Electronic Journal of Statistics
Vol. 12 (2018) 3736–3757
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1490

On the post selection inference constant under restricted isometry properties

François Bachoc

Institut de Mathématiques de Toulouse;
UMR 5219; Université de Toulouse; CNRS
UPS, F-31062 Toulouse Cedex 9, France
e-mail: francois.bachoc@math.univ-toulouse.fr

Gilles Blanchard

Universität Potsdam, Institut für Mathematik
Karl-Liebknecht-Straße 24-25 14476 Potsdam, Germany
e-mail: gilles.blanchard@math.uni-potsdam.de

Pierre Neuvial

Institut de Mathématiques de Toulouse;
UMR 5219; Université de Toulouse; CNRS
UPS, F-31062 Toulouse Cedex 9, France
e-mail: pierre.neuvial@math.univ-toulouse.fr

Abstract: Uniformly valid confidence intervals post model selection in regression can be constructed based on Post-Selection Inference (PoSI) constants. PoSI constants are minimal for orthogonal design matrices, and can be upper bounded in function of the sparsity of the set of models under consideration, for generic design matrices.

In order to improve on these generic sparse upper bounds, we consider design matrices satisfying a Restricted Isometry Property (RIP) condition. We provide a new upper bound on the PoSI constant in this setting. This upper bound is an explicit function of the RIP constant of the design matrix, thereby giving an interpolation between the orthogonal setting and the generic sparse setting. We show that this upper bound is asymptotically optimal in many settings by constructing a matching lower bound.

MSC 2010 subject classifications: 62J05, 62J15, 62F25.
Keywords and phrases: Inference post model-selection, confidence intervals, PoSI constants, linear regression, high-dimensional inference, sparsity, restricted isometry property.

Received April 2018.

1. Introduction

Fitting a statistical model to data is often preceded by a model selection step. The construction of valid statistical procedures in such post model selection situations is quite challenging (cf. [21, 22, 23], [17] and [25], and the references given in that literature), and has recently attracted a considerable amount of attention. Among various recent references in this context, we can mention

those addressing sparse high dimensional settings with a focus on lasso-type model selection procedures [4, 5, 29, 31], those aiming for conditional coverage properties for polyhedral-type model selection procedures [14, 19, 20, 27, 28] and those achieving valid post selection inference universally over the model selection procedure [1, 2, 6].

In this paper, we shall focus on the latter type of approach and adopt the setting introduced in [6]. In that work, a linear Gaussian regression model is considered, based on an $n \times p$ design matrix X. A model $M \subset \{1, ..., p\}$ is defined as a subset of indices of the p covariates. For a family $\mathcal{M} \subset \{M | M \subset \{1, \ldots, p\}\}$ of admissible models, it is shown in [6] that a universal coverage property is achievable (see Section 2) by using a family of confidence intervals whose sizes are proportional to a constant $K(X, \mathcal{M}) > 0$. This constant $K(X, \mathcal{M})$ is called a PoSI (Post-Selection Inference) constant in [6]. This setting was later extended to prediction problems in [1] and to misspecified non-linear settings in [2].

The focus of this paper is on the order of magnitude of the PoSI constant $K(X, \mathcal{M})$ for large p. We shall consider $n \geq p$ for simplicity of exposition in the rest of this section (and asymptotics $n, p \to \infty$). It is shown in [6] that $K(X, \mathcal{M}) = \Omega(\sqrt{\log(p)})$; this rate is reached in particular when X has orthogonal columns. On the other hand, in full generality $K(X, \mathcal{M}) = O(\sqrt{p})$ for all X. It can also be shown, as discussed in an intermediary version of [32], that when $\mathcal{M}$ is composed of s-sparse submodels, the sharper upper bound $K(X, \mathcal{M}) = O(\sqrt{s \log(p/s)})$ holds. Hence, intuitively, design matrices that are close to orthogonal and consideration of sparse models yield smaller PoSI constants.

In this paper, we obtain additional quantitative insights for this intuition, by considering design matrices X satisfying restricted isometry property (RIP) conditions. RIP conditions have become central in high dimensional statistics and compressed sensing [8, 10, 15]. In the s-sparse setting and for design matrices X that satisfy a RIP property of order s with RIP constant $\delta \to 0$, we show that $K(X, \mathcal{M}) = O(\sqrt{\log(p)} + \delta\sqrt{s \log(p/s)})$. This corresponds to the intuition that for such matrices, any subset of s columns of X is "approximately orthogonal". Thus, under the RIP condition we improve the upper bound of [32] for the s-sparse case, by up to a factor $\delta \to 0$. We show that our upper bound is complementary to the bounds recently proposed in [18]. In addition, we obtain lower bounds on $K(X, \mathcal{M})$ for a class of design matrices that extends the equicorrelated design matrix in [6]. From these lower bounds, we show that the new upper bound we provide is optimal, in a large range of situations.

While the main interest of our results is theoretical, our suggested upper bound can be practically useful in cases where it is computable whereas the PoSI constant $K(X, \mathcal{M})$ is not. The only challenge for computing our upper bound is to find a value δ for which the design matrix X satisfies a RIP property. While this is currently challenging in general for large p, we discuss, in this paper, specific cases where it is feasible.

The rest of the paper is organized as follows. In Section 2 we introduce in more details the setting and the PoSI constant $K(X, \mathcal{M})$. In Section 3 we introduce the RIP condition, provide the upper bound on $K(X, \mathcal{M})$ and discuss

its theoretical comparison with [18] and its applicability. In Section 4 we provide the lower bound and the optimality result for the upper bound. All the proofs are given in the appendix.

2. Settings and notation

2.1. PoSI confidence intervals

We consider and review briefly the framework introduced by [6] for which the so-called PoSI constant plays a central role. The goal is to construct post-model selection confidence intervals that are agnostic with respect to the model selection mehod used. The authors of [6] assume a Gaussian vector of observations

$$Y = \mu + \epsilon, \tag{1}$$

where the $n \times 1$ mean vector μ is fixed and unknown, and ϵ follows the $\mathcal{N}(0, \sigma^2 I_n)$ distribution where $\sigma^2 > 0$ is unknown. Consider an $n \times p$ fixed design matrix X, whose columns correspond to explanatory variables for μ. It is not necessarily assumed that μ belongs to the image of X or that $n \geq p$.

A model M corresponds to a subset of selected variables in $\{1, \ldots, p\}$. A set of models of interest $\mathcal{M} \subset \mathcal{M}_{all} = \{M | M \subset \{1, \ldots, p\}\}$ is supposed to be given. Following [6], for any $M \in \mathcal{M}$, the projection based vector of regression coefficients β_M is a target of inference, with

$$\beta_M := \underset{\beta \in \mathbb{R}^{|M|}}{\text{Arg Min}} \|\mu - X_M \beta\|^2 = (X_M^t X_M)^{-1} X_M^t \mu, \tag{2}$$

where X_M is the submatrix of X formed of the columns of X with indices in M, and where we assume that for each $M \in \mathcal{M}$, X_M has full rank and M is non-empty. We refer to [6] for an interpretation of the vector β_M and a justification for considering it as a target of inference. In [6], a family of confidence intervals $(\text{CI}_{i,M}; i \in M \in \mathcal{M})$ for β_M is introduced, containing the targets $(\beta_M)_{M \in \mathcal{M}}$ simultaneously with probability at least $1 - \alpha$. The confidence intervals take the form

$$\text{CI}_{i,M} := (\hat{\beta}_M)_{i.M} \pm \hat{\sigma} \|v_{M,i}\| K(X, \mathcal{M}, \alpha, r); \tag{3}$$

the different quantities involved, which we now define, are standard ingredients for univariate confidence intervals for regression coefficients in the Gaussian model, except for the last factor (the "PoSI constant") which will account for multiplicity of covariates and models, and their simultaneous coverage. The confidence interval is centered at $\hat{\beta}_M := (X_M^t X_M)^{-1} X_M^t Y$, the ordinary least squares estimator of β_M; also, if $M = \{j_1, \ldots, j_{|M|}\}$ with $j_1 < \ldots < j_{|M|}$, for $i \in M$ we denote by $i.M$ the number $k \in \mathbb{N}$ for which $j_k = i$, that is, the rank of the i-th element in the subset M. The quantity $\hat{\sigma}^2$ is an unbiased estimator of σ^2, more specifically it is assumed that it is an observable random variable, such that $\hat{\sigma}^2$ is independent of $P_X Y$ and is distributed as σ^2/r times a chi-square distributed random variable with r degrees of freedom (P_X denoting

the orthogonal projection onto the column space of X). We allow for $r = \infty$ corresponding to $\hat{\sigma} = \sigma$, i.e., the case of known variance (also called Gaussian limiting case). In [6], it is assumed that $\hat{\sigma}$ exists and it is shown that this indeed holds in some specific situations. A further analysis of the existence of $\hat{\sigma}$ is provided in [1, 2].

