Journal of Machine Learning Research 22 (2021) 1-32 Submitted 5/20; Revised 5/21; Published 10/21

Estimating the Lasso's Effective Noise

Johannes Lederer JOHANNES.LEDERER@RUB.DE
Department of Mathematics
Ruhr-University Bochum
44801 Bochum, Germany

Michael Vogt M.VOGT@UNI-ULM.DE
Institute of Statistics
Department of Mathematics and Economics
Ulm University
89081 Ulm, Germany

Editor: Massimiliano Pontil

Abstract

Much of the theory for the lasso in the linear model $Y = \boldsymbol{X}\beta^* + \varepsilon$ hinges on the quantity $2\|\boldsymbol{X}^\top\varepsilon\|_\infty/n$, which we call the lasso's effective noise. Among other things, the effective noise plays an important role in finite-sample bounds for the lasso, the calibration of the lasso's tuning parameter, and inference on the parameter vector β^*. In this paper, we develop a bootstrap-based estimator of the quantiles of the effective noise. The estimator is fully data-driven, that is, does not require any additional tuning parameters. We equip our estimator with finite-sample guarantees and apply it to tuning parameter calibration for the lasso and to high-dimensional inference on the parameter vector β^*.

Keywords: high-dimensional regression, lasso, finite-sample guarantees, tuning parameter calibration, high-dimensional inference

1. Introduction

Consider the high-dimensional linear model $Y = \boldsymbol{X}\beta^* + \varepsilon$ with response vector $Y \in \mathbb{R}^n$, design matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, target vector $\beta^* \in \mathbb{R}^p$, and random noise $\varepsilon \in \mathbb{R}^n$. We allow for a dimension p that is of the same order or even much larger than the sample size n, and we assume a target vector β^* that is sparse. A popular estimator of β^* in this framework is the lasso (Tibshirani, 1996)

$$\hat{\beta}_\lambda \in \mathop{\arg\min}_{\beta\in\mathbb{R}^p}\left\{\frac{1}{n}\|Y - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_1\right\}, \tag{1}$$

where $\lambda \in [0,\infty)$ is a tuning parameter. The lasso estimator satisfies the well-known prediction bound

$$\lambda \geq \frac{2\|\boldsymbol{X}^\top\varepsilon\|_\infty}{n} \quad \Longrightarrow \quad \frac{1}{n}\|\boldsymbol{X}(\beta^* - \hat{\beta}_\lambda)\|_2^2 \leq 2\lambda\|\beta^*\|_1, \tag{2}$$

which is an immediate consequence of the basic inequality for the lasso (Bühlmann and van de Geer, 2011, Lemma 6.1) and Hölder's inequality. This simple bound highlights that a crucial quantity in the analysis of the lasso estimator is $2\|\boldsymbol{X}^\top\varepsilon\|_\infty/n$. We call this quantity henceforth the *effective noise*.

The effective noise does not only play a central role in the stated prediction bound but rather in almost all known finite-sample bounds for the lasso. Such bounds, called oracle inequalities, are generally of the form (Bühlmann and van de Geer, 2011; Giraud, 2014; Hastie et al., 2015; Lederer, 2021)

$$\lambda \geq (1+\delta)\,\frac{2\|\boldsymbol{X}^\top\varepsilon\|_\infty}{n} \quad\Longrightarrow\quad \|\beta^*-\hat{\beta}_\lambda\| \leq \kappa\lambda \tag{3}$$

with some constant $\delta \in [0,\infty)$, a factor $\kappa = \kappa(\beta^*)$ that may depend on β^*, and a (pseudo-)norm $\|\cdot\|$. Oracle inequalities of the form (3) are closely related to tuning parameter calibration for the lasso: they suggest to control the loss $L(\beta^*,\hat{\beta}_\lambda) = \|\beta^*-\hat{\beta}_\lambda\|$ of the lasso estimator $\hat{\beta}_\lambda$ by taking the smallest tuning parameter λ for which the bound $\|\beta^*-\hat{\beta}_\lambda\| \leq \kappa\lambda$ holds with probability at least $1-\alpha$ for some given $\alpha \in (0,1)$. Denoting the $(1-\alpha)$-quantile of the effective noise $2\|\boldsymbol{X}^\top\varepsilon\|_\infty/n$ by λ^*_α, we immediately derive from the oracle inequality (3) that

$$\mathbb{P}\Big(\|\beta^*-\hat{\beta}_{(1+\delta)\lambda}\| \leq \kappa(1+\delta)\lambda\Big) \geq 1-\alpha \tag{4}$$

for $\lambda \geq \lambda^*_\alpha$. Stated differently, $\lambda = (1+\delta)\lambda^*_\alpha$ is the smallest tuning parameter for which the oracle inequality (3) yields the finite-sample bound $\|\beta^*-\hat{\beta}_\lambda\| \leq \kappa\lambda$ with probability at least $1-\alpha$. Importantly, the tuning parameter choice $\lambda = (1+\delta)\lambda^*_\alpha$ is not feasible in practice, since the quantile λ^*_α of the effective noise is not observed. An immediate question is, therefore, whether the quantile λ^*_α can be estimated.

The effective noise is also closely related to high-dimensional inference. To give an example, we consider testing the null hypothesis $H_0 : \beta^* = 0$ against the alternative $H_1 : \beta^* \neq 0$. Testing this hypothesis corresponds to an important question in practice: do the regressors in the model $Y = \boldsymbol{X}\beta^* + \varepsilon$ have any effect on the response at all? A test statistic for the hypothesis H_0 is given by $T = 2\|\boldsymbol{X}^\top Y\|_\infty/n$. Under H_0, it holds that $T = 2\|\boldsymbol{X}^\top\varepsilon\|_\infty/n$, that is, T is the effective noise. A test based on the statistic T can thus be defined as follows: reject H_0 at the significance level α if $T > \lambda^*_\alpha$. Since the quantile λ^*_α is not observed, this test is not feasible in practice, which brings us back to the question of whether the quantile λ^*_α can be estimated.

In this paper, we devise a novel estimator of the quantile λ^*_α based on bootstrap. Besides the level $\alpha \in (0,1)$, it does not depend on any free parameters, which means that it is fully data-driven. The estimator can be used to approach a number of statistical problems in the context of the lasso. We focus on two such problems: (i) tuning parameter calibration for the lasso and (ii) inference on the parameter vector β^*. The idea of using an estimator of the quantile λ^*_α to approach statistical issues such as (i) and (ii) is very natural and by no means new. Belloni and Chernozhukov (2013), for example, choose the tuning parameter of the lasso based on an estimator of λ^*_α. Similar procedures for the square-root lasso and the Dantzig selector are considered in Belloni et al. (2011) and Chernozhukov et al. (2013), respectively. However, these methods are quite limited as they presume that either the noise distribution or a good initial guess for the lasso's tuning parameter is known. Even though our estimator builds on ideas from the aforementioned papers, it is markedly different from the methods considered there and goes beyond them in important aspects. We discuss this in detail in Section 3 after introducing our estimator.

We now briefly summarize the main contributions of our paper with regards to the two statistical problems (i) and (ii).

(i) *Tuning parameter calibration for the lasso.* Our estimator $\hat{\lambda}_\alpha$ of the quantile λ^*_α can be used to calibrate the lasso with essentially optimal finite-sample guarantees. Specifically, we derive finite-sample statements of the form

$$\mathbb{P}\Big(\|\beta^* - \hat{\beta}_{(1+\delta)\hat{\lambda}_\alpha}\| \leq \kappa(1+\delta)\lambda^*_{\alpha-\nu_n}\Big) \geq 1-\alpha-\eta_n, \tag{5}$$

where $0 < \nu_n \leq Cn^{-K}$ and $0 < \eta_n \leq Cn^{-K}$ for some positive constants C and K. Statement (5) shows that calibrating the lasso with the estimator $\hat{\lambda}_\alpha$ yields almost the same finite-sample bound on the loss $L(\beta^*, \beta) = \|\beta^* - \beta\|$ as calibrating it with the oracle parameter λ^*_α. In particular, (5) is almost as sharp as the oracle bound $\mathbb{P}(\|\beta^* - \hat{\beta}_{(1+\delta)\lambda^*_\alpha}\| \leq \kappa(1+\delta)\lambda^*_\alpha) \geq 1-\alpha$, which is obtained by plugging $\lambda = \lambda^*_\alpha$ into (4).

Finite-sample guarantees for the practical calibration of the lasso's tuning parameter are scarce. Exceptions include finite-sample bounds for Adaptive Validation (AV) (Chichignoud et al., 2016) and Cross-Validation (CV) (Chetverikov et al., 2021). One advantage of our approach via the effective noise is that it yields finite-sample guarantees not only for a specific loss but for any loss for which an oracle inequality of the type (3) is available. Another advantage is that it does not depend on secondary tuning parameters that are difficult to choose in practice; the only parameter it depends on is the level $1-\alpha$, which plays a similar role as the significance level of a test and, therefore, can be chosen in the same vein in practice.

(ii) *Inference on the parameter vector β^*.* Our estimator $\hat{\lambda}_\alpha$ of the quantile λ^*_α can also be used to test hypotheses on the parameter vector β^* in the model $Y = \boldsymbol{X}\beta^* + \varepsilon$. Consider again the problem of testing $H_0 : \beta^* = 0$ against $H_1 : \beta^* \neq 0$. Our approach motivates the following test: reject H_0 at the significance level α if $T > \hat{\lambda}_\alpha$. We prove under mild regularity conditions that this test has the correct level α under H_0 and is consistent against alternatives that are not too close to H_0. Moreover, we show that the test can be generalized readily to more complex hypotheses.

High-dimensional inference based on the lasso has turned out to be a very difficult problem. Some of the few advances that have been made in recent years include tests for the significance of small, fixed groups of parameters (Belloni et al., 2013; Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014; Gold et al., 2020), tests for the significance of parameters entering the lasso path (Lockhart et al., 2014), rates for confidence balls for the entire parameter vector (and infeasibility thereof) (Nickl and van de Geer, 2013; Cai and Guo, 2018), and methods for inference after model selection (Lee et al., 2016; Tibshirani et al., 2016). In stark contrast to most other methods for high-dimensional inference, our tests are completely free of tuning parameters and, therefore, dispense with any fine-tuning (such as the calibration of multiple lasso tuning parameters in the first group of papers cited above).

