Electronic Journal of Statistics
Vol. 11 (2017) 2519–2546
ISSN: 1935-7524
DOI: 10.1214/17-EJS1287

Optimal two-step prediction in regression

Didier Chételat

Department of Decision Sciences
HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal, Canada
e-mail: didier.chetelat@hec.ca

Johannes Lederer

Departments of Statistics and Biostatistics
University of Washington
Box 354322
Seattle, WA 98195-4322
e-mail: ledererj@uw.edu

and

Joseph Salmon

LTCI, Télécom ParisTech,
Université Paris-Saclay,
75013, Paris, France
e-mail: first.last@telecom-paristech.fr

Abstract: High-dimensional prediction typically comprises two steps: variable selection and subsequent least-squares refitting on the selected variables. However, the standard variable selection procedures, such as the lasso, hinge on tuning parameters that need to be calibrated. Cross-validation, the most popular calibration scheme, is computationally costly and lacks finite sample guarantees. In this paper, we introduce an alternative scheme, easy to implement and both computationally and theoretically efficient.

MSC 2010 subject classifications: Primary 62G08; secondary 62J07.
Keywords and phrases: High-dimensional prediction, tuning parameter selection, lasso.

Received May 2016.

Contents

1. Introduction

Variable selection has become a basic tool for estimating linear models on large data sets. The most popular method for variable selection is the lasso [37], which minimizes the sum of squares errors under an ℓ_1-penalty. Although efficient at selecting variables when properly tuned, the lasso has the disadvantage that all coefficients are shrunk towards zero. To mitigate this bias, practitioners typically rely on a two-step estimation of the coefficients by computing a least-squares estimate on the variables selected by the lasso.

For illustration, consider the leukemia micro-array data set of [16], which consists of $n = 38$ bone marrow samples analyzed with $p = 7129$ probes from several thousand human genes. A particular interest is to predict the type of leukemia (AML or ALL) present in a patient. The data set also contains an independent test set of 34 observations that are used for assessment of the predictive performance.

In this problem, there are many more variables (7129 features) than available observations (38 samples), and in such a context, a least-squares fitting is not appropriate. A standard solution is to perform variable selection using the lasso, with tuning parameter chosen by cross-validation on the prediction loss. However, since the lasso is known to involve a bias, practitioners commonly refit a least-squares estimate on the selected variables. If the lasso tuning parameter is chosen using 10-fold cross-validation, this approach, called lassoCV in the following, yields a prediction risk of 0.36 on the test set, computed in 463 seconds.

Although common among practitioners, this approach is suboptimal, because the cross-validation does not take into account the least-squares refitting. Another alternative is to tune the cross-validation for the entire two-step procedure. On the test set, this approach with 10-fold cross-validation, called lslassoCV in the following, yields a prediction risk of 0.45 computed in 499 seconds.

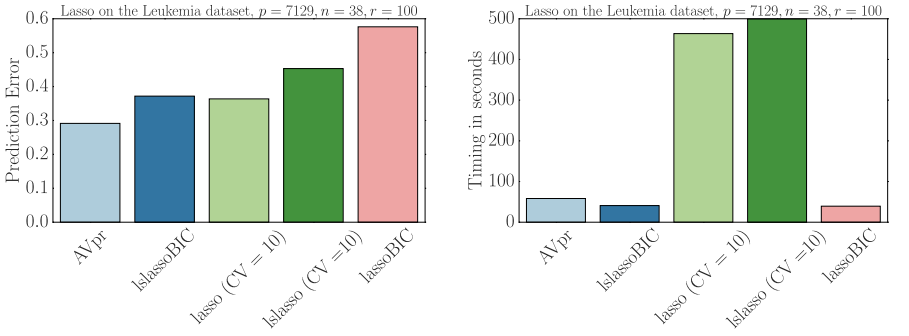

FIG 1. *Prediction error and computing times of the* $\mathrm{AV}_{\mathrm{Pr}}$*, lslassoBIC [2], lassoCV, lslassoCV, and lassoBIC procedures. The bars represent the prediction error on the 34 left out observations. Note that the grid for the tuning parameter contains the same* 50 *values for all methods.*

This adjusted approach is natural, yet suffers from two drawbacks. First, every cross-validation fold must fit a least-squares on each subset selected on the lasso path, which becomes computationally intensive once larger data sets are considered. Second, the method does not come with theoretical guarantees, an issue shared by most cross-validation procedures.

To address these problems, we propose Adaptive Validation for Prediction, ($\mathrm{AV}_{\mathrm{Pr}}$), a novel variable selection scheme. A pseudo-code description of the algorithm is given as Algorithm 1. Our proposal is closely related to the recently introduced ℓ_∞-Adaptive Validation (AV_∞) scheme [12], which is based on tests inspired by isotropic versions of Lepski's method [11, 24, 25]. This approach has been shown to provide fast and optimal calibration of the lasso for (one-step) estimation with respect to ℓ_∞-loss. For the two-step prediction considered in this paper, however, a considerably different and more technical approach inspired by non-isotropic tests is required.

As a practical example, Figure 1 compares $\mathrm{AV}_{\mathrm{Pr}}$ and standard methods on the Leukemia dataset discussed above. The methods under consideration to select the lasso tuning parameter are $\mathrm{AV}_{\mathrm{Pr}}$, 10-fold cross-validation, Bayesian Information Criterion (lassoBIC), and an estimator obtained by selecting with BIC a least-square estimator over the supports geneterated by a the lasso path (lslassoBIC) following [2] (see the Appendix for further information about the implementation of the latter approach). As can be seen, $\mathrm{AV}_{\mathrm{Pr}}$ is faster (39 seconds) to compute than cross-validation, and it is nearly as fast as the lassoBIC (40 seconds) and lslassoBIC (56 seconds). At the same time, it rivals the predictive performance of all competing approaches. Note at this point that the lassoBIC is a variable selection method rather than a predictive method; two goals that can be considerably different from each other.

The organization of this article is as follows. In the next section, we introduce the algorithm and prove that $\mathrm{AV}_{\mathrm{Pr}}$ predictions satisfy an oracle inequality, that is, are optimal up to a constant factor. In Section 3, we show that on simulations,

$\mathrm{AV}_{\mathrm{Pr}}$ is substantially faster than cross-validation while being comparable in accuracy.

All proofs are deferred to the Supplementary Material.

Framework and notation

Let us describe the framework and the notation. We are interested in linear regression models of the form

$$Y = X\beta + \varepsilon, \tag{1.1}$$

where $Y \in \mathbb{R}^n$ is the data, $X \in \mathbb{R}^{n\times p}$ the design matrix, $\beta \in \mathbb{R}^p$ the regression vector, and $\varepsilon \in \mathbb{R}^n$ the random noise. For ease of exposition, we assume that the noise is Gaussian with unknown variance σ^2, that is $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. We assume that the columns of the design matrix $X_1, \ldots, X_p \in \mathbb{R}^n$ have been standardized to have Euclidean norm $\|X_j\|_2 = \sqrt{n}$, but we otherwise allow for arbitrary correlations between the columns and noise distributions. We are mainly motivated by (but not limited to) high-dimensional settings with sparse regression vectors, where the number of parameters p can rival or even exceed the number of samples n. We finally denote by $S := \mathrm{supp}[\beta] := \{j \in [p] : \beta_j \neq 0\}$ the true support, whose cardinality is usually smaller than n and p (where throughout the paper $[d]$ stands for the set $\{1, \ldots, d\}$).

A standard approach to find a vector $\hat{\beta}$ with small prediction loss $\|X\hat{\beta} - X\beta\|_2^2/n$ is performing a least-squares refitting to the lasso. After reducing the initially large set of variables to a small number of relevant ones, the subsequent refitting aims to lessen the bias associated with the lasso. For a fixed tuning parameter λ, the lasso $\hat{\beta}^\lambda$ is defined via the minimization of objective function

$$\hat{\beta}^\lambda \in \operatorname*{arg\,min}_{\theta\in\mathbb{R}^p} \left\{ \|Y - X\theta\|_2^2 + 2\lambda\|\theta\|_1 \right\} . \tag{1.2}$$

For simplicity, we will assume that the support of the minimizer equals the equicorrelation set (see Supplementary Material for details). The subsequent least-squares refitting is defined as a minimizer of

$$\bar{\beta}^\lambda \in \operatorname*{arg\,min}_{\mathrm{supp}[\theta]=\mathrm{supp}[\hat{\beta}^\lambda]} \|Y - X\theta\|_2^2 . \tag{1.3}$$

We call this estimator the least-squares lasso (lslasso). This two-step procedure is very popular as it has smaller bias than the lasso for a range of models [3, 21].

Our goal is to find optimal tuning parameters for the lslasso (1.3) in terms of prediction. In practice, only finitely many estimators can be computed. Therefore, we consider finite sets of tuning parameters $\Lambda = \{\lambda_1, \ldots, \lambda_r\}$, $r \in \mathbb{N}$ and the associated supports $(\hat{S}^1, \ldots, \hat{S}^r)$, $\hat{S}^i := \mathrm{supp}[\hat{\beta}^{\lambda_i}]$. We denote the collection of supports by $\mathcal{S} := \{\hat{S}^i : i \in [r]\}$. Finally, we introduce surrogate sets $\hat{S}^{i,j} := \hat{S}^i \cup \hat{S}^j$ and corresponding estimators

$$\overline{\beta}^{i,j} \in \operatorname*{arg\,min}_{\mathrm{supp}[\xi]\subset\hat{S}^{i,j}} \|Y - X\xi\|_2^2. \tag{1.4}$$

In the special case $i = j$, it holds that $\hat{S}^{i,j} = \hat{S}^i$, and hence, $\overline{\beta}^i := \overline{\beta}^{i,i} = \overline{\beta}^{\lambda_i}$.

The lasso is only one out of many variable selection procedures. Our algorithms and derivations can be easily adapted to other procedures, such as the square-root lasso [1, 4, 9], scaled-lasso variants [28, 34, 35] or thresholded ridge regression [33], combined with subsequent least-squares refitting. Note for instance that by one-to-one correspondence, our results also hold for the square-root lasso. However, due to its popularity, we focus here only on the lasso.

Related literature

Besides the references to the papers that are most closely connected with our study, we provide some additional pointers to related literature. A discussion of multi-stage methods for regression can be found in [41]. Approaches to tuning parameter calibration in the single-stage setting include [8, 10, 15, 22, 27, 30, 32]. Related papers that appeared recently include [40], which contains an alternative to least-squares refitting, and [2], which discusses BIC-type selection as well as Q-aggregation approaches to model selection over the lasso path. The latter contains sparse oracle inequalities as well as prediction and estimation bounds under the standard restricted eigenvalue condition [5] - both for (a refitted) BIC-type procedure and for a Q-aggregation procedure. These methods enjoy similar theoretical guarantees as the ones we provide for $\mathrm{AV}_{\mathrm{Pr}}$, and they are also subject to the same issue, namely, that a preliminary estimate of the noise level is required.

2. $\mathbf{AV_{Pr}}$ and its properties

2.1. The $\mathbf{AV_{Pr}}$ *algorithm*

The $\mathrm{AV}_{\mathrm{Pr}}$ scheme is summarized in Algorithm 1. As inputs, it takes the data (Y, X), a set of tuning parameters Λ, and a constant $a > 0$ specified in the following section. It then conducts simple tests along the tuning parameter path of the lasso until a stopping criterion is met. It returns the index of the current tuning parameter $\overline{i}$ as well as the corresponding two-stage estimator $\overline{\beta}^{\overline{i}}$.