The next quantity to define is

$$v_{M,i}^t := (e_{i.M}^{|M|})^t G_M^{-1} X_M^t \in \mathbb{R}^n, \tag{4}$$

where e_a^b is the a-th base column vector of $\mathbb{R}^b$; and $G_M := X_M^t X_M$ is the $|M| \times |M|$ Gram matrix formed from the columns of X_M. Observe that $v_{M,i}$ is nothing more than the row corresponding to covariate i in the estimation matrix $G_M^{-1} X_M^t$, in other words $(\hat{\beta}_M)_{i.M} = v_{M,i}^t Y$.

Finally, $K(X, \mathcal{M}, \alpha, r)$ is called a PoSI constant and we turn to its definition. We shall occasionally write for simplicity $K(X, \mathcal{M}, \alpha, r) = K(X, \mathcal{M})$. Furthermore, if the value of r is not specified in $K(X, \mathcal{M})$, it is implicit that $r = \infty$.

Definition 2.1. Let $\mathcal{M} \subset \mathcal{M}_{all}$ for which each $M \in \mathcal{M}$ is non-empty, and so that X_M has full rank. Let also

$$w_{M,i} = \begin{cases} v_{M,i}/\|v_{M,i}\|, & \text{if } \|v_{M,i}\| \neq 0; \\ 0 \in \mathbb{R}^n & \text{else.} \end{cases}$$

Let ξ be a Gaussian vector with zero mean vector and identity covariance matrix on $\mathbb{R}^n$. Let N be a random variable, independent of ξ, and so that rN^2 follows a chi-square distribution with r degrees of freedom. If $r = \infty$, then we let $N = 1$. For $\alpha \in (0, 1)$, $K(X, \mathcal{M}, \alpha, r)$ is defined as the $1 - \alpha$ quantile of

$$\gamma_{\mathcal{M},r} := \frac{1}{N} \max_{M \in \mathcal{M}, i \in M} \left| w_{M,i}^t \xi \right|. \tag{5}$$

We remark that $K(X, \mathcal{M}, \alpha, r)$ is the same as in [6]. For $j = 1, \ldots, p$, let X_j be the column j of X. We also remark, from [6], that the vector $v_{M,i}/\|v_{M,i}\|^2$ in (4) is the residual of the regression of X_i with respect to the variables $\{j | j \in M \setminus \{i\}\}$; in other words, it is the component of the vector X_i orthogonal to $\mathrm{Span}\{X_j | j \in M \setminus \{i\}\}$. It is shown in [6] that we have, with probability larger than $1 - \alpha$,

$$\forall M \in \mathcal{M}, \quad \forall i \in M, \quad (\beta_M)_{i.M} \in \mathrm{CI}_{i,M}. \tag{6}$$

Hence, the PoSI confidence intervals guarantee a simultaneous coverage of all the projection-based regression coefficients, over all models M in the set $\mathcal{M}$.

For a square symmetric non-negative matrix A, we let

$$\mathrm{corr}(A) = (\mathrm{diag}(A)^\dagger)^{1/2} A (\mathrm{diag}(A)^\dagger)^{1/2},$$

where $\mathrm{diag}(A)$ is obtained by setting all the non-diagonal elements of A to zero and where $B^\dagger$ is the Moore-Penrose pseudo-inverse of B. Then we show in the following lemma that $K(X, \mathcal{M})$ depends on X only through $\mathrm{corr}(X^t X)$.

Lemma 2.2. *Let X and Z be two $n \times p$ and $m \times p$ matrices satisfying the relation* $\mathrm{corr}(X^t X) = \mathrm{corr}(Z^t Z)$. *Then* $K(X, \mathcal{M}, \alpha, r) = K(Z, \mathcal{M}, \alpha, r)$.

2.2. Order of magnitude of the PoSI constant

The confidence intervals in (3) are similar in form to the standard confidence intervals that one would use for a single fixed model M and a fixed $i \in M$. For a standard interval, $K(X, \mathcal{M})$ would be replaced by a standard Gaussian or Student quantile. Of course, the standard intervals do not account for multiplicity and do not have uniform coverage over $i \in M \in \mathcal{M}$ (see [1, 2]). Hence $K(X, \mathcal{M})$ is the inflation factor or correction over standard intervals to get uniform coverage; it must go to infinity as $p \to \infty$ [6]. Studying the asymptotic order of magnitude of $K(X, \mathcal{M})$ is thus an important problem, as this order of magnitude corresponds to the price one has to pay in order to obtain universally valid post model selection inference.

We now present the existing results on the asymptotic order of magnitude of $K(X, \mathcal{M})$. Let us define

$$\gamma_{\mathcal{M},\infty} := \max_{M \in \mathcal{M}, i \in M} \left| w_{M,i}^t \xi \right|, \tag{7}$$

so that $\gamma_{\mathcal{M},r} = \gamma_{\mathcal{M},\infty}/N$, where we recall that rN^2 follows a chi-square distribution with r degrees of freedom.

We can relate the quantiles of $\gamma_{\mathcal{M},r}$ (which coincide with the PoSI constants $K(X, \mathcal{M})$) to the expectation $\mathbb{E}[\gamma_{\mathcal{M},\infty}]$ by the following argument based on Gaussian concentration (see Appendix A):

Proposition 2.3. *Let $T(\mu, r, \alpha)$ denote the α-quantile of a noncentral T distribution with r degrees of freedom and noncentrality parameter μ. Then*

$$K(X, \mathcal{M}, \alpha, r) \leq T(\mathbb{E}[\gamma_{\mathcal{M},\infty}], r, 1 - \alpha/2).$$

To be more concrete, we observe that we can get a rough estimate of the latter quantile via

$$T(\mathbb{E}[\gamma_{\mathcal{M},\infty}], r, 1 - \alpha/2) \leq \frac{\mathbb{E}[\gamma_{\mathcal{M},\infty}] + \sqrt{2\log(4/\alpha)}}{(1 - 2\sqrt{2\log(4/\alpha)/r})_+};$$

furthermore, as $r \to +\infty$, this quantile reduces to the $(1 - \alpha/2)$ quantile of a Gaussian distribution with mean $\mathbb{E}[\gamma_{\mathcal{M},\infty}]$ and unit variance.

The point of the above estimate is that the dependence in the set of models $\mathcal{M}$ is only present through $\mathbb{E}[\gamma_{\mathcal{M},\infty}]$. Therefore, we will focus in this paper on the problem of bounding $\mathbb{E}[\gamma_{\mathcal{M},\infty}]$, which is nothing more than the Gaussian width [15, chapter 9] of the set $\Gamma_{\mathcal{M}} = \{\pm w_{M,i} | M \in \mathcal{M}, i \in M\}$.

When $n \geq p$, it is shown in [6] that $\mathbb{E}[\gamma_{\mathcal{M},\infty}]$ is no smaller than $\sqrt{2\log(2p)}$ and asymptotically no larger than $\sqrt{p}$. These two lower and upper bound are reached by respectively orthogonal design matrices and equi-correlated design matrices (see [6]).

We now concentrate on s-sparse models. For $s \leq p$, let us define $\mathcal{M}_s = \{M | M \subset \{1, \ldots, p\}, |M| \leq s\}$. In this case, using a direct argument based on cardinality, one gets the following generic upper bound (proved in Appendix B).

Lemma 2.4. *For any $s, n, p \in \mathbb{N}$, with $s \leq n$, we have*

$$\mathbb{E}[\gamma_{\mathcal{M}_s,\infty}] \leq \sqrt{2s\log(6p/s)}\,. \tag{8}$$

We remark that an asymptotic version of the bound in Lemma 2.4 (as p and s go to infinity) appears in an intermediary version of [32].