The paper is organized as follows. In Section 2, we detail the modeling framework. Our estimator of the quantiles of the effective noise is developed in Section 3. In Section 4, we apply the estimator to tuning parameter calibration and inference for the lasso. Our theoretical analysis is complemented by a simulation study in Section 5, which investigates the finite-sample performance of our methods.

2. Model Setting

We consider the standard linear model

$$Y = \boldsymbol{X}\beta^* + \varepsilon, \tag{6}$$

where $Y = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ is the response vector, $\boldsymbol{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n \times p}$ is the design matrix with the vectors $X_i = (X_{i1}, \ldots, X_{ip})^\top$, $\beta^* = (\beta_1^*, \ldots, \beta_p^*)^\top \in \mathbb{R}^p$ is the parameter vector, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$ is the noise vector. We are particularly interested in high-dimensional versions of the model, that is, $p \approx n$ or even $p \gg n$. Throughout the paper, we assume the design matrix $\boldsymbol{X}$ to be random, but our results carry over readily to fixed design matrices. We impose the following regularity conditions on the model (6):

(C1) The random variables (X_i, ε_i) are independent across i.

(C2) The covariates X_{ij} have bounded support, that is, $|X_{ij}| \le C_X$ for all i, j and some sufficiently large constant $C_X < \infty$. Moreover, $n^{-1} \sum_{i=1}^n \mathbb{E}[X_{ij}^2] \ge c_X^2$ for some constant $c_X > 0$.

(C3) The noise variables ε_i are such that $\mathbb{E}[\varepsilon_i | X_i] = 0$ and $\mathbb{E}[|\varepsilon_i|^\theta] \le C_\theta < \infty$ for some $\theta > 4$ and all i. Moreover, the conditional noise variance $\sigma^2(X_i) = \mathbb{E}[\varepsilon_i^2 | X_i]$ satisfies $0 < c_\sigma^2 \le \sigma^2(\cdot) \le C_\sigma^2 < \infty$ with some suitable constants c_σ and C_σ.

(C4) It holds that $p \le C_r n^r$, where $r > 0$ is an arbitrarily large but fixed constant and $C_r > 0$.

(C5) There exist a constant $C_\beta < \infty$ and some small $\delta_\beta > 0$ such that $\|\beta^*\|_1 \le C_\beta n^{1/2 - \delta_\beta}$.

Condition (C1) stipulates independence across the observations, but the observations need not be identically distributed. The assumption about the boundedness of the covariates X_{ij} in (C2) makes the derivations more lucid but can be relaxed to sufficiently strong moment conditions on the variables X_{ij}. Assumption (C3) on the moments of the noise terms ε_i is quite mild: only a bit more than the first four moments are required to exist. Condition (C4) on the relationship between n and p is mild as well: p is allowed to grow as any polynomial of n. Condition (C5) imposes sparsity on the parameter vector β^* in an ℓ_1-sense. One could also replace it by a similar assumption in terms of the ℓ_0-norm. However, an advantage of the ℓ_1-version is that it allows for approximate sparsity—see e.g. Section 3.2 in van de Geer and Lederer (2013) or Section 2.8 in van de Geer (2016).

3. Estimating the Effective Noise

3.1 Definition of the Estimator

Let λ_α^* be the $(1-\alpha)$-quantile of the effective noise $2\|\boldsymbol{X}^\top \varepsilon\|_\infty / n$, which is formally defined as $\lambda_\alpha^* = \inf\{q : \mathbb{P}(2\|\boldsymbol{X}^\top \varepsilon\|_\infty / n \le q) \ge 1 - \alpha\}$. We estimate λ_α^* as follows: for any λ, let

$\hat{\varepsilon}_\lambda = Y - \boldsymbol{X}\hat{\beta}_\lambda$ be the residual vector that results from fitting the lasso with the tuning parameter λ, and let $e = (e_1, \ldots, e_n)^\top$ be a standard normal random vector independent of the data $(\boldsymbol{X}, Y)$. Define the criterion function

$$\hat{Q}(\lambda, e) = \max_{1 \le j \le p} \Big| \frac{2}{n} \sum_{i=1}^{n} X_{ij} \hat{\varepsilon}_{\lambda,i} e_i \Big|,$$

and let $\hat{q}_\alpha(\lambda)$ be the $(1-\alpha)$-quantile of $\hat{Q}(\lambda, e)$ conditionally on $\boldsymbol{X}$ and Y. Formally, $\hat{q}_\alpha(\lambda) = \inf\{q : \mathbb{P}_e(\hat{Q}(\lambda, e) \le q) \ge 1 - \alpha\}$, where we use the shorthand $\mathbb{P}_e(\,\cdot\,) = \mathbb{P}(\,\cdot \mid \boldsymbol{X}, Y)$. Our estimator of λ^*_α is defined as

$$\hat{\lambda}_\alpha = \inf \big\{ \lambda > 0 \ : \ \hat{q}_\alpha(\lambda') \le \lambda' \text{ for all } \lambda' \ge \lambda \big\}. \tag{7}$$

In practice, $\hat{\lambda}_\alpha$ can be computed by the following algorithm:

Step 1: For some large natural number M, specify a grid of points $0 < \lambda_1 < \ldots < \lambda_M = \overline{\lambda}$, where $\overline{\lambda} = 2\|\boldsymbol{X}^\top Y\|_\infty / n$ is the smallest tuning parameter λ for which $\hat{\beta}_\lambda$ equals zero. Simulate L samples $e^{(1)}, \ldots, e^{(L)}$ of the standard normal random vector e.

Step 2: For each grid point $1 \le m \le M$, compute the values of the criterion function $\{\hat{Q}(\lambda_m, e^{(\ell)}) : 1 \le \ell \le L\}$ and calculate the empirical $(1-\alpha)$-quantile $\hat{q}_{\alpha,\text{emp}}(\lambda_m)$ from them.

Step 3: Approximate $\hat{\lambda}_\alpha$ by $\hat{\lambda}_{\alpha,\text{emp}} := \hat{q}_{\alpha,\text{emp}}(\lambda_{\hat{m}})$, where $\hat{m} = \min\{m : \hat{q}_{\alpha,\text{emp}}(\lambda_{m'}) \le \lambda_{m'}$ for all $m' \ge m\}$ if $\hat{q}_{\alpha,\text{emp}}(\lambda_M) \le \lambda_M$ and $\hat{m} = M$ otherwise.

The values of M and L in this algorithm can be chosen large without excessive load on the computations: the dependence of the computational complexity on M can be reduced by computing the lasso with warm starts along the tuning parameter path; the influence of L can be reduced through basic parallelization. Hence, the algorithm is computationally feasible even when n and p are very large.

3.2 Heuristic Idea of the Estimator

Before analyzing the estimator $\hat{\lambda}_\alpha$ mathematically, we describe the heuristic idea behind it: for every $\lambda \in (0, \infty)$, the criterion function $\hat{Q}(\lambda, e)$ can be regarded as a multiplier bootstrap version of the effective noise $2\|\boldsymbol{X}^\top \varepsilon\|_\infty / n = 2 \max_{1 \le j \le p} |\sum_{i=1}^n X_{ij} \varepsilon_i| / n$, where e is the vector of bootstrap multipliers. Consequently, $\hat{q}_\alpha(\lambda)$ can be interpreted as a bootstrap estimator of the $(1-\alpha)$-quantile λ^*_α of the effective noise. Since the quality of the estimator $\hat{q}_\alpha(\lambda)$ hinges on the choice of λ, the question is how to select an estimator $\hat{q}_\alpha(\lambda)$ from the family $\{\hat{q}_\alpha(\lambda) : \lambda > 0\}$ that is a good approximation of λ^*_α. Our selection rule (7) is motivated by the following two heuristic claims which are justified below: with high probability, it holds that

$$\hat{q}_\alpha(\lambda) \approx \lambda \quad \text{for} \quad \lambda \in [\lambda^*_\alpha - \delta, \lambda^*_\alpha + \delta] \tag{8}$$

$$\hat{q}_\alpha(\lambda) < \lambda \quad \text{for} \quad \lambda > \lambda^*_\alpha + \delta \tag{9}$$

with some small $\delta > 0$. Equation (8) suggests that the function $\lambda \mapsto \hat{q}_\alpha(\lambda)$ has a fixed point near λ^*_α, whereas equation (9) tells us that there should not be any fixed point for

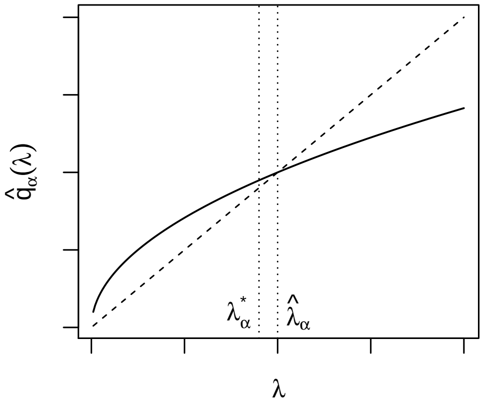


Figure 1: Graphical illustration of the estimator $\hat{\lambda}_\alpha$. The solid black line is the function $\lambda \mapsto \hat{q}_\alpha(\lambda)$, the dashed line is the 45-degree line, and the two vertical dotted lines indicate the values of λ^*_α and $\hat{\lambda}_\alpha$, respectively.

values $\lambda > \lambda^*_\alpha + \delta$. Taken together, (8) and (9) suggest approximating λ^*_α by solving the fixed point equation $\hat{q}_\alpha(\lambda) = \lambda$ and picking the largest such fixed point $\lambda = \hat{\lambda}_\alpha$. This is the heuristic idea which underlies the formal definition of our estimator $\hat{\lambda}_\alpha$ in (7). A graphical illustration is provided in Figure 1.

Discussion of the heuristic claim (8). To start with, we bound the criterion function $\hat{Q}(\lambda, e)$ from below and above by

$$Q(e) - r_\lambda(e) \le \hat{Q}(\lambda, e) \le Q(e) + r_\lambda(e),$$

where

$$\begin{aligned} Q(e) &= \max_{1 \le j \le p} \Big| \frac{2}{n} \sum_{i=1}^n X_{ij} \varepsilon_i e_i \Big| \\ r_\lambda(e) &= \max_{1 \le j \le p} \Big| \frac{2}{n} \sum_{i=1}^n X_{ij} X_i^\top (\beta^* - \hat{\beta}_\lambda) e_i \Big|. \end{aligned}$$

Here, $Q(e)$ is a multiplier bootstrap version of the effective noise which is based on the true noise terms ε rather than the residuals $\hat{\varepsilon}_\lambda$. The term $r_\lambda(e)$ is a remainder that captures the estimation error $\hat{\varepsilon}_\lambda - \varepsilon$ produced by the lasso $\hat{\beta}_\lambda$. Let λ_α be the $(1-\alpha)$-quantile of $Q(e)$ conditionally on $\boldsymbol{X}$ and Y. Theory for the multiplier bootstrap in high dimensions (Chernozhukov et al., 2013) suggests that the quantile λ_α gives a good approximation to λ^*_α. If the remainder $r_\lambda(e)$ tends to be small for a certain choice of λ, then the criterion function $\hat{Q}(\lambda, e)$ tends to be close to $Q(e)$, which in turn suggests that the quantile $\hat{q}_\alpha(\lambda)$ is close to λ_α. Since λ_α gives a good approximation to λ^*_α, we expect $\hat{q}_\alpha(\lambda)$ to be an accurate estimate of λ^*_α as well.

Standard prediction bounds for the lasso suggest that the tuning parameter choice $\lambda = \lambda^*_\alpha$ produces a precise model fit $\boldsymbol{X}\hat{\beta}_{\lambda^*_\alpha}$. The prediction bound (2), for example, implies that with probability at least $1-\alpha$, we have $\|\boldsymbol{X}(\beta^* - \hat{\beta}_{\lambda^*_\alpha})\|_2^2/n \le 2\lambda^*_\alpha \|\beta^*\|_1$, where $2\lambda^*_\alpha\|\beta^*\|_1 =$

$O(\|\beta^*\|_1\sqrt{\log(p)/n}) = o(1)$ under our technical conditions. Hence, we expect the remainder term $r_{\lambda^*_\alpha}(e)$ to be small. From the considerations in the previous paragraph, it follows that $\hat{q}_\alpha(\lambda^*_\alpha)$ should be a suitable estimate of λ^*_α, that is, $\hat{q}_\alpha(\lambda^*_\alpha) \approx \lambda^*_\alpha$. Since $\hat{q}_\alpha(\lambda) \approx \hat{q}_\alpha(\lambda^*_\alpha)$ for values of λ close to λ^*_α (which is due to the continuity of the solution path of the lasso), we further expect that

$$\hat{q}_\alpha(\lambda) \approx \lambda \quad \text{for} \quad \lambda \in [\lambda^*_\alpha - \delta, \lambda^*_\alpha + \delta]$$

with some small $\delta > 0$, which is the heuristic claim (8).

Discussion of the heuristic claim (9). As we gradually increase λ from λ^*_α to larger values, the lasso estimator $\hat{\beta}_\lambda$ tends to become more biased towards zero, implying that the residual vector $\hat{\varepsilon}_\lambda$ gets a less accurate proxy of the noise vector ε. As a consequence, we expect the remainder term $r_\lambda(e)$ and thus the criterion function $\hat{Q}(\lambda, e)$ to increase as λ gets larger. This in turn suggests that the quantile $\hat{q}_\alpha(\lambda)$ gets larger with increasing λ, thus overestimating λ^*_α more and more strongly. On the other hand, one can formally prove that the remainder $r_\lambda(e)$ grows quite slowly with λ. In particular, one can show that with high probability, $r_\lambda(e) \le C\{(\log n)^2/n^{1/4}\}\sqrt{\lambda}$ for all $\lambda \ge \lambda^*_\alpha$. A formalized version of this statement is given in Lemma A.2 in the Appendix. Since $\hat{Q}(\lambda, e) \le Q(e) + r_\lambda(e)$, this implies that the criterion function $\hat{Q}(\lambda, e)$ and thus its $(1-\alpha)$-quantile $\hat{q}_\alpha(\lambda)$ grow fairly slowly with increasing λ. In particular, we expect $\hat{q}_\alpha(\lambda)$ to grow more slowly than λ, that is,

$$\hat{q}_\alpha(\lambda) < \lambda \quad \text{for} \quad \lambda > \lambda^*_\alpha + \delta$$

with some small $\delta > 0$. This is the heuristic claim (9).

3.3 Theoretical Analysis of the Estimator

We now analyze the theoretical properties of the estimator $\hat{\lambda}_\alpha$. To do so, we use the following notation. By C_1, K_1, C_2 and K_2, we denote positive real constants that depend only on the set of model parameters $\Theta = \{c_X, C_X, c_\sigma, C_\sigma, C_\theta, \theta, C_r, r, C_\beta, \delta_\beta\}$ defined in (C1)–(C5). The constants C_1, K_1, C_2 and K_2 are thus in particular independent of the sample size n and the dimension p. Moreover, we let

$$\mathcal{T}_\lambda = \Big\{\frac{2}{n}\|\boldsymbol{X}^\top\varepsilon\|_\infty \le \lambda\Big\}$$

be the event that the effective noise $2\|\boldsymbol{X}^\top\varepsilon\|_\infty/n$ is smaller than λ. The following theorem, which is the main result of the paper, formally relates the estimator $\hat{\lambda}_\alpha$ to the quantiles of the effective noise. In the sequel, we will use this theorem to derive results on optimal tuning parameter choice and inference for the lasso.

Theorem 1 *Let (C1)–(C5) be satisfied. There exist an event $\mathcal{A}_n$ with $\mathbb{P}(\mathcal{A}_n) \ge 1 - C_1 n^{-K_1}$ for some positive constants C_1 and K_1 and a sequence of real numbers ν_n with $0 < \nu_n \le C_2 n^{-K_2}$ for some positive constants C_2 and K_2 such that the following holds: on the event $\mathcal{T}_{\lambda^*_{\alpha+\nu_n}} \cap \mathcal{A}_n$,*

$$\lambda^*_{\alpha+\nu_n} \le \hat{\lambda}_\alpha \le \lambda^*_{\alpha-\nu_n}$$

for every $\alpha \in (a_n, 1-a_n)$ with $a_n = 2\nu_n + (n \vee p)^{-1}$.

The proof of Theorem 1 is given in the Appendix. Precise definitions of $\mathcal{A}_n$ and ν_n are provided in equations (A.2) and (A.6), respectively. It is important to note that the bounds $\lambda^*_{\alpha+\nu_n}$ and $\lambda^*_{\alpha-\nu_n}$ in Theorem 1 are design-specific, that is, they depend on the distribution of the design vectors X_i (as well as on the distribution of the noise variables ε_i). Among other things, the bounds tend to get smaller as the design gets more correlated, that is, as the correlation between the covariates X_{ij} increases.

Since $\mathbb{P}(\mathcal{T}_{\lambda^*_{\alpha+\nu_n}} \cap \mathcal{A}_n) \ge 1 - \alpha - Cn^{-K}$ with some constants C and K that only depend on the model parameters Θ, Theorem 1 immediately implies that

$$\mathbb{P}\big(\lambda^*_{\alpha+\nu_n} \le \hat{\lambda}_\alpha \le \lambda^*_{\alpha-\nu_n}\big) \ge 1 - \alpha - Cn^{-K}.$$

Hence, with probability at least $1 - \alpha - Cn^{-K} = 1 - \alpha - o(1)$, our estimator $\hat{\lambda}_\alpha$ gives a good approximation to λ^*_α in the sense that $\lambda^*_{\alpha+\nu_n} \le \hat{\lambda}_\alpha \le \lambda^*_{\alpha-\nu_n}$. Another immediate consequence of Theorem 1 is that $|\hat{\lambda}_\alpha - \lambda^*_\alpha| \le \lambda^*_{\alpha-\nu_n} - \lambda^*_{\alpha+\nu_n}$ on the event $\mathcal{T}_{\lambda^*_{\alpha+\nu_n}} \cap \mathcal{A}_n$. From this, we obtain the deviation inequality

$$\mathbb{P}\big(|\hat{\lambda}_\alpha - \lambda^*_\alpha| \le \rho_{n,\boldsymbol{X},\varepsilon,\alpha}\big) \ge 1 - \alpha - Cn^{-K},$$

where $\rho_{n,\boldsymbol{X},\varepsilon,\alpha} = \lambda^*_{\alpha-\nu_n} - \lambda^*_{\alpha+\nu_n}$. With the help of Proposition A.9 from the Appendix, it is further possible to replace the bound $\rho_{n,\boldsymbol{X},\varepsilon,\alpha}$ by some kind of Gaussian version: Let $G = (G_1, \ldots, G_p)^\top$ be a Gaussian random vector with $\mathbb{E}[G_j] = 0$ for all j and the covariances

$$\mathbb{E}[G_j G_k] = \mathbb{E}\left[\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{ij}\varepsilon_i\Big)\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{ik}\varepsilon_i\Big)\right]$$

for $1 \le j \le k \le p$. Moreover, let γ^G_α be the $(1-\alpha)$-quantile of $\max_{1\le j\le p}|G_j|$. Then

$$\mathbb{P}\big(|\hat{\lambda}_\alpha - \lambda^*_\alpha| \le \rho^G_{n,\boldsymbol{X},\varepsilon,\alpha}\big) \ge 1 - \alpha - Cn^{-K}$$

with $\rho^G_{n,\boldsymbol{X},\varepsilon,\alpha} = 2(\gamma^G_{\alpha-2\nu_n} - \gamma^G_{\alpha+2\nu_n})/\sqrt{n}$.

Remark 2 An interesting question is the following: how large is the distance $\gamma^G_{\alpha-\nu_n} - \gamma^G_{\alpha+\nu_n}$ and thus the bound $\rho^G_{n,\boldsymbol{X},\varepsilon,\alpha}$? By definition, γ^G_α is the $(1-\alpha)$-quantile of the maximum $\max_{1\le j\le p}|G_j|$ of p Gaussian random variables with a general, potentially very complicated covariance structure. It is highly non-trivial to characterize the distribution of maxima of Gaussian random variables with a general correlation structure. Hence, finding precise bounds on the quantiles γ^G_α (and thus their distance) is a hard problem in general. In some special cases, however, it is possible to obtain suitable bounds. Consider in particular the situation that the variables G_j are i.i.d., which occurs for example when (i) the design variables X_{ij} are normalized such that $\mathbb{E}[X_{ij}] = 0$ and $\mathbb{E}[X^2_{ij}] = 1$ for all i and j, (ii) the design is uncorrelated (i.e., $\mathbb{E}[X_{ij}X_{ik}] = 0$ for all $j \ne k$), and (iii) the noise ε_i is homoskedastic (i.e., $\sigma^2(X_i) \equiv$ const.). In this case, one can show that

$$\gamma^G_{\alpha-2\nu_n} - \gamma^G_{\alpha+2\nu_n} \le \frac{C}{\sqrt{\log p}} \tag{10}$$

via classic extreme value theory, where C is a sufficiently large constant independent of n and p. A brief sketch of the proof is included in the Supplementary Material for completeness.

3.4 Relationship of the Estimator to Existing Methods

Roughly speaking, existing methods for estimating the quantiles of the effective noise fall into two categories:

(A) If the distribution of the error vector ε is known, it is trivial to construct an approximation of λ^*_α. To fix ideas, let $\varepsilon \sim N(0, \sigma^2 \boldsymbol{I})$ with known variance parameter σ^2 and consider a fixed design $\boldsymbol{X}$ for simplicity. (Otherwise, assume that ε is independent of $\boldsymbol{X}$ and condition on the latter.) In this case, the distribution of the effective noise $2\|\boldsymbol{X}^\top \varepsilon\|_\infty / n$ and thus its $(1-\alpha)$-quantile λ^*_α is known and can be computed by Monte Carlo simulations in practice. If the distribution of ε is not known exactly but is known to belong to a small family of distributions $\mathcal{F}$, it is further possible to compute (finite sample and asymptotic) upper bounds on λ^*_α under certain conditions as detailed in Belloni et al. (2011).

(B) In the more interesting situation where the distribution of ε is unknown, a simple way to estimate λ^*_α is as follows: Let $\lambda^{[0]}$ be a preliminary choice of the lasso's tuning parameter. Plug $\lambda^{[0]}$ into $\hat{q}_\alpha(\cdot)$ and use the resulting value $\hat{q}_\alpha(\lambda^{[0]})$ as an estimator of λ^*_α.—This plug-in approach is not very satisfactory: The quality of the estimator $\hat{q}_\alpha(\lambda^{[0]})$ obviously hinges on the precise choice of $\lambda^{[0]}$. In particular, as discussed in our heuristic considerations of Section 3.2, we can expect the following: If $\lambda^{[0]}$ is close to λ^*_α, then the estimator $\hat{q}_\alpha(\lambda^{[0]})$ will tend to be close to λ^*_α as well. In contrast, if $\lambda^{[0]}$ happens to be far away from λ^*_α, then $\hat{q}_\alpha(\lambda^{[0]})$ may also be far off. Hence, simply plugging a preliminary choice $\lambda^{[0]}$ into $\hat{q}_\alpha(\cdot)$ does not solve the problem of estimating λ^*_α but merely shifts it to the choice of $\lambda^{[0]}$: if we want to make sure that $\hat{q}_\alpha(\lambda^{[0]})$ is a good estimator of λ^*_α, we need to make sure that the same holds for the preliminary estimator $\lambda^{[0]}$.

Estimators of category (A) are for example considered in Belloni et al. (2011) and Belloni and Chernozhukov (2013), estimators of category (B) can be found in Chernozhukov et al. (2013) (in the context of the Dantzig selector rather than the lasso). It is important to emphasize that the problem of estimating the quantile λ^*_α is not the focus but only a very minor aspect of the aforementioned papers. This is presumably the reason why only the two simple approaches (A) and (B) have been considered there. Indeed, we are not aware of any article whose main focus is the estimation of the quantiles of the effective noise.

One way to improve on the plug-in method from (B) is to iterate it: Given some starting value $\lambda^{[0]}$, one computes the update $\lambda^{[r]} = \hat{q}_\alpha(\lambda^{[r-1]})$ for $r = 1, 2, \ldots$ until some convergence criterion is satisfied. The idea behind this iterative procedure is to find a fixed point $\lambda = \hat{q}_\alpha(\lambda)$ of the function $\hat{q}_\alpha(\cdot)$. Hence, it relies on the same heuristic as our method. The main contribution of our paper is (i) to devise an estimation approach which formalizes the fixed point heuristic and (ii) to derive finite sample theory for it. An important practical advantage of our fixed point method over the plug-in method from (B) is that it is free of tuning parameters: unlike the plug-in method, it does not require a preliminary estimator $\lambda^{[0]}$. Its only free parameter is the value $\alpha \in (0, 1)$.

4. Statistical Applications

4.1 Tuning Parameter Choice

A major challenge when implementing the lasso estimator $\hat{\beta}_\lambda$ is to choose the regularization parameter λ. As already discussed in the Introduction, the lasso satisfies the prediction bound (2), which can be rephrased as follows:

$$\text{On the event } \mathcal{T}_\lambda, \|\boldsymbol{X}(\beta^* - \hat{\beta}_{\lambda'})\|_2^2/n \le 2\lambda'\|\beta^*\|_1 \text{ for every } \lambda' \ge \lambda. \tag{11}$$

To control the prediction error, we would like to choose the smallest tuning parameter λ such that the bound $\|\boldsymbol{X}(\beta^* - \hat{\beta}_\lambda)\|_2^2/n \le 2\lambda\|\beta^*\|_1$ holds with high probability. Formally speaking, we may consider

$$\lambda_\alpha^{\text{oracle}} = \inf\{\lambda > 0 : \mathbb{P}(\mathcal{T}_\lambda) \ge 1 - \alpha\}$$

with some $\alpha \in (0,1)$ as the optimal tuning parameter. We call $\lambda_\alpha^{\text{oracle}}$ the oracle tuning parameter. It immediately follows from (11) that for every $\lambda \ge \lambda_\alpha^{\text{oracle}}$,

$$\mathbb{P}\biggl(\frac{1}{n}\|\boldsymbol{X}(\beta^* - \hat{\beta}_\lambda)\|_2^2 \le 2\lambda\|\beta^*\|_1\biggr) \ge 1 - \alpha,$$

whereas this probability bound is not guaranteed for any other $\lambda < \lambda_\alpha^{\text{oracle}}$. Consequently, $\lambda_\alpha^{\text{oracle}}$ is the smallest tuning parameter for which the prediction bound (11) yields the finite-sample guarantee

$$\frac{1}{n}\|\boldsymbol{X}(\beta^* - \hat{\beta}_{\lambda_\alpha^{\text{oracle}}})\|_2^2 \le 2\lambda_\alpha^{\text{oracle}}\|\beta^*\|_1 \tag{12}$$

with probability at least $1 - \alpha$. Importantly, the oracle tuning parameter $\lambda_\alpha^{\text{oracle}}$ is nothing else than the $(1 - \alpha)$-quantile λ_α^* of the effective noise, that is, $\lambda_\alpha^{\text{oracle}} = \lambda_\alpha^*$ for every $\alpha \in (0,1)$. Our estimator $\hat{\lambda}_\alpha$ can thus be interpreted as an approximation of the oracle parameter $\lambda_\alpha^{\text{oracle}}$. With the help of Theorem 1, we can show that implementing $\hat{\beta}_\lambda$ with the estimator $\lambda = \hat{\lambda}_\alpha$ produces almost the same finite-sample guarantee as (12).

Proposition 3 *Let the conditions of Theorem 1 be satisfied. With probability* $\ge 1 - \alpha - \nu_n - C_1 n^{-K_1} = 1 - \alpha + o(1)$, *it holds that*

$$\frac{1}{n}\|\boldsymbol{X}(\beta^* - \hat{\beta}_{\hat{\lambda}_\alpha})\|_2^2 \le 2\lambda_{\alpha-\nu_n}^{\text{oracle}}\|\beta^*\|_1.$$

For completeness, a short proof is provided in the Appendix. The upper bound $2\lambda_{\alpha-\nu_n}^{\text{oracle}}\|\beta^*\|_1$ in Proposition 3 is almost as sharp as the bound $2\lambda_\alpha^{\text{oracle}}\|\beta^*\|_1$ in (12); the only difference is that the $(1-\alpha)$-quantile $\lambda_\alpha^{\text{oracle}}$ is replaced by the somewhat larger $(1 - \{\alpha - \nu_n\})$-quantile $\lambda_{\alpha-\nu_n}^{\text{oracle}}$. There are improved versions of the prediction bound (2) (Lederer et al., 2019) as well as other types of prediction bounds (Dalalyan et al., 2017; Hebiri and Lederer, 2012; van de Geer and Lederer, 2013) that can be treated in the same way.

Our method does not only allow us to obtain finite-sample bounds on the prediction loss. It can also be used to equip the lasso with finite-sample guarantees for other losses. We consider the ℓ_∞-loss $L_\infty(\beta^*, \beta) = \|\beta^* - \beta\|_\infty$ as an example. Analogous considerations apply to any other loss for which an oracle inequality of the form (3) is available, such as the ℓ_1- and ℓ_2-losses. Let $S = \{j : \beta^*_j \neq 0\}$ be the active set of β^*. Moreover, for any vector $v = (v_1, \ldots, v_p)^\top \in \mathbb{R}^p$, let $v_S = (v_j \mathbf{1}(j \in S))_{j=1}^p$ and define $v_{S^\complement}$ analogously with $S^\complement = \{1 \ldots, p\} \setminus S$. The design matrix $\boldsymbol{X}$ is said to fulfill the ℓ_∞-restricted eigenvalue condition (Chichignoud et al., 2016) with the constants $\phi > 0$ and $\delta > 0$ if

$$\frac{\|\boldsymbol{X}^\top \boldsymbol{X} v\|_\infty}{n} \geq \phi \|v\|_\infty \quad \text{for all } v \in \mathbb{C}_\delta(S), \tag{13}$$

where $\mathbb{C}_\delta(S)$ is the double cone

$$\mathbb{C}_\delta(S) = \left\{ v \in \mathbb{R}^p : \|v_{S^\complement}\|_1 \leq \frac{2+\delta}{\delta} \|v_S\|_1 \right\}.$$

Under condition (13), we obtain the following oracle inequality, whose proof is provided in the Supplementary Material.

Lemma 4 *Suppose that* $\boldsymbol{X}$ *satisfies the restricted eigenvalue condition* (13)*. On the event* $\mathcal{T}_\lambda$*, it holds that*

$$\|\hat{\beta}_{\lambda'} - \beta^*\|_\infty \leq \kappa \lambda' \tag{14}$$

for every $\lambda' \geq (1+\delta)\lambda$ *with* $\kappa = 2/\phi$.

Whereas this ℓ_∞-oracle inequality is valid under condition (13), different conditions are needed to obtain oracle inequalities for other losses—see van de Geer and Bühlmann (2009) for a discussion of different assumptions. In the ℓ_2-loss case, for instance, an ℓ_2-restricted eigenvalue condition is usually imposed, which is somewhat different (and less restrictive) than (13). Moreover, in the prediction loss case considered above, no conditions on the design (in particular, no restricted eigenvalue conditions) are needed at all. Hence, condition (13) is not an assumption imposed by our method, it is rather inflicted by the oracle inequality of the ℓ_∞-loss.

Let $\mathcal{B}_n$ be the event that $\boldsymbol{X}$ satisfies the restricted eigenvalue condition (13) and note that $\mathbb{P}(\mathcal{B}_n) \to 1$ for certain classes of random design matrices $\boldsymbol{X}$ (van de Geer and Muro, 2014). The oracle inequality of Lemma 4 can be rephrased as follows: on the event $\mathcal{T}_\lambda \cap \mathcal{B}_n$, it holds that $\|\hat{\beta}_{\lambda'} - \beta^*\|_\infty \leq \kappa\lambda'$ for any $\lambda' \geq (1+\delta)\lambda$. The oracle parameter $\lambda_\alpha^{\text{oracle}}$ yields the following finite-sample guarantee: on the event $\mathcal{T}_{\lambda_\alpha^{\text{oracle}}} \cap \mathcal{B}_n$, that is, with probability $\geq 1 - \alpha - P(\mathcal{B}_n^\complement)$, it holds that

$$\|\hat{\beta}_{(1+\delta)\lambda_\alpha^{\text{oracle}}} - \beta^*\|_\infty \leq (1+\delta)\kappa\lambda_\alpha^{\text{oracle}}. \tag{15}$$

Theorem 1 implies that we can approximately recover this finite-sample guarantee when replacing the oracle parameter $\lambda_\alpha^{\text{oracle}}$ with the estimator $\hat{\lambda}_\alpha$.

Proposition 5 *Let the conditions of Theorem 1 be satisfied. With probability* $\geq 1 - \alpha - \mathbb{P}(\mathcal{B}_n^\complement) - \nu_n - C_1 n^{-K_1} = 1 - \alpha - \mathbb{P}(\mathcal{B}_n^\complement) + o(1)$*, it holds that*

$$\|\hat{\beta}_{(1+\delta)\hat{\lambda}_\alpha} - \beta^*\|_\infty \leq (1+\delta)\kappa\lambda_{\alpha-\nu_n}^{\text{oracle}}.$$

A proof of Proposition 5 can be found in the Appendix. It is important to note that the ℓ_∞-bound of Proposition 5 entails finite-sample guarantees for variable selection. Specifically, it implies that with probability $\geq 1-\alpha-\mathbb{P}(\mathcal{B}_n^\complement)+o(1)$, the lasso estimator $\hat{\beta}_{(1+\delta)\hat{\lambda}_\alpha}$ recovers all non-zero components of β^* that are larger in absolute value than $(1+\delta)\kappa\lambda_{\alpha-\nu_n}^{\text{oracle}}$. From Lemma A.3 and Proposition A.9 in the Appendix, it follows that $\lambda_{\alpha-\nu_n}^{\text{oracle}} \leq C\sqrt{\log(n \vee p)/n}$ with some sufficiently large constant C. Hence, with probability $\geq 1-\alpha-\mathbb{P}(\mathcal{B}_n^\complement)+o(1)$, the lasso estimator $\hat{\beta}_{(1+\delta)\hat{\lambda}_\alpha}$ in particular recovers all non-zero entries of β^* that are of larger order than $O(\sqrt{\log(n \vee p)/n})$.

4.2 Inference for the Lasso

Inference for the lasso is a notoriously difficult problem: the distribution of the lasso has a complicated limit and is hardly useful for statistical inference (Knight and Fu, 2000; Leeb and Pötscher, 2005). For this reason, inferential methods for the lasso are quite rare. Some exceptions are tests for the significance of small, fixed groups of parameters (Belloni et al., 2013; Zhang and Zhang, 2014; van de Geer et al., 2014; Javanmard and Montanari, 2014; Gold et al., 2020) and tests for the significance of parameters entering the lasso path (Lockhart et al., 2014). In what follows, we show that our method enables us to construct tuning-parameter-free tests for certain hypotheses of interest.

We first consider testing the null hypothesis $H_0 : \beta^* = 0$ against the alternative $H_1 : \beta^* \neq 0$, which was briefly discussed in the Introduction. Our test statistic of H_0 is defined as

$$T = \frac{2\|\boldsymbol{X}^\top Y\|_\infty}{n},$$

which implies that $T = 2\|\boldsymbol{X}^\top \varepsilon\|_\infty/n$ under H_0, that is, T is the effective noise under H_0. This observation suggests to define a test of H_0 as follows: reject H_0 at the significance level α if $T > \hat{\lambda}_\alpha$, where $\hat{\lambda}_\alpha$ estimates the $(1-\alpha)$-quantile λ_α^* of T under H_0. This test has the following theoretical properties.

Proposition 6 *Let the conditions of Theorem 1 be satisfied. Under the null hypothesis $H_0 : \beta^* = 0$, it holds that*

$$\mathbb{P}(T \leq \hat{\lambda}_\alpha) \geq 1 - \alpha + o(1).$$

Moreover, under any alternative $\beta^* \neq 0$ that satisfies the condition $\mathbb{P}(\|\boldsymbol{X}^\top \boldsymbol{X}\beta^*\|_\infty/n \geq c\sqrt{\log(n \vee p)/n}) \to 1$ for every fixed $c > 0$, it holds that

$$\mathbb{P}(T > \hat{\lambda}_\alpha) = 1 - o(1).$$

The proof is deferred to the Appendix. Proposition 6 ensures that the proposed test is of level α asymptotically and has asymptotic power 1 against any alternative $\beta^* \neq 0$ that satisfies the condition $\mathbb{P}(\|\boldsymbol{X}^\top \boldsymbol{X}\beta^*\|_\infty/n \geq c\sqrt{\log(n \vee p)/n}) \to 1$ for every $c > 0$. Such a condition is inevitable: in the model $Y = \boldsymbol{X}\beta^* + \varepsilon$, it is not possible to distinguish between vectors $\beta^* \neq 0$ that satisfy $\boldsymbol{X}\beta^* = 0$ and the null vector. Hence, a test can only have power against alternatives $\beta^* \neq 0$ that satisfy $\boldsymbol{X}\beta^* \neq 0$, that is, against alternatives $\beta^* \neq 0$ that do not lie in the kernel $\text{Ker}(\boldsymbol{X}) = \text{Ker}(\boldsymbol{X}^\top \boldsymbol{X}/n)$ of the linear mapping $\boldsymbol{X}$. By imposing the condition $\mathbb{P}(\|\boldsymbol{X}^\top \boldsymbol{X}\beta^*\|_\infty/n \geq c\sqrt{\log(n \vee p)/n}) \to 1$, we restrict attention to alternatives $\beta^* \neq 0$ that have enough signal outside the kernel of $\boldsymbol{X}$.

We now generalize the discussed test procedure in a way that allows to handle more complex hypotheses. Specifically, we generalize it such that a low-dimensional linear model can be tested against a high-dimensional alternative. To do so, we partition the design matrix $\boldsymbol{X}$ into two parts according to $\boldsymbol{X} = (\boldsymbol{X}_A, \boldsymbol{X}_B)$, where $A \dot{\cup} B = \{1, \ldots, p\}$, $\boldsymbol{X}_A$ is the part of the design matrix that contains the observations on the regressors in the set A, and $\boldsymbol{X}_B$ contains the observations on the regressors in the set B. We also partition the parameter vector β^* accordingly into two parts $\beta^*_A \in \mathbb{R}^{|A|}$ and $\beta^*_B \in \mathbb{R}^{|B|}$ such that $\beta^* = ((\beta^*_A)^\top, (\beta^*_B)^\top)^\top$. The linear model (6) can then be written as

$$Y = \boldsymbol{X}_A \beta^*_A + \boldsymbol{X}_B \beta^*_B + \varepsilon. \tag{16}$$

In practice, regression is often based on simple, low-dimensional models of the form $Y = \boldsymbol{X}_A \beta^*_A + w$, where w is the error term, and the number of regressors $|A|$ is small. Quite frequently, however, the question arises whether important explanatory variables are missing from these simple models. This question can formally be checked by a statistical test of the low-dimensional model $Y = \boldsymbol{X}_A \beta^*_A + w$ against a high-dimensional alternative of the form (16) that contains a large number $|B|$ of controls. More precisely speaking, a test of the null hypothesis $H_{0,B} : \beta^*_B = 0$ against the alternative $H_{1,B} : \beta^*_B \neq 0$ is required. Note that setting $A = \emptyset$ and $B = \{1, \ldots, p\}$ nests the previously discussed problem of testing H_0 against H_1 as a special case.

We construct a test of $H_{0,B}$ as follows: let $\mathcal{P} = \boldsymbol{I} - \boldsymbol{X}_A(\boldsymbol{X}_A^\top \boldsymbol{X}_A)^{-1} \boldsymbol{X}_A^\top$ be the projection matrix onto the orthogonal complement of the column space of $\boldsymbol{X}_A$. Applying $\mathcal{P}$ to both sides of the model equation (16) gives

$$\mathcal{P}Y = \mathcal{P}\boldsymbol{X}_B \beta^*_B + u \tag{17}$$

with $u = \mathcal{P}\varepsilon$, which is itself a high-dimensional linear model with response $\mathcal{P}Y$ and design matrix $\mathcal{P}\boldsymbol{X}_B$. In order to test whether the parameter vector β^*_B in model (17) is equal to 0, we use the same strategy as for the simpler problem of testing H_0: our test statistic is given by

$$T_B = \frac{2\|(\mathcal{P}\boldsymbol{X}_B)^\top \mathcal{P}Y\|_\infty}{n},$$

which implies that $T_B = 2\|(\mathcal{P}\boldsymbol{X}_B)^\top u\|_\infty / n$ under $H_{0,B}$. The quantiles of the statistic $2\|(\mathcal{P}\boldsymbol{X}_B)^\top u\|_\infty / n$ can be approximated by our method developed in Section 3: define the criterion function

$$\hat{Q}_B(\lambda, e) = \max_{j \in B} \Big| \frac{2}{n} \sum_{i=1}^n (\mathcal{P}\boldsymbol{X}_B)_{ij} \, \hat{u}_{\lambda,i} \, e_i \Big|,$$

where $(\mathcal{P}\boldsymbol{X}_B)_{ij}$ is the (i,j)-th element of the matrix $\mathcal{P}\boldsymbol{X}_B$, $\hat{u}_\lambda = \mathcal{P}Y - \mathcal{P}\boldsymbol{X}_B \hat{\beta}_{B,\lambda}$ is the residual vector which results from fitting the lasso with tuning parameter λ to the model (17), and $e = (e_1, \ldots, e_n)^\top$ is a standard normal random vector independent of the data $(\boldsymbol{X}, Y)$. Moreover, let $\hat{q}_{\alpha,B}(\lambda)$ be the $(1-\alpha)$-quantile of $\hat{Q}_B(\lambda, e)$ conditionally on $(\boldsymbol{X}, Y)$. As described in Section 3, we estimate the $(1-\alpha)$-quantile $\lambda^*_{\alpha,B}$ of $2\|(\mathcal{P}\boldsymbol{X}_B)^\top u\|_\infty / n$ by

$$\hat{\lambda}_{\alpha,B} = \inf\big\{ \lambda > 0 : \hat{q}_{\alpha,B}(\lambda') \leq \lambda' \text{ for all } \lambda' \geq \lambda \big\}.$$

Our test of the hypothesis $H_{0,B}$ is now carried out as follows: reject $H_{0,B}$ at the significance level α if $T_B > \hat{\lambda}_{\alpha,B}$.

To derive the formal properties of the test, we define $\vartheta^{(j)} = \arg\min_{\vartheta \in \mathbb{R}^{|A|}} \mathbb{E}[(X_{ij} - X_{i,A}^\top \vartheta)^2]$ with $X_{i,A} = (X_{ij} : j \in A)$. Put differently, we define $X_{i,A}^\top \vartheta^{(j)}$ to be the L_2-projection of X_{ij} onto the linear subspace spanned by the elements of $X_{i,A}$. We assume that $\min_{j \in B} \mathbb{E}[(X_{ij} - X_{i,A}^\top \vartheta^{(j)})^2] \geq c_\vartheta > 0$ for some constant c_ϑ. Such an assumption is to be expected: it essentially says that the random variables X_{ij} with $j \in B$ cannot be represented by a linear combination of the random variables X_{ij} with $j \in A$. The assumption is also mild; in particular, it is much weaker than irrepresentable-type conditions that are usually imposed in the context of variable selection for the lasso (van de Geer and Bühlmann, 2009). We can now summarize the formal properties of the test.

Proposition 7 *Let the conditions of Theorem 1 be satisfied, suppose for simplicity that the random variables (X_i, ε_i) are identically distributed across i, and let $|A|$ be a fixed number that does not grow with the sample size n. In addition, assume that the $|A| \times |A|$ matrix $\Psi_A = (\mathbb{E}[X_{ij}X_{ik}] : j,k \in A)$ is positive definite and that $\min_{j \in B} \mathbb{E}[(X_{ij} - X_{i,A}^\top \vartheta^{(j)})^2] \geq c_\vartheta > 0$. Under the null hypothesis $H_{0,B} : \beta_B^* = 0$, it holds that*

$$\mathbb{P}(T_B \leq \hat{\lambda}_{\alpha,B}) \geq 1 - \alpha + o(1).$$

Moreover, under any alternative $\beta_B^* \neq 0$ with the property that $\mathbb{P}(\|(\mathcal{P}\boldsymbol{X}_B)^\top(\mathcal{P}\boldsymbol{X}_B)\beta_B^*\|_\infty / n \geq c\sqrt{\log(n \vee p)/n}) \to 1$ for every $c > 0$, it holds that

$$\mathbb{P}(T_B > \hat{\lambda}_{\alpha,B}) = 1 - o(1).$$

This result shows that the proposed procedure is an asymptotic level-α-test that has asymptotic power 1 against any alternative $\beta_B^* \neq 0$ with the property that $\mathbb{P}(\|(\mathcal{P}\boldsymbol{X}_B)^\top(\mathcal{P}\boldsymbol{X}_B)\beta_B^*\|_\infty / n \geq c\sqrt{\log(n \vee p)/n}) \to 1$ for any $c > 0$. The latter condition parallels the one in Proposition 6. The proof of Proposition 7 is provided in the Appendix.

5. Simulations

In this section, we corroborate our results through Monte Carlo experiments. We simulate data from the linear regression model (6) with sample size $n = 500$ and dimension $p \in \{250, 500, 1000\}$. The covariate vectors $X_i = (X_{i1}, \ldots, X_{ip})^\top$ are independently sampled from a p-dimensional normal distribution with mean 0 and covariance matrix $(1-\kappa)\boldsymbol{I} + \kappa\boldsymbol{E}$, where $\boldsymbol{I}$ is the $p \times p$ identity matrix, $\boldsymbol{E} = (1, \ldots, 1)^\top(1, \ldots, 1) \in \mathbb{R}^{p \times p}$, and $\kappa \in [0, 1)$ is the correlation between the entries of the covariate vector X_i. We show the simulation results for $\kappa = 0.25$ unless indicated differently, but we obtained similar results for other values of κ as well. The noise variables ε_i are drawn i.i.d. from a normal distribution with mean 0 and variance $\sigma^2 = 1$. The target vector β^* has the form $\beta^* = (c, \ldots, c, 0, \ldots, 0)^\top$, where the first 5 entries are set to c and the remaining ones to 0. The value of c is chosen such that one obtains a prespecified value for the signal-to-noise ratio $\text{SNR} = \sqrt{\|\boldsymbol{X}\beta^*\|_2^2/n}/\sigma = \sqrt{\|\boldsymbol{X}\beta^*\|_2^2/n}$. We set SNR $= 1$ except when we analyze the hypothesis tests from Section 4.2: there, we consider the value SNR $= 0$, which corresponds to the null hypothesis, and the values SNR $\in \{0.1, 0.2\}$, which correspond to two different alternatives. We implement

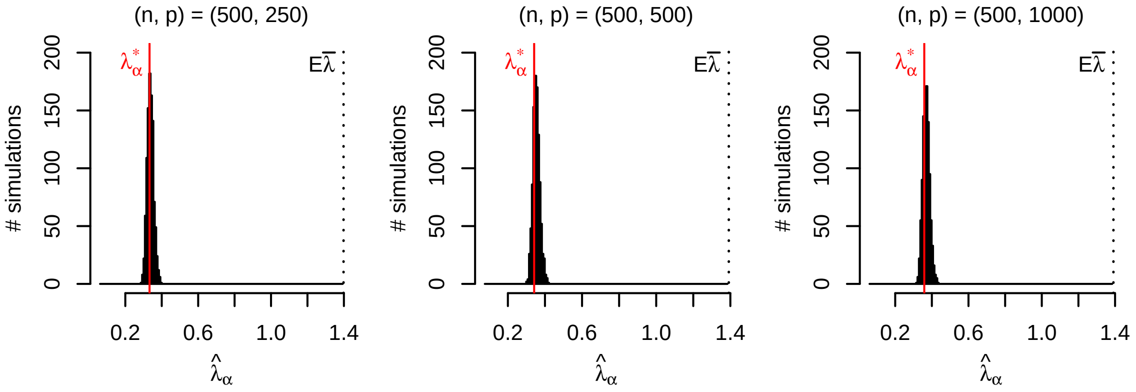


Figure 2: Histograms of the estimates $\hat{\lambda}_\alpha$ for different values of n and p. The red vertical lines indicate the values of the oracle parameter λ_α^*; the dotted vertical lines give the values of $\mathbb{E}[\overline{\lambda}]$, where $\overline{\lambda} = 2\|\boldsymbol{X}^\top Y\|_\infty/n$ is the smallest λ for which $\hat{\beta}_\lambda = 0$.

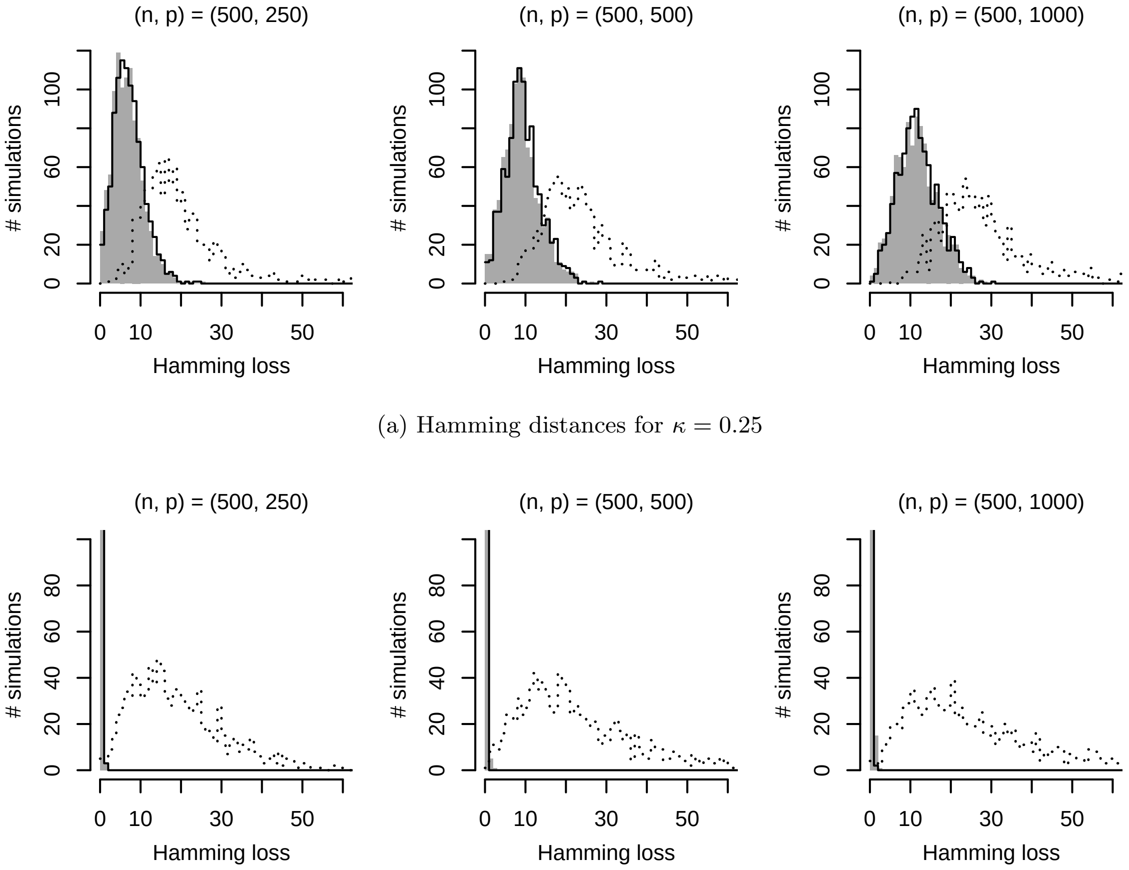
our estimation method with $L = 100$ bootstrap replicates, which seems sufficient across a wide variety of settings. The lasso paths are computed through `glmnet` (Friedman et al., 2010) version 2.2.1 with an equidistant grid of λ-values and $M = 100$, that is, $\lambda \in \{1 \cdot 2\|\boldsymbol{X}^\top Y\|_\infty/(100n), 2 \cdot 2\|\boldsymbol{X}^\top Y\|_\infty/(100n), \ldots\}$. All Monte Carlo experiments are based on $N = 1000$ simulation runs. The implementations are in `R` version 3.5.1.

5.1 Approximation Quality

We first examine how well the estimator $\hat{\lambda}_\alpha$ approximates the quantile λ_α^*. Figure 2 contains histograms of the $N = 1000$ estimates of $\hat{\lambda}_\alpha$ for $\alpha = 0.05$ and different values of n and p. The red vertical line in each plot indicates the value of the quantile λ_α^*, which is computed by simulating 1000 values of the effective noise $2\|\boldsymbol{X}^\top\varepsilon\|_\infty/n$ and then taking their empirical $(1-\alpha)$-quantile. The x-axis covers the interval $[0, \mathbb{E}\overline{\lambda}]$ in each plot, where $\overline{\lambda} = 2\|\boldsymbol{X}^\top Y\|_\infty/n$ is the smallest tuning parameter for which the lasso estimator is constantly equal to zero. This range is motivated as follows: varying the tuning parameter λ in the interval $[0, \overline{\lambda}]$ produces all possible lasso solutions. It is thus natural to measure the approximation quality of $\hat{\lambda}_\alpha$ by the deviation $|\hat{\lambda}_\alpha - \lambda_\alpha^*|$ relative to the length of the interval $[0, \overline{\lambda}]$ rather than by the absolute deviation $|\hat{\lambda}_\alpha - \lambda_\alpha^*|$. This, in turn, suggests that the right scale to plot histograms of the estimates $\hat{\lambda}_\alpha$ is the interval $[0, \overline{\lambda}]$. Since this interval is stochastic, we let the x-axis of our plots span the interval $[0, \mathbb{E}\overline{\lambda}]$ instead. The histogram plots of Figure 2 can be regarded as an empirical illustration of the deviation inequalities derived after Theorem 1. They demonstrate that the estimates $\hat{\lambda}_\alpha$ approximate the oracle quantile λ_α^* accurately.

5.2 Tuning Parameter Calibration

We next investigate the performance of our method for calibrating the tuning parameter of the lasso. Our estimator of β^* is defined as $\hat{\beta} := \hat{\beta}_{\hat{\lambda}_\alpha}$, where we use the estimator $\hat{\lambda}_\alpha$ with $\alpha = 0.05$ as the tuning parameter. Our main interest is a comparison between $\hat{\beta}$ and



(a) Hamming distances for $\kappa = 0.25$

(b) Hamming distances for $\kappa = 0$

Figure 3: Histograms of the Hamming distances produced by the estimators $\hat{\beta}$, $\hat{\beta}_{\text{oracle}}$, and $\hat{\beta}_{\text{CV}}$. The solid black lines indicate the histograms of $\Delta_H(\hat{\beta}, \beta^*)$, the gray-shaded areas indicate the histograms of $\Delta_H(\hat{\beta}_{\text{oracle}}, \beta^*)$, and the dotted lines indicate the histograms of $\Delta_H(\hat{\beta}_{\text{CV}}, \beta^*)$. The histograms of our estimator $\hat{\beta}$ and the oracle $\hat{\beta}_{\text{oracle}}$ in Subfigure (b) essentially consist of only one bin at the value 0 that goes up to almost 1000 (which is the total number of simulation runs); to make the histograms of the cross-validated estimator visible, we cut the y-axis of the plots in Subfigure (b) at the value 100.

the oracle estimator $\hat{\beta}_{\text{oracle}} := \hat{\beta}_{\lambda_\alpha^*}$, which is tuned with the oracle parameter λ_α^* rather than its estimate $\hat{\lambda}_\alpha$. This comparison allows us to investigate whether $\hat{\beta}$ is as accurate as suggested by our theory. To highlight the practical performance of our estimator further, we also compare $\hat{\beta}$ to the lasso estimator $\hat{\beta}_{\text{CV}} := \hat{\beta}_{\hat{\lambda}_{\text{CV}}}$, where $\hat{\lambda}_{\text{CV}}$ is the tuning parameter chosen by 10-fold cross-validation (which is performed on the same grid of λ-values as our method). Of course, there are many other tuning parameter calibration schemes besides cross-validation, but a comprehensive comparison of all calibration schemes is beyond the scope of this paper, and, therefore, we focus on the arguably most popular representative.

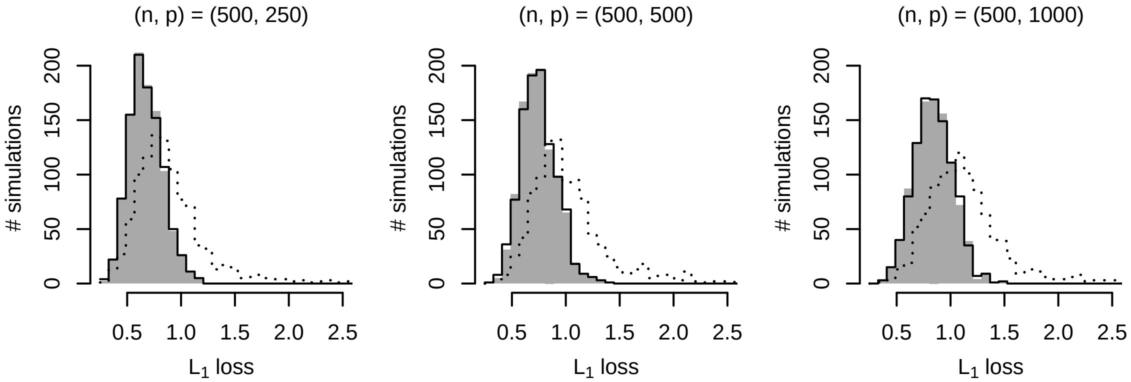


Figure 4: Histograms of the ℓ_1-loss produced by the estimators $\hat{\beta}$, $\hat{\beta}_{\text{oracle}}$, and $\hat{\beta}_{\text{CV}}$. The format of the plots is the same as in Figure 3.

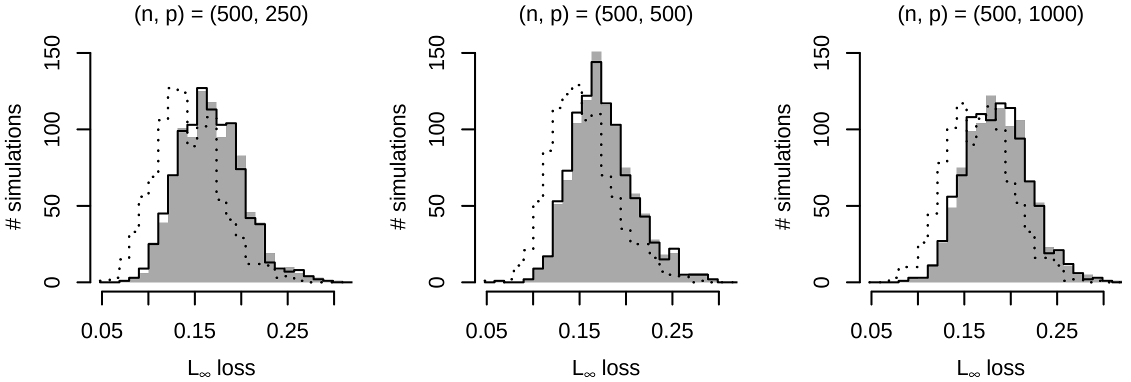


Figure 5: Histograms of the ℓ_∞-loss produced by the estimators $\hat{\beta}$, $\hat{\beta}_{\text{oracle}}$, and $\hat{\beta}_{\text{CV}}$. The format of the plots is the same as in Figure 3.

We use four error measures to compare vectors $\beta \in \mathbb{R}^p$ to β^*: the Hamming distance $\Delta_H(\beta, \beta^*) = \sum_{j=1}^{p} |\mathbf{1}(\beta_j = 0) - \mathbf{1}(\beta_j^* = 0)|$, the ℓ_1-distance $\Delta_1(\beta, \beta^*) = \|\beta - \beta^*\|_1$, the ℓ_∞-distance $\Delta_\infty(\beta, \beta^*) = \|\beta - \beta^*\|_\infty$, and the prediction error $\Delta_{\text{pr}}(\beta, \beta^*) = \|\boldsymbol{X}(\beta - \beta^*)\|_2^2/n$. The Hamming distance allows us to investigate the variable selection properties of the estimators $\hat{\beta}$, $\hat{\beta}_{\text{oracle}}$ and $\hat{\beta}_{\text{CV}}$: the quantity $\Delta_H(\beta, \beta^*)$ counts the number of false-negative and false-positive entries in the vector β, where the entry j is defined to be a false negative if $\beta_j^* \neq 0$ but $\beta_j = 0$ and a false positive if $\beta_j^* = 0$ but $\beta_j \neq 0$. The ℓ_p-loss with $\ell \in \{1, \infty\}$ and the mean-squared prediction error Δ_{pr}, on the other hand, allow us to investigate the estimators' estimation and prediction properties, respectively.

The simulation results for the Hamming distance are reported in Figure 3a for our usual value $\kappa = 0.25$ of the correlation and in Figure 3b for $\kappa = 0$. The black line in each plot depicts the histogram of the Hamming distances $\Delta_H(\hat{\beta}, \beta^*)$ that are produced by our estimator $\hat{\beta}$ over the $N = 1000$ simulation runs, the gray-shaded area depicts the histogram

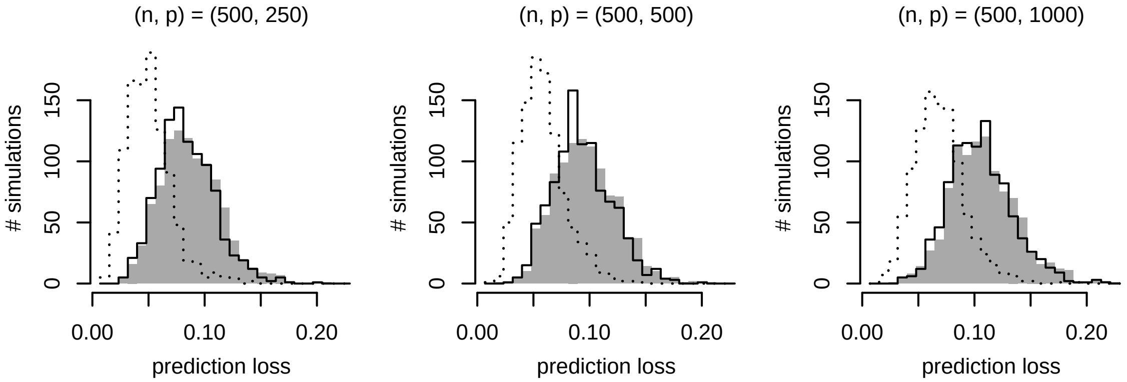


Figure 6: Histograms of the prediction loss produced by the estimators $\hat{\beta}$, $\hat{\beta}_{\text{oracle}}$, and $\hat{\beta}_{\text{CV}}$. The format of the plots is the same as in Figure 3.

of $\Delta_H(\hat{\beta}_{\text{oracle}}, \beta^*)$ produced by the oracle $\hat{\beta}_{\text{oracle}}$, and the dotted line depicts the histogram of $\Delta_H(\hat{\beta}_{\text{CV}}, \beta^*)$ produced by the cross-validated estimator $\hat{\beta}_{\text{CV}}$.

Comparing Figures 3a and 3b, we find that both the oracle and our estimator provide more accurate variable selection for smaller correlations—in line with theories for the lasso (Zhao and Yu, 2006). We also find that both the oracle and our estimator provide more accurate variable selection than cross-validation—in line with the well-known fact that cross-validation typically overselects. Finally, we find that the histograms of our estimator are virtually the same as the ones of the oracle estimator—in line with our theory.

The simulation results for the ℓ_1-norm are reported in Figure 4 and for the ℓ_∞-norm in Figure 5. We find again that for both the ℓ_1- and the ℓ_∞-loss, the histograms produced by our estimator $\hat{\beta}$ are extremely close to those of the oracle $\hat{\beta}_{\text{oracle}}$, meaning that the performance of our procedure matches the performance of the oracle. We also find that our estimator improves on cross-validation in terms of the ℓ_1-norm but slightly loses in terms of the ℓ_∞-norm. The reason for this difference is that $\hat{\lambda}_{\text{CV}}$ tends to be much smaller than $\hat{\lambda}_\alpha$ and λ^*_α; this induces an accumulation of small, spurious parameters, which affects the ℓ_1-norm more than the ℓ_∞-norm.

The simulation results for the prediction error are reported in Figure 6. Once more, the histograms of our estimator are extremely close to those of the oracle. Cross-validation performs best, which is no surprise in view of it being specifically designed for this task.

The two main conclusions from the simulations are that our method (i) exhibits virtually the same performance as the oracle and (ii) rivals cross-validation in terms of variable selection and estimation but not necessarily prediction.

5.3 Inference

We finally explore the empirical performance of the tests developed in Section 4.2. We focus on the simpler test $H_0 : \beta^* = 0$ against $H_1 : \beta^* \neq 0$, where we reject H_0 at the significance level α if $T = 2\|\boldsymbol{X}^\top Y\|_\infty/n > \hat{\lambda}_\alpha$. We compare this test with an oracle version that rejects H_0 if $T > \lambda^*_\alpha$. Similarly as before, this comparison allows us to investigate if our practical test matches its theoretical (and in practice infeasible) analog as suggested by our theory.

(a) empirical size under $H_0 : \beta^* = 0$

	feasible test			oracle test		
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
$(n, p) = (500, 250)$	0.024	0.057	0.110	0.021	0.056	0.087
$(n, p) = (500, 500)$	0.018	0.050	0.097	0.008	0.064	0.116
$(n, p) = (500, 1000)$	0.015	0.044	0.082	0.010	0.050	0.095

(b) empirical power under the alternative with SNR = 0.1

	feasible test			oracle test		
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
$(n, p) = (500, 250)$	0.151	0.304	0.440	0.118	0.341	0.458
$(n, p) = (500, 500)$	0.148	0.293	0.433	0.089	0.341	0.456
$(n, p) = (500, 1000)$	0.122	0.284	0.409	0.090	0.293	0.417

(c) empirical power under the alternative with SNR = 0.2

	feasible test			oracle test		
	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
$(n, p) = (500, 250)$	0.644	0.850	0.923	0.664	0.890	0.940
$(n, p) = (500, 500)$	0.631	0.840	0.909	0.579	0.880	0.926
$(n, p) = (500, 1000)$	0.599	0.811	0.904	0.600	0.867	0.918

Table 1: Empirical size under the null and power against different alternatives.

The simulation setup is as described before, including the mentioned variations over the signal-to-noise ratio SNR: the value SNR $= 0$ specifies the null hypothesis $H_0 : \beta^* = 0$; the values SNR $\in \{0.1, 0.2\}$ specify the alternative (the larger SNR, the further the setup deviates from the null).

Table 1a reports the empirical size of our feasible test and of its oracle version under the null for different values of the nominal size α, sample size n, and dimension p. The empirical size is defined as the number of rejections divided by the total number of simulation runs. We find that the size both of our feasible test and of the oracle test is close to the target α in all considered scenarios.

Tables 1b and 1c report the empirical power of the tests for the signal-to-noise ratios SNR $= 0.1$ and SNR $= 0.2$, respectively. The empirical power is again defined as the number of rejections divided by the total number of simulation runs. We find that the power increases when the signal-to-noise ratio SNR goes up, as expected. We further find that the power of our test is very similar to the one of the oracle test. Moreover, the power can be seen to be quite substantial despite the small signal-to-noise ratios.

We conclude that our test has (i) similar performance as its oracle version, (ii) sizes close to the nominal ones, and (iii) considerable power against alternatives.

5.4 Robustness Checks

In Section S.2 of the Supplementary Material, we carry out some robustness checks. We in particular examine how our simulation results are affected by different distributions of the noise ε_i and the design X_i, how our method for tuning parameter calibration is influenced by the choice of α, and what is the effect of the number of bootstrap iterations L and the grid size M on the simulation results.

6. Future Research Directions

In this paper, we have focused on the lasso in high-dimensional linear regression. However, the theoretical analysis of many other high-dimensional estimators involves terms similar to the effective noise. Examples are nuclear-norm penalized estimators for trace regression (Koltchinskii et al., 2011) and lasso-type estimators for settings with structured sparsity (Micchelli et al., 2013; Maurer et al., 2012). We believe that our methods can be extended to such estimators as well, but we leave the detailed analysis for future research.

Acknowledgments

We thank the editor and two anonymous reviewers for their insightful comments.

Appendix A. Proofs

In what follows, we prove the main theoretical results of the paper. We assume throughout that the technical conditions (C1)–(C5) are fulfilled.

A.1 Notation

Throughout the Appendix, the symbols B, c, C, D and K denote generic constants that may take a different value on each occurrence. Moreover, the symbols B_j, c_j, C_j, D_j and K_j with subscript j (which may be either a natural number or a letter) are specific constants that are defined in the course of the Appendix. Unless stated differently, the constants B, c, C, D, K, B_j, c_j, C_j, D_j and K_j depend neither on the sample size n nor on the dimension p. For ease of notation, we let $\Theta = \{c_X, C_X, c_\sigma, C_\sigma, C_\theta, \theta, C_r, r, C_\beta, \delta_\beta\}$ be the list of model parameters specified in (C1)–(C5). For a, $b \in \mathbb{R}$, we write $a \vee b = \max\{a, b\}$. The random variables $\boldsymbol{X}$, ε and e are assumed to be defined on the same probability space $(\Omega, \mathcal{A}, \mathