correlation design;
- FCD (t-OLS): FCD is a thresholded ordinary least squares when $n \geq p$

Figure 7: Comparison of the number of true and false positives for procedures: GtSt-BH, Knockoff (with SDP optimisation designs in Knockoff1 and equi-correlation designs in Knockoff2), FDC and Slope. In the procedures, the FDR aimed at is $\alpha = 20\%$. Blue indicates protease positions that appear in the TSM panel for the PI class of treatments, given in [Rhee et al., 2006, Tabel 1], while orange indicates positions selected by the method that do not appear in the TSM list. The total number of HIV-1 protease positions appearing in the TSM list is 34.

Supplement to
Multiple Testing and Variable Selection along the path of the Least Angle Regression

5. Representing the LARS knots

5.1 *The equivalent formulations of the LARS algorithm*

We present here three equivalent formulations of the LARS that are a consequence of the analysis provided in Appendices 5 and 6. One new formulation is given by Algorithm 4.

Algorithm 2 LARS algorithm (standard formulation)

```
Data: Correlations vector $\overline{Z}$ and variance-covariance matrix $\overline{R}$.
Result: Sequence $((\lambda_k, \bar{\imath}_k, \varepsilon_k))_{k \geqslant 1}$ where $\lambda_1 \geqslant \lambda_2 \geqslant \ldots > 0$ are the knots, and $\bar{\imath}_1, \bar{\imath}_2, \ldots$ are the variables that enter the model with signs $\varepsilon_1, \varepsilon_2, \ldots$ ($\varepsilon_k = \pm 1$).

/* Initialize computing $(\lambda_1, \bar{\imath}_1, \varepsilon_1)$ and defining a 'residual' $\overline{N}^{(1)}$.  */
1 Set $k = 1$, $\lambda_1 := \max|\overline{Z}|$, $\bar{\imath}_1 := \arg\max|\overline{Z}|$ and $\varepsilon_1 = \overline{Z}_{\bar{\imath}_1}/\lambda_1 \in \pm 1$, and $\overline{N}^{(1)} := \overline{Z}$.

/* Note that $((\lambda_\ell, \bar{\imath}_\ell, \varepsilon_\ell))_{1 \leqslant \ell \leqslant k-1}$ and $\overline{N}^{(k-1)}$ have been defined at the previous step.  */
2 Set $k \leftarrow k + 1$ and compute the least-squares fit
      $\overline{\theta}_j := \big(\overline{R}_{j,\bar{\imath}_1} \cdots \overline{R}_{j,\bar{\imath}_{k-1}}\big) M^{-1}_{\bar{\imath}_1,\ldots,\bar{\imath}_{k-1}} (\varepsilon_1, \ldots, \varepsilon_{k-1}), \quad j = 1, \ldots, p,$
  where $M_{\bar{\imath}_1,\ldots,\bar{\imath}_{k-1}}$ is the sub-matrix of $\overline{R}$ keeping the columns and the rows indexed by $\{\bar{\imath}_1, \ldots, \bar{\imath}_{k-1}\}$.
3 For $0 < \lambda \leqslant \lambda_{k-1}$ 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
      $\lambda_k := \max\big\{\beta > 0;\ \exists j \notin \{\bar{\imath}_1, \ldots, \bar{\imath}_{k-1}\},\ \text{s.t. } |\overline{N}^{(k)}_j(\beta)| = \beta\big\}$ and $\bar{\imath}_k := \underset{j \notin \{\bar{\imath}_1,\ldots,\bar{\imath}_{k-1}\}}{\arg\max} |\overline{N}^{(k)}_j(\lambda_k)|,$
      $\varepsilon_k := \overline{N}^{(k)}_{\bar{\imath}_k}(\lambda_k)/\lambda_k \in \pm 1$ and $\overline{N}^{(k)} := \overline{N}^{(k)}(\lambda_k).$
  Then, iterate from **2**.
```

Algorithm 3 LARS algorithm (“projected” formulation)

```
Data: Correlations vector $\overline{Z}$ and variance-covariance matrix $\overline{R}$.
Result: Sequence $((\lambda_k, \bar{\imath}_k, \varepsilon_k))_{k\geqslant 1}$ where $\lambda_1 \geqslant \lambda_2 \geqslant \ldots > 0$ are the knots, and $\bar{\imath}_1, \bar{\imath}_2, \ldots$ are the variables that enter the model with signs $\varepsilon_1, \varepsilon_2, \ldots$ $(\varepsilon_k = \pm 1)$.
    /* Initialize computing $(\lambda_1, \bar{\imath}_1, \varepsilon_1)$.                         */
1 Define $Z = (\overline{Z}, -\overline{Z})$ and $R$ as in (2.4), and set $k = 1$, $\lambda_1 := \max Z$ , $\widehat{\imath}_1 := \arg\max Z$, $\bar{\imath}_1 = \widehat{\imath}_1 \mod p$ and $\varepsilon_1 = 1 - 2(\widehat{\imath}_1 - \bar{\imath}_1)/p \in \pm 1$.
    /* Note that $((\lambda_\ell, \widehat{\imath}_\ell))_{1\leqslant \ell\leqslant k-1}$ have been defined at the previous
    step/loop.                                                                                    */
2 Set $k \leftarrow k+1$ and compute
    $\lambda_k = \max_{\{j:\, \theta_j(\widehat{\imath}_1,\ldots,\widehat{\imath}_{k-1})<1\}} \left\{ \frac{Z_j - \Pi_{\widehat{\imath}_1,\ldots,\widehat{\imath}_{k-1}}(Z_j)}{1 - \theta_j(\widehat{\imath}_1,\ldots,\widehat{\imath}_{k-1})} \right\}$ and $\widehat{\imath}_k = \arg\max_{\{j:\, \theta_j(\widehat{\imath}_1,\ldots,\widehat{\imath}_{k-1})<1\}} \left\{ \frac{Z_j - \Pi_{\widehat{\imath}_1,\ldots,\widehat{\imath}_{k-1}}(Z_j)}{1 - \theta_j(\widehat{\imath}_1,\ldots,\widehat{\imath}_{k-1})} \right\}$,
  where
    $\Pi_{\widehat{\imath}_1,\ldots,\widehat{\imath}_{k-1}}(Z_j) := \big(R_{j,\widehat{\imath}_1} \cdots R_{j,\widehat{\imath}_{k-1}}\big) M^{-1}_{\widehat{\imath}_1,\ldots,\widehat{\imath}_{k-1}} (Z_{\widehat{\imath}_1}, \ldots, Z_{\widehat{\imath}_{k-1}})$
    $\theta_j(\widehat{\imath}_1,\ldots,\widehat{\imath}_{k-1}) := \big(R_{j,\widehat{\imath}_1} \cdots R_{j,\widehat{\imath}_{k-1}}\big) M^{-1}_{\widehat{\imath}_1,\ldots,\widehat{\imath}_{k-1}} (1, \ldots, 1)$
  and set $\bar{\imath}_k = \widehat{\imath}_k \mod p$ and $\varepsilon_k = 1 - 2(\widehat{\imath}_k - \bar{\imath}_k)/p \in \pm 1$. Then, iterate from 2.
```

5.2 *A new formulation of Least Angle Regression algorithm*

The Least Angle Regression (LARS) algorithm has been introduced in the seminal article Efron et al. [2004]. In the context of linear regression in high dimensions, the LARS algorithm can be used to identify a subset of potential covariates. The LARS outputs a piecewise affine solutions path, and the *knots* $\lambda_1 \geqslant \lambda_2 \geqslant \cdots > 0$ are the change points of the LARS path that are built by tracking the ℓ_∞ of the residual. At each knot, the LAR algorithm adds to the active set of variables the covariate the most correlated with the actual residual. In that way, the descent direction is always equiangular to all variables present in the current active set. . This sequence of knots is closely related to the sequence of knots of LASSO [Tibshirani, 1996], 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 [Tibshirani et al., 2015, Page 120] or [Efron et al., 2004, Theorem 1] for instance.

THEOREM 5.1 *Let n, p be integers. Given a vector $Y \in \mathbb{R}^n$ and matrix $X \in \mathbb{R}^{n\times p}$ of rank r then Algorithm 2 (LARS standard formulation), Algorithm 3 (LARS projected formulation) and Algorithm 4 (LARS recursive formulation) output the same sequence $((\lambda_k, \bar{\imath}_k, \varepsilon_k))_{k=1}^r$ from the input given by $\overline{Z} = X^\top Y$ and $\overline{R} = X^\top X$, where $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_2 \ldots \geqslant 0$ are the knots, and $\bar{\imath}_1, \bar{\imath}_2, \ldots, \bar{\imath}_r$ are the variables that enter the model with signs $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_r$ $(\varepsilon_k = \pm 1)$.*

The formulation of Algorithm 2 (LARS standard formulation), Algorithm 3 (LARS projected formulation) and the proof of Theorem 5.1 are given in Appendices 5 and 6.

ALGORITHM 4
LARS algorithm ("recursive" formulation)

```
/* Given a response $Y$ and a design $X$, set $\overline{Z}=X^\top Y$ and $\overline{R}=X^\top X$ */
Data: Correlations vector $\overline{Z}$ and variance-covariance matrix $\overline{R}$.
Result: Sequence $((\lambda_k, \bar{\imath}_k, \varepsilon_k))_{k \geqslant 1}$ where $\lambda_1 \geqslant \lambda_2 \geqslant \ldots > 0$ are the knots, and $\bar{\imath}_1, \bar{\imath}_2, \ldots$ 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
        $\lambda = \max_{\{j : T_j < 1\}} \left\{ \frac{Z_j}{1 - T_j} \right\}$,   $\mathbf{i} = \arg \max_{\{j : T_j < 1\}} \left\{ \frac{Z_j}{1 - T_j} \right\}$,   $\mathbf{x} = \frac{R_{\mathbf{i}}}{R_{\mathbf{ii}}}$.
    Update
        $R \leftarrow R - \mathbf{x} R_{\mathbf{i}}^\top$,   $Z \leftarrow Z - \mathbf{x} Z_{\mathbf{i}}$,   $T \leftarrow T + \mathbf{x}(1 - T_{\mathbf{i}})$.
    return ($R$, $Z$, $T$, $\lambda$, $\mathbf{i}$)
Set $k = 0$, $T = 0$, $Z = (\overline{Z}, -\overline{Z})$ and $R = \begin{bmatrix} \overline{R} & -\overline{R} \\ -\overline{R} & \overline{R} \end{bmatrix}$.
Update $k \leftarrow k + 1$ and compute
    $(R, Z, T, \lambda_k, \widehat{\imath}_k) = \mathrm{Rec}\,(R, Z, T)$
Set $\bar{\imath}_k = \widehat{\imath}_k \mod p$ and $\varepsilon_k = 1 - 2(\widehat{\imath}_k - \bar{\imath}_k)/p \in \pm 1$.
```

5.3 *Initialization: First Knot*

The first step of the LARS 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 $N^{(1)} := (\overline{N}^{(1)}, -\overline{N}^{(1)})$. We define the first knot $\lambda_1 > 0$ as

$$\lambda_1 = \max Z \quad \text{and} \quad \widehat{\imath}_1 = \arg\max Z\,.$$

One may see that this definition is consistent with λ_1 in Algorithm 2 and note that $\widehat{\imath}_1$ and $(\bar{\imath}_1, \varepsilon_1)$ are related as in (2.1).