The algorithm requires the computation of a single lasso path and least-squares estimators along this path. The computation of the paths can be conducted with readily available, easy-to-use, and highly efficient software such as glmnet (in R) or scikit-learn(in Python) [14, 29]. For the computation of the least-squares estimators, off-the-shelf solvers can be used since the number of active variables of the second step is typically small.

In Section 2.3, $\mathrm{AV}_{\mathrm{Pr}}$ is shown to satisfy an optimal finite sample prediction bound, and the practical performance of $\mathrm{AV}_{\mathrm{Pr}}$ is illustrated in Section 3.

2.2. Assumption 2.1

Let us first introduce and motivate an assumption that ensures a certain stability of lasso solution. In general, if an estimator is unstable for data very close to

Data: $Y, X, \Lambda = \{\lambda_1, \ldots, \lambda_r\}, a$
Result: $\bar{i} \in [r], \overline{\beta} \in \mathbb{R}^p$
Initialize index: $i \leftarrow 1$
Compute $\overline{\beta}^1, \ldots, \overline{\beta}^r$
If needed, re-sort the estimators such that $|\hat{S}^1| \leq \cdots \leq |\hat{S}^r|$
while $i \leq r - 1$ **do**
  Initialize stopping criterion: $TestFailure \leftarrow False$
  Initialize comparisons: $j \leftarrow i + 1$
  while *(*$j \leq r$*) and (TestFailure==False)* **do**
    Compute $\hat{S}^{i,j}$ and $\overline{\beta}^{i,j}$
    if $\|X\overline{\beta}^i - X\overline{\beta}^{i,j}\|_2^2 \leq a|\hat{S}^i| + a|\hat{S}^{i,j}|$ **then**
      $j \leftarrow j + 1$
    else
      $TestFailure \leftarrow True$
  if $TestFailure == True$ **then**
    $i \leftarrow i + 1$
  else
    break
Set output: $\bar{i} \leftarrow i$ and $\overline{\beta} \leftarrow \overline{\beta}^{\bar{i}}$

Algorithm 1: $\mathrm{AV}_{\mathrm{Pr}}$

the (noiseless) underlying truth, accurate estimation and inference hardly seem realistic. For the goal of refitting, we thus introduce an assumption that ensures the stability of supports. In the case of the lasso, this means that we restrict $X\beta$ from being too close to hyperplanes generated by the geometric arrangement of the columns in X. Figure 2 contains a schematic picture of this: $X\beta$ needs to lie outside of small neighborhoods (depicted in orange) around the black boundaries that represent the geometry of X. Most importantly, we stress the assumption does not imply restrictions on the correlations of the design, and does not require estimated supports to be accurate.

While the assumption concerns the model, it is most convenient to put the precise formulation in terms of the lasso itself. For this, recall that for a fixed X, the support of the lasso evaluated at a vector $z \in \mathbb{R}^n$ is determined by which "zone" of $\mathbb{R}^n$ the vector z falls into [18, 38]. These zones exactly correspond to the zones in Figure 2 that are separated by the black lines. Importantly, note that we do not require additional variable selection guarantees for the lasso, but merely that the selection is unambiguous. We now define

$$\begin{aligned} D: \ & \mathbb{R}^n \to [0, \infty) \\ D(z) := \inf\{ & \|z - z'\|_\infty / \sqrt{n} \ : \ z' \in \mathbb{R}^n \text{ s.t. for some } \lambda \in \Lambda, \\ & \operatorname{supp}[\hat{\beta}^\lambda(z)] \neq \operatorname{supp}[\hat{\beta}^{\lambda'}(z')] \text{ for all } \lambda' \in \Lambda\} \ . \end{aligned}$$

The function D quantifies how far away a signal can be from the zone boundaries. The factor $1/\sqrt{n}$ in the definition reflects our normalization of the design matrix. We also stress that the function involves lasso solutions only at fixed, non-random vectors z, z'; in particular, D is independent of ε and Y.

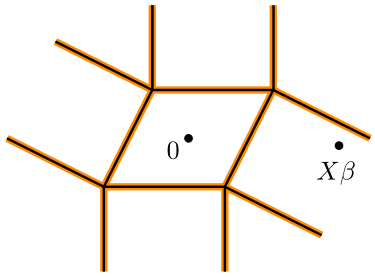

FIG 2. *An illustration of Assumption 2.1:* $X\beta$ *needs to be separated from the boundaries of the zones that determine the active set of the lasso.*

Assumption 2.1. *There is an integer* N *such that for all* $n \geq N$*, it holds that*

$$D(X\beta) > \sqrt{\frac{6\sigma^2 \log n}{n}} \ .$$

This assumption now ensures that $X\beta$ is sufficiently far from the zone boundaries. Note that the assumption is very different from restricted eigenvalues [7] or similar hypothesis in the theory for the lasso [13, 39]. While the latter assumptions need to be strict to ensure a good performance of the lasso, our assumption only requires that the estimates are unambiguous. In the specific case where $X = \mathrm{I}_{n\times n}$, some insight can be obtained, since the quantity $D(\beta)$ can be computed. With the convention: $|\beta_{(s)}|$ is the s-th largest amplitude of the vector $|\beta|$, $D(\beta)$ represents the smallest difference $|\beta_{(s)}| - |\beta_{(s+1)}|$ where s is a support size of a Lasso solution applied on β (*i.e.,* a soft-thresholded version of β) for a threshold $\lambda \in \Lambda$. In such a case the assumption is then mostly on bounded by below such differences, and is therefore an assumption on the underlying signal itself.

To motivate this assumption further, we finally show that a slightly weaker version of Assumption 2.1 automatically holds for all $X\beta$ up to a set of measure zero.

Theorem 2.2. *For all* $X\beta \in \mathbb{R}^n$ *up to a set of Lebesgue measure zero, the lasso satisfies*

$$D(X\beta) > 0 \ .$$

Theorem 2.2 does not completely exclude cases that violate Assumption 2.1. However, together with the above discussion, the result indicates that these cases are hardly generic and of limited relevance in applications.

2.3. Oracle inequality

Oracle inequalities are bounds for the risk of an estimator. More precisely, they compare the risk of an estimator with the risk of an oracle estimator, an estimator that has knowledge of the best model [6, 20].

In this section, we show that our estimator $\mathrm{AV}_{\mathrm{Pr}}$ satisfies such an oracle inequality. To this end, we first introduce the oracle set.

Definition 2.1 (Oracle)**.** *The oracle set $S^* \in \mathcal{S}$ is the set $S^* := \hat{S}^{i^*}$ with index*

$$i^* := \min\left\{i \in [r] : \hat{S}^i \supset S\right\} .$$

and the associated oracle estimator is $\beta^* := \overline{\beta}^{i^*}$.

In other words, the oracle set contains the true support S and has minimal cardinality among all such sets. The oracle set can therefore be viewed as the best possible approximation of S in $\mathcal{S}$; in particular, $S^* = S$ whenever $S \in \mathcal{S}$.

We implicitly assume that the oracle set exists, that is, the true support set is a subset of an estimated support along the path. However, one can easily generalize the definition to avoid this assumption. Let S^* be an arbitrary set and P^* the projection onto the space spanned by the columns with indexes in S^*. Adding this projection in our proofs (cf. (A.2) for example) yields the same results as below except for an additional term $\|(\mathrm{I} - P^*)\beta\|_2^2$ in the bounds. However, as the above definition exists in generic cases (since the lasso supports tend to be very exhaustive for small tuning parameters), and as it provides a concise formulation of the results, we do not consider the extended version in the following.

Now, given the oracle, we can state a bound for the two-step lasso procedure with the *optimal* tuning parameter, that is, the tuning parameter that leads to the oracle set. Throughout this section we invoke Assumption 2.1, which helps us rule out ambiguous design settings.

Proposition 2.1. *Under Assumption 2.1, for any $\alpha > 0$, there exist constants $t, N, R > 0$ such that for all $n \geq N$ and $r \geq R$, the oracle estimator satisfies with probability at least $1 - \alpha$ the bound*

$$\frac{\|X\beta^* - X\beta\|_2^2}{n} \leq (1 + t \log r)\frac{\sigma^2 |S^*|}{n} .$$

This is a bound for the lasso with refitting - under the assumption that the oracle set S^* is known and incorporated in the selection of the tuning parameter. The constants t, N, and R are specified in the proof section.

In practice, we do not have access to the oracle set S^*. Therefore, we hope to find a procedure that does not require its knowledge and still satisfies the bound (up to constants) stated in Proposition 2.1. The following result shows that $\mathrm{AV}_{\mathrm{Pr}}$ provides this.

Theorem 2.3 (Oracle inequality for $\mathrm{AV}_{\mathrm{Pr}}$)**.** *If Assumption 2.1 holds, for any $\alpha > 0$, there exist constants $t, N, R > 0$ such that for all $n \geq N$ and all $r > R$, our estimator $\mathrm{AV}_{\mathrm{Pr}}$ with $a \geq 2\sigma^2(1 + t \log r)$ satisfies with probability at least $1 - \alpha$ the bounds*

$$|\hat{S}| \leq |S^*| \tag{i}$$

$$\text{and} \quad \frac{\|X\overline{\beta} - X\beta\|_2^2}{n} \leq \left[6a + 4\sigma^2(1 + t \log r)\right]\frac{|S^*|}{n} . \tag{ii}$$

This proves optimality of $\mathrm{AV}_{\mathrm{Pr}}$: indeed, if $a \gtrsim 2\sigma^2(1 + t \log r)$, $\mathrm{AV}_{\mathrm{Pr}}$ satisfies the same bound (up to constants) as the two-step approach that is based on the knowledge of the oracle set S^*. Explicit constants can be found in the proofs section, though we did not attempt to optimize them.

Theorem 2.3 holds for any sufficiently large a. The question is now how to choose a in practice. Theorem 2.3 entails precise guidance for this choice. In view of the bounds, one should select the smallest a that is still allowed, that is, $a = 2\sigma^2(1+t \log r)$. However, since σ^2 is typically unknown in practice, we suggest to replace it with a rough estimate $\hat{\sigma}$. Moreover, we argue that the term $2(1+t \log r)$ is an artifact of our proof technique rather than a fundamental aspect of the bound. We thus suggest the simple choice $a = \hat{\sigma}^2$, see the empirical section below. Consequently, the bounds above provide a solid theoretical foundation for $\mathrm{AV}_{\mathrm{Pr}}$; however, there is still a gap between theory and practice that deserves to be studied further.

We note that our approach is very different from just replacing the unknown noise variance in the existing theoretical tuning parameters. Standard oracle inequalities for the lasso hold true with probability t for tuning parameters of the form $\mathrm{const}_t\, \sigma \sqrt{(\log p)/n}$, where const_t is a factor involving the level t, see [6] and references herein. Thus, one might be tempted to use these tuning parameters with an estimate of σ. However, the above form is valid only for Gaussian noise, while we aim at more general calibration. Moreover, even for Gaussian noise, the above form is known to be suboptimal both in the near orthogonal case (because p could be replaced by p/s, where s is the true sparsity level) and in the correlated case (where much smaller tuning parameters might be favored), we refer to [6, 13, 19] and references therein. Finally, even if the above form were optimal in terms of the standard oracle inequalities for prediction, estimation, and variable selection, there are no guarantees on their performance in terms of refitting.

3. Experiments

3.1. General setup

We measure the numerical performance and the computational speed of $\mathrm{AV}_{\mathrm{Pr}}$ in two-step prediction. The methods of comparison are cross-validation with 2, 5, 10, and 20 number of folds, which are typically regarded as the standard calibration schemes.

Variable selection is performed with the lasso. We emphasize that the motivation of this work is not to compare different variable selection methods, but instead, to compare different calibration schemes in two-step prediction.

The data are generated according to a linear regression model as in (1.1) with $n = p = 100, 200$. The first 10 entries of the regression vector β are set to 1, while all other entries are set to 0. The components of the noise vector are independently sampled from a univariate standard normal distribution with mean 0 and variance 1. The rows of the design matrix X are independently

Data: Y, X, δ
Result: $\widehat{\sigma}$
Initialize tuning parameter and variance: $\lambda_0 \leftarrow \sqrt{2n \log p}$ and $\widehat{\sigma} \leftarrow 1$
repeat
  Save $\widehat{\sigma}' \leftarrow \widehat{\sigma}$
  Update $\widehat{\sigma}$:
    Set $\lambda \leftarrow \widehat{\sigma}\lambda_0$
    Compute $\hat{\beta}^{\lambda}$ as the lasso with tuning parameter λ according to (1.2)
    Set $\widehat{\sigma} \leftarrow \|Y - X\hat{\beta}^{\lambda}\|_2/\sqrt{n}$
until $|\widehat{\sigma} - \widehat{\sigma}'| \leq \delta$

Algorithm 2: Scaled lasso algorithm with early stopping, cf. [35]

sampled from a multivariate normal distribution with mean 0 and covariance matrix Σ that is set to $\Sigma_{ij} = 1$ for $i = j$ and to $\Sigma_{ij} = \rho$ for $i \neq j$ with $\rho = 0.5$. Subsequently, the columns of X are normalized to Euclidean norm $\sqrt{n}$. For all experiments, we perform 50 repetitions.

In addition to the described parameter settings, we tested various other settings, including different correlation coefficients ρ, regression vectors β, and tuning parameter grids. As the conclusions were similar across all settings, we restrict our presentation to the ones described. All computations are conducted with the standard implementations of the lasso from Python *scikit-learn* (version 0.16) [29], and our code is available at `https://github.com/josephsalmon/AVp`.

3.2. Practical choice of a

We follow the suggestions after Theorem 2.3. Specifically, if σ^2 is known, we recommend using Algorithm 1 with $a = \sigma^2$ as suggested by Theorem 2.3 (regarding the term $t \log r$ as a superfluous term coming from our proof technique). In practice, however, the noise variance σ^2 is often unknown. We then advocate using $a = \widehat{\sigma}^2$ with a rough estimate $\widehat{\sigma}^2$ of σ^2. Such a rough estimate can be easily obtained by using a (very) small number of iterations of the algorithms for the square-root lasso [9] or the scaled lasso [36]. For our simulations, we have opted for the latter, which consists of an alternating minimization for estimating both the regression parameter and the noise level. Algorithm 2 states our concrete implementation. We set the tolerance to $\delta = 10^{-2}$, which typically leads to less than five iterations of the loop in the algorithm and therefore, as illustrated below, to very low computational costs.

3.3. Choice of the tuning parameter grids

The tuning parameter grid is chosen as a default grid in the lasso function in *scikit-learn*. More precisely, we take a geometric grid of size $r = 100$ starting from $\lambda_{\max} = \|X^\top Y\|_\infty$, the smallest tuning parameter that leads to a lasso solution of all zeros, and ending at $\lambda_{\max}/1000$.

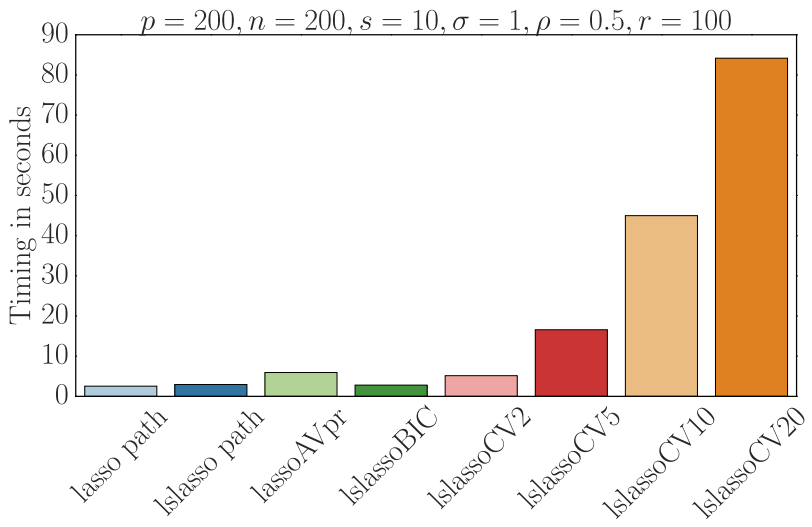

FIG 3. *Computation times of the lasso path, lslasso path, lslassoCV, and lasso*$\mathrm{AV}_{\mathrm{Pr}}$ *(with* $\widehat{\sigma}$*). Cross-validation is performed using a refitting step (lslassoCV) for different numbers of folds.*

3.4. Computational and statistical performance

We first report in Figure 3 the computational times for each of the following:

lasso path:	Computation of one tuning parameter path of lasso.
lslasso path:	Computation of one tuning parameter path of lasso with least-squares refitting.
lassoAV$_{\mathrm{Pr}}$:	Least-squares refitted lasso with tuning parameter selected by $\mathrm{AV}_{\mathrm{Pr}}$ with $a = \widehat{\sigma}^2$ as detailed above.
lslassoBIC:	Least-squares refitted lasso with tuning parameter selected by a BIC-type procedure [2], detailed in Appendix B.
lassoCV:	Least-squares refitted lasso with tuning parameter selected by cross-validation on the estimates of the (one-step) lasso.
lslassoCV:	Least-squares refitted lasso with tuning parameter selected by cross-validation on the estimates of least-squares refitted lasso.

We then also report in Figure 4 the prediction performances of the last three methods.

In conclusion, our simulations demonstrate that $\mathrm{AV}_{\mathrm{Pr}}$ is competitive both in computational speed and in prediction performance.

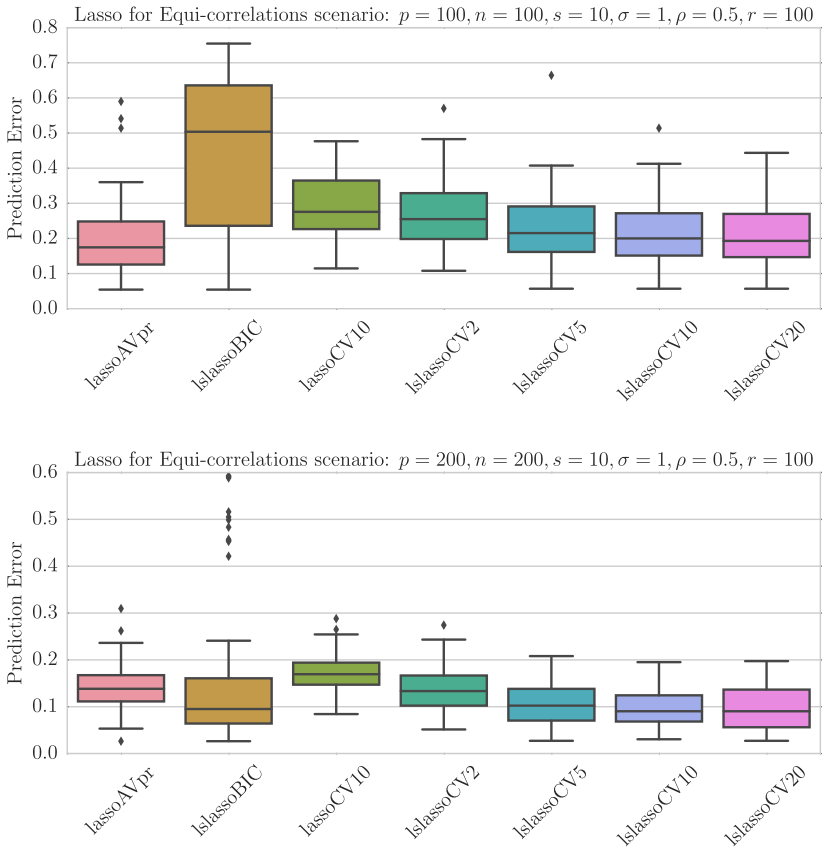

FIG 4. *Prediction errors of the lassoCV, lslassoCV, and lasso*$\mathrm{AV}_{\mathrm{Pr}}$ *(with* $\widehat{\sigma}$*). Cross-validation is performed using (lslassoCV) or not using (lassoCV) a refitting step and using different numbers of folds.*

4. Discussion

The standard scheme for calibrating the lasso is cross-validation. However, cross-validation entails two main deficiencies: it is computationally inefficient and lacks finite sample guarantees. In contrast, $\mathrm{AV}_{\mathrm{Pr}}$ is fast and satisfies optimal bounds for prediction with the refitted lasso. We therefore propose $\mathrm{AV}_{\mathrm{Pr}}$ as an alternative scheme for prediction with lasso followed by refitting. Moreover, our work can be readily extended to the square-root lasso and to ridge regression with thresholding.

A direction for further research could be sharpening the theoretical bounds. The current result contains a term that grows logarithmically in the number of tuning parameters under consideration. Such artifacts are common in the non-parametric literature. In a sense, one might consider our approach as a non-parametric version of [11]. Thus, it is not surprising that such a term ap-

pears. Empirically, however, there are no indications that this term is needed. An improved understanding of the bound could especially strengthen the connections between the theoretical and the practical choice of a.

The exact specification of the method is an issue that appears more generally in tuning parameter calibration. In our case, there is flexibility in how to estimate the noise variance; in cross-validation, one has to specify the number of folds; in BIC-type approaches, the constant in front of the log penalty needs to be adjusted; in Q-aggregation, there is a trade-off between the KL regularization and the quadratic term, ... We believe that although out of the scope for this contribution, a comprehensive sensitivity analysis studying the selections in each of the methods would be of interest.

Appendix A: Proofs

A.1. Definitions

A subtlety in the definition of the lasso variable selection scheme is that it is defined as the solution to the minimization problem (1.2), but the solution is not necessarily unique. Different lasso algorithms can yield different solutions to the lasso problem, and all could reasonably called the lasso estimator.

Define the lasso equicorrelation set [38] to be

$$E[\lambda; Y] := \left\{ i \in [p] \,:\, |X_i^\top (Y - X\hat{\beta}^\lambda)| = \lambda \right\}.$$

This set is unique and contains the support of any lasso solution $\hat{\beta}^\lambda$. In common cases, there is at least one lasso solution $\hat{\beta}^\lambda$ whose support equals the equicorrelation set. This property turns out to be quite valuable in our analysis and consequently, we will always assume from now on that the support set equals the equicorrelation set.

Denote the sets of lasso outputs that lead to the same sign vectors $\eta \in \{1, 0, -1\}^p$ by $W^\lambda(\eta) := \{Y \in \mathbb{R}^n \,:\, \operatorname{sgn}[\hat{\beta}^\lambda] = \eta\}$, cf. [23]. The closures of the sets of equal sign vector will be called regions, and the collection of all regions will be written $\mathcal{V} = \{\operatorname{cl} W^1(\eta) \,:\, \eta \subset \{-1, 0, 1\}^p\}$. To simplify notation, we will write the target as $\xi = X\beta$. When relevant, this will also be written ξ_n to emphasize the dependence on n.

A.2. Lemmas

Recall that in convex geometry, a polyhedron is a finite intersection of closed half-spaces – more details can be found in Appendix A.

Lemma A.1. *The lasso $\hat{\beta}$ fulfills the following:*

1. *it is scale-symmetric, in the sense that* $\operatorname{supp}[\hat{\beta}^\lambda(Y)] = \operatorname{supp}[\hat{\beta}^1(Y/\lambda)]$ *for all* $\lambda \in \Lambda$, $Y \in \mathbb{R}^n$, *and* $X \in \mathbb{R}^{n \times p}$;

2. *for all* $\lambda \in \Lambda$ *and* $\eta \in \{1, 0, -1\}^p$*, the closure of its regions of equal sign vector* $cl[\mathcal{W}^\lambda(\eta)]$ *are polyhedra.*

Proof of Lemma A.1. We prove each condition in order.
Part i) The scale-symmetry follows from consideration of the dual problem to (1.2). Let $\hat{\beta}_\lambda$ be a lasso solution whose active set equals the equicorrelation set $E[\lambda; Y]$. Let C_λ stand for the polyhedron $\{x \in \mathbb{R}^p : \|X^\top x\|_\infty \leq \lambda\}$, and let P_{C_λ} denote the Euclidean projection on this set. Notice that for any $x \in C_\lambda$, we have $x/\lambda \in C_1$ and therefore

$$\|\lambda P_{C_1}\left(\frac{Y}{\lambda}\right) - Y\|_2 \leq \lambda \|P_{C_1}\left(\frac{Y}{\lambda}\right) - \frac{Y}{\lambda}\|_2 \leq \lambda \|\frac{x}{\lambda} - \frac{Y}{\lambda}\|_2 \leq \|x - Y\|_2.$$

Since this is true for all $x \in C_\lambda$, and that $\lambda P_{C_1}\left(\frac{Y}{\lambda}\right) \in C_\lambda$, we conclude that $\lambda P_{C_1}\left(\frac{Y}{\lambda}\right) = P_{C_\lambda}(Y)$. As shown in [38], the residual from the lasso satisfies

$$Y - X\hat{\beta}_\lambda = P_{C_\lambda}(Y).$$

Therefore, for any $\lambda > 0$ the active set of $\hat{\beta}_\lambda$, which is the equicorrelation set here, satisfies $E[\lambda; Y] = E[1; Y/\lambda]$ since

$$E[\lambda; Y] = \left\{i \in [p] : \left|X_i^\top P_{C_\lambda}(Y)\right| = \lambda\right\} = \left\{i \in [p] : \left|X_i^\top P_{C_1}\left(\frac{Y}{\lambda}\right)\right| = 1\right\},$$

as desired.
Part ii) The polyhedron $C_1 = \{x \in \mathbb{R}^n : \|X^\top x\|_\infty \leq 1\}$ has an irreducible decomposition into half-spaces

$$C_1 = \left(\bigcap_{i=1}^p \{x \in \mathbb{R}^n : X_i^\top x - 1 \leq 0\}\right) \bigcap \left(\bigcap_{i=1}^p \{x \in \mathbb{R}^n : -X_i^\top x - 1 \leq 0\}\right),$$

so the facets of C_1 are $F_i^\pm = C_1 \cap \{x \in \mathbb{R}^n : \pm X_i^\top x - 1 \leq 0\}$ – see [17, Sec. 2.6]. Since by assumption, the active set of the lasso estimate coincides with the equicorrelation set, by the Karush-Kuhn-Tucker conditions and $Y - X\hat{\beta}_1 = P_{C_1}Y$, see [38, Equations (13)-(14) and Lemma 3], we have

$$X_i^\top P_{C_1}(Y) = \operatorname{sgn}\hat{\beta}_i \in \{-1, 1\} \;\Leftrightarrow\; P_{C_1}Y \in F_i^{\operatorname{sgn}\hat{\beta}_i}$$

for $i \in E[1; Y]$. Moreover,

$$|X_i^\top P_{C_1}(Y)| < 1 \;\Leftrightarrow\; P_{C_1}Y \in \left(C_1 \backslash F_i^+\right) \cap \left(C_1 \backslash F_i^-\right)$$

for $i \notin E[1; Y]$. In light of this, $\operatorname{sgn}\hat{\beta} = \eta$ if and only if

$$P_{C_1}(Y) \in \bigcap_{\substack{i \in [p] \\ \eta_i = 1}} F_i^+ \cap \bigcap_{\substack{i \in [p] \\ \eta_i = -1}} F_i^- \cap \bigcap_{\substack{i \in [p] \\ \eta_i \neq 0}} C_1 \backslash F_i^+ \cap C_1 \backslash F_i^-$$

$$= \operatorname{relint}\left[\left(\bigcap_{\substack{i\in[p]\\ \eta_i=1}} F_i^+\right)\cap\left(\bigcap_{\substack{i\in[p]\\ \eta_i=\text{-}1}} F_i^-\right)\right].$$

So, $W(\eta) = P_{C_1}^{-1}(\operatorname{relint} F_\eta)$ for the face $F_\eta = \left(\bigcap_{\substack{i\in[p]\\ \eta_i=1}} F_i^+\right)\cap\left(\bigcap_{\substack{i\in[p]\\ \eta_i=\text{-}1}} F_i^-\right)$.

Let $V \in \mathcal{V}$ - then by definition of $\mathcal{V}$, there must be an $\eta \subset \{-1,0,1\}^p$ such that $V = \operatorname{cl} W(\eta) = \operatorname{cl} P_{C_1}^{-1}(\operatorname{relint} F_\eta)$. According to [31, Equation (2.3) and Page 83], we have $N(C_1, F_\eta) + \operatorname{relint} F_\eta = P_{C_1}^{-1}(\operatorname{relint} F_\eta) = W(\eta)$, where $N(C_1, F_\eta)$ is the normal cone of C_1 to the face F_η.

Now, since $\operatorname{cl} A + \operatorname{cl} B \subset \operatorname{cl}(A+B)$ for any sets A, B,

$$N(C_1, F_\eta) + \operatorname{relint} F_\eta \subset N(C_1, F_\eta) + F_\eta \subset \operatorname{cl} P_{C_1}^{-1}(\operatorname{relint} F_\eta).$$

But $N(C_1, F_\eta) + F_\eta$ is the sum of two polyhedra, hence a polyhedron, hence closed. Thus $N(C_1, F_\eta) + F_\eta = \operatorname{cl} P_{C_1}^{-1}(\operatorname{relint} F_\eta) = V$ is a polyhedron.

Notice that $N(C_1, F_\eta) + \operatorname{relint} F_\eta$ is convex as a sum of convex sets, and thus

$$\begin{aligned}\operatorname{relint} W(\eta) =& \operatorname{relint} P_C^{-1}(\operatorname{relint} F_\eta) = \operatorname{relint}(N(C_1, F_\eta) + \operatorname{relint} F_\eta)\\ =& \operatorname{relint}\operatorname{cl}(N(C_1, F_\eta) + \operatorname{relint} F_\eta) = \operatorname{relint} V = \operatorname{int} V,\end{aligned}$$

since V is n-dimensional. Thus $\operatorname{int} V \subset W(\eta)$, and $\operatorname{sgn}\hat{\beta}$ is constant over $\operatorname{int} V$, as desired. □

Theorem A.2. *For any $t \geq 6$, there exists a deterministic integer N such that with probability at least $1 - t^{-1} - R_n^{-1}\sqrt{1 + t\log R_n}$,*

$$\left\|X\bar{\beta}^{i,j} - X\beta\right\|_2^2 \leq \sigma^2(1 + t\log r)|\hat{S}^{i,j}| + \left\|(\mathrm{I}_n - P_{\hat{S}^{i,j}})X\beta\right\|_2^2 \tag{A.1}$$

for all i, j and all $n \geq N$.

Proof of Lemma A.2. By (1.1), the loss of these estimators can be broken down as

$$\begin{aligned}\left\|X\overline{\beta}^{i,j} - X\beta\right\|_2^2 &= \left\|P_{\hat{S}^{i,j}}(X\overline{\beta}^{i,j} - X\beta)\right\|_2^2 + \left\|(\mathrm{I}_n - P_{\hat{S}^{i,j}})(X\overline{\beta}^{i,j} - X\beta)\right\|_2^2\\ &= \left\|P_{\hat{S}^{i,j}}(XX^+_{\hat{S}^{i,j}}Y - Y + \varepsilon)\right\|_2^2 + \left\|(\mathrm{I}_n - P_{\hat{S}^{i,j}})(XX^+_{\hat{S}^{i,j}}Y - X\beta)\right\|_2^2\\ &= \left\|P_{\hat{S}^{i,j}}\varepsilon\right\|_2^2 + \left\|(\mathrm{I}_n - P_{\hat{S}^{i,j}})X\beta\right\|_2^2,\end{aligned} \tag{A.2}$$

where $X^+_{\hat{S}^{i,j}}$ is the Moore-Penrose pseudo-inverse of the submatrix of X comprising the columns with indexes in $\hat{S}^{i,j}$. Our goal is to control the noise term $\|P_{\hat{S}^{i,j}}\varepsilon\|_2^2$ in (A.2). We do this in four parts. We first show that on an appropriate scale, the response Y must be close to the target $X\beta$ with high probability. This is then shown to imply, using the scale-symmetry property of the lasso, that the ordered active sets must be unique with high probability. This allows us to control the noise term by showing that the projected noise behaves like a chi-square distribution, construct an appropriate event, and bound its probability. Finally, on this event the inequality of the theorem is shown to hold.

1. We first use a Gaussian tail bound to show that Y is close to the target $X\beta$; precisely, we show that the event

$$\bar{\Omega}_1 := \bigcap_{n=n_1}^{\infty} \left\{ \|Y - X\beta\|_\infty \leq \sqrt{6\sigma^2 \log n} \right\}$$

fulfills the bound

$$\mathbb{P}\left[\bar{\Omega}_1\right] \geq 1 - 1/t, \tag{A.3}$$

where $n_1 := \min\{n \,:\, \sqrt{6\sigma^2 \log n} > 1/\sqrt{2\pi}\} \vee \lceil 4t \rceil$.
For this, write $\xi_n := X\beta$ (the subscript n highlights the sample size dependence) and define the event $\Omega_n := \left\{ \|Y - \xi_n\|_\infty \leq \sqrt{6\sigma^2 \log n} \right\}$ for ease of notation. Using a union bound, the Gaussian tail bound $\mathbb{P}[\mathrm{N}(0,1) > t] \leq e^{-t^2/2}/\sqrt{2\pi}t < e^{-t^2/2}$ (note that $t > 6 > 1/\sqrt{2\pi}$), and the definition of n_1, we find that for the complements Ω_n^C of the sets Ω_n,

$$\begin{aligned}\sum_{n=n_1}^{\infty} \mathbb{P}\left[\Omega_n^C\right] &\leq \sum_{n=n_1}^{\infty} 2ne^{-3\log n} = 2\sum_{n=n_1}^{\infty} \frac{1}{n^2} \\ &\leq \frac{2}{n_1^2} + 2\int_{n_1}^{\infty} \frac{1}{w^2} dw = \frac{2}{n_1^2} + \frac{2}{n_1} \leq \frac{4}{n_1} \leq \frac{1}{t}.\end{aligned}$$

From this and the definition of $\bar{\Omega}_1$, the bound (A.3) follows.

2. We now use Part 1 to deduce that the active sets are deterministic if the response is close to the target, or more specifically, we derive that on $\bar{\Omega}_1$,

$$\hat{S}^i[Y] = \hat{S}^i[\xi_n] \tag{A.4}$$

for $1 \leq i \leq r$.
From Part 1, we deduce that on $\bar{\Omega}_1$ and for $n \geq n_1$,

$$\frac{\|Y - \xi_n\|_\infty}{\sqrt{n}} \leq \sqrt{\frac{6\sigma^2 \log n}{n}} = \frac{\sqrt{6\sigma^2 \log n/n}}{D(\xi_n)} D(\xi_n).$$

Then, by Assumption 2.1, there must be an n_2, without loss of generality satisfying $n_2 \geq n_1$, such that $\|Y - \xi_n\|_\infty/\sqrt{n} < D(\xi_n)$. But by definition of $D(\xi_n)$, this means that for all $\lambda > 0$, there must be a $\lambda' > 0$ such that $\mathrm{supp}[\hat{\beta}^1(Y/\lambda)] = \mathrm{supp}[\hat{\beta}^1(\xi_n/\lambda')]$. For a given $y \in \mathbb{R}^n$, we now define the collection of active sets by $\hat{Q}[y] := \{\mathrm{supp}[\hat{\beta}^1(y/\lambda)] : \lambda > 0\}$. That is, $\hat{Q}[Y]$ is the collection of active sets along the tuning parameter path of the estimator for given data (Y, X). There are at most 2^p different subsets of $[p]$, so these are always finite sets. We therefore obtain

$$\mathbb{P}\left[\hat{Q}[Y] = \hat{Q}[\xi_n] \;;\; \forall n \geq n_2 \;:\; \bar{\Omega}_1\right] = 1.$$

In particular, the random cardinality $r := |\hat{Q}[Y]|$ and deterministic cardinality $\bar{r} := |\hat{Q}[\xi_n]|$ coincide almost surely on this event, that is,

$$\mathbb{P}\left[r = \bar{r} \;;\; \forall n \geq n_2 \;:\; \bar{\Omega}_1\right] = 1.$$

Because we follow a fixed ordering rule, we must then have

$$\mathbb{P}\Big[(\hat{S}^1[Y],...,\hat{S}^r[Y]) = (\hat{S}^1[\xi_n],...,\hat{S}^r[\xi_n]) \quad \forall n \geq n_2 \ : \bar{\Omega}_1\Big] = 1.$$

This finishes the proof of Equation (A.4).

3. Let us define the sets $\hat{S}^{i,j} := \hat{S}^{i,j}[Y] := \hat{S}^i[Y] \cup \hat{S}^j[Y]$, the random ranks $r^{i,j} := \operatorname{rk} X_{\hat{S}^{i,j}}$ and deterministic ranks $\bar{r}^{i,j} := \operatorname{rk} X_{\hat{S}^{i,j}[\xi_n]}$. Our next step is to show that Part 2 provides a chi-square bound for the noise part in (A.2), that is, we prove on $\bar{\Omega}_1$ the relations

$$\|P_{\hat{S}^{i,j}}\varepsilon\|_2^2 \sim \sigma^2\chi^2_{\bar{r}^{i,j}} \text{ if } \bar{r}^{i,j} \geq 1 \ \text{ and } \ \|P_{\hat{S}^{i,j}}\varepsilon\|_2^2 = 0 \text{ if } \ \bar{r}^{i,j} = 0. \tag{A.5}$$

To show this, we apply (A.2) to our two-step method $\bar{\beta}^{i,j}$ to get

$$\big\|X\bar{\beta}^{i,j} - X\beta\big\|_2^2 = \big\|P_{\hat{S}^{i,j}}\varepsilon\big\|_2^2 + \big\|(\mathrm{I}_n - P_{\hat{S}^{i,j}})X\beta\big\|_2^2.$$

According to Part 2, the sets $\hat{S}^{i,j}$ satisfy

$$\mathbb{P}\left[\hat{S}^{i,j} = \hat{S}^{i,j}[\xi_n] \quad \forall n \geq n_2 \ : \bar{\Omega}_1\right] = 1$$

for all $1 \leq i,j \leq r$. This has two consequences. First, the random ranks $r^{i,j}$ equal the deterministic ranks $\bar{r}^{i,j}$ almost surely on the event $\bar{\Omega}_1$:

$$\mathbb{P}\Big[r^{i,j} = \bar{r}^{i,j} \quad \forall n \geq n_2 \ : \bar{\Omega}_1\Big] = 1.$$

Second, the matrices $P_{\hat{S}^{i,j}}$ are indeed projection matrices on $\bar{\Omega}_1$ of rank $r^{i,j}$, since on this event the active sets $\hat{S}^{i,j}$ (and, therefore, the subspaces spanned by $X_{\hat{S}^{i,j}}$) are constant. Formally,

$$\mathbb{P}\left[P_{\hat{S}^{i,j}} = P_{\hat{S}^{i,j}[\xi_n]} \quad \forall n \geq n_2 \ : \bar{\Omega}_1\right] = 1.$$

Combining this with $\varepsilon \sim \mathcal{N}(0, \sigma^2\,\mathrm{I}_n)$ yields the results in (A.5). We can now control the noise part in (A.2) with a chi-square Chernoff bound. More specifically, we obtain the bound

$$\mathbb{P}[\bar{\Omega}_2] \geq 1 - t^{-1} - R_n^{-1}\sqrt{1 + t\log R_n} \tag{A.6}$$

for $\bar{\Omega}_2 := \bar{\Omega}_1 \cap \big\{\|P_{\hat{S}^{i,j}}\varepsilon\|_2^2 \leq (1 + t\log r)\bar{r}^{i,j} \text{ for all } 1 \leq i,j \leq r\big\}$. To this end, recall that by definition, the integers $\bar{r}^{i,j} = |\hat{S}^{i,j}[\xi_n]|$ and $\bar{r} = |\hat{Q}[\xi_n]|$ are deterministic. According to Part 2, it also holds that on $\bar{\Omega}_1$, we have $\bar{r}^{i,j} = |\hat{S}^{i,j}[Y]|$ and $\bar{r} = r$. We use this, result (A.5), a union bound, and the bound $\mathbb{P}[\bar{\Omega}_1] \geq 1 - 1/t$ stated in (A.3) to deduce

$$\mathbb{P}\left[\bar{\Omega}_2\right] = \mathbb{P}\left[\bar{\Omega}_1 \cap \left\{\max_{\substack{1\leq i,j\leq \bar{r}\\ \bar{r}^{i,j}\geq 1}} \frac{\chi^2_{\bar{r}^{i,j}}}{\bar{r}^{i,j}} \leq 1 + t\log\bar{r}\right\}\right]$$

$$\geq 1 - t^{-1} - \sum_{\substack{1\leq i,j\leq \bar{r} \\ \bar{r}^{i,j}\geq 1}} \mathbb{P}\left[\chi^2_{\bar{r}^{i,j}} > \bar{r}^{i,j}(1 + t\log\bar{r})\right].$$

Now, using the chi-square Chernoff bound $\mathbb{P}[\chi^2_k > k(1+a)] < [(1+a)e^{-a}]^{k/2}$, we obtain

$$\mathbb{P}\left[\bar{\Omega}_2\right] \geq 1 - t^{-1} - \sum_{\substack{1\leq i,j\leq \bar{r} \\ \bar{r}^{i,j}\geq 1}} \left[(1 + t\log\bar{r})e^{-t\log\bar{r}}\right]^{\bar{r}^{i,j}/2}.$$

As $\bar{r} \geq 1$, $(1 + t\log\bar{r})e^{-t\log\bar{r}} \leq 1$ and so

$$\begin{aligned}\mathbb{P}\left[\bar{\Omega}_2\right] &\geq 1 - t^{-1} - \sum_{\substack{1\leq i,j\leq \bar{r} \\ \bar{r}^{i,j}\geq 1}} \left[(1 + t\log\bar{r})e^{-t\log\bar{r}}\right]^{1/2} \\ &\geq 1 - t^{-1} - \bar{r}^{2-t/2}\sqrt{1 + t\log\bar{r}}.\end{aligned}$$

We now use that that $2 - t/2 \leq -1$ for all $t \geq 6$ and the fact that $\bar{r} \geq R_n$ to find that

$$\mathbb{P}\left[\bar{\Omega}_2\right] \geq 1 - t^{-1} - R_n^{-1}\sqrt{1 + t\log R_n},$$

which concludes Part 3.

4. We finally collect the pieces to deduce that with probability at least $1 - t^{-1} - R_n^{-1}\sqrt{1 + t\log R_n}$, the bound

$$\left\|X\bar{\beta}^{i,j} - X\beta\right\|_2^2 \leq \sigma^2(1 + t\log r)|\hat{S}^{i,j}| + \left\|(\mathrm{I}_n - P_{\hat{S}^{i,j}})X\beta\right\|_2^2 \qquad \text{(A.7)}$$

holds for all $n \geq n_2$.
For this, we assume that indeed $n \geq n_2$ and then combine the initial bound (A.2) and the results of Part 3 to find that on $\bar{\Omega}_2$,

$$\left\|X\bar{\beta}^{i,j} - X\beta\right\|_2^2 \leq \sigma^2(1 + t\log r)\bar{r}^{i,j} + \left\|(\mathrm{I}_n - P_{\hat{S}^{i,j}})X\beta\right\|_2^2.$$

Recalling that $\bar{r}^{i,j} = |\hat{S}^{i,j}|$ according to Part 2, the desired bound (A.7) now follows from Inequality (A.6) derived in Part 3. □

A.3. Proofs of results from Section 2

Recall that the path of active sets is $|\hat{S}^1| \leq |\hat{S}^2| \leq \cdots \leq |\hat{S}^r|$. The cardinality r is typically random, but we can always bound it almost surely by some constant R_n, so that $1 \leq R_n \leq r$ almost surely. This constant should be independent of the data but can be chosen to vary with n and p. For example, we might have agreed *a priori* with considering 50 sets, or our variable selection method might always yield at least $\min(n, p)$ different sets by construction.

Corollary A.3 (Oracle benchmark)**.** *Say the oracle set exists, that the design satisfies Assumption 2.1. Then, for any $t \geq 6$, there exists a deterministic integer N such that with probability at least $1 - t^{-1} - R_n^{-1}\sqrt{1 + t \log R_n}$, the oracle estimator satisfies*

$$\|X\beta^* - X\beta\|_2^2 \leq \sigma^2(1 + t \log r)|S^*|$$

for all $n \geq N$.

Proof of Corollary A.3. The oracle estimator is $\beta^* = \overline{\beta}^{i^*}$, the refitted estimator on the oracle set $S^* = \hat{S}^{i^*}$. The result therefore follows immediately from Theorem A.2 applied to $i = j = i^*$. □

Corollary A.4 (Oracle inequality for $\mathrm{AV}_{\mathrm{Pr}}$)**.** *Say the oracle set exists, that the design satisfies Assumption 2.1, and that the AV_p parameter a is such that $a \geq 2\sigma^2(1 + t \log r)$. Then, for any $t \geq 6$, there exists a deterministic integer N such that with probability at least $1 - t^{-1} - R_n^{-1}\sqrt{1 + t \log R_n}$, it holds that $|\hat{S}| \leq |S^*|$ and*

$$\|X\overline{\beta} - X\beta\|_2^2 \leq \Big[6a + 4\sigma^2(1 + t \log r)\Big]|S^*|$$

for all $n \geq N$.

Proof of Corollary A.4. By Theorem A.2, there exists an N such that the event

$$\Omega = \Big\{ \big\|X\bar{\beta}^{i,j} - X\beta\big\|_2^2 \leq \sigma^2(1 + t \log r)|\hat{S}^{i,j}| + \big\|(\mathrm{I}_n - P_{\hat{S}^{i,j}})X\beta\big\|_2^2, \forall i, j \text{ and } n \geq N \Big\}$$

holds with probability at least $1 - t^{-1} - R_n^{-1}\sqrt{1 + t \log R_n}$.

Claim (i): On Ω, it holds that $\bar{i} \leq i^* = \min\big\{i \in [r] \big| \hat{S}^i \supset S\big\}$.
We prove this claim by contradiction and, therefore, assume that $\bar{i} > i^*$. Then, by the definition of our estimator, there must be an integer $k \in [r]$ such that $|\hat{S}^k| \geq |S^*|$ and

$$\|X\beta^* - X\overline{\beta}^{k,i^*}\|_2^2 > a|S^*| + a|\hat{S}^k \cup S^*| \,. \tag{A.8}$$

The fact that $|\hat{S}^k| \geq |S^*| \geq |S|$, together with the bound (A.1) and $S^* \supset S$, yields

$$\begin{aligned}
&\|X\beta^* - X\overline{\beta}^{k,i^*}\|_2^2 \\
&\leq 2\|X\beta^* - X\beta\|_2^2 + 2\|X\beta - X\overline{\beta}^{k,i^*}\|_2^2 \\
&\leq 2\sigma^2(1 + t \log r)|S^*| + 2\|(\mathrm{I}_n - P_{S^*})X\beta\|_2^2 + 2\sigma^2(1 + t \log r)|\hat{S}^k \cup S^*| \\
&\quad + 2\|(\mathrm{I}_n - P_{\hat{S}^k \cup S^*})X\beta\|_2^2
\end{aligned}$$

$$\leq 2\sigma^2(1+t\log r)|S^*| + 2\|(\mathrm{I}_n - P_S)X\beta\|_2^2 + 2c|\hat{S}^k \cup S^*| + 2\|(\mathrm{I}_n - P_S)X\beta\|_2^2$$
$$= 2\sigma^2(1+t\log r)|S^*| + 0 + 2\sigma^2(1+t\log r)|\hat{S}^k \cup S^*| + 0\,.$$

Since $a \geq 2\sigma^2(1+t\log r)$, this contradicts (A.8) and, therefore, concludes the proof of Claim (i).

Claim (ii): On Ω, it holds that $\|X\overline{\beta} - X\beta\|_2^2 \leq (6a + 4\sigma^2(1+t\log r))|S^*|$.
To prove this claim, we note that by Claim 1, we have $\bar{\imath} \leq i^*$ and, therefore, $|\bar{S}| \leq |S^*|$. Hence, the definition of the estimator implies for $\bar{\imath} = r$ that

$$\|X\overline{\beta} - X\overline{\beta}^{\bar{\imath},i^*}\|_2^2 = \|X\overline{\beta} - X\overline{\beta}^{\bar{\imath},\bar{\imath}}\|_2^2 = 0$$

and otherwise, if $\bar{\imath} < r$, that (recall that $S^* \supset S$)

$$\|X\overline{\beta} - X\overline{\beta}^{\bar{\imath},i^*}\|_2^2 \leq a|\bar{S}| + a|\bar{S} \cup S^*| \leq 3a|S^*|\,.$$

The bound (A.1), on the other hand, yields

$$\|X\overline{\beta}^{\bar{\imath},i^*} - X\beta\|_2^2 \leq \sigma^2(1+t\log r)|\bar{S} \cup S^*| + 0 \leq 2\sigma^2(1+t\log r)|S^*|\,.$$

Combining these two inequalities, we finally obtain

$$\|X\overline{\beta} - X\beta\|_2^2 \leq 2\|X\overline{\beta} - X\overline{\beta}^{\bar{\imath},i^*}\|_2^2 + 2\|X\overline{\beta}^{\bar{\imath},i^*} - X\beta\|_2^2$$
$$\leq (6a + 4\sigma^2(1+t\log r))|S^*|\,,$$

which concludes the proof of Claim (ii). □

In particular, this yields the results of Section 2.

Proof of Proposition 2.1. Let $t = \max(\frac{2}{\alpha}, 6)$, and let R be large enough so that

$$t^{-1} + R^{-1}\sqrt{1+t\log R} \;\leq\; \frac{\alpha}{2} + \sqrt{\frac{1}{R^2} + \frac{2}{\alpha}\frac{\log R}{R^2}} \;\leq\; \alpha.$$

Using Corollary A.3 with t and $R_n = R$ gives the result. □

Proof of Theorem 2.3. Let $t = \max(\frac{2}{\alpha}, 6)$, and let R be large enough so that

$$t^{-1} + R^{-1}\sqrt{1+t\log R} \;\leq\; \frac{\alpha}{2} + \sqrt{\frac{1}{R^2} + \frac{2}{\alpha}\frac{\log R}{R^2}} \;\leq\; \alpha.$$

Using Corollary A.4 with t and $R_n = R$ gives the result. □

A.4. Proof of Theorem 2.2

The proof of Theorem 2.2 rely on various convex geometry notions, so we first remind the reader of some background on the subject.

The affine hull of a set A, denoted $\operatorname{aff} A$ is the intersection of all affine spaces that contain A, or alternatively, the unique affine set of minimal dimension that contains A. We write, by extension, $\dim A = \dim \operatorname{aff} A$. We denote the interior, closure, and boundary of A by $\operatorname{int} A$, $\operatorname{cl} A$, and ∂A, respectively. The relative interior and boundary of A, denoted $\operatorname{relint} A$ and $\operatorname{relbd} A$, are respectively the interior and the boundary when A is seen as a subset of its affine hull. We write $A \subset B$ if A is a (not necessarily strict) subset of B.

A half-space H^+ is a set of the form $\{x \in \mathbb{R}^n : \alpha^\top x \leq b\}$ for $\alpha \in \mathbb{R}^n$, $b \in \mathbb{R}$. Its boundary $H = \partial H^+ = \{x \in \mathbb{R}^n : \alpha^\top x = b\}$ is a hyperplane in $\mathbb{R}^n$. A polyhedron is a finite intersection of half-spaces, $X = \bigcap_{i \in I} H_i^+$. Such decompositions are usually not unique; we call a decomposition irreducible if $\bigcap_{j \neq i} H_j^+ \neq X$ for all $i \in I$. Given an irreducible decomposition, a facet of X is a set $F_i = X \cap H_i$. The faces are the intersections of (potentially many) facets. The normal cone to a point $x_0 \in X$ is $N(X, x) = \{y : y^\top (x - x_0) \leq 0 \text{ for all } x \in X\}$. One can show [31, Page 83] that for a given face F, all $x_0 \in F$ have the same normal cone; hence we define the normal cone to F to be $N(X, F) = N(X, x_0)$ for any $x_0 \in F$.

Next, we make the following remarks. Recall that $\mathcal{V}$ is the collection of regions, namely the closures of sets of points that have the same sign vector under the lasso.

By Lemma A.1, then the regions must have disjoint interiors. Indeed, for $V \neq V' \in \mathcal{V}$ we must have $V = \operatorname{cl} W^1(\eta)$, $V' = \operatorname{cl} W^1(\eta')$ for some $\eta \neq \eta' \subset \{-1, 0, 1\}^p$. Since $\operatorname{sgn} \hat{\beta}$ is constant on $\operatorname{int} V$, we conclude that $\operatorname{int} V \subset W^1(\eta)$, $\operatorname{int} V' \subset W^1(\eta')$. But $W^1(\eta) \cap W(\eta') = \{Y : \eta = \operatorname{sgn} \hat{\beta}^1 = \eta'\} = \emptyset$, so $\operatorname{int} V \cap V' = \emptyset$.

Take $V \neq V' \in \mathcal{V}$ again: being polyhedra, their boundaries ∂V, $\partial V'$ can be partitioned by the relative interiors of their proper faces [31, Theorem 2.1.2]. Let $\mathcal{F}(V)$ denote the set of proper faces of a polyhedron V and $\mathcal{C} = \{\operatorname{relint} F \cap \operatorname{relint} F' : F \in \mathcal{F}(V), F' \in \mathcal{F}(V'), V \neq V' \in \mathcal{V}\}$ be the collection of "boundary pieces". We enumerate, for reference, two properties of $\mathcal{C}$:

i) For two distinct $V, V' \in \mathcal{V}$ and an $x \in \partial V \cap \partial V'$, there is a unique $B \in \mathcal{C}$ such that $x \in B$.
ii) Each $B \in \mathcal{C}$ has dimension at most $n - 1$.

Indeed, for the first statement we notice that by partitioning, there exists unique faces $F \in \mathcal{F}(V)$ and $F' \in \mathcal{F}(V')$ such that $x \in \operatorname{relint} F \cap \operatorname{relint} F'$, hence a unique $B \in \mathcal{C}$ such that $x \in B$. For the second, since the interiors of V, V' are disjoint, for $B = \operatorname{relint} F \cap \operatorname{relint} F' \neq \emptyset$ we must have $F \in \partial V$,$F' \in \partial V'$, hence $\dim B \leq \dim \operatorname{relint} F \wedge \dim \operatorname{relint} F' \leq n - 1$. In addition to these observations, we will need the following two lemmas and one supporting proposition.

Lemma A.5. *If $B \in \mathcal{C}$ is of dimension at most $n-2$, then $\mathbb{R}_+ B$ is of dimension at most $n-1$.*

Proof. Let $S = \operatorname{aff} B$ be the affine hull of B, of dimension at most $n-2$. Being an affine subspace, it can be written $S = \{x : Ax + b = 0\}$ for some matrix $A \in \mathbb{R}^{n \times n}$ with $\operatorname{rank} A \leq n-2$ and some vector $b \in \mathbb{R}^n$. Now, $B \subset S$ implies $\mathbb{R}_+ B \subset \mathbb{R}S$, and $\mathbb{R}S = \{x : \exists t \text{ s.t. } Atx + b = 0\} = [I_n\, 0]\,\{(x,t) : Ax + tb = 0\} = [I_n\, 0] \operatorname{Ker}[A\, b]$. Now, since $\operatorname{rank} A \leq n-2$, $[A\, b]$ has rank at most $n-1$, and thus $\operatorname{Ker}[A\, b]$ is a subspace of $\mathbb{R}^{n+1}$ of rank at most $n-1$. Hence, $S = [I_n\, 0] \operatorname{Ker}[A\, b]$ has dimension at most $n-1$. Since $\mathbb{R}_+ B \subset \mathbb{R}S$, and $\mathbb{R}S$ is affine (actually, a subspace), then $\operatorname{aff} \mathbb{R}_+ B \subset \mathbb{R}S$. Consequently, $\dim \mathbb{R}_+ B \leq \dim \mathbb{R}S \leq n-1$, as desired. □

Proposition A.1. *Let $K \subset \mathbb{R}^n$ be a polyhedron of full dimension n with irreducible decomposition $K = \bigcap_j H_j^+$. Denote the hyperplanes by $H_i = \partial H_i^+$ as usual. Then for any facet $F_i = K \cap H_i$ of K and any point $x \in \operatorname{relint} F_i$, there exists an $\varepsilon > 0$ such that $\operatorname{int} H_i^+ \cap B(x,\varepsilon) \subset \operatorname{int} K$, $H_i \cap B(x,\varepsilon) \subset F_i$ and $H_i^{+C} \cap B(x,\varepsilon) \subset K^C$.*

Proof. With the irreducible decomposition, $x \in K \cap H_j^+$ for all j; say for $j \neq i$ we had $x \in F_j = K \cap H_j$. As argued in [17, Page 27], $x \in F_i \cap F_j$ implies that x is in a facet of F_i, hence in ∂F_i, a contradiction with $x \in \operatorname{relint} F_i$. Thus $x \in K \cap \operatorname{int} H_j^+$ for all $j \neq i$.

Since $x \in \bigcap_{j \neq i} \operatorname{int} H_j^+$, we can find $\varepsilon > 0$ such that $B(x,\varepsilon) \subset \bigcap_{j \neq i} \operatorname{int} H_j^+$. Then $H_i^+ \cap B(x,\varepsilon) \subset \bigcap_j \operatorname{int} H_j^+ \subset \operatorname{int} K$, $H_i \cap B(x,\varepsilon) \subset H_i \cap \bigcap_{j \neq i} \operatorname{int} H_j^+ \subset H_i \cap K = F_i$, and finally $H_i^{+C} \cap B(x,\varepsilon) \subset \bigcup_j H_j^{+C} = K^C$, as desired. □

Lemma A.6. *Let $B \in \mathcal{C}$ be of dimension $n-1$ and L be a ray centered at the origin such that $B \cap L \neq \emptyset$. Then either $\mathbb{R}_+ B$ has dimension $n-1$, or $L \cap \operatorname{int} V \neq \emptyset$ and $L \cap \operatorname{int} V' \neq \emptyset$.*

Proof. Write $B = \operatorname{relint} F \cap \operatorname{relint} F'$. Since $\operatorname{relint} F$, $\operatorname{relint} F'$ both have dimension at most $n-1$, and B has dimension $n-1$, then they must have exactly dimension $n-1$. They are thus facets of their respective polyhedra $V, V' \in \mathcal{V}$, which must have dimension n, and have exactly one supporting hyperplane. Thus $\operatorname{aff} B = \operatorname{aff} F = \operatorname{aff} F'$, which we might denote S. Let $b \in L \cap S$, write $S = \langle a_1, ..., a_{n-1} \rangle + b$ for some linearly independent vectors $a_1, ..., a_{n-1}$, and let a_n be such that $L = \mathbb{R}_+ a_n$. Consider the affine transformation

$$\phi(x) = [a_1, ..., a_{n-1}, a_n]\, x + b = Ax + b,$$

which maps vectors $(x_1, ..., x_{n-1}, 0)$ bijectively to S, and vectors $(0, ..., 0, x_n)$ with $x_n \geq -\|b\|/\|a_n\|$ bijectively to L. (Recall that $b \in S \cap L$.) There are then two possibilities.

Case i) A has rank $n-1$. Then $a_n \subset \langle a_1, ...a_{n-1} \rangle$ and $L \subset S$. But this means that $0 \in L \subset S$, that is, that S is a subspace. Then $\mathbb{R}_+ B \subset \mathbb{R}S = S$, that is, $\mathbb{R}_+ B$ has dimension $n-1$.

Case ii) A has rank n. Then ϕ is bijective and $L \cap S$ is the singleton $\{b\}$. Recall that V and V' are polyhedra of dimension n, and let $V = \bigcap_i H_i^+$, $V' = \bigcap_i H_i'^+$ be irreducible representations into half-spaces H_i^+, $H_j'^+$ with boundary H_i, H_j'. By irreducibility, there are unique indices i, j such that $\operatorname{aff} F = H_i = \operatorname{aff} F' = H_j' = S$. Since $b \in \operatorname{relint} F \cap \operatorname{relint} F'$, by Proposition A.1 there must be an $\varepsilon > 0$ small enough so that $\operatorname{int} H_i^+ \cap B(b,\varepsilon) \subset \operatorname{int} V$, $H_i^{+C} \cap B(b,\varepsilon) \subset V^C$, $\operatorname{int} H_j' \cap B(b,\varepsilon) \subset \operatorname{int} V'$ and $H_j'^{+C} \cap B(b,\varepsilon) \subset V'^C$.

But clearly $H_i^+ \neq H_j'^+$, as otherwise $\emptyset \neq B(b,\varepsilon) \cap \operatorname{int} H_i^+ = B(b,\varepsilon) \cap \operatorname{int} H_j'^+ \subset \operatorname{int} V \cap \operatorname{int} V' = \emptyset$, a contradiction since interiors of regions are disjoint. Thus it must hold that $\operatorname{int} H_i^+ = H_j'^{+C}$. In light of this, we may simplify the notation by writing $S^+ = \operatorname{int} H_i^+$ and $S^- = H_j'^{+C}$. The state of affairs is then that $B(b,\varepsilon) \cap S^+ \subset \operatorname{int} V$, $B(b,\varepsilon) \cap S \subset B$ and $B(b,\varepsilon) \cap S^- \subset \operatorname{int} V'$.

Write $\mathbb{R}^{(n-1)+} = \{x \in \mathbb{R}^n : x_n > 0\}$ and $\mathbb{R}^{(n-1)-} = \{x \in \mathbb{R}^n : x_n < 0\}$. Since ϕ is open, it must map connected components to connected components, and being bijective it must hold that $\phi(\mathbb{R}^{(n-1)+}) = S^+$ and $\phi(\mathbb{R}^{(n-1)-}) = S^-$, or vice versa. Fix the former by considering $x \mapsto -Ax + b$ instead of $x \mapsto Ax + b$ if necessary. Since $\phi^{-1}(B(b,\varepsilon))$ is an open set, there must be an ε' such that $B(0,\varepsilon') \subset \phi^{-1}(B(b,\varepsilon))$.

Let $\varepsilon'' = \varepsilon' \wedge \|b\|_2 / \|a_n\|_2$, so that $|t| < \varepsilon''$ implies $ta_n + b \in L$. Then, we have that $\emptyset \neq \phi(\{(0,...,0,t), t \in (0,\varepsilon'')\}) \subset S^+ \cap L \cap B(b,\varepsilon) \subset L \cap \operatorname{int} V$ and $\emptyset \neq \phi(\{(0,...,0,t), t \in (-\varepsilon'',0)\}) \subset S^- \cap L \cap B(b,\varepsilon) \subset L \cap \operatorname{int} V'$. Thus, $L \cap \operatorname{int} V$ and $L \cap \operatorname{int} V'$ are non-empty, as desired. □

We may now turn to the proof of the theorem.

Proof of Theorem 2.2. Every region $V \in \mathcal{V}$ is a polyhedron, so has a decomposition $V = P(V) + C(V)$ into a polytope $P(V)$ and a cone $C(V)$ by Minkowski's theorem [42, Theorem 1.2]. Define $\mathcal{V}_0 = \{V \in \mathcal{V} : \operatorname{int} V \cap \mathbb{R}_+ \xi \neq \emptyset\}$, $T = \bigcup_{V \in \mathcal{V}_0} V$ and

$$R = \bigcup_{\substack{B \in \mathcal{C} \\ \dim \mathbb{R}_+ B \\ \leq n-1}} \mathbb{R}_+ B \quad \cup \bigcup_{\substack{V \in \mathcal{V} \\ \dim C(V) \\ \leq n-1}} C(V) \quad \cup \bigcup_{V \in \mathcal{V}} \partial C(V).$$

The set R is a finite union of closed sets of dimension at most $n-1$, so is closed and has measure zero. We argue that if $\xi \in R^C$, then $\mathbb{R}_+ \xi \subset \operatorname{int} T$. Indeed, say that $t\xi \in \partial T$ for some $t > 0$. Since $\partial T \subset \bigcup_{V \in \mathcal{V}_0} \partial V$, there is a $V_0 \in \mathcal{V}_0$ such that $t\xi \in \partial V_0$. Since $\mathbb{R}^n = \bigcup_{V \in \mathcal{V}} V$ and $t\xi \in \partial T$, we have $t\xi \in \operatorname{cl}(T^C) = \bigcup_{V \notin \mathcal{V}_0} V$. Thus there must also be a $V_1 \neq V_0$, $V_1 \notin \mathcal{V}_0$ such that $t\xi \in V_1$. But since the interiors are disjoint, if $t\xi \in \operatorname{int} V_1$ there would be a contradiction with $t\xi \in \partial V_0$; hence $t\xi \in \partial V_1$. Thus $t\xi \in \partial V_0 \cap \partial V_1$ and there must be a unique $B \in \mathcal{C}$ such that $t\xi \in B$. That piece, like all elements of $\mathcal{C}$, must be of dimension $n-1$ or lower.

We argue that $B \subset R$. If it has dimension $n-2$, then $\dim \mathbb{R}_+ B \leq n-1$ by Lemma A.5, so B is indeed a subset of R. Now say instead it has dimension $n-1$ and recall that $t\xi \in \mathbb{R}_+ \xi \cap B$. Let $F \in \mathcal{F}(V_0)$ and $F' \in \mathcal{F}(V_1)$ be such that

$B = \operatorname{relint} F \cap \operatorname{relint} F'$. By Lemma A.6, we must have either $\dim \mathbb{R}_+ B = n-1$, or $\mathbb{R}_+\xi \cap \operatorname{int} V_0 \neq \emptyset$ and $\mathbb{R}_+\xi \cap \operatorname{int} V_1 \neq \emptyset$. But if the latter was the case, then $V_1 \subset T$ by definition, which is impossible; thus we must have $\dim \mathbb{R}_+ B = n-1$, hence $B \subset R$ again.

Thus in all cases, $B \subset R$. Let $d(x, A) := \inf_{y\in A} \|x-y\|_2$ denote the Euclidean distance between a point x and a set A. Since $t\xi \in B$, we have $d(t\xi, R) = 0$. But at the same time, since $\xi \in R^C$ and R is closed we must have $d(\xi, R) > 0$, and since R is invariant under positive multiplication,

$$d(t\xi, R) = \inf_{y\in R} \|t\xi - y\|_2 = t \inf_{y/t\in R} \|\xi - y\|_2 = t \inf_{y\in R} \|\xi - y\|_2 = td(\xi, R) > 0.$$

This is a contradiction, and we conclude that $\mathbb{R}_+\xi \subset \operatorname{int} T$.

Now, $\mathcal{V}_0$ is finite, since C has only a finite number of faces. Let h be the continuous map $t \mapsto t\xi$, and consider for each $V \in \mathcal{V}_0$ the closed set $h^{-1}(V)$. This set must be convex, since for $s, t \in h^{-1}(V)$, $h(\gamma s + (1-\gamma)t) = [\gamma s + (1-\gamma)t]\xi = \gamma[s\xi] + (1-\gamma)[t\xi] \in V$ by convexity of V for any $\gamma \in [0, 1]$. The only closed, convex sets of $\mathbb{R}$ are the closed intervals: thus $h^{-1}(V) = [\alpha, \beta]$ for some $\alpha \leq \beta$.

Enumerate arbitrarily the $V \in \mathcal{V}_0$ as $V_1, ..., V_m$, and for $V_i \in \mathcal{V}_0$ let $[\alpha_i, \beta_i] = h^{-1}(V_i)$. Now, $\mathbb{R}_+\xi \subset T = \bigcup_{i=1}^m V_i$, so $h^{-1}(T) = \bigcup_{i=1}^m [\alpha_i, \beta_i] = \mathbb{R}_+$. Then some β_i must equal ∞, otherwise the union would be bounded. Moreover, since the interiors of the V's are disjoint, $(\alpha_i, \beta_i) \cap (\alpha_j, \beta_j) = \emptyset$ for $i \neq j$ and the $\beta_i = \infty$ must be unique, all the others finite. By reordering if necessary, take $0 = \alpha_1 < \beta_1 \leq \alpha_2 < \beta_2 \leq ... \leq \alpha_m < \beta_m = \infty$.

The region V_m is a polyhedron, so has a decomposition as $V_m = P(V_m) + C(V_m)$ for some polytope $P(V_m)$ and cone $C(V_m)$ by Minkowski's theorem [42, Theorem 1.2]. Fix a point $t_0 \in (\alpha_m, \infty)$; then since $t_0\xi + \mathbb{R}_+\xi = (t_0, \infty)\xi \subset \operatorname{int} V_m$, by [17, 2.5.1], we conclude that $\mathbb{R}_+\xi \subset C(V_m)$, so $t_0\xi, \xi \in C(V_m)$. If $C(V_m)$ has dimension at most $n-1$, or $t_0\xi \in \partial C(V_m) \Leftrightarrow \xi \in \partial C(V_m)$, then $\xi \in R$, which contradicts $\xi \in R^C$ – thus $C(V_m)$ must have dimension n and $t_0\xi, \xi \in \operatorname{int} C(V_m)$. Let ε_1 be small enough so that $B(\xi, \varepsilon_1) \subset \operatorname{int} C(V_m)$. Then for any $s > 0$, $B([t_0 + s]\xi, s\varepsilon_1) = t_0\xi + sB(\xi, \varepsilon_1) \subset \operatorname{int} C_m \subset \operatorname{int} V_m$. Thus for all $s > 2t_0\|\xi\|_2/\varepsilon_1$, $B(s\xi, s\varepsilon_1/2) \subset B([t_0 + s]\xi, s\varepsilon_1) \subset \operatorname{int} V_m \subset \operatorname{int} T$.

Next, notice that the segment $[0, 2t_0\|\xi\|_2/\varepsilon_1]\xi$ is compact and in $\operatorname{int} T$. This implies that $d([0, 2t_0\|\xi\|_2/\varepsilon_1]\xi, \operatorname{int} T^C)$ is strictly positive and also that there must be an ε_2-neighborhood such that $B([0, 2t_0\|\xi\|_2/\varepsilon_1]\xi, \varepsilon_2) \subset \operatorname{int} T$. Hence, for all $s \in [0, 2t_0\|\xi\|_2/\varepsilon_1]$, it holds $B(s\xi, s\varepsilon_2\varepsilon_1/2t_0\|\xi\|_2) \subset B(s\xi, \varepsilon_2) \subset \operatorname{int} T$.

Finally, let $\varepsilon = \min(\varepsilon_1/2, \varepsilon_1\varepsilon_2/2t_0\|\xi\|_2)$. Then for all $s \geq 0$, we have $B(s\xi, s\varepsilon) \subset \operatorname{int} T$. Let $|y - \xi| < \varepsilon$ and define $\eta = \operatorname{sgn} \hat{\beta}(sy)$. Then $\operatorname{cl} W(\eta) = V_0$ for some $V_0 \in \mathcal{V}_0$, since otherwise $\operatorname{cl} W(\eta) \subset \operatorname{cl} T^C = \bigcup_{V\notin\mathcal{V}_0} V$, which would contradict $sy \in \operatorname{int} T$. But there is a $t > 0$ such that $t\xi \in \operatorname{int} V_0$, since $V_0 \in \mathcal{V}_0$, and since $\operatorname{sgn} \hat{\beta}$ is constant over $\operatorname{int} V_0$, $\operatorname{sgn} \hat{\beta}(t\xi) = \operatorname{sgn} \hat{\beta}(sy)$. Thus in particular $\hat{S}[sy] = \hat{S}[t\xi]$. Since this is true for all $|y-\xi| < \varepsilon$, we conclude that $D(\xi) \geq \varepsilon > 0$, as desired. □

Appendix B: Description of the lslassoBIC

In this section, we provide details for the lslassoBIC implementation that we have used. This method is similar to applying a BIC procedure over the refitted models obtained by a lasso path, and was recently analyzed in [2].

We consider the same collection of tuning parameters $\Lambda = \{\lambda_1, \dots, \lambda_r\}$ as before, and we denote the associated supports by $(\hat{S}^1, \dots, \hat{S}^r)$, where $\hat{S}^i := \mathrm{supp}[\hat{\beta}^{\lambda_i}]$. Following Equation (1.3), we write $(\bar{\beta}^{\lambda_1}, \dots, \bar{\beta}^{\lambda_r})$ for the estimated least-squares over these supports. Let us introduce for each $j \in [r]$ a prior π_j via

$$\pi_j = \left(H_p \binom{p}{|\hat{S}^j|} \exp(|\hat{S}^j|) \right)^{-1}$$

with $H_p = (e - e^{-p})/(e-1)$. Then, the lslassoBIC is defined by

$$\bar{\beta}^{\mathrm{lslassoBIC}} = \bar{\beta}^{\lambda_{j^\star}} \quad \text{with} \quad j^\star \in \operatorname*{arg\,min}_{j \in [r]} \left(\|Y - X\bar{\beta}^{\lambda_j}\|_2^2 + 14\hat{\sigma}^2 \log \frac{1}{\pi_j} \right) ,$$

where $\hat{\sigma}$ is a standard deviation estimate of the (Gaussian) noise. As the practical estimation of σ is not discussed further in [2], we have used the same estimator as for $\mathrm{AV}_{\mathrm{Pr}}$, namely the one defined in Algorithm 2.

Note that we have adapted the method proposed by [2] to the case of a predetermined number of lasso parameters. This is because the number of kinks over the lasso path can be as large as $(3^p + 1)/2$ [26], making an estimator based on the entire collection of kinks intractable.

Acknowledgements

We thank Michaël Chichignoud for the insightful remarks and for the inspiring discussions, and we thank Pierre Bellec for providing us with valuable insights about numerical aspects of his work. We also thank the editors and reviewers for the valuable comments that have improved the paper.

References

[1] A. Antoniadis. Comments on: ℓ_1-penalization for mixture regression models. *TEST*, 19(2):257–258, 2010. MR2677723

[2] P. Bellec. Aggregation of supports along the Lasso path. In *COLT*, pages 488–529, 2016.

[3] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013. MR3037163

[4] A. Belloni, V. Chernozhukov, and L. Wang. Square-root Lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011. MR2860324

[5] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009. MR2533469

[6] P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications. MR2807761

[7] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194 (electronic), 2007. MR2312149

[8] F. Bunea, Y. She, H. Ombao, A. Gongvatana, K. Devlin, and R. Cohen. Penalized least squares regression methods and applications to neuroimaging. *Neuroimage*, 55, 2011.

[9] F. Bunea, J. Lederer, and Y. She. The group square-root Lasso: Theoretical properties and fast algorithms. *IEEE Trans. Inf. Theory*, 60(2):1313–1325, 2014. MR3164977

[10] S. Chatterjee and J. Jafarov. Prediction error of cross-validated lasso. *arXiv:1502.06291*, 2015.

[11] M. Chichignoud and J. Lederer. A robust, adaptive M-estimator for pointwise estimation in heteroscedastic regression. *Bernoulli*, 20(3):1560–1599, 2014. MR3217454

[12] M. Chichignoud, J. Lederer, and M. Wainwright. Tuning Lasso for sup-norm optimality. *J. Mach. Learn. Res.*, 17, 2016. MR3595165

[13] A. S. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the Lasso. *Bernoulli*, 23(1):552–581, 2017. MR3556784

[14] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1):1–22, 2010.

[15] C. Giraud, S. Huet, and N. Verzelen. High-dimensional regression with unknown variance. *Statist. Sci.*, 27(4):500–518, 2012. MR3025131

[16] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

[17] B. Grünbaum. *Convex Polytopes*. Springer-Verlag, New York, second edition, 2003. MR1976856

[18] N. Harris and A. Sepehri. The accessible lasso models. *arXiv:1501.02559*, 2015.

[19] M. Hebiri and J. Lederer. How correlations influence Lasso prediction. *IEEE Transactions on Information Theory*, 59:1846–1854, 2013. MR3030757

[20] V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Heidelberg, 2011. MR2829871

[21] J. Lederer. Trust, but verify: benefits and pitfalls of least-squares refitting in high dimensions. *arXiv:1306.0113 [stat.ME]*, 2013.

[22] J. Lederer and C. Müller. Don't fall for tuning parameters: Tuning-free variable selection in high dimensions with the trex. In *Proceed-*

ings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[23] D. Lee, J.and Sun and Y. Sun. Exact post-selection inference, with applications to the lasso. *Preprint arXiv:1311.6238v5*, 2015.

[24] O. Lepski. On a problem of adaptive estimation in gaussian white noise. *Theory Probab. Appl.*, 35(3):454–466, 1990. MR1091202

[25] O. Lepski, E. Mammen, and V. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.*, 25(3):929–947, 1997. MR1447734

[26] J. Mairal and B. Yu. Complexity analysis of the lasso regularization path. *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[27] N. Meinshausen and P. Bühlmann. Stability selection. *J. Roy. Statist. Soc. Ser. B*, 72(4):417–473, 2010. MR2758523

[28] A. Owen. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics*, 443:59–72, 2007. MR2433285

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011. MR2854348

[30] J. Sabourin, W. Valdar, and A. Nobel. A permutation approach for selecting the penalty parameter in penalized model selection. *Biometrics*, 71:1185–1194, 2015. MR3436744

[31] R. Schneider. *Convex bodies: the Brunn–Minkowski theory*, volume 151 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, second edition, 2013. MR3155183

[32] R. Shah and R. Samworth. Variable selection with error control: another look at stability selection. *J. Roy. Statist. Soc. Ser. B*, 75(1):55–80, 2013. MR3008271

[33] J. Shao and X. Deng. Estimation in high-dimensional linear models with deterministic design matrices. *Ann. Statist.*, 40(2):812–831, 2012. MR2933667

[34] N. Städler, P. Bühlmann, and Sara s van de Geer. ℓ_1-penalization for mixture regression models. *TEST*, 19(2):209–256, 2010. MR2677722

[35] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012. MR2999166

[36] T. Sun and C.-H. Zhang. Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res.*, 14:3385–3418, 2013. MR3144466

[37] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996. MR1379242

[38] R. J. Tibshirani and J. Taylor. Degrees of freedom in lasso problems. *Ann. Statist.*, 40(2):1198–1232, 2012. MR2985948

[39] S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392, 2009. MR2576316

[40] X. Wang, D. Dunson, and C. Leng. No penalty no tears: Least squares in high-dimensional linear models. *arXiv:1506.02222*, 2015.

[41] L. Wasserman and K. Roeder. High dimensional variable selection. *Ann. Stat.*, 37(5A):2178, 2009. MR2543689

[42] G. M. Ziegler. *Lectures on polytopes*, volume 152. Springer, 1995. MR1311028