3. Upper bound under RIP conditions

3.1. Main result

We recall the definition and a property of the RIP constant $\kappa(X, s)$ associated to a design matrix X and a sparsity condition s given in [15, Chap.6]:

$$\kappa(X, s) = \sup_{|M|\leq s} \left\| X_M^t X_M - I_{|M|} \right\|_{op}. \tag{9}$$

Letting $\kappa = \kappa(X, s)$, we have for any subset $M \subset \{1, \ldots, p\}$ such that $|M| \leq s$:

$$\forall \beta \in \mathbb{R}^{|M|}, \qquad (1-\kappa)_+ \|\beta\|^2 \leq \|X_M\beta\|^2 \leq (1+\kappa)\|\beta\|^2\,. \tag{10}$$

Remark 3.1. The RIP condition may also be stated between norms instead of squared norms in (10). Following [15, Chap.6] we will consider the formulation in terms of squared norms, which is more convenient here.

Since the PoSI constant $K(X, \mathcal{M})$ only depends on $\mathrm{corr}(X^tX)$ (see Lemma 2.2), we shall rather consider the RIP constant associated to $\mathrm{corr}(X^tX)$. We let

$$\delta(X, s) = \sup_{|M|\leq s} \left\| \mathrm{corr}(X_M^t X_M) - I_{|M|} \right\|_{op}. \tag{11}$$

Any upper bound for $\kappa(X, s)$ yields an upper bound for $\delta(X, s)$ as shown in the following lemma.

Lemma 3.2. *Let $\kappa = \kappa(X, s)$. If $\kappa \in [0, 1)$, then*

$$\delta(X, s) \leq \frac{2\kappa}{1-\kappa}.$$

The next theorem is the main result of the paper. It provides a new upper bound on the PoSI constant, under RIP conditions and with sparse submodels. We remark that in this theorem, we do not necessarily assume that $n \geq p$.

Theorem 3.3. *Let X be a $n \times p$ matrix with $n, p \in \mathbb{N}$. Let $\delta = \delta(X, s)$. We have*

$$\mathbb{E}[\gamma_{\mathcal{M}_s,\infty}] \leq \sqrt{2\log(2p)} + 2\delta\left(\frac{\sqrt{1+\delta}}{1-\delta}\right)\sqrt{2s\log(6p/s)}.$$

This upper bound is of the form

$$U_{\mathrm{RIP}}(p, s, \delta) = U_{\mathrm{orth}}(p) + 2\delta c(\delta) U_{\mathrm{sparse}}(p, s),$$

where:

- $U_{\mathrm{orth}}(p) = \sqrt{2\log(2p)}$ is the upper bound in the orthogonal case;
- $U_{\mathrm{sparse}}(p,s)$ is the right-hand side of (8) corresponding to the cardinality-based upper bound in the sparse case;
- $c(\delta) = \sqrt{1+\delta}/(1-\delta)$ satisfies: $c(\delta) \geq 0$, $c(\delta) \to 1$ as $\delta \to 0$, and c is increasing.

We observe that if $\delta \to 0$, our bound U_{RIP} is $o(U_{\mathrm{sparse}})$. Moreover, when $\delta\sqrt{s}\sqrt{1 - \log s/\log p + 1/\log p} \to 0$, then U_{RIP} is even asymptotically equivalent to U_{orth}. In particular, this is the case if $\delta\sqrt{s} \to 0$.

We now consider the specific case where X is a subgaussian random matrix, that is, X has independent subgaussian entries [15, Definition 9.1]. We discuss in which situations $\delta = \delta(X,s) \to 0$. The estimate of κ in [15, Theorem 9.2] combined with Lemma 3.2 yields

$$\delta = O_P\left(\sqrt{s\log(ep/s)/n}\right), \tag{12}$$

so that $\delta \to 0$ as soon as $n/(s\log(ep/s)) \to +\infty$.

3.2. Comparison with upper bounds based on Euclidean norms

We now compare our upper bound in Theorem 3.3 to upper bounds recently and independently obtained in [18]. Recall the notation Y, μ, β_M and $\hat{\beta}_M$ from Section 2 and let $r = \infty$ for simplicity of exposition. The authors in [18] address the case where X is random (random design) and consider deviations of $\widehat{\beta}_M$ to $\bar{\beta}_M = \mathbb{E}[X_M^t X_M]^{-1}\mathbb{E}[X_M^t Y]$, the population version of the regression coefficients β_M, assuming that the rows of X are independent random vectors in dimension p. They derive uniform bounds over $M \in \mathcal{M}_s$ for $\|\bar{\beta}_M - \hat{\beta}_M\|_2$. They also consider briefly (Remark 4.3 in [18]) the fixed design case with $\beta_M = (X_M^t X_M)^{-1} X_M^t \mu$ as in the present paper. This target β_M can be interpreted as the random design model conditional to X. They assume that the individual coordinates of X and Y have exponential moments bounded by a constant independently from n, p (thus their setting is more general than the Gaussian regression setting, but for the purpose of this discussion we assume Gaussian noise).

Let us additionally assume that the RIP property $\kappa(X/\sqrt{n}, s) \leq \kappa$ is satisfied (on an event of probability tending to 1) and for κ restricted to a compact of $[0,1)$ independently of n, p; note that we used the rescaling of X by $\sqrt{n}$, which is natural in the random design case. Then some simple estimates obtained as a consequence of Theorems[1] 3.1 and 4.1 in [18] lead to

$$\sup_{M \in \mathcal{M}_s} \|\beta_M - \hat{\beta}_M\|_2 = O_P\left(\sigma\sqrt{\frac{s\log(ep/s)}{n}}\right), \tag{13}$$

[1]The technical conditions assumed by [18] imply a slightly weaker version of the RIP property $\kappa(X/\sqrt{n}, s) \leq \kappa < 1$.

as $p, n \to \infty$ and assuming $s \log^2 p = o(n)$. On our side, under the same assumptions we have that

$$\sup_{M \in \mathcal{M}_s, i \in M} \left(\left(\frac{X_M^t X_M}{n} \right)^{-1} \right)_{i.Mi.M}$$

is bounded on an event of probability tending to 1. This leads to $\|v_{i.M}\| = O_P(1/\sqrt{n})$ uniformly for all $M \in \mathcal{M}_s, i \in M$. Hence, from Theorem 3.3, (3), (6), we obtain

$$\sup_{M \in \mathcal{M}_s} \left\| \beta_M - \hat{\beta}_M \right\|_\infty = O_P \left(\sigma \left(\sqrt{\frac{\log(p)}{n}} + \delta \sqrt{\frac{s \log(ep/s)}{n}} \right) \right). \tag{14}$$

Thus, if $\delta = \Omega(1)$, since the Euclidean norm upper bounds the supremum norm, the results of [18] imply ours (at least in the sense of these asymptotic considerations). On the other hand, in the case where $\delta \to 0$, which is the case we are specifically interested in, we obtain a sharper bound (in the weaker supremum norm).

In particular, if X is a subgaussian random matrix (as discussed in the previous section), due to (12) we obtain

$$\sup_{M \in \mathcal{M}_s} \left\| \beta_M - \hat{\beta}_M \right\|_\infty = O_P \left(\sigma \left(\sqrt{\frac{\log(p)}{n}} + \frac{s \log(ep/s)}{n} \right) \right). \tag{15}$$

This improves over the estimate deduced from (13) as soon as $s \log(ep/s) = o(n)$, which corresponds to the case where (13) tends to 0. Conversely, in this situation our bound (15) yields for the Euclidean norm (using $\|w\|_2 \leq \|w\|_0 \|w\|_\infty$):

$$\sup_{M \in \mathcal{M}_s} \left\| \beta_M - \hat{\beta}_M \right\|_2 = O_P \left(\sigma \left(\sqrt{\frac{s \log(p)}{n}} + \frac{s^{3/2} \log(ep/s)}{n} \right) \right). \tag{16}$$

Assuming $s = O(p^\lambda)$ for some $\lambda < 1$ for ease of interpretation, we see that (16) is of the same order as (13) when $s^2 \log(p) = O(n)$, and is of a strictly larger order otherwise. In this sense, it seems that (14) and (13) are complementary to each other since we are using a weaker norm, but obtain a sharper bound in the case $\delta \to 0$.

3.3. Applicability

While the main interest of our results is theoretical, we now discuss the applicability of our bound. For any $\delta \geq \delta(X, s)$, Theorem 3.3 combined with Proposition 2.3 provides a bound of the form $\overline{U}_{\mathrm{RIP}}(p, s, \delta) \geq K(X, \mathcal{M}_s)$, with

$$\overline{U}_{\mathrm{RIP}}(p, s, \delta) = T \left(\sqrt{2 \log(2p)} + 2\delta \left(\frac{\sqrt{1+\delta}}{1-\delta} \right) \sqrt{2s \log(6p/s)}, r, 1 - \alpha/2 \right).$$

This bound can be used in practice in situations where $\delta(X,s)$ (or an upper bound of it) can be computed, whereas $K(X,\mathcal{M}_s)$ cannot because the number of inner products in (5) is too large. Indeed, for a given δ, it is immediate to compute $\overline{U}_{\mathrm{RIP}}(p,s,\delta)$.

Upper bounding the RIP constant When $n \geq p$, we have $\delta(X,s) \leq \delta(X,p)$ and $\delta(X,p)$ can be computed in practice for a given X. Specifically, $\delta(X,p)$ is the largest eigenvalue of $\mathrm{corr}(X^tX) - I_p$ in absolute value. When X is a subgaussian random matrix, $\delta(X,p) \sim \sqrt{p/n}$ [3, 24]. Thus, if n is large enough compared to p, the computable upper bound $\overline{U}_{\mathrm{RIP}}(p,s,\delta(X,p))$ will improve on the sparsity-based upper bound $\overline{U}_{\mathrm{sparse}}(p,s) = T((2s\log(6p/s))^{1/2}, r, 1-\alpha/2) \geq K(X,\mathcal{M}_s)$, see Proposition 2.3 and Lemma 2.4.

On the other hand, when $n < p$, it is typically too costly to compute $\delta(X,s)$ (or an upper bound of it) for a large p. Nevertheless, if one knows that X is a subgaussian random matrix, they can compute an upper bound $\tilde{\delta}$ satisfying $\tilde{\delta} \geq \delta(X,s)$ with high probability, as in [15, Chapter 9]. We remark that using the values of $\tilde{\delta}$ currently available in the literature, one would need n to be very large for $\overline{U}_{\mathrm{RIP}}(p,s,\tilde{\delta})$ to improve on $\overline{U}_{\mathrm{sparse}}(p,s)$.

Alternative upper bound on the PoSI constant For any $\delta \geq \delta(X,s)$, we now show how to compute an alternative bound of the form $\tilde{U}_{\mathrm{RIP}}(p,s,\delta) \geq K(X,\mathcal{M}_s)$. Our numerical experiments suggest that this alternative bound is generally sharper than $\overline{U}_{\mathrm{RIP}}(p,s,\delta)$. For $q,r,\rho \in \mathbb{N}$ and $\ell \in (0,1)$, let $B_\ell(q,r,\rho)$ be defined as the smallest $t > 0$ so that

$$\mathcal{H}_{q,\rho}(t) := \mathbb{E}_G\big(\min\big(1, \rho\big[1 - F_{\mathrm{Beta},1/2,(q-1)/2}(t^2/G^2)\big]\big)\big) \leq \ell,$$

where G^2/q follows a Fisher distribution with q and r degrees of freedom, and $F_{\mathrm{Beta},a,b}$ denotes the cumulative distribution function of the Beta(a,b) distribution. In the case $r = +\infty$, B_ℓ is also defined and further described in [2, Section 2.5.2].

It can be seen from the proof of Theorem 3.3 (see specifically (22) which also holds without the expectation operators), and from the arguments in [1], that we have

$$K(X,\mathcal{M}_s,\alpha) \leq B_{t\alpha}(n \wedge p, r, p) + 2\delta c(\delta) B_{(1-t)\alpha}(n \wedge p, r, |\mathcal{M}_s|)$$

for any $t \in (0,1)$. This upper bound can be minimized with respect to t, yielding $\tilde{U}_{\mathrm{RIP}}(p,s,\delta)$.

The quantity $B_\ell(q,r,\rho)$ can be easily approximated numerically, as it is simply the quantile of the tail distribution $\mathcal{H}_{q,\rho}$, which only involves standard distributions. Algorithm E.3 in the supplementary materials of [1] can be used to compute $B_\ell(q,r,\rho)$. An implementation of this algorithm in R [26] is available in Appendix C. Hence, the upper bound $\tilde{U}_{\mathrm{RIP}}(p,s,\delta)$ can be computed for large values of p for a given δ.

4. Lower bound

4.1. Equi-correlated design matrices

The goal of this section is to find a matching lower bound for Theorem 3.3. For this we extend ideas of [6, Example 6.2] and, following that reference, we restrict our study to design matrices X for which $n \geq p$. The lower bound is based on the $p \times p$ matrix $Z^{(c,k)} = (e_1^p, e_2^p, \ldots, e_{p-1}^p, x_k(c))$, where

$$x_k(c) = (\underbrace{c, c, \ldots c}_{k}, \underbrace{0, 0, \ldots 0}_{p-1-k}, \underbrace{\sqrt{1-kc^2}}_{1})^t,$$

where we assume $k < p$, and the constant c satisfies $c^2 < 1/k$, so that $Z^{(c,k)}$ has full rank. By definition, the correlation between any of the first k columns of $Z^{(c,k)}$ and the last one is c, and $Z^{(c,k)}$ restricted to its first $p-1$ columns is the identity matrix I_{p-1}. The case where $k = p-1$ is studied in [6, Example 6.2]: Theorem 6.2 in [6] implies that the PoSI constant $K(X, \mathcal{M})$, where X is a $n \times p$ matrix such that $X^t X = (Z^{(c,k)})^t Z^{(c,k)}$, is of the order of $\sqrt{p}$ when $k = p-1$ and $\mathcal{M} = \mathcal{M}_{all}$. The Gram matrix of $Z^{(c,k)}$ is the 3×3 block matrix with sizes $(k, p-k-1, 1) \times (k, p-k-1, 1)$ defined by

$$(Z^{(c,k)})^t Z^{(c,k)} = \begin{bmatrix} I_k & [0] & [c] \\ [0] & I_{p-k-1} & [0] \\ [c] & [0] & 1 \end{bmatrix}, \tag{17}$$

where $[a]$ means that all the entries of the corresponding block are identical to a. We begin by studying the RIP coefficient $\delta(X, s)$ for design matrices X yielding the Gram matrix (17). Since this Gram matrix has full rank p, there exists a design matrix satisfying this condition if and only if $n \geq p$.

Lemma 4.1. *Let X be a $n \times p$ matrix for which $X^t X$ is given by (17) with $kc^2 < 1$. Then for $s \leq k \leq p-1$, we have $\kappa(X, s) = \delta(X, s) \leq c\sqrt{s-1}$.*

4.2. A matching lower bound

In the following proposition, we provide a lower bound of $K(X, \mathcal{M}_s)$ for matrices X yielding the Gram matrix (17).

Proposition 4.2. *For any $s \leq k < p$, $c^2 < 1/k$ and $\alpha \leq \frac{1}{2}$, let X be a $n \times p$ matrix for which $X^t X$ is given by (17) with $kc^2 < 1$. We have*

$$K(X, \mathcal{M}_s, \alpha, \infty) \geq A \frac{c(s-1)}{\sqrt{1-(s-1)c^2}} \sqrt{\log\lfloor k/s \rfloor} - \sqrt{2\log 2},$$

where $A > 0$ is a universal constant.

From the previous lemma, we now show that the upper bound of Theorem 3.3 is optimal (up to a multiplicative constant) for a large range of behavior of s and δ relatively to p. As discussed after Theorem 3.3, in the case where $\delta\sqrt{s}\sqrt{1-\log s/\log p+1/\log p}=O(1)$, the upper bound we obtain is optimal, since it can be written as $O(\sqrt{\log p})$. In the next Corollary, we show that the upper bound of Theorem 3.3 is also optimal when $\delta\sqrt{s}\sqrt{1-\log s/\log p+1/\log p}$ tends to $+\infty$, and when $\delta=O(p^{-\lambda})$ for some $\lambda>0$.

Corollary 4.3 (Optimality of the RIP-PoSI bound)**.** *Let $(s_p,\delta_p)_{p\geq 0}$ be sequences of values such that $s_p<p$, $\delta_p>0$, $\delta_p\to 0$ and satisfying:*

$$\lim_{p\to\infty}\delta_p\sqrt{s_p}\sqrt{1-\log s_p/\log p+1/\log p}=+\infty.$$

Then Theorem 3.3 implies

$$\sup_{\substack{n\in\mathbb{N}\\ s\leq s_p, X\in\mathbb{R}^{n\times p}\\ s.t.\ \delta(X,s)\leq\delta_p}} K(X,\mathcal{M}_{s_p})\leq B\delta_p\sqrt{s_p}\sqrt{\log(6p/s_p)}, \tag{18}$$

where B is a constant. Moreover, there exists a sequence of design matrices X_p such that $\delta(X_p,s_p)\leq\delta_p$ and

$$K(X_p,\mathcal{M}_{s_p})\geq A\delta_p\sqrt{s_p}\sqrt{\log\left(\min(1/\delta_p^2,\lfloor(p-1)/s_p\rfloor)\right)}, \tag{19}$$

where A is a constant.

In particular, if $\delta_p=O(p^{-\lambda})$ for some $\lambda>0$ and if $\lfloor(p-1)/s_p\rfloor\geq 2$, then the above upper and lower bounds have the same rate.

Therefore, the upper bound in Theorem 3.3 is optimal in most configurations of s_p and δ_p, except if δ_p goes to 0 slower than any inverse power of p.

5. Concluding remarks

In this paper, we have proposed an upper bound on PoSI constants in s-sparse situations where the $n\times p$ design matrix X satisfies a RIP condition. As the value of the RIP constant δ increases from 0, this upper bound provides an interpolation between the case of an orthogonal X and an existing upper bound only based on sparsity and cardinality. We have shown that our upper bound is asymptotically optimal for many configurations of (s,δ,p) by giving a matching lower bound. In the case of random design matrices with independent entries, since δ decreases with n, our upper bound compares increasingly more favorably to the cardinality-based upper bound as n gets larger. It is also complementary to the bounds recently proposed in [18]. The interest and various applications of the RIP property are well-known in the high-dimensional statistics literature, in particular for statistical risk analysis or support recovery. Our analysis puts into light an additional interest of the RIP property for agnostic post-selection inference (uncertainty quantification).

The PoSI constant corresponds to confidence intervals on β_M in (2). In section 3.2 we also mention another target of interest in the case of random X, $\bar{\beta}_M = \mathbb{E}[X_M^t X_M]^{-1}\mathbb{E}[X_M^t Y]$. This quantity depends on the distribution of X rather than on its realization, which is a desirable property as discussed in [1, 18] where the same target has also been considered. In [1], it is shown that valid confidence intervals for β_M are also asymptotically valid for $\bar{\beta}_M$, provided that p is fixed. These results require that μ belongs to the column space of X and hold for models M such that μ is close to the column space of X_M. It would be interesting to study whether assuming RIP conditions on X enables to alleviate these assumptions.

The purpose of post-selection inference based on the PoSI constant $K(X, \mathcal{M})$ is to achieve the coverage guarantee (6). The guarantee (6) implies that, for any model selection procedure $\hat{M} : \mathbb{R}^n \to \mathcal{M}$, with probability larger than $1 - \alpha$, for all $i \in \hat{M}$, $(\hat{M})_{i.\hat{M}} \in \mathrm{CI}_{i,\hat{M}}$. Hence, there is in general no need to make assumptions about the model selection procedure when using PoSI constants. On the other hand, the RIP condition that we study here is naturally associated to specific model selection procedures, namely the lasso or the Dantzig selector [9, 10, 30, 33]. Hence, it is natural to ask whether the results in this paper could help post-selection inference specifically for such procedures. We believe that the answer could be positive in some situations. Indeed, if the lasso model selector is used in conjunction with a design matrix X satisfying a RIP property, then asymptotic guarantees exist on the sparsity of the selected model [8]. Thus, one could investigate the combination of bounds on the size of selected models (of the form $|\hat{M}| \leq S$ and holding with high probability) with our upper bound, by replacing s by S.

In the case of the lasso model selector, we have referred, in the introduction section, to the post-selection intervals achieving conditional coverage [19], specifically for the lasso model selector. These intervals are simple to compute (when the conditioning is on the signs, see [19]). Generally speaking, in comparison with confidence intervals based on PoSI constants, the confidence intervals of [19] have the benefit of guaranteeing a coverage level conditionally on the selected model. On the other hand the confidence intervals in [19] can be large, and can provide small coverage rates when the regularization parameter of the lasso is data-dependent [1]. It would be interesting to study whether these general conclusions would be modified in the special case of design matrices satisfying RIP properties.

Finally, the focus of this paper is on PoSI constants in the context of linear regression. Recently, [2] extended the PoSI approach to more general settings (for instance generalized linear models), provided a joint asymptotic normality property holds between model dependent targets and estimators. This extension was suggested in the case of asymptotics for fixed dimension and fixed number of models. In the high-dimensional case, an interesting direction would be to apply the results of [12], that provide Gaussian approximations for maxima of sums of high-dimensional random vectors. This opens the perspective of applying our results to various high-dimensional post model selection settings, beyond linear regression.

Acknowledgements

This work has been supported by ANR-16-CE40-0019 (SansSouci). The second author acknowledges the support from the german DFG, under the Research Unit FOR-1735 "Structural Inference in Statistics – Adaptation and Efficiency", and under the Collaborative Research Center SFB-1294 "Data Assimilation".

Appendix

Appendix A: Gaussian concentration

To relate the expectation of a supremum of Gaussian variables to its quantiles, we use the following classical Gaussian concentration inequality [13] (see e.g. [16], Section B.2.2. for a short exposition):

Theorem A.1 (Cirel'son, Ibragimov, Sudakov)**.** *Assume that $F : \mathbb{R}^d \to \mathbb{R}$ is a 1-Lipschitz function (w.r.t. the Euclidean norm of its input) and Z follows the $\mathcal{N}(0, \sigma^2 I_d)$ distribution. Then, there exists two one-dimensional standard Gaussian variables ζ, ζ' such that*

$$\mathbb{E}[F(Z)] - \sigma|\zeta'| \leq F(Z) \leq \mathbb{E}[F(Z)] + \sigma|\zeta|. \tag{20}$$

It is known that in certain situations one can expect an even tighter concentration, through the phenomenon known as superconcentration [11]. While such situations are likely to be relevant for the setting considered in this paper, we leave such improvements as an open issue for future work.

We use the previous property in our setting as follows:

Proposition A.2. *Let $\mathcal{C}$ be finite a family of unit vectors of $\mathbb{R}^n$, ξ a standard Gaussian vector in $\mathbb{R}^n$ and N an independent nonnegative random variable so that rN^2 follows a chi-squared distribution with r degrees of freedom. Define the random variable*

$$\gamma_{\mathcal{C},r} := \frac{1}{N} \max_{v \in \mathcal{C}} |v^t \xi|.$$

Then the $(1-\alpha)$ quantile of $\gamma_{\mathcal{C},r}$ is upper bounded by the $(1-\alpha/2)$ quantile of a noncentral T distribution with r degrees of freedom and noncentrality parameter $\mathbb{E}[\max_{v \in \mathcal{C}} |v^t \xi|]$.

Proof. Observe that $\xi \mapsto \max_{v \in \mathcal{C}} |v^t \xi|$ is 1-Lipschitz since the vectors of $\mathcal{C}$ are unit vectors. Therefore we conclude by Theorem A.1 that there exists a standard normal variable ζ (which is independent of N since N is independent of ξ) so that the following holds:

$$\gamma_{\mathcal{C}} \leq \frac{1}{N} \Big(\mathbb{E}\Big[\max_{v \in \mathcal{C}} |v^t \xi| \Big] + |\zeta| \Big).$$

We can represent the above right-hand side as $\max(T_+, T_-)$ where

$$T_\pm = \frac{1}{N} \Big(\mathbb{E}\Big[\max_{v \in \mathcal{C}} |v^t \xi| \Big] \pm \zeta \Big),$$

i.e. T_+, T_- are two (dependent) noncentral t distributions with r degrees of freedom and noncentrality parameter $\mathbb{E}[\max_{v\in\mathcal{C}}|v^t\xi|]$. Finally since

$$\mathbb{P}[\max(T_+, T_-) > t] \leq \mathbb{P}[T_+ > t] + \mathbb{P}[T_- > t] = 2\mathbb{P}[T_+ > t],$$

we obtain the claim. □

Since a noncentral distribution is (stochastically) increasing in its noncentrality parameter, any bound obtained for $\mathbb{E}[\max_{v\in\mathcal{C}}|v^t\xi|]$ will result in a corresponding bound on the quantiles of the corresponding noncentral T distribution and therefore of those of $\gamma_{\mathcal{C}}$. In the limit $r \to \infty$, the quantiles of the noncentral T distribution reduce to those of a shifted Gaussian distribution with unit variance.

Here is a naive bound on (some) quantiles of a noncentral T:

Lemma A.3. *The $1-\alpha$ quantile of a noncentral T distribution with r degrees of freedom and noncentrality parameter $\mu \geq 0$ is upper bounded by:*

$$(\mu + \sqrt{2\log(2/\alpha)}/(1 - 2\sqrt{2\log(2/\alpha)/r})_+.$$

Proof. Let

$$T = \frac{\mu + \zeta}{\sqrt{V/r}},$$

where $\zeta \sim \mathcal{N}(0,1)$ and $V \sim \chi^2(r)$. We have (as a consequence of e.g. [7], Lemma 8.1), for any $\eta \in (0,1]$:

$$\mathbb{P}\left[\sqrt{V} \leq \sqrt{r} - 2\sqrt{2\log\eta^{-1}}\right] \leq \eta,$$

as well as the classical bound

$$\mathbb{P}\left[\zeta \geq \sqrt{2\log\eta^{-1}}\right] \leq \eta.$$

It follows that

$$\mathbb{P}\left[T \geq (\mu + \sqrt{2\log\eta^{-1}})/(1 - 2\sqrt{2\log(\eta^{-1})/r})_+\right] \leq 2\eta.$$

The claimed estimate follows. □

Appendix B: Proofs

Proof of Lemma 2.2. With the notation of Definition 2.1, $K(X, \mathcal{M}, \alpha, r)$ is the $1-\alpha$ quantile of $(1/N)\|z\|_\infty$ where $z = (z_{M,i}, M \in \mathcal{M}, i \in \mathcal{M})$ is a Gaussian vector, independent of N, with mean vector zero and covariance matrix $\mathrm{corr}(\Sigma)$, where Σ is defined by, for $i \in M \in \mathcal{M}$ and $i' \in M' \in \mathcal{M}$,

$$\begin{aligned}\Sigma_{(M,i),(M',i')} &= v_{M,i}^t v_{M',i'} \\ &= (e_{i.M}^{|M|})^t (X_M^t X_M)^{-1} X_M^t X_{M'} (X_{M'}^t X_{M'})^{-1} e_{i'.M'}^{|M'|}.\end{aligned}$$

Hence, Σ depends on X only through X^tX. Also, if X is replaced by XD, where D is a diagonal matrix with positive components, Σ becomes the matrix Λ with for $i \in M \in \mathcal{M}$ and $i' \in M' \in \mathcal{M}$,

$$\begin{aligned}\Lambda_{(M,i),(M',i')} &= (e_{i.M}^{|M|})^t D_{M,M}^{-1}(X_M^tX_M)^{-1}X_M^tX_{M'}(X_{M'}^tX_{M'})^{-1}D_{M',M'}^{-1}e_{i'.M'}^{|M'|}\\ &= D_{i,i}^{-1}D_{i',i'}^{-1}\Sigma_{(M,i),(M',i')}.\end{aligned}$$

Hence, $\mathrm{corr}(\Sigma) = \mathrm{corr}(\Lambda)$. This shows that Σ depends on X only through $\mathrm{corr}(X^tX)$ (we remark that because $\cup_{\mathcal{M}} M = \{1,\dots,p\}$ and each $X_M^tX_M$ is invertible we have that $\|X_i\| > 0$ for $i = 1,\dots,p$). Hence $K(X,\mathcal{M},\alpha,r)$ depends on X only through $\mathrm{corr}(X^tX)$. □

Proof of Lemma 2.4. Using a direct cardinality-based bound we have the well-known inequality $\mathbb{E}[\gamma_{\mathcal{M}_s,\infty}] \le \sqrt{2\log(2|\mathcal{M}_s|)}$, hence

$$\mathbb{E}[\gamma_{\mathcal{M}_s,\infty}] \le \sqrt{2\log\left(2\sum_{i=1}^{s} i\binom{p}{i}\right)},$$

moreover

$$\sum_{i=1}^{s} i\binom{p}{i} \le s\sum_{i=0}^{s}\binom{p}{i} \le s\Big(\frac{pe}{s}\Big)^s,$$

the last inequality being classical and due to

$$\Big(\frac{s}{p}\Big)^s\sum_{i=0}^{s}\binom{p}{i} \le \sum_{i=0}^{s}\Big(\frac{s}{p}\Big)^i\binom{p}{i} \le \Big(1+\frac{s}{p}\Big)^p \le e^s.$$

Since $\log s \le s/e$, and using $e^{1+2/e} \le 6$, we obtain

$$\log\left(2\sum_{i=1}^{s} i\binom{p}{i}\right) \le \log 2s + s\log\Big(\frac{pe}{s}\Big) \le s\log\Big(\frac{p}{s}e^{1+2/e}\Big) \le s\log\left(\frac{6p}{s}\right),$$

implying (8). □

Proof of Lemma 3.2. Put $\kappa = \kappa(X,s) < 1$. Then, $\|X_i\| \ge (1-\kappa)^{1/2}$ for $i = 1,\dots,p$ so that for $i \in M \in \mathcal{M}_s$, $\mathrm{corr}(X_M^tX_M) = D_MX_M^tX_MD_M$ where D_M is a $|M|\times|M|$ matrix defined by $[D_M]_{i.M,i.M} = 1/\|X_i\|$. Hence $\|D_M\|_{op} \le 1/\sqrt{1-\kappa}$. We have, by applications of the triangle inequality and since $\|.\|_{op}$ is a matrix norm,

$$\begin{aligned}&\big\|\mathrm{corr}(X_M^tX_M) - I_{|M|}\big\|_{op}\\ &\quad= \big\|(D_M - I_{|M|})X_M^tX_MD_M + X_M^tX_M(D_M - I_M) + X_M^tX_M - I_{|M|}\big\|_{op}\\ &\quad\le \big\|D_M - I_{|M|}\big\|_{op}\big\|X_M^tX_M\big\|_{op}\|D_M\|_{op} + \big\|D_M - I_{|M|}\big\|_{op}\big\|X_M^tX_M\big\|_{op}\\ &\qquad+ \big\|X_M^tX_M - I_{|M|}\big\|_{op}\\ &\quad= \big\|D_M - I_{|M|}\big\|_{op}\big\|X_M^tX_M\big\|_{op}\Big(\|D_M\|_{op}+1\Big) + \big\|X_M^tX_M - I_{|M|}\big\|_{op}. \qquad (21)\end{aligned}$$

From (9)-(10), we have for all $M \in \mathcal{M}_s$: $\|X_M^t X_M\|_{op} \leq 1+\kappa$, as well as

$$\begin{aligned}\left\|D_M - I_{|M|}\right\|_{op} &\leq \max_{i=1,\ldots,p}\left|\frac{1}{\|X_i\|} - 1\right| \\ &\leq \max\left(1 - \frac{1}{\sqrt{1+\kappa}}, \frac{1}{\sqrt{1-\kappa}} - 1\right) \\ &= \frac{1}{\sqrt{1-\kappa}} - 1.\end{aligned}$$

Plugging this into (21), we obtain

$$\begin{aligned}\delta(X,s) &\leq \left(\frac{1}{\sqrt{1-\kappa}} - 1\right)(1+\kappa)\left(\frac{1}{\sqrt{1-\kappa}} + 1\right) + \kappa \\ &= \frac{2\kappa}{1-\kappa}.\end{aligned}$$

□

Proof of Theorem 3.3. From Lemma 2.2, it is sufficient to treat the case where, for any M, $G_M = X_M^t X_M$ has ones on the diagonal; in that case $\delta(X,s) = \kappa(X,s)$. We have

$$\begin{aligned}v_{M,i}^t &= (e_{i.M}^{|M|})^t G_M^{-1} X_M^t \\ &= (e_{i.M}^{|M|})^t I_{|M|} X_M^t + (e_{i.M}^{|M|})^t \left(G_M^{-1} - I_{|M|}\right) X_M^t \\ &= X_i^t + r_{M,i}^t,\end{aligned}$$

say. We have

$$\begin{aligned}r_{M,i}^t r_{M,i} &= (e_{i.M}^{|M|})^t \left(G_M^{-1} - I_{|M|}\right) G_M \left(G_M^{-1} - I_{|M|}\right) e_{i.M}^{|M|} \\ &\leq \left\|e_{i.M}^{|M|}\right\|^2 \left\|G_M^{-1} - I_{|M|}\right\|_{op}^2 \|G_M\|_{op}.\end{aligned}$$

From (10), the eigenvalues of G_M are all between $(1-\delta)$ and $(1+\delta)$, hence we have

$$r_{M,i}^t r_{M,i} \leq \left(\frac{\delta}{1-\delta}\right)^2 (1+\delta),$$

so that letting $c(\delta) = \sqrt{1+\delta}/(1-\delta)$

$$\|r_{M,i}\| \leq \delta c(\delta),$$

and

$$\begin{aligned}\|w_{M,i} - X_i\| = \left\|\frac{v_{M,i}}{\|v_{M,i}\|} - X_i\right\| &= \left\|\frac{v_{M,i}}{\|v_{M,i}\|}(1 - \|v_{M,i}\|) + v_{M,i} - X_i\right\| \\ &\leq 2\|r_{M,i}\|,\end{aligned}$$

from two applications of the triangle inequality, and using that $\|X_i\| = 1$ since we assumed that G_M has ones on its diagonal for all M. Hence, we have

$$\begin{aligned}
\mathbb{E}[\gamma_{\mathcal{M}_s,\infty}] &= \mathbb{E}\left[\sup_{M\in\mathcal{M}_s; i\in M} |w_{M,i}^t \xi|\right] \\
&\leq \mathbb{E}\left[\sup_{M\in\mathcal{M}_s; i\in M} |X_i^t \xi|\right] + \mathbb{E}\left[\sup_{M\in\mathcal{M}_s; i\in M} \left|(w_{M,i} - X_i)^t \xi\right|\right] \\
&\leq \mathbb{E}\left[\sup_{i=1,\ldots,p} |X_i^t \xi|\right] \\
&\quad + 2\delta c(\delta)\mathbb{E}\left[\sup_{M\in\mathcal{M}_s; i\in M} \left|\left(\frac{w_{M,i} - X_i}{\|w_{M,i} - X_i\|}\right)^t \xi\right|\right] \qquad (22) \\
&\leq \sqrt{2\log(2p)} + 2\delta c(\delta)\sqrt{2s\log(6p/s)},
\end{aligned}$$

where in the last step we have used Lemma 2.4. □

Proof of Lemma 4.1. Since $\|X_i\| = 1$ for $i = 1, ..., p$ we have $\mathrm{corr}(X^t X) = X^t X$ and so $\kappa(X, s) = \delta(X, s)$. The Gram matrix in (17) can be written as $I_p + cU_{p,k}$, where $U_{p,k}$ is the 3×3 block matrix with sizes $(k, p-k-1, 1) \times (k, p-k-1, 1)$ defined by

$$U_{p,k} = \begin{bmatrix} [0] & [0] & [1] \\ [0] & [0] & [0] \\ [1] & [0] & 0 \end{bmatrix}.$$

Consider a model M with $|M| = s \leq k \leq p-1$, and denote by G_M its Gram matrix. If $p \notin M$, then $G_M = I_s$ and $\|G_M - I_s\|_{op} = 0$. If $p \in M$, then $G_M = I_s + cU_{s,m}$, where $m = m(M) = |(M \setminus \{p\}) \cap \{1, \ldots k\}| \leq s-1$. The operator norm of $G_M - I_s$ is the square root of the largest eigenvalue of $(cU_{s,m})^2$, where $U_{s,m}^2$ is a 3×3 block matrix with sizes $(m, s-m-1, 1) \times (m, s-m-1, 1)$ defined by

$$U_{s,m}^2 = \begin{bmatrix} [1] & [0] & [0] \\ [0] & [0] & [0] \\ [0] & [0] & m \end{bmatrix}.$$

The first block is a $m \times m$ matrix with all entries equal to 1, hence its only non-null eigenvalue is m. This is also the (only) eigenvalue of the last block (an 1×1 matrix). Thus, the largest eigenvalue of $U_{s,m}^2$ is m. Therefore, as $m \leq s-1$, we have $\|G_M - I_s\|_{op} = c\sqrt{s-1}$ for all M such that $|M| = s \leq k \leq p-1$, which concludes the proof. □

Proof of Proposition 4.2. Without loss of generality (by Lemma 2.2) we can assume that $X = Z^{(c,k)}$, where $Z^{(c,k)}$ is the $p \times p$ matrix defined as the beginning of Section 4.1. The proof is an extension of the proof of [6, Theorem 6.2]. For $m \geq 0$, consider a model M such that $M \ni p$, $M \cap \{k+1, \ldots, p-1\} = \emptyset$, and $|M| = m+1$; in other words, $M = \{i_1, \ldots, i_m, p\}$ such that $i_1, \ldots, i_m$ are elements of $\{1, \ldots, k\}$. Denote as $\mathcal{M}_{m:k}^{+p}$ the set of all such models. Let

$u_{M,p} = Z_p - P_{M\setminus\{p\}}(Z_p)$, where Z_p is the last column of $Z^{(c,k)}$, and where $P_{M\setminus\{p\}}(Z_p)$ is the orthogonal projection of Z_p onto the span of the columns with indices $M \setminus \{p\}$. Observe that the column i_j of $Z^{(c,k)}$ is the i_j-th base column vector of $\mathbb{R}^p$ that we write e_{i_j}, therefore

$$P_{M\setminus\{p\}}(Z_p) = \sum_{j=1}^{m} (e_{i_j}^t Z_p) e_{i_j} = c(e_{i_1} + \ldots + e_{i_m}).$$

Hence, we have, for $M \in \mathcal{M}_{m:k}^{+p}$,

$$\left[u_{M,p}\right]_j = \begin{cases} 0 & \text{for } j = k+1, \ldots, p-1, \\ 0 & \text{for } j = 1, \ldots, k; j \in M, \\ c & \text{for } j = 1, \ldots, k; j \notin M, \\ \sqrt{1-kc^2} & \text{for } j = p. \end{cases}$$

Recall that we have $w_{M,p} = u_{M,p}/\|u_{M,p}\|$. Hence, for $M \in \mathcal{M}_{m:k}^{+p}$,

$$\left[w_{M,p}\right]_j = \begin{cases} 0 & \text{for } j = k+1, \ldots, p-1, \\ 0 & \text{for } j = 1, \ldots, k; j \in M, \\ c/\sqrt{1-mc^2} & \text{for } j = 1, \ldots, k; j \notin M, \\ \sqrt{1-kc^2}/\sqrt{1-mc^2} & \text{for } j = p. \end{cases}$$

Hence, we have

$$\begin{aligned} \mathbb{E}[\gamma_{\mathcal{M}_s,\infty}] &= \mathbb{E}\left[\max_{|M|\le s, i\in M} |w_{M,i}^t \xi|\right] \\ &\ge \mathbb{E}\left[\max_{M\in\mathcal{M}_{(s-1):k}^{+p}} w_{M,p}^t \xi\right] \\ &= \mathbb{E}\left[\frac{\sqrt{1-kc^2}}{\sqrt{1-(s-1)c^2}}\xi_p + \frac{c}{\sqrt{1-(s-1)c^2}} \sum_{j=1}^{k-s+1} \xi_{k-j:k}\right], \end{aligned}$$

where $\xi_{1:k} \le \ldots \le \xi_{k:k}$ are the order statistics of $\xi_1, \ldots, \xi_k$. Hence, since $s-1 < k$, we obtain

$$\begin{aligned} \mathbb{E}[\gamma_{\mathcal{M}_s,\infty}] &\ge 0 + \frac{c}{\sqrt{1-(s-1)c^2}} \mathbb{E}\left[\sum_{j=1}^{k} \xi_j - \sum_{j=1}^{s-1} \xi_{j:k}\right] \\ &= \frac{c}{\sqrt{1-(s-1)c^2}} \mathbb{E}\left[\sum_{j=1}^{s-1} \xi_{k-j:k}\right] \\ &\ge \frac{c}{\sqrt{1-(s-1)c^2}} \mathbb{E}\left[\sum_{j=1}^{s-1} \max_{l=1,\ldots,\lfloor k/s\rfloor} \xi_{(j-1)\lfloor k/s\rfloor + l}\right]. \end{aligned}$$

In the above display, each maximum has mean value larger than $A\sqrt{\log\lfloor k/s\rfloor}$, with $A>0$ a universal constant (see e.g. Lemma A.3 in [11]). Hence, we have

$$\mathbb{E}[\gamma_{\mathcal{M}_s,\infty}] \geq A\frac{c(s-1)}{\sqrt{1-(s-1)c^2}}\sqrt{\log\lfloor k/s\rfloor}.$$

Finally, a consequence of Gaussian concentration (Theorem A.1) is that mean and median of $\gamma_{\mathcal{M}_s,\infty}$ are within $\sqrt{2\log 2}$ of each other. Since we assumed $\alpha\leq\frac{1}{2}$, $K(Z^{(c,k)},\mathcal{M}_s,\alpha,\infty)\geq\mathbb{E}[\gamma_{\mathcal{M}_s,\infty}]-\sqrt{2\log 2}$, which concludes the proof. □

Proof of Corollary 4.3. When $\delta_p\sqrt{s_p}\sqrt{1-\log s_p/\log p+1/\log p}\to\infty$, one can see that in Theorem 3.3, the first term is negligible compared to the second one. Since $\delta_p\to 0$, the first result (18) follows from Theorem 3.3.

We now apply Proposition 4.2 with $c_p=\delta_p/\sqrt{s_p-1}$ and $k_p=\min(p-1,\lfloor 1/c_p^2-1\rfloor)$. From Lemma 4.1, $\delta(Z^{(c_p,k_p)},s_p)\leq c_p\sqrt{s_p-1}=\delta_p$. We then have, with two positive constants A' and A,

$$\begin{aligned}K(Z^{(c_p,k_p)},\mathcal{M}_s,\alpha,\infty)\geq&A'\delta_p\sqrt{s_p}\sqrt{\log\left(\left\lfloor\frac{\min(p-1,\lfloor 1/c_p^2-1\rfloor)}{s_p}\right\rfloor\right)}\\ \geq&A\delta_p\sqrt{s_p}\sqrt{\log\big(\min(\lfloor(p-1)/s_p\rfloor,1/\delta_p^2)\big)}.\end{aligned}$$

This concludes the proof of (19). □

Appendix C: Code for computing $B_\ell(q,r,\rho)$

```
Bl <- function(q, r, rho, l, I = 1000) {
    ##
    ## Compute an upper bound for the quantile 1-l of
    ## max_{i=1,...,rho} (1/N) | w_i' V |
    ## where:
    ##   - the w_1,...w_{rho} are unit vectors
    ##   - V follows N(0,I_q)
    ##   - N^2/r follows X^2(r)
    ##
    ## Adapted from K4 in Bachoc, Leeb, Poetscher 2018
    ##
    ## Parameters:
    ## q.......: dimension of the Gaussian vector
    ## r.......: degrees of freedom for the variance estimator
    ## rho.....: number of unit vectors
    ## l.......: type I error rate (1 - confidence level)
    ## I.......: numerical precision
    ##
    ## Value:
    ##   A numerical approximation of the upper bound
```

```
    ##
    ##  vector of quantiles of Beta distribution:
    vC <-  qbeta(p = seq(from = 0, to = 1/rho, length = I),
                 shape1 = 1/2, shape2 = (q-1)/2,
                 lower.tail = FALSE)
    ## Monte-Carlo evaluation of confidence level
    ## for a constant K
    fconfidence <- function(K){
        prob <- pf(q = K^2/vC/q, df1 = q,
                   df2 = r, lower.tail = FALSE)
        mean(prob) - l
    }
    quant <- qf(p = l, df1 = q, df2 = r, lower.tail = FALSE)
    Kmax <- sqrt(quant) * sqrt(q)
    uniroot(fconfidence, interval = c(1, 2*Kmax))$root
}
```

References

[1] F. Bachoc, H. Leeb, and B. M. Pötscher. Valid confidence intervals for post-model-selection predictors. *The Annals of Statistics (forthcoming)*, 2018.

[2] F. Bachoc, D. Preinerstorfer, and L. Steinberger. Uniformly valid confidence intervals post-model-selection. arXiv:1611.01043, 2016.

[3] Z. Bai and J. W. Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010. MR2567175

[4] A. Belloni, V. Chernozhukov, and C. Hansen. Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics. 10th World Congress of the Econometric Society, Volume III*, pages 245–295, 2011.

[5] A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, **81**:608–650, 2014. MR3207983

[6] R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013. MR3099122

[7] L. Birgé. An alternative point of view on Lepski's method. In *State of the art in probability and statistics (Leiden, 1999)*, volume 36 of *IMS Lecture Notes Monogr. Ser.*, pages 113–133. Inst. Math. Statist., 2001. MR1836557

[8] P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011. MR2807761

[9] E. Candes, T. Tao, et al. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007. MR2382644

[10] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005. MR2243152
[11] S. Chatterjee. *Superconcentration and related topics*. Springer, 2014. MR3157205
[12] V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819, 2013. MR3161448
[13] B. S. Cirel'son, I. A. Ibragimov, and V. N. Sudakov. Norm of Gaussian sample functions. In *Proceedings of the 3rd Japan-U.S.S.R. Symposium on Probability Theory (Tashkent, 1975)*, volume 550 of *Lecture Notes in Mathematics*, pages 20–41. Springer, 1976. MR0458556
[14] W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection. arXiv:1410.2597, 2015.
[15] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. Basel: Birkhäuser, 2013. MR3100033
[16] C. Giraud. *Introduction to high-dimensional statistics.*, volume 139 of *Monographs on Statistics and Applied Probability*. CRC Press, 2015. MR3307991
[17] P. Kabaila and H. Leeb. On the large-sample minimal coverage probability of confidence intervals after model selection. *Journal of the American Statistical Association*, 101:619–629, 2006. MR2256178
[18] A. K. Kuchibhotla, L. D. Brown, A. Buja, E. I. George, and L. Zhao. A model free perspective for linear regression: Uniform-in-model bounds for post selection inference. arXiv:1802.05801, 2018.
[19] J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016. MR3485948
[20] J. D. Lee and J. E. Taylor. Exact post model selection inference for marginal screening. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 136–144. Curran Associates, Inc., 2014.
[21] H. Leeb and B. M. Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21:21–59, 2005. MR2153856
[22] H. Leeb and B. M. Pötscher. Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory*, 22:69–97, 2 2006. MR2212693
[23] H. Leeb and B. M. Pötscher. Model selection. In T. G. Andersen, R. A. Davis, J.-P. Kreiß, and T. Mikosch, editors, *Handbook of Financial Time Series*, pages 785–821, New York, NY, 2008. Springer.
[24] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967. MR0221550
[25] B. M. Pötscher. Confidence sets based on sparse estimators are necessarily large. *Sankhya*, 71:1–18, 2009. MR2579644
[26] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[27] R. J. Tibshirani, A. Rinaldo, R. Tibshirani, and L. Wasserman. Uniform asymptotic inference and the bootstrap after model selection. *The Annals of Statistics, forthcoming*, 2015. MR3798003

[28] R. J. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016. MR3538689

[29] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, **42**:1166–1202, 2014. MR3224285

[30] S. A. Van De Geer, P. Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009. MR2576316

[31] C.-H. Zhang and S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society B*, **76**:217–242, 2014. MR3153940

[32] K. Zhang. Spherical cap packing asymptotics and rank-extreme detection. *IEEE Transactions on Information Theory*, 63(7), 2017. MR3666977

[33] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563, 2006. MR2274449