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

$$N^{(2)}(\lambda) = N^{(1)} - (\lambda_1 - \lambda)\theta(\widehat{\imath}_1)\,, \quad 0 < \lambda \leqslant \lambda_1\,, \tag{5.1}$$

and one can check that $N^{(2)}(\lambda) = (\overline{N}^{(2)}(\lambda), -\overline{N}^{(2)}(\lambda))$ where $\overline{N}(\lambda)$ is defined in Algorithm 2. It is clear that the coordinate $\widehat{\imath}_1$ of $N^{(2)}(\lambda)$ is equal to λ. On the other hand $N^{(1)} = Z$ attains its maximum at the single point $\widehat{\imath}_1$. By continuity this last property is kept for λ in a left neighborhood of λ_1. We search for the first value of λ such that this property is not met, *i.e.* the largest value of λ such that

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

as in Step **3** of Algorithm 2. We call this value λ_2 and one may check that this definition is consistent with λ_2 in Algorithm 2.

Now, we can be more explicit about the expression of λ_2. Indeed, we make the following discussion on the values of $\theta_j(\widehat{\imath}_1)$.

- If $\theta_j(\widehat{\imath}_1) \geqslant 1$, since $N_j^{(1)} < N_{\widehat{\imath}_1}^{(1)}$ for $j \neq \widehat{\imath}_1$ there is no hope to achieve the equality between $N_j^{(2)}(\lambda)$ and $N_{\widehat{\imath}_1}^{(2)}(\lambda) = \lambda$ for $0 < \lambda \leqslant \lambda_1$ in view of (5.1).

- Thus we limit our attention to the j's such that $\theta_j(\widehat{\imath}_1) < 1$. We have equality $N_j^{(2)}(\lambda) = \lambda$ when

$$\lambda = \frac{N_j^{(1)} - \lambda_1 \theta_j(\widehat{\imath}_1)}{1 - \theta_j(\widehat{\imath}_1)}\,.$$

So we can also define the second knot λ_2 of the LARS as

$$\lambda_2 = \max_{j:\theta_j(\widehat{\imath}_1)<1} \left\{ \frac{Z_j - \Pi_{\widehat{\imath}_1}(Z_j)}{1 - \theta_j(\widehat{\imath}_1)} \right\}.$$

where $\Pi_{i_1}(Z_j) := Z_{i_1}\theta_j(i_1)$. Remark that $\Pi_{i_1}(Z_j) = \mathbb{E}(Z_j \mid Z_{i_1})$ is the regression of Z_j on Z_{i_1} when $\mathbb{E}Z = 0$.

5.4 *Recursion: Next Knots*

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

Assume that $k \geqslant 2$ and we have build $\lambda_1, \ldots, \lambda_{k-1}$ and selected the "signed" variables $\widehat{\imath}_1, \ldots, \widehat{\imath}_{k-1}$. Introduce $N^{(k-1)} := (\overline{N}^{(k-1)}, -\overline{N}^{(k-1)})$ and define

$$N^{(k)}(\lambda) = N^{(k-1)} - (\lambda_{k-1} - \lambda)\theta(\widehat{\imath}_1, \ldots, \widehat{\imath}_{k-1})\,, \quad 0 < \lambda \leqslant \lambda_{k-1}\,.$$

Check that $\theta_j(\widehat{\imath}_1, \ldots, \widehat{\imath}_{k-1}) = (\overline{\theta}_j, -\overline{\theta}_j)$ where we recall that we define

$$\overline{\theta}_j := \left(\overline{R}_{j,\bar{\imath}_1} \cdots \overline{R}_{j,\bar{\imath}_{k-1}}\right) M^{-1}_{\bar{\imath}_1,\ldots,\bar{\imath}_{k-1}}(\varepsilon_1, \ldots, \varepsilon_{k-1})\,, \quad j = 1, \ldots, p\,,$$

at Step **2** and it holds that $\widehat{\imath}_\ell$ and $(\bar{\imath}_\ell, \varepsilon_\ell)$ are related as in (2.1). From this equality, we deduce that it holds $N^{(k)}(\lambda) = (\overline{N}^{(k)}(\lambda), -\overline{N}^{(k)}(\lambda))$. One may also check that the coordinates $\widehat{\imath}_1, \ldots, \widehat{\imath}_{k-1}$ of $N^{(k)}(\lambda)$ are equal to λ.

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 $\theta_j(\widehat{\imath}_1, \ldots, \widehat{\imath}_{k-1}) < 1$. Solving this latter equality yields to

$$\lambda_k = \max_{j:\theta_j(\widehat{\imath}_1,\ldots,\widehat{\imath}_{k-1})<1} \left\{ \frac{N_j^{(k-1)} - \lambda_{k-1}\theta_j(\widehat{\imath}_1, \ldots, \widehat{\imath}_{k-1})}{1 - \theta_j(\widehat{\imath}_1, \ldots, \widehat{\imath}_{k-1})} \right\}.$$

This expression is consistent with λ_k in Algorithm 2.

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

$$N^{(k)} = N^{(k-1)} - (\lambda_{k-1} - \lambda_k)\theta(\widehat{\imath}_1, \dots, \widehat{\imath}_{k-1}), \tag{5.2}$$

and that $N_j^{(k-1)} = \lambda_{k-1}$ for $j = \widehat{\imath}_1, \dots, \widehat{\imath}_{k-1}$. The following lemma permits a drastic simplification of the expression of the knots. Its proof is given in Appendix 7.6.

LEMMA 5.1 It holds

$$N^{(k-1)} - \lambda_{k-1}\theta(\widehat{\imath}_1, \dots, \widehat{\imath}_{k-1}) = Z - \varPi_{\widehat{\imath}_1,\dots,\widehat{\imath}_{k-1}}(Z)$$

where we denote $\varPi_{i_1,\dots,i_{k-1}}(Z) = (\varPi_{i_1,\dots,i_{k-1}}(Z_1), \dots, \varPi_{i_1,\dots,i_{k-1}}(Z_{2p}))$ and note that, for all $j \in [2p]$, one has $\varPi_{i_1,\dots,i_{k-1}}(Z_j) = \big(R_{j,i_1} \cdots R_{j,i_{k-1}}\big) M^{-1}_{i_1,\dots,i_{k-1}}(Z_{i_1}, \dots, Z_{i_{k-1}})$.

Using Lemma 5.1 we deduce that λ_k in Algorithm 2 is consistent with

$$\lambda_k = \max_{j:\theta_j(\widehat{\imath}_1,\dots,\widehat{\imath}_{k-1})<1} \left\{ \frac{Z_j - \varPi_{\widehat{\imath}_1,\dots,\widehat{\imath}_{k-1}}(Z_j)}{1 - \theta_j(\widehat{\imath}_1, \dots, \widehat{\imath}_{k-1})} \right\}. \tag{5.3}$$

where $\varPi_{\widehat{\imath}_1,\dots,\widehat{\imath}_{k-1}}(Z_j) = \big(R_{j,\widehat{\imath}_1} \cdots R_{j,\widehat{\imath}_{k-1}}\big) M^{-1}_{\widehat{\imath}_1,\dots,\widehat{\imath}_{k-1}}(Z_{\widehat{\imath}_1}, \dots, Z_{\widehat{\imath}_{k-1}})$. When $\mathbb{E}Z = 0$, one may remark that $\varPi_{i_1,\dots,i_{k-1}}(Z_j)$ is the regression of Z_j on the vector $(Z_{i_1}, \cdots, Z_{i_{k-1}})$ whose variance-covariance matrix is $M_{i_1,\dots,i_{k-1}}$. This analysis leads to an equivalent formulation of the LARS algorithm (Algorithm 2). We present this formulation in Algorithm 3.

REMARK 5.1 *Note that Algorithm 2 implies that $\widehat{\imath}_1, \dots, \widehat{\imath}_k$ are pairwise different, but also that they differ modulo p.*

6. First Steps to Derive the Joint Law of the LARS knots

6.1 *Law of the First Knot*

One has the following lemma governing the law of λ_1.

LEMMA 6.1 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 $\{\widehat{\imath}_1 = i_1\} = \{\lambda_2^{(i_1)} \leqslant Z_{i_1}\}$,
- If $\theta_j(i_1) < 1$ for all $j \neq i_1$ then, conditional on $\{\widehat{\imath}_1 = i_1\}$ and λ_2, λ_1 is a truncated Gaussian random variable with mean $\mathbb{E}(Z_{i_1})$ and variance $\rho_1^2 := R_{\widehat{\imath}_1,\widehat{\imath}_1}$ subject to be greater than λ_2.

Proof. The first point is a consequence or the properties of Gaussian regression. Now, observe that

$$\begin{aligned}
\{\lambda_2^{(i_1)} \leqslant Z_{i_1}\} &\Leftrightarrow \Big\{\forall j \neq i_1\,,\ \frac{Z_j - Z_{i_1}\theta_j(i_1)}{1 - \theta_j(i_1)} \leqslant Z_{i_1}\Big\} \\
&\Leftrightarrow \{\forall j \neq i_1\,,\ Z_j - Z_{i_1}\theta_j(i_1) \leqslant Z_{i_1} - Z_{i_1}\theta_j(i_1)\} \\
&\Leftrightarrow \{\forall j \neq i_1\,,\ Z_j \leqslant Z_{i_1}\} \\
&\Leftrightarrow \{\widehat{\imath}_1 = i_1\}\,,
\end{aligned}$$

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

6.2 *Recursive Formulation of the LARS*

One has the following proposition whose proof can be found in Section 7.7. As we will see in this section, this intermediate result as a deep consequence, the LARS algorithm can be stated in a recursive way applying the same function repeatedly, as presented in Algorithm 4.

PROPOSITION 6.1 Set

$$\tau_{j,i_k} := \frac{R_{j,i_k} - \big(R_{j,i_1} \cdots R_{j,i_{k-1}}\big) M^{-1}_{i_1,\ldots,i_{k-1}} \big(R_{i_k,i_1}, \cdots, R_{i_k,i_{k-1}}\big)}{(1-\theta_j(i_1,\ldots,i_{k-1}))(1-\theta_{i_k}(i_1,\ldots,i_{k-1}))},$$

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

$$\frac{\tau_{j,i_k}}{\tau_{i_k,i_k}} = 1 - \frac{1-\theta_j(i_1,\ldots,i_k)}{1-\theta_j(i_1,\ldots,i_{k-1})} \tag{6.1}$$

and

$$\forall j \neq i_1,\ldots,i_k, \quad Z_j^{(i_1,\ldots,i_k)} = \frac{Z_j^{(i_1,\ldots,i_{k-1})} - Z_{i_k}^{(i_1,\ldots,i_{k-1})} \tau_{j,i_k}/\tau_{i_k,i_k}}{1-\tau_{j,i_k}/\tau_{i_k,i_k}}. \tag{6.2}$$

Now, we present Algorithm 4. Define $R(0) := R$, $Z(0) = Z$ and $T(0) = 0$. For $k \geqslant 1$ and $i_1,\ldots,i_k \in [2p]$, introduce

$$\begin{aligned} R(k) &:= \Big(R_{j,\ell} - \big(R_{j,i_1} \cdots R_{j,i_k}\big) M^{-1}_{i_1,\ldots,i_k} \big(R_{\ell,i_1}, \cdots, R_{\ell,i_k}\big)\Big)_{j,\ell} \\ Z(k) &:= Z - \Pi_{i_1,\ldots,i_k}(Z) \\ T(k) &:= (\theta_j(i_1,\ldots,i_k))_j, \end{aligned}$$

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

$$v := P^{\perp}_{(v_1,\ldots,v_{k-1})} v_k,$$

the projection of v_k orthogonally to $v_1,\ldots,v_{k-1}$.Then

$$P^{\perp}_{(v_1,\ldots,v_k)} u = P^{\perp}_v P_{(v_1,\ldots,v_{k-1})} u.$$

Using this result we deduce that

$$\begin{aligned} Z(k) &= \Pi^{\perp}_{i_1,\ldots,i_k}(Z) \\ &= \Pi^{\perp}_{i_k}(\Pi^{\perp}_{i_1,\ldots,i_{k-1}}(Z)) \\ &= \Pi^{\perp}_{i_k}(Z(k-1)) \\ &= Z(k-1) - \Pi_{i_k}(Z(k-1)) \\ &= Z(k-1) - \mathbf{x}(k) Z(k-1), \end{aligned} \tag{6.3}$$

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

$$R(k) = R(k-1) - \mathbf{x}(k) R_{i_k}(k-1)^{\top}. \tag{6.4}$$

Using (6.1) (or (7.10)), remark that

$$T(k) = T(k-1) - \mathbf{x}(k)(1 - T_{i_k}(k-1)). \tag{6.5}$$

These relations give a recursive formulation of the LARS as presented in Algorithm 4.

7. Proofs

7.1 *Proof of Proposition 3.2*

First and third points: The first point works by induction. The initialization of the proof is given by the second point of Lemma 6.1. We will use Proposition 6.1 to prove the first point. We have

$$
\begin{aligned}
&\lambda_{k+1}^{(i_1,\dots,i_k)} \leqslant Z_{i_k}^{(i_1,\dots,i_{k-1})}\\
&\Leftrightarrow \forall j \neq i_1,\dots,i_k\,,\ Z_j^{(i_1,\dots,i_k)} \leqslant Z_{i_k}^{(i_1,\dots,i_{k-1})}\\
&\Leftrightarrow \forall j \neq i_1,\dots,i_k\,,\ Z_j^{(i_1,\dots,i_{k-1})} - Z_{i_k}^{(i_1,\dots,i_{k-1})}\,\frac{\tau_{j,i_k}}{\tau_{i_k,i_k}} \leqslant Z_{i_k}^{(i_1,\dots,i_{k-1})} - Z_{i_k}^{(i_1,\dots,i_{k-1})}\,\frac{\tau_{j,i_k}}{\tau_{i_k,i_k}}\\
&\Leftrightarrow \forall j \neq i_1,\dots,i_k\,,\ Z_j^{(i_1,\dots,i_{k-1})} \leqslant Z_{i_k}^{(i_1,\dots,i_{k-1})} \qquad\qquad (7.1)\\
&\Leftrightarrow \lambda_k^{(i_1,\dots,i_{k-1})} = Z_{i_k}^{(i_1,\dots,i_{k-1})}\,.
\end{aligned}
$$

using (6.2) and that $1-\tau_{j,i_k}/\tau_{i_k,i_k} > 0$ (which is a consequence of (6.1) and ($\mathscr{A}_{\mathrm{Irr.}}$)) in (7.1). By induction and using (7.1), it holds that

$$
\begin{aligned}
&\big\{\lambda_{k+1}^{(i_1,\dots,i_k)} \leqslant Z_{i_k}^{(i_1,\dots,i_{k-1})} \leqslant Z_{i_{k-1}}^{(i_1,\dots,i_{k-2})} \leqslant \ldots \leqslant Z_{i_2}^{(i_1)} \leqslant Z_{i_1}\big\}\\
&\quad\Leftrightarrow \big\{\forall j \neq i_1,\dots,i_k\,,\ Z_j^{(i_1,\dots,i_{k-1})} \leqslant Z_{i_k}^{(i_1,\dots,i_{k-1})} \leqslant Z_{i_{k-1}}^{(i_1,\dots,i_{k-2})} \leqslant \ldots \leqslant Z_{i_2}^{(i_1)} \leqslant Z_{i_1}\big\}\\
&\quad\Leftrightarrow \big\{\forall j \neq i_1,\dots,i_{k-1}\,,\ Z_j^{(i_1,\dots,i_{k-1})} \leqslant Z_{i_{k-1}}^{(i_1,\dots,i_{k-2})} \leqslant \ldots \leqslant Z_{i_2}^{(i_1)} \leqslant Z_{i_1}\\
&\qquad\quad \text{and } \forall j \neq i_1,\dots,i_k\,,\ Z_j^{(i_1,\dots,i_{k-1})} \leqslant Z_{i_k}^{(i_1,\dots,i_{k-1})}\big\}\\
&\quad\Leftrightarrow \big\{\lambda_k^{(i_1,\dots,i_{k-1})} \leqslant Z_{i_{k-1}}^{(i_1,\dots,i_{k-2})} \leqslant \ldots \leqslant Z_{i_2}^{(i_1)} \leqslant Z_{i_1} \text{ and } \lambda_k^{(i_1,\dots,i_{k-1})} = Z_{i_k}^{(i_1,\dots,i_{k-1})}\big\}\\
&\quad\ \vdots\\
&\quad\Leftrightarrow \big\{\lambda_a^{(i_1,\dots,i_{a-1})} \leqslant Z_{i_{a-1}}^{(i_1,\dots,i_{a-2})} \leqslant \ldots \leqslant Z_{i_2}^{(i_1)} \leqslant Z_{i_1}\\
&\qquad\quad \text{and } \lambda_k^{(i_1,\dots,i_{k-1})} = Z_{i_k}^{(i_1,\dots,i_{k-1})} \leqslant \ldots \leqslant \lambda_a^{(i_1,\dots,i_{a-1})} = Z_{i_a}^{(i_1,\dots,i_{a-1})}\big\} \qquad\qquad (s_a)\\
&\quad\ \vdots\\
&\quad\Leftrightarrow \big\{\widehat{\imath}_1 = i_1,\dots,\widehat{\imath}_k = i_k\big\}\,.
\end{aligned}
$$

Now, observe that $\widehat{i}_{k+1}$ is the (unique) arg max of $\lambda_{k+1}^{(i_1,\dots,i_k)}$ on the event $\big\{\widehat{\imath}_1 = i_1,\dots,\widehat{\imath}_k = i_k\big\}$.
It yields that

$$
\big\{\widehat{\imath}_1 = i_1,\dots,\widehat{\imath}_{k+1} = i_{k+1}\big\} = \big\{\lambda_{k+1}^{(i_1,\dots,i_k)} = Z_{i_{k+1}}^{(i_1,\cdots,i_k)} \leqslant Z_{i_k}^{(i_1,\cdots,i_{k-1})} \leqslant \cdots \leqslant Z_{i_2}^{(i_1)} \leqslant Z_{i_1}\big\}\,,
$$

as claimed. Stopping at a as in (s_a) gives the second part of the statement. The third point of the proposition is a direct consequence of the first point.

Second point: The proof of the second point can be lead by induction. The initialization of the proof is given by the first point of Lemma 6.1. Now, observe that $Z_{i_k}^{(i_1,\dots,i_{k-1})},\cdots,Z_{i_2}^{(i_1)},Z_{i_1}$ are measurable functions of $(Z_{i_1},\dots,Z_{i_k})$ and one may check that $(Z_{i_1},\dots,Z_{i_k})$ is independent of $(Z_j^{(i_1,\dots,i_k)})_{j\neq i_1,\dots,i_k}$.

We deduce that $Z_{i_k}^{(i_1,\dots,i_{k-1})},\cdots,Z_{i_2}^{(i_1)},Z_{i_1}$ are independent of $(Z_j^{(i_1,\dots,i_k)})_{j\neq i_1,\dots,i_k}$. One can also check that $Z_{i_{k+1}}^{(i_1,\dots,i_k)}$ is independent of $(Z_j^{(i_1,\dots,i_k)})_{j\neq i_1,\dots,i_{k+1}}$. Deduce that the variables $Z_{i_K}^{(i_1,\dots,i_{K-1})},\cdots,Z_{i_2}^{(i_1)},Z_{i_1}$ are independent and independent of $(Z_j^{(i_1,\dots,i_K)})_{j\neq i_1,\dots,i_K}$. This gives the first part of the second point of the proposition.

Now observe that $(Z_j^{(i_1,\dots,i_K)})_{j\neq i_1,\dots,i_K}$ is a linear function of $P_K^\perp Y$. More precisely, consider the vector $\widetilde{W} := (Z_j^{(i_1,\dots i_K)})_{j\in[p]}$ and note that

$$(\widehat{\sigma}^{i_1,\dots,i_K})^2 = \frac{\|P_K^\perp(Y)\|_2^2}{n-K} \quad \text{and} \quad \widetilde{W} = \operatorname{Diag}(\mathbb{1}_p - \theta^K)^{-1} \times X^\top \big(P_K^\perp(Y)\big)\,,$$

where $\theta^K := (\theta_j(i_1,\dots,i_K))_{j\in[p]}$, $P_K^\perp = \mathrm{Id}_n - P^{(i_1,\dots,i_K)}$, and with the convention $0/0=0$. The key remark is that the direction and the norm of a centered Gaussian vector are independent, namely $\|P_K^\perp(Y)\|$ is independent of $P_K^\perp(Y)/\|P_K^\perp(Y)\|$ when $\mathbb{E}P_K^\perp(Y)=0$. Observe that $\mathbb{E}P_K^\perp(Y) = P_K^\perp(X\beta^0)$ and $\mathbb{E}P_K^\perp(Y)=0$ is equivalent to $X\beta^0 \in H_K$. We deduce that if $X\beta^0\in H_K$ then $\widehat{\sigma}^{i_1,\dots,i_K}$ is independent of $\widetilde{W}/\sigma^{i_1,\dots,i_K}$, as claimed.

7.2 *Proof of Proposition 3.4*

Fix a such that $0\leqslant a\leqslant K-1$ and consider any selection procedure $\widehat{m}$ satisfying ($\mathscr{A}_{\text{Stop}}$). From Theorem 1.1 (more precisely (3.15)) we know that the density of $(\lambda_{a+1},\lambda_{a+2},\dots,\lambda_K)$ conditional on

$$\mathscr{F} := \big\{\widehat{\imath}_1 = i_1,\dots,\widehat{\imath}_K = i_K, \lambda_a, \lambda_{K+1}\big\}$$

is given by

$$(const)\Big(\prod_{k=a+1}^{K} \varphi_{m_k,v_k^2}(\ell_k)\Big)\,\mathbb{1}_{\lambda_a\leqslant\ell_{a+1}\leqslant\cdots\leqslant\ell_K\leqslant\lambda_{K+1}}\,.$$

From Proposition 3.1, conditional on $\mathscr{F}$ and under the null hypothesis described in Proposition 3.4, we know that $m_{a+1}=\ldots=m_K=0$. It implies that Φ_k is the CDF of λ_k for $a<k\leqslant K$.

From the definition of a stopping time given by ($\mathscr{A}_{\text{Stop}}$) and on the event $\mathscr{F}$, we know that $\mathbb{1}_{\{\widehat{m}=a\}}$ is a measurable function of $\lambda_1,\dots,\lambda_{a-1}$ which are respectively equal to $Z_{i_1},\dots,Z_{i_{a-1}}^{(i_1,\dots,i_{a-2})}$ on $\mathscr{F}$ by (3.5) (as proven in Appendix 5.4 and Eq. (5.3)). By Proposition 3.2 (more precisely (3.10)), we also know that this function is independent of $(\lambda_{a+1},\lambda_{a+2},\dots,\lambda_K)$ conditional on $\mathscr{F}$. Remark that its is also independent of $\widehat{\sigma}^{i_1,\dots,i_K}$ conditional of $\mathscr{F}$ for the same reason (it would be useful later, when we will build testing procedures when the variance is unknown). We deduce that the conditional density above is also the conditional density on the event

$$\mathscr{G} := \big\{\widehat{m}=a, \widehat{\imath}_1 = i_1,\dots,\widehat{\imath}_K = i_K, \lambda_a, \lambda_{K+1}\big\}\,.$$

From $F_k = \Phi_k(\lambda_k)$ (*i.e.,* applying the CDF) we deduce by a change of variables that conditional on the selection event $\mathscr{G}$, the vector $(F_{a+1},\dots,F_K)$ is uniformly distributed on

$$\begin{aligned}\mathscr{D}_{a+1,K} :=\big\{&(f_{a+1},\dots,f_K)\in\mathbb{R}^{K-a} :\\ &\mathscr{P}_{a+1,a}(F_a)\geqslant f_{a+1}\geqslant \mathscr{P}_{a+1,a+2}(f_{a+2})\geqslant\cdots\geqslant \mathscr{P}_{a+1,K}(f_K)\geqslant \mathscr{P}_{a+1,K+1}(F_{K+1})\big\}\,,\end{aligned}$$

where $\mathscr{P}_{i,j}$ are described in (3.16).

7.3 *Proof of Proposition 3.5*

By Proposition 3.4, a simple integration shows that

$$\mathbb{P}\big[\lambda_b \leqslant t \mid \widehat{m} = a, \lambda_a, \lambda_c, \widehat{\imath}_1, \dots, \widehat{\imath}_a, \widehat{\imath}_{a+1}, \dots, \widehat{\imath}_{c-1}, \widehat{\imath}_c, \dots, \widehat{\imath}_K\big] = \frac{\mathbb{F}_{abc}(t)}{\mathbb{F}_{abc}(\lambda_a)}\,,$$

under the null hypothesis of Proposition 3.5 (which implies that $m_{a+1} = \ldots = m_K = 0$). Then note that the function $\mathbb{F}_{abc}$ is defined by $\sigma, \lambda_a, \lambda_c, \widehat{\imath}_{a+1}, \dots, \widehat{\imath}_{c-1}$ only. We deduce that we can de-condition on $\widehat{m} = a, \widehat{\imath}_1, \dots, \widehat{\imath}_a, \widehat{\imath}_c, \dots, \widehat{\imath}_K$, which gives the result.

7.4 *Proof of Proposition 3.6*

Let us fix some values $i_1, \dots, i_{K+1}$. Recall that the frozen values of the knots

$$\lambda_1^f := Z_{i_1}, \dots, \lambda_K^f := Z_{i_K}^{(i_1,\dots i_{K-1})}, \lambda_{K+1}^f := Z_{i_{K+1}}^{(i_1,\dots i_K)},$$

are Gaussian, independent, and $\widehat{\sigma}^{i_1,\dots,i_K} \perp\!\!\!\perp \big(Z_{i_{K+1}}^{(i_1,\dots,i_K)}/\widehat{\sigma}^{i_1,\dots,i_K}\big) \perp\!\!\!\perp Z_{i_K}^{(i_1,\dots,i_{K-1})} \perp\!\!\!\perp \cdots \perp\!\!\!\perp Z_{i_2}^{(i_1)} \perp\!\!\!\perp Z_{i_1}$, see Proposition 3.2 and (3.10). Let us condition by $\{\widehat{\imath}_1 = i_1, \dots, \widehat{\imath}_K = i_K, \lambda_{K+1} = \ell_{K+1}\}$. Note that, on this event, $\lambda_h^f = \lambda_h$, $h \in [K]$ and by Proposition 3.2 this event is equivalent to

$$\{Z_{i_1} > \cdots > Z_{i_K}^{(i_1,\dots i_{K-1})} > \max_j Z_j^{(i_1,\dots i_K)} = \lambda_{K+1} = \ell_{K+1}\}. \tag{7.2}$$

Because of the independence above, we get the conditional independence stated in the first and second point of the proposition.

For the last point, consider $\widetilde{W} := (Z_j^{(i_1,\dots i_K)})_{j\in[p]}$ and note that

$$(\widehat{\sigma}^{i_1,\dots,i_K})^2 = \frac{\|P_K^\perp(Y)\|_2^2}{n-K} \quad \text{and} \quad \widetilde{W} = \mathrm{Diag}(\mathbb{1}_p - \theta^K)^{-1} \times X^\top\big(P_K^\perp(Y)\big)\,,$$

where $\theta^K := (\theta_j(i_1, \dots, i_K))_{j\in[p]}$, $P_K^\perp = \mathrm{Id}_n - P^{(i_1,\dots,i_K)}$, and with the convention $0/0 = 0$. Now, let $U \in \mathbb{R}^{n\times(n-K)}$ be any matrix such that $UU^\top = P_K^\perp$ and define $W := U^\top Y/\sigma$, then

$$\Big(\frac{\widehat{\sigma}^{i_1,\dots,i_K}}{\sigma}\Big)^2 = \frac{\|W\|_2^2}{n-K} \quad \text{and} \quad \widetilde{W} = \sigma \times \mathrm{Diag}(\mathbb{1}_p - \theta^K)^{-1} \times X^\top U W\,.$$

Because of the independence above and (7.2), the distribution of $\widehat{\sigma}$ is independent of the other variables and such that

$$\Big(\frac{\widehat{\sigma}}{\sigma}\Big)^2 = \frac{\|W\|_2^2}{n-K} \quad \text{with } W \text{ s.t. } \|\mathrm{Diag}(\mathbb{1}_p - \theta^K)^{-1} \times X^\top U W\|_\infty = \ell_{K+1}/\sigma\,.$$

This implies that the conditional distribution is the one claimed.

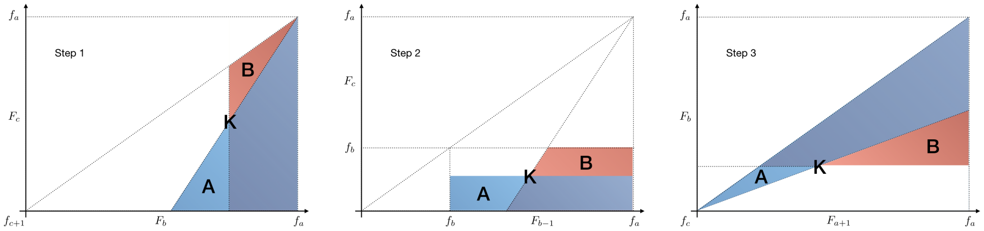


Figure 8: Rejection domains associated to the different comparison sets appearing in steps of the proof of Theorem 1.4.

7.5 *Orthogonal Case: Proof of Theorem 1.4*

Let $\mathscr{I}$ the set of admissible indexes

$$\mathscr{I} := \{a,b,c : a_0 \leqslant a < b < c \leqslant K+1\}\,.$$

○ **Step 1:** We prove that, when the considered indexes such that $c+1 \leqslant K+1$ belong to $\mathscr{I}$, $\mathscr{S}_{a,b,c+1}$ is more powerful than $\mathscr{S}_{a,b,c}$. Our proof is conditional to $F_a = f_a, F_{c+1} = f_{c+1}$. Note that $(F_{a+1},\dots,F_c)$ has for distribution the uniform distribution on the simplex

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

This implies by direct calculations that

$$\mathscr{I}_{ab}(s,t) = \frac{(s-t)^{b-a-1}}{(b-a-1)!}$$

and that

$$\frac{\mathbb{F}_{abc}(\lambda_b)}{\mathbb{F}_{abc}(\lambda_a)} = \boldsymbol{F}_{\beta((b-a),(c-b))}\left(\frac{F_b - F_c}{f_a - F_c}\right). \tag{7.3}$$

where $\boldsymbol{F}_\beta$ is the cumulative distribution of the Beta distribution in reference. Using monotony of this function the $\mathscr{S}_{abc}$ test has for rejection region

$$(F_b - F_c) \geqslant z_1(f_a - F_c) \Leftrightarrow F_b \geqslant z_1 f_a + (1-z_1)F_c, \tag{7.4}$$

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

Similarly $\mathscr{S}_{ab(c+1)}$ has for rejection region

$$F_b \geqslant z_2(f_a - f_{c+1}) + f_{c+1}, \tag{7.5}$$

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

LEMMA 7.1 Let $c \leqslant K$. The density h_μ of $f_1,\dots,f_c$, conditional on F_{c+1} with respect of the Lebesgue measure under the alternative is coordinate-wise non-decreasing and given by (7.6).

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

$$\lambda_{c+1}^{i_1,\dots,i_c} = \max_{j\in[p]\,,\ j\neq\bar{\imath}_1,\dots,j\neq\bar{\imath}_c} |Z_j|.$$

Thus its density $p_{\mu^0,i_1,\dots,i_c}$ does not depend on $\mu^0_{\bar{i}_1},\dots,\mu^0_{\bar{i}_c}$. As a consequence the following variables have the same distribution ; $\lambda_{c+1}^{i_1+\varepsilon_1 p,\dots,i_c+\varepsilon_c p}$, where $\varepsilon_1,\dots,\varepsilon_c$ take the value 0 or 1 and indices are taken modulo p.

Because of the independence of the different variables, the joint density, under the alternative hypothesis, of $\lambda_1,\dots,\lambda_{c+1}$ taken at $\ell_1,\dots,\ell_{c+1}$, on the domain $\{\lambda_1 > \cdots > \lambda_{c+1}\}$ takes the value

$$(Const)\sum{}^{'} \big(\varphi(\ell_1-\mu^0_{j_1})+\varphi(\ell_1+\mu^0_{j_1})\big),\dots,\big(\varphi(\ell_c-\mu^0_{j_c})+\varphi(\ell_c+\mu^0_{j_c})\big)p_{\mu^0,j_1,\dots,j_K}(\ell_{k+1}).$$

Here the sum $\sum{}^{'}$ is taken over all different $j_1,\dots,j_c$ belonging to $[\![1,p]\!]$.

Then the density, conditional on $F_{c+1}=f_{c+1}$, of $F_1,\dots,F_c$ at $f_1,\dots,f_c$ takes the value

$$(\mathrm{const})\sum{}^{'}\cosh(\mu_{j_1}f_1)\dots\cosh(\mu_{j_c}f_c)\mathbb{1}_{f_1>\dots>f_c>F_{c+1}}, \tag{7.6}$$

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

LEMMA 7.2 Let ν_0 the image on the plane (F_b,F_c) on the uniform probability on $\mathscr{S}$: it is the distribution under the null of (F_b,F_c). The two rejection regions : $\mathscr{R}_1$ associated to (7.4) and $\mathscr{R}_2$ associated to (7.5) have of course the same probability α under ν_0. Let η_{μ^0} the density w.r.t. ν_0 of the distribution of (F_b,F_c) under the alternative. Then η_{μ^0} is non decreasing coordinate-wise.

Proof. Integration yields that density of ν_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 ν_{μ^0} w.r.t. Lebesgue measure is

$$\int_{f_b}^{f_a} df_{a+1}\dots\int_{f_b}^{f_b-2} df_{a+1}\int_{f_b}^{f_b-2} df_{b-1}\int_{f_c}^{f_b} df_{b+1}\dots\int_{f_c}^{f_{c-2}} df_{c-1}h_{\mu^0}(f_a,\dots,f_c). \tag{7.7}$$

Thus η_{μ^0} which is the quotient of these two quantities is just a mean value of h_{μ^0} on the domain of integration $\mathscr{D}_{f_b,f_c}$ in (7.7).

Suppose that f_b and f_c increase, then all the borns of the domain $\mathscr{D}_{f_b,f_c}$ increase also. By Lemma 7.1 the mean value increases. □

We finish now the proof of Step 1: For a given level α 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

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

see Figure 8. These two regions have the same ν_0 measure. By elementary geometry there exist a point $K=(K_b,K_c)$ in the plane such that

- For every point of A, $F_b \leqslant K_b$, $F_c \leqslant K_c$,

- For every point of B, $F_b \geqslant K_b$, $F_c \geqslant K_c$,

By transport of measure there exists a transport function $\mathscr{T}$ that preserve the measure ν_0 and that is one-to one $A \to B$. As a consequence the transport by $\mathscr{T}$ improve the probability under the alternative: the power of $\mathscr{S}_{a,b,c+1}$ is larger than that of $\mathscr{S}_{a,b,c}$.

○ **Step 2:** We prove that, when the considered indexes belong to $\mathscr{I}$ such that $a < b-1$, $\mathscr{S}_{a,(b-1),c}$ is more powerful than $\mathscr{S}_{a,b,c}$. Our proof is conditional on $F_a = f_a, F_b = f_b$ and is located in the plane (F_{b-1}, F_c).
The rejection region $R_{a,b,c}$ takes the form $F_c \leqslant \frac{1}{1-z_1} f_b - \frac{z_1}{1-z_1} f_a$ for some threshold z_1 belonging to $(0,1)$.
The rejection region $R_{a,(b-1),c}$ takes the form $F_c \leqslant \frac{1}{1-z_2} F_{b-1} - \frac{z_2}{1-z_2} f_a$ 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 8.

Transport of measure and the convenient modification of Lemma 7.2 imply that the power of the test $\mathscr{S}_{a,(b-1),c}$ is greater of equal than that of $\mathscr{S}_{a,b,c}$.

○ **Step 3:** We prove that, when the considered indexes belong to $\mathscr{I}$ such that $a+1 < b$, $\mathscr{S}_{a,b,c}$ is more powerful than $\mathscr{S}_{(a+1),b,c}$. Our proof is conditional on $F_a = f_a, F_c = f_c$ and is located in the plane F_{a+1}, F_b.
The rejection region $R_{a,b,c}$ takes the form $F_b \geqslant z_1 f_a + (1-z_1) f_c$ for some threshold $z_1$belonging to $(0,1)$.
The rejection region $R_{a+1,b,c}$ takes the form $F_b \geqslant z_2 F_{a+1} + (1-z_2) f_c$ 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 8.

Transport of measure and the convenient modification of Lemma 7.2 imply that the power of $\mathscr{S}_{a,b,c}$ is greater of equal that that of $\mathscr{S}_{(a+1),b,c}$.

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

7.6 *Proof of Lemma 5.1*

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

$$N^{(1)} - \lambda_1 \theta(\widehat{\imath}_1) = Z - Z_{\widehat{\imath}_1} \theta(\widehat{\imath}_1) = Z - \Pi_{\widehat{\imath}_1}(Z) .$$

Now, let $k \geqslant 3$. First, the three perpendicular theorem implies that for every $j, i_1, \dots, i_{k-1}$,

$$\begin{aligned} \theta_j(i_1,\dots,i_{k-2}) &= \big(R_{j,i_1} \cdots R_{j,i_{k-1}}\big) M^{-1}_{i_1,\dots,i_{k-1}} (\theta_{i_1}(i_1,\dots,i_{k-2}),\dots,\theta_{i_{k-1}}(i_1,\dots,i_{k-2})) , \\ \text{and } \Pi_{i_1,\dots,i_{k-2}}(Z_j) &= \big(R_{j,i_1} \cdots R_{j,i_{k-1}}\big) M^{-1}_{i_1,\dots,i_{k-1}} (\Pi_{i_1,\dots,i_{k-2}}(Z_{i_1}),\dots,\Pi_{i_1,\dots,i_{k-2}}(Z_{i_{k-1}})) . \end{aligned}$$

By induction, using (5.2), we get that

$$\begin{aligned} N^{(k-1)} &= N^{(k-2)} - (\lambda_{k-2} - \lambda_{k-1}) \theta(\widehat{\imath}_1,\dots,\widehat{\imath}_{k-2}) , \\ &= (N^{(k-2)} - \lambda_{k-2} \theta(\widehat{\imath}_1,\dots,\widehat{\imath}_{k-2})) + \lambda_{k-1} \theta(\widehat{\imath}_1,\dots,\widehat{\imath}_{k-2}) , \\ &= Z - \Pi_{\widehat{\imath}_1,\dots,\widehat{\imath}_{k-2}}(Z) + \lambda_{k-1} \theta(\widehat{\imath}_1,\dots,\widehat{\imath}_{k-2}) . \end{aligned} \tag{7.8}$$

Then, recall that $N_j^{(k-1)} = \lambda_{k-1}$ for $j = \widehat{\imath}_1, \dots, \widehat{\imath}_{k-1}$ and remark that

$$\lambda_{k-1}\theta_j(\widehat{\imath}_1,\dots,\widehat{\imath}_{k-1}) = \big(R_{j,\widehat{\imath}_1} \cdots R_{j,\widehat{\imath}_{k-1}}\big) M^{-1}_{\widehat{\imath}_1,\dots,\widehat{\imath}_{k-1}} (N^{(k-1)}_{\widehat{\imath}_1}, \dots, N^{(k-1)}_{\widehat{\imath}_{k-1}})\,.$$

Using (7.8) at indices $j = \widehat{\imath}_1, \dots, \widehat{\imath}_{k-1}$, we deduce that

$$\begin{aligned}
&\lambda_{k-1}\theta_j(\widehat{\imath}_1,\dots,\widehat{\imath}_{k-1})\\
&\qquad= \big(R_{j,\widehat{\imath}_1} \cdots R_{j,\widehat{\imath}_{k-1}}\big) M^{-1}_{\widehat{\imath}_1,\dots,\widehat{\imath}_{k-1}} (N^{(k-1)}_{\widehat{\imath}_1}, \dots, N^{(k-1)}_{\widehat{\imath}_{k-1}})\\
&\qquad= \big(R_{j,\widehat{\imath}_1} \cdots R_{j,\widehat{\imath}_{k-1}}\big) M^{-1}_{\widehat{\imath}_1,\dots,\widehat{\imath}_{k-1}} (Z_{\widehat{\imath}_1}, \dots, Z_{\widehat{\imath}_{k-1}})\\
&\qquad\quad - \big(R_{j,\widehat{\imath}_1} \cdots R_{j,\widehat{\imath}_{k-1}}\big) M^{-1}_{\widehat{\imath}_1,\dots,\widehat{\imath}_{k-1}} (\mathit{\Pi}_{\widehat{\imath}_1,\dots,\widehat{\imath}_{k-2}}(Z_{\widehat{\imath}_1}), \dots, \mathit{\Pi}_{\widehat{\imath}_1,\dots,\widehat{\imath}_{k-2}}(Z_{\widehat{\imath}_{k-1}}))\\
&\qquad\quad + \lambda_{k-1}\big(R_{j,\widehat{\imath}_1} \cdots R_{j,\widehat{\imath}_{k-1}}\big) M^{-1}_{\widehat{\imath}_1,\dots,\widehat{\imath}_{k-1}} (\theta_{\widehat{\imath}_1}(\widehat{\imath}_1,\dots,\widehat{\imath}_{k-2}), \dots, \theta_{\widehat{\imath}_{k-1}}(\widehat{\imath}_1,\dots,\widehat{\imath}_{k-2}))\\
&\qquad= \mathit{\Pi}_{\widehat{\imath}_1,\dots,\widehat{\imath}_{k-1}}(Z_j) - \mathit{\Pi}_{\widehat{\imath}_1,\dots,\widehat{\imath}_{k-2}}(Z_j) + \lambda_{k-1}\theta_j(\widehat{\imath}_1,\dots,\widehat{\imath}_{k-2})\,,
\end{aligned}$$

Namely

$$\mathit{\Pi}_{\widehat{\imath}_1,\dots,\widehat{\imath}_{k-1}}(Z) - \lambda_{k-1}\theta(\widehat{\imath}_1,\dots,\widehat{\imath}_{k-1}) = \mathit{\Pi}_{\widehat{\imath}_1,\dots,\widehat{\imath}_{k-2}}(Z) - \lambda_{k-1}\theta(\widehat{\imath}_1,\dots,\widehat{\imath}_{k-2})\,.$$

Using again (7.8) we get that

$$\begin{aligned}
N^{(k-1)} &= Z - \mathit{\Pi}_{\widehat{\imath}_1,\dots,\widehat{\imath}_{k-2}}(Z) + \lambda_{k-1}\theta(\widehat{\imath}_1,\dots,\widehat{\imath}_{k-2})\,,\\
&= Z - \mathit{\Pi}_{\widehat{\imath}_1,\dots,\widehat{\imath}_{k-1}}(Z) + \lambda_{k-1}\theta(\widehat{\imath}_1,\dots,\widehat{\imath}_{k-1})\,,
\end{aligned}$$

as claimed.

7.7 *Proof of Proposition 6.1*

We denote

$$\begin{aligned}
R_j &:= \big(R_{j,i_1}, \dots, R_{j,i_{k-1}}\big)\,,\\
R_{i_k} &:= \big(R_{i_k,i_1}, \dots, R_{i_k,i_{k-1}}\big)\,,\\
M &:= M_{i_1,\dots,i_{k-1}}\,,\\
\overline{M} &:= M_{i_1,\dots,i_k} = \begin{bmatrix} M & R_{i_k} \\ R_{i_k}^\top & R_{i_k,i_k} \end{bmatrix}\,,\\
\overline{R} &:= \big(R_{j,i_1}, \dots, R_{j,i_k}\big)\,,\\
x &:= \frac{1-\theta_j(i_1,\dots,i_{k-1})}{1-\theta_{i_k}(i_1,\dots,i_{k-1})}\,\frac{\tau_{j,i_k}}{\tau_{i_k,i_k}}\,,
\end{aligned}$$

and observe that

$$\begin{aligned}
x &= \frac{R_{j,i_k} - R_j^\top M^{-1} R_{i_k}}{R_{i_k,i_k} - R_{i_k}^\top M^{-1} R_{i_k}}\,,\\
\overline{M}^{-1} &= \begin{bmatrix} \mathrm{Id}_{k-1} & -M^{-1}R_{i_k} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} M^{-1} & 0 \\ 0 & \big(R_{i_k,i_k} - R_{i_k}^\top M^{-1} R_{i_k}\big)^{-1} \end{bmatrix} \begin{bmatrix} \mathrm{Id}_{k-1} & 0 \\ -R_{i_k}^\top M^{-1} & 1 \end{bmatrix}\,,\\
\overline{M}^{-1}\overline{R} &= \begin{bmatrix} M^{-1}\big(R_j - xR_{i_k}\big) \\ x \end{bmatrix}\,,
\end{aligned} \tag{7.9}$$

using Schur complement of block M of the matrix $\overline{M}$ and a LU decomposition. Note also that

$$\frac{Z_j^{(i_1,\dots,i_{k-1})} - Z_{i_k}^{(i_1,\dots,i_{k-1})}\tau_{j,i_k}/\tau_{i_k,i_k}}{1-\tau_{j,i_k}/\tau_{i_k,i_k}} = \frac{Z_j - \Pi_{i_1,\dots,i_{k-1}}(Z_j) - x\,(Z_{i_k} - \Pi_{i_1,\dots,i_{k-1}}(Z_{i_k}))}{1-\theta_j(i_1,\dots,i_{k-1}) - x(1-\theta_{i_k}(i_1,\dots,i_{k-1}))}\,.$$

To prove (6.2), it suffices to show that the R.H.S term above is equal to the following R.H.S term

$$Z_j^{(i_1,\dots,i_k)} = \frac{Z_j - \Pi_{i_1,\dots,i_k}(Z_j)}{1-\theta_j(i_1,\dots,i_k)}\,.$$

We will prove that numerators are equal and that denominators are equal. For denominators,

$$\begin{aligned}
&1-\theta_j(i_1,\dots,i_{k-1}) - x(1-\theta_{i_k}(i_1,\dots,i_{k-1}))\\
&= 1-\theta_j(i_1,\dots,i_{k-1}) - x + x\,\theta_{i_k}(i_1,\dots,i_{k-1})\\
&= 1-(\underbrace{1\cdots 1}_{k\text{ times}})\begin{bmatrix} M^{-1}\big(R_j - xR_{i_k}\big)\\ x\end{bmatrix}\\
&= 1-\theta_j(i_1,\dots,i_k)\,,
\end{aligned} \tag{7.10}$$

using (7.9). Furthermore, it proves (6.1). For the numerators, we use that

$$\begin{aligned}
&Z_j - \Pi_{i_1,\dots,i_{k-1}}(Z_j) - x\,(Z_{i_k} - \Pi_{i_1,\dots,i_{k-1}}(Z_{i_k}))\\
&= Z_j - \Pi_{i_1,\dots,i_{k-1}}(Z_j) - xZ_{i_k} + x\Pi_{i_1,\dots,i_{k-1}}(Z_{i_k})\\
&= Z_j - (Z_{i_1}\cdots Z_{i_k})\begin{bmatrix} M^{-1}\big(R_j - xR_{i_k}\big)\\ x\end{bmatrix}\\
&= Z_j - \Pi_{i_1,\dots,i_k}(Z_j)\,.
\end{aligned}$$

using (7.9).

7.8 *Proof of Theorem 3.8*

We rely on the **Weak Positive Regression Dependency (WPRDS) property** to prove the result, one may consult [Giraud, 2014, Page 173] for instance. We say that a function $g : [0,1]^K \to \mathbb{R}^+$ is *nondecreasing* if for any $p,q \in [0,1]^K$ such that $p_k \geqslant q_k$ for every $k = 1,\dots,K$, we have $g(p) \geqslant g(q)$. We say that a Borel set $\Gamma \in [0,1]^K$ is *nondecreasing* if $g = \mathbb{1}_\Gamma$ is nondecreasing. In other words if $y \in \gamma$ and if $z \geqslant 0$, then $y+z \in \gamma$. We say that the p-values $\big(\widehat{p}_1 = \widehat{\alpha}_{0,1,K+1},\dots,\widehat{p}_K = \widehat{\alpha}_{K-1,K,K+1}\big)$ satisfy the WPRDS property if for any nondecreasing set Γ and for all $k^0 \in I_0$, the function

$$u \mapsto \mathbb{P}_{\mu^0}\big[(\widehat{p}_1,\dots,\widehat{p}_K) \in \Gamma \big| \widehat{p}_{k^0} \leqslant u\big] \text{ is nondecreasing}$$

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

$$I_0 = \big\{k \in [K]\ :\ \mathbb{H}_{0,k} \text{ is true}\big\}\,.$$

To prove Theorem 3.8, note that it is sufficient [Giraud, 2014, Chapter 8] to prove that

$$u \mapsto \overline{\mathbb{P}}\big[(\widehat{p}_1,\dots,\widehat{p}_K) \in \Gamma \big| \widehat{p}_{k^0} \leqslant u\big] \text{ is nondecreasing} \tag{7.11}$$

where $\overline{\mathbb{E}}, \overline{\mathbb{P}}$ will denote that expectations and probabilities are conditional on $\{\bar\imath_1, \dots, \bar\imath_K, \lambda_{K+1}\}$ and under the hypothesis that $\mu^0 = X^\top X \beta^0$. Note that one can integrate in λ_{K+1} to get the statement of Theorem 3.8.

○ **Step 1:** We start by giving the joint law of the LARS knots under the alternative in the orthogonal design case. Lemma 7.1 and (7.6) show that, conditional on $\{\bar\imath_1, \dots, \bar\imath_K, \lambda_{K+1}\}$, $(\lambda_1, \dots, \lambda_K)$ is distributed on the set $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_K \geqslant \lambda_{K+1}$ 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 $\rho_k^2 = 1$ implying that $F_k = \Phi(\lambda_k)$ $\mathscr{P}_{i,j} = \Phi_i \circ \Phi_j^{-1} = \mathrm{Id}$. We deduce that, conditional on $\{\bar\imath_1, \dots, \bar\imath_K, F_{K+1}\}$, $(F_1, \dots, F_K)$ is distributed on the set

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

it has an **explicit density** given by (7.6), and we denote it by h_{μ^0}. By the change of variables $G_k := \frac{F_k - F_{K+1}}{F_{k-1} - F_{K+1}}$ one obtains that the distribution of $(G_1, \dots, G_K)$ is supported on $[0,1]^K$. More precisely, define

$$\psi(f_1, \dots, f_K) := (g_1, \dots, g_K) := \Big(\frac{f_1 - F_{K+1}}{1 - F_{K+1}}, \dots, \frac{f_K - F_{K+1}}{f_{K-1} - F_{K+1}}\Big)$$
$$\psi^{-1}(g_1, \dots, g_K) := \Big((1 - F_{K+1})g_1 + F_{K+1}, \dots, (1 - F_{K+1})g_1 g_2 \dots g_K + F_{K+1}\Big)\,,$$

whose inverse Jacobian determinant is

$$\det\Big[\frac{\partial \psi}{\partial f_1} \cdots \frac{\partial \psi}{\partial f_K}\Big]^{-1} = \prod_{k=1}^{K} (f_{k-1} - F_{K+1}) = (1 - F_{K+1})^K \prod_{k=1}^{K} g_k^{K-k}\,.$$

We deduce that the density of $(G_1, \dots, G_K) | \{\bar\imath_1, \dots, \bar\imath_K, F_{K+1}\}$ at point g with respect to Lebesgue measure is

$$\mathbf{p}(g) := (\mathrm{const}) \mathbb{1}_{g \in (0,1)^K} \prod_{k=1}^{K} g_k^{K-k} \cosh\big[\mu^0_{\bar\imath_k}((1 - F_{K+1}) \prod_{\ell=1}^{k} g_\ell + F_{K+1})\big]\,, \tag{7.12}$$

where we have used (7.6). From (1.3) and (7.3), one has

$$\widehat{p}_k = 1 - \boldsymbol{F}_{\beta(1,K-k+1)}\Big(\frac{F_k - F_{K+1}}{F_{k-1} - F_{K+1}}\Big) = 1 - \boldsymbol{F}_{\beta(1,K-k+1)}(G_k) \tag{7.13}$$

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

$$\widehat{p}_k = v \Leftrightarrow (G_1, \dots, G_K) \in [0,1]^K \cap \{\boldsymbol{F}^{-1}_{\beta(1,K-k+1)}(1 - v) = G_k\}\,,$$

so that

$$\overline{\mathbb{P}}\big[(\widehat{p}_1, \dots, \widehat{p}_K) \in \Gamma \big| \widehat{p}_{k^0} \leqslant u\big] = \overline{\mathbb{P}}\big[(G_1, \dots, G_K) \in \overline{\Gamma} \big| G_{k^0} \geqslant \boldsymbol{F}^{-1}_{\beta(1,K-\ell+1)}(1 - u)\big]\,, \tag{7.14}$$

where $\overline{\Gamma}$ can be proved to be a **nonincreasing Borel set** from (7.13).

○ **Step 2:** Let $0 < x < y < 1$ and denote by μ_x the following conditional law

$$\mu_x := \mathrm{law}\big[(G_1, \dots, G_K) | \{\bar\imath_1, \dots, \bar\imath_K, F_{K+1}, G_{k^0} \geqslant x\}\big]\,.$$

Remark that if there exists a measurable $T : [0,1]^K \mapsto [0,1]^K$ such that

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

then it holds

- $\mathbb{1}_{\{T(g)\in\overline{\Gamma}\}} \leqslant \mathbb{1}_{\{g\in\overline{\Gamma}\}}$;
- $\mathrm{law}\big[T(G)|\{\bar{\imath}_1,\ldots,\bar{\imath}_K,F_{K+1},G_{k^0}\geqslant x\}\big] = \mathrm{law}\big[G|\{\bar{\imath}_1,\ldots,\bar{\imath}_K,F_{K+1},G_{k^0}\geqslant y\}\big]$ where $G=(G_1,\ldots,G_K)$.

In this case, we deduce that

$$\overline{\mathbb{P}}\big[G\in\overline{\Gamma}\big|G_{k^0}\geqslant x\big] \geqslant \overline{\mathbb{P}}\big[T(G)\in\overline{\Gamma}\big|G_{k^0}\geqslant x\big] = \overline{\mathbb{P}}\big[G\in\overline{\Gamma}\big|G_{k^0}\geqslant y\big]\,.$$

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

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

and, in view of (7.14), it proves (7.11). 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 μ_x toward μ_y following the order

$$k^0 \to k^0+1 \to \cdots \to K \to k^0-1 \to k^0-2 \to \cdots \to 1\,.$$

It is based on a sequence of conditional quantile transforms defined following the ordering above. Its construction is presented for instance in [Santambrogio, 2015, Sec.2.3, P.67] or [Villani, 2008, P.20]. The transport T is defined as follows. Given $z,z'\in[0,1]^K$ such that $z'=T(z)$ it holds

$$\begin{aligned}
z'_{k^0} &= T^{(k^0)}(z_{k^0});\\
z'_{k^0+1} &= T^{(k^0+1)}(z_{k^0+1},z'_{k^0});\\
&\;\,\vdots\\
z'_K &= T^{(K)}(z_K,z'_{K-1},\ldots,z'_{k^0});\\
z'_{k^0-1} &= T^{(k^0-1)}(z_{k^0-1},z'_K,\ldots,z'_{k^0});\\
&\;\,\vdots\\
z'_1 &= T^{(1)}(z_1,z'_2,\ldots,z'_{k^0-1},z'_K,\ldots,z'_{k^0});
\end{aligned}$$

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 $g\in[0,1]^K$, $T(g)\geqslant g$;
- T is such that push-forward of μ_x by T gives μ_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.

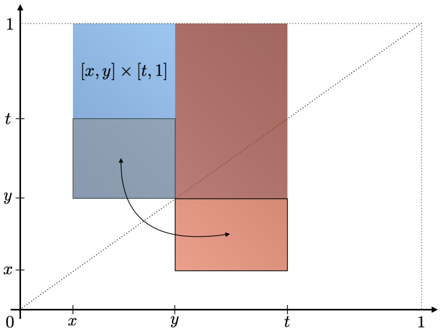


Figure 9: Note that, by symmetry the two boxed regions have same $\mathbf{p} \otimes \mathbf{p}$ measure. The blue region is $[x,t] \times [y,1]$, its measure is the measure of the red region (namely $cD_y(t) \times \mathscr{D}_x(1)$) more the bluest upper left corner (namely $[x,y] \times [t,1]$).

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

$$\mu_x^{(k^0)} := \mathrm{law}\big[G_{k^0} | \{\bar\imath_1, \ldots, \bar\imath_K, F_{K+1}, G_{k^0} \geqslant x\}\big] ,$$

and $\mathbb{F}_x^{(k^0)}$ its cdf. Note that the Knothe-Rosenblatt construction gives $T^{(k^0)} = (\mathbb{F}_y^{(k^0)})^{-1} \circ \mathbb{F}_x^{(k^0)}$. We would like to prove that $T^{(k^0)}(t) \geqslant t$ for all $z \in (0,1)$. This is equivalent to prove that it holds $\mathbb{F}_x^{(k^0)} \geqslant \mathbb{F}_y^{(k^0)}$. For $t \leqslant y$, $\mathbb{F}_y^{(k^0)}(t) = 0$ and it implies that $\mathbb{F}_x^{(k^0)}(t) \geqslant \mathbb{F}_y^{(k^0)}(t)$. Let $t > y$, using the conditional density $\mathbf{p}$ defined in (7.12), note that

$$\begin{aligned} \mathbb{F}_x^{(k^0)}(t) \geqslant \mathbb{F}_y^{(k^0)}(t) \quad &\Leftrightarrow \quad \frac{\int_x^t \mathbf{p}}{\int_x^1 \mathbf{p}} \geqslant \frac{\int_y^t \mathbf{p}}{\int_y^1 \mathbf{p}} \\ &\Leftrightarrow \quad \int_x^t \int_y^1 \mathbf{p} \otimes \mathbf{p} \geqslant \int_y^t \int_x^1 \mathbf{p} \otimes \mathbf{p}, \end{aligned}$$

where, for example

$$\int_x^t \text{ means the integral over the hyper rectangle } [x,t] := \Big\{ (g_1, \ldots, g_K) \in [0,1]^K \ : x \leqslant g_{k^0} \leqslant t \Big\}.$$

A simple calculation (see also Figure 9) gives that

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

and it proves that $\mathbb{F}_x^{(k^0)} \geqslant \mathbb{F}_y^{(k^0)}$.

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

$$\mu_{z_{k^0}}^{(k^0+1)} := \mathrm{law}\big[G_{k^0+1} | \{\bar\imath_1, \ldots, \bar\imath_K, F_{K+1}, G_{k^0} = z_{k^0}\}\big] ,$$

and $\mathbb{F}^{(k^0+1)}_{z_{k^0}}$ its cdf. Let $z'_{k^0} := T^{(k^0)}(z'_{k^0})$ and denote $\mu^{(k^0+1)}_{z'_{k^0}}$ the following conditional law

$$\mu^{(k^0+1)}_{z'_{k^0}} := \mathrm{law}\big[G_{k^0+1}|\{\bar\imath_1,\dots,\bar\imath_K,F_{K+1},G_{k^0}=z'_{k^0}\}\big]\,,$$

and $\mathbb{F}^{(k^0+1)}_{z'_{k^0}}$ its cdf. Note that $x < z_{k^0} \leqslant z'_{k^0} = T^{(k^0)}(z_{k^0}) \leqslant 1$. Again, we would like to prove that $\mathbb{F}^{(k^0+1)}_{z_{k^0}} \geqslant \mathbb{F}^{(k^0+1)}_{z'_{k^0}}$ which implies that the transport map $T^{(k^0+1)} := \big(\mathbb{F}^{(k^0+1)}_{z'_{k^0}}\big)^{-1} \circ \mathbb{F}^{(k^0+1)}_{z_{k^0}}$ satisfies $T^{(k^0+1)}(u) \geqslant u$ for all $u \in (0,1)$.

Recall that the conditional density $\mathbf{p}$ of $G|\{\bar\imath_1,\dots,\bar\imath_K,F_{K+1}\}$ is given by (7.12) and recall that $k^0 \in I_0$. Observe that $\mu^0_{k^0} = 0$, so that the conditional density of $G|\{\bar\imath_1,\dots,\bar\imath_K,F_{K+1},G_{k^0}=\mathbf{z}\}$ is

$$\begin{aligned}(\mathrm{const})\,\mathbb{1}_{g\in(0,1)^K}\mathbb{1}_{g_{k^0}=\mathbf{z}}\prod_{k<k^0} g_k^{K-k}\cosh\Big[\mu^0_{\bar\imath_k}\big((1-F_{K+1})\prod_{\ell=1}^{k} g_\ell + F_{K+1}\big)\Big]&\\ \times\prod_{k>k^0} g_k^{K-k}\cosh\Big[\mu^0_{\bar\imath_k}\big((1-F_{K+1})\,\mathbf{z}\prod_{1\leqslant\ell\neq k^0\leqslant k} g_\ell + F_{K+1}\big)\Big]\,.&\end{aligned}\tag{7.15}$$

Set $\tau := z'_{k^0}/z_{k^0} \geqslant 1$ 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 $G' := (G_1,\dots,G_{k^0-1},G'_{k^0+1},G_{k^0+2},\dots,G_K) \in (0,1)^{k^0-1}\times(0,\tau)\times(0,1)^{K-k^0-1}$ and note that the conditional density of $G'|\{\bar\imath_1,\dots,\bar\imath_K,F_{K+1},G_{k^0}=\tau\mathbf{z}\}$ is

$$\begin{aligned}(\mathrm{const})\,\mathbb{1}_{g\in(0,1)^{k^0-1}\times(0,\tau)\times(0,1)^{K-k^0-1}}\prod_{k<k^0} g_k^{K-k}\cosh\Big[\mu^0_{\bar\imath_k}\big((1-F_{K+1})\prod_{\ell=1}^{k} g_\ell + F_{K+1}\big)\Big]&\\ \times\prod_{k>k^0} g_k^{K-k}\cosh\Big[\mu^0_{\bar\imath_k}\big((1-F_{K+1})\,\mathbf{z}\prod_{1\leqslant\ell\neq k^0\leqslant k} g_\ell + F_{K+1}\big)\Big]\,,&\end{aligned}$$

which, up to some normalising constant, is the same as (7.15) up to the following change of support

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

By an abuse of notation, we denote by $\mathbf{p}$ this function, namely

$$\begin{aligned}\mathbf{p}(g) = \prod_{k<k^0} g_k^{K-k}\cosh\Big[\mu^0_{\bar\imath_k}\big((1-F_{K+1})\prod_{\ell=1}^{k} g_\ell + F_{K+1}\big)\Big]&\\ \times\prod_{k>k^0} g_k^{K-k}\cosh\Big[\mu^0_{\bar\imath_k}\big((1-F_{K+1})\,\mathbf{z}\prod_{1\leqslant\ell\neq k^0\leqslant k} g_\ell + F_{K+1}\big)\Big]\,.&\end{aligned}$$

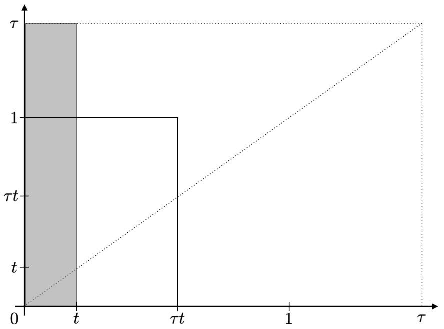


Figure 10: The two boxed rectangles have Lebesgue measure, namely τt. The $\mathbf{p}\otimes\mathbf{p}$ measure of the grey box is greater than the $\mathbf{p}\otimes\mathbf{p}$ measure of the white box.

We deduce that

$$\begin{aligned}
\mathbb{F}_{z_{k^0}}^{(k^0+1)}(t) \geqslant \mathbb{F}_{z'_{k^0}}^{(k^0+1)}(t) \quad &\Leftrightarrow \quad \overline{\mathbb{P}}(G_{k^0+1} \leqslant t | G_{k^0} = z_{k^0}) \geqslant \overline{\mathbb{P}}(G_{k^0+1} \leqslant t | G_{k^0} = z'_{k^0}) \\
&\Leftrightarrow \quad \overline{\mathbb{P}}(G_{k^0+1} \leqslant t | G_{k^0} = z_{k^0}) \geqslant \overline{\mathbb{P}}(G'_{k^0+1} \leqslant \tau t | G_{k^0} = \tau z_{k^0}) \\
&\Leftrightarrow \quad \frac{\int_{\mathscr{D}(t)} \mathbf{p}}{\int_{\mathscr{D}(1)} \mathbf{p}} \geqslant \frac{\int_{\mathscr{D}(\tau t)} \mathbf{p}}{\int_{\mathscr{D}(\tau)} \mathbf{p}} \\
&\Leftrightarrow \quad \int_{\mathscr{D}(t)\times\mathscr{D}(\tau)} \mathbf{p}\otimes\mathbf{p} \geqslant \int_{\mathscr{D}(\tau t)\times\mathscr{D}(1)} \mathbf{p}\otimes\mathbf{p}, \qquad (7.16)
\end{aligned}$$

where

$$\mathscr{D}(s) := \Big\{(g_1,\dots,g_{k^0-1},g_{k^0+1}\dots,g_K) \in (0,1)^{K-1} \ : \ 0 < g_{k^0+1} \leqslant s\Big\}.$$

We now present an inequality on the to conclude. Observe that we are integrating on domains depicted in Figure 10. The two boxes have same area for the uniform measure and we would like to compare their respective measure for the $\mathbf{p}\otimes\mathbf{p}$ measure. We start by the next lemma whose proof is omitted.

LEMMA 7.3 Let $a,b \geqslant 0$. The function

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

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

Now, let $(g_1,\dots,g_{k^0-1},g_{k^0+2}\dots,g_K) \in (0,1)^{K-1}$ be fixed in the integrals (7.16). 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

$$\begin{aligned}
w(h_1,h_2) =&\mathrm{C}_1 h_1^{K-k^0-1}\cosh\big[\mu^0_{\bar\imath_k}((1-F_{K+1})\,z_{k^0}\prod_{1\leqslant\ell<k^0} g_\ell\times h_1+F_{K+1})\big]\\
&\times\prod_{k>k^0+1}\cosh\big[\mu^0_{\bar\imath_k}((1-F_{K+1})\,z_{k^0}\prod_{1\leqslant\ell\neq k^0,k^0+1\leqslant k} g_\ell\times h_1+F_{K+1})\big]\\
&\times h_2^{K-k^0-1}\cosh\big[\mu^0_{\bar\imath_k}((1-F_{K+1})\,z_{k^0}\prod_{1\leqslant\ell<k^0} g_\ell\times h_2+F_{K+1})\big]\\
&\times\prod_{k>k^0+1}\cosh\big[\mu^0_{\bar\imath_k}((1-F_{K+1})\,z_{k^0}\prod_{1\leqslant\ell\neq k^0,k^0+1\leqslant k} g_\ell\times h_2+F_{K+1})\big]\,,
\end{aligned}$$

where the constant C_1 depends on $(g_1,\ldots,g_{k^0-1},g_{k^0+2}\cdots,g_K)\in(0,1)^{K-1}$. By the change of variables $h'_1=h_3/t$ and $h'_2=th_4$, the right hand term of (7.16) 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

$$\begin{aligned}
w'(h'_1,h'_2) =&\mathrm{C}_1 {h'_1}^{K-k^0-1}\cosh\big[\mu^0_{\bar\imath_k}((1-F_{K+1})\,z_{k^0}\prod_{1\leqslant\ell<k^0} g_\ell\times t\times h'_1+F_{K+1})\big]\\
&\times\prod_{k>k^0+1}\cosh\big[\mu^0_{\bar\imath_k}((1-F_{K+1})\,z_{k^0}\prod_{1\leqslant\ell\neq k^0,k^0+1\leqslant k} g_\ell\times t\times h'_1+F_{K+1})\big]\\
&\times {h'_2}^{K-k^0-1}\cosh\big[\mu^0_{\bar\imath_k}((1-F_{K+1})\,z_{k^0}\prod_{1\leqslant\ell<k^0} g_\ell\times h'_2/t+F_{K+1})\big]\\
&\times\prod_{k>k^0+1}\cosh\big[\mu^0_{\bar\imath_k}((1-F_{K+1})\,z_{k^0}\prod_{1\leqslant\ell\neq k^0,k^0+1\leqslant k} g_\ell\times h'_2/t+F_{K+1})\big]\,.
\end{aligned}$$

Now, invoke Lemma 7.3 with

$$\begin{aligned}
&a=\mu^0_{\bar\imath_k}(1-F_{K+1})\,z_{k^0}\prod_{1\leqslant\ell<k^0} g_\ell\times h\\
&b=\mu^0_{\bar\imath_k}F_{K+1}\\
&z=t\geqslant 1\,,
\end{aligned}$$

where $h=h_1$ or h_2, to get that $w'\geqslant w$ and so

$$\int_{\mathscr{D}(t)\times\mathscr{D}(\tau)}\mathbf{p}\otimes\mathbf{p}\geqslant\int_{\mathscr{D}(\tau t)\times\mathscr{D}(1)}\mathbf{p}\otimes\mathbf{p}\,,$$

which concludes this part of the proof.

○ *Step 3.3:* We continue by induction with the other transport maps in Knothe-Rosenblatt's construction. Assume that we have built $z':=(z'_k,\ldots,z'_{k^0})$ and $z:=(z_k,\ldots,z_{k^0})$ for some $k>k^0$. Denote $\mu_z^{(k+1)}$ the following conditional law

$$\mu_z^{(k+1)}:=\mathrm{law}\big[G_{k+1}|\{\bar\imath_1,\ldots,\bar\imath_K,F_{K+1},\underbrace{G_k=z_k,\ldots,G_{k^0}=z_{k^0}}_{\text{denoted } G^{[k,k^0]}=z}\}\big]\,,$$

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

$$\mu_{z'}^{(k+1)} := \mathrm{law}\big[G_{k+1} | \{\bar\imath_1, \ldots, \bar\imath_K, F_{K+1}, \underbrace{G_k = z'_k, \ldots, G_{k^0} = z'_{k^0}}_{G^{[k,k^0]}=z'}\}\big]\,,$$

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

For $\mathbf{z} \in (0,1)^{k-k^0} \times (x,1)$, the conditional density of $G | \{\bar\imath_1, \ldots, \bar\imath_K, F_{K+1}, G^{[k,k^0]} = \mathbf{z}\}$ is

$$\begin{aligned}(\mathrm{const})\, \mathbb{1}_{g\in(0,1)^K} \mathbb{1}_{g^{[k,k^0]}=\mathbf{z}} \prod_{m<k^0} g_m^{K-m} \cosh\big[\mu^0_{\bar\imath_m}((1-F_{K+1})\prod_{\ell=1}^{m} g_\ell + F_{K+1})\big]\big] \\ \times \prod_{k^0\leqslant m\leqslant k} \mathbf{z}_m^{K-m} \cosh\big[\mu^0_{\bar\imath_m}((1-F_{K+1}) \prod_{1\leqslant\ell<k^0} g_\ell \prod_{n=k^0}^{m} \mathbf{z}_n + F_{K+1})\big] \\ \times \prod_{k<m} g_m^{K-m} \cosh\big[\mu^0_{\bar\imath_m}((1-F_{K+1}) \prod_{1\leqslant\ell<k^0} g_\ell \prod_{n=k^0}^{k} \mathbf{z}_n \prod_{k<\ell\leqslant m} g_\ell + F_{K+1})\big]\,.\end{aligned}$$

Set $\tau := \prod_{n=k^0}^k z'_n / \prod_{n=k^0}^k z_n \geqslant 1$ and $G'_k = \tau G_{k^0+1}$ so that

$$\Big[\prod_{n=k^0}^{k} z'_n\Big] G_{k+1} = \Big[\prod_{n=k^0}^{k} z_n\Big] G'_{k+1}\,.$$

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

○ *Step 3.4:* This is the last step of the proof. Assume that we have built $z' := (z'_K, \ldots, z'_{k^0})$ and $z := (z_K, \ldots, z_{k^0})$. Denote $\mu_z^{(k^0-1)}$ the following conditional law

$$\mu_z^{(k^0-1)} := \mathrm{law}\big[G_{k^0-1} | \{\bar\imath_1, \ldots, \bar\imath_K, F_{K+1}, G^{[K,k^0]} = z\}\big]\,,$$

and $\mathbb{F}_z^{(k^0-1)}$ its cdf. Denote $\mu_{z'}^{(k^0-1)}$ the following conditional law

$$\mu_{z'}^{(k^0-1)} := \mathrm{law}\big[G_{k^0-1} | \{\bar\imath_1, \ldots, \bar\imath_K, F_{K+1}, G^{[K,k^0]} = z'\}\big]\,,$$

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

For $\mathbf{z} \in (0,1)^{K-k^0} \times (x,1)$, the conditional density of $G | \{\bar\imath_1, \ldots, \bar\imath_K, F_{K+1}, G^{[K,k^0]} = \mathbf{z}\}$ is

$$\begin{aligned}(\mathrm{const})\, \mathbb{1}_{g\in(0,1)^K} \mathbb{1}_{g^{[K,k^0]}=\mathbf{z}} \prod_{m<k^0} g_m^{K-m} \cosh\big[\mu^0_{\bar\imath_m}((1-F_{K+1})\prod_{\ell=1}^{m} g_\ell + F_{K+1})\big] \\ \times \prod_{k^0\leqslant m\leqslant K} \mathbf{z}_m^{K-m} \cosh\big[\mu^0_{\bar\imath_m}((1-F_{K+1}) \prod_{1\leqslant\ell<k^0} g_\ell \prod_{n=k^0}^{m} \mathbf{z}_n + F_{K+1})\big]\,.\end{aligned}$$

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

$$\forall g\in(0,1),\quad w_z(g):=g^{K-k^0+1}\cosh\big[\mu^0_{\bar\imath_{k^0-1}}((1-F_{K+1})\prod_{\ell=1}^{k^0-2}g_\ell\times g+F_{K+1})\big]$$
$$\times\prod_{k^0\leqslant m\leqslant K}\mathbf{z}_m^{K-m}\cosh\big[\mu^0_{\bar\imath_m}((1-F_{K+1})\prod_{n=k^0}^{m}\mathbf{z}_n\prod_{\ell=1}^{k^0-2}g_\ell\times g+F_{K+1})\big]\,.$$

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

$$\int_0^t\Big(\int_0^1 w_z(g)w_{z'}(g')\mathrm{d}g'\Big)\mathrm{d}g\geqslant\int_0^1\Big(\int_0^t w_z(g)w_{z'}(g')\mathrm{d}g'\Big)\mathrm{d}g\,.$$

Substituting

$$\int_0^t\Big(\int_0^t w_z(g)w_{z'}(g')\mathrm{d}g'\Big)\mathrm{d}g$$

on both parts, one is reduced to prove that

$$\int_0^t\Big(\int_t^1 w_z(g)w_{z'}(g')\mathrm{d}g'\Big)\mathrm{d}g\geqslant\int_0^t\Big(\int_t^1 w_{z'}(g)w_z(g')\mathrm{d}g'\Big)\mathrm{d}g\,.$$

Observe that $g\leqslant g'$ in the last two integrals. Now, we have this lemma whose proof is omitted.

LEMMA 7.4 Let $0<a\leqslant a'$ and $b>0$. The function

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

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

Let $g\leqslant g'$. From Lemma 7.4, we deduce that $\cosh(ag+b)\cosh(a'g'+b)\geqslant\cosh(ag'+b)\cosh(a'g+b)$, proving that $w_z(g)w_{z'}(g')\geqslant w_{z'}(g)w_z(g')$. It proves that $T^{(k^0-1)}(u)\geqslant u$ for all $u\in(0,1)$.

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

8. A Quasi Monte Carlo (QMC) method: Cubature 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 $x^{(i)},i=1,\dots,n$ is the so-called *lattice rule*: from the first point $x^{(1)}$ we deduce the others $x^{(i)}$ by

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

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 [Nuyens and Cools, 2006] 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 $\mathscr{E}$ of some *RKHS*, namely a tensorial product of *Koborov spaces*. In addition it gives an expression of its minimax error, namely

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

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 estimate of the error is obtained by adding a *Monte-Carlo layer* as in Genz [1992] for instance. This can be done as follows. Let U a **unique** uniform variable on $[0,1]^d$, we define

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

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.

References

Azaïs, J.-M., De Castro, Y., and Mourareau, S. (2018). Power of the spacing test for least-angle regression. *Bernoulli*, 24(1):465–492.

Bachoc, F., Blanchard, G., Neuvial, P., et al. (2018). On the post selection inference constant under restricted isometry properties. *Electronic Journal of Statistics*, 12(2):3736–3757.

Barber, R. F., Candès, E. J., et al. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085.

Bellec, P. C., Lecué, G., Tsybakov, A. B., et al. (2018). Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46(6B):3603–3642.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L., et al. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837.

Bickel, P. J., Ritov, Y., Tsybakov, A. B., et al. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Blanchard, G., Neuvial, P., and Roquain, E. (2017). Post hoc inference via joint family-wise error rate control. *arXiv preprint arXiv:1703.02307*.

Blanchard, G., Roquain, E., et al. (2008). Two simple sufficient conditions for fdr control. *Electronic journal of Statistics*, 2:963–992.

Bogdan, M., Van Den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015). Slope—adaptive variable selection via convex optimization. *The annals of applied statistics*, 9(3):1103.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg. Methods, theory and applications.

Candès, E. J., Romberg, J., and Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61 (electronic).

De Castro, Y. (2021). github:ydecastro/lar_testing: GtSt experiments on real and simulated data, doi:10.5281/zenodo.507976.

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

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of computational and graphical statistics*, 1(2):141–149.

Genz, A. and Bretz, F. (2009). *Computation of multivariate normal and t probabilities*, volume 195. Springer Science & Business Media.

Giraud, C. (2014). *Introduction to high-dimensional statistics*. Chapman and Hall/CRC.

Javanmard, A., Javadi, H., et al. (2019). False discovery rate control via debiased lasso. *Electronic Journal of Statistics*, 13(1):1212–1253.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909.

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

Nuyens, D. and Cools, R. (2006). Fast algorithms for component-by-component construction of rank-1 lattice rules in shift-invariant reproducing kernel hilbert spaces. *Mathematics of Computation*, 75(254):903–920.

Rhee, S.-Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. L., and Shafer, R. W. (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103(46):17355–17360.

Roquain, E. (2011). Type i error rate control for testing many hypotheses: a survey with proofs. *Journal de la Société Française de Statistique*, 152(2):3–38.

Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, 55:58–63.

Taylor, J. and Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634.

Tian, X., Loftus, J. R., and Taylor, J. E. (2018). Selective inference with unknown variance via the square-root lasso. *Biometrika*, 105(4):755–768.

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

Tibshirani, R., Wainwright, M., and Hastie, T. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Monographs on Statistics & Applied Probability. Chapman and Hall/CRC press.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620.

van de Geer, S. (2016). Estimation and testing under sparsity. *Lecture Notes in Mathematics*, 2159.

Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer Science & Business Media.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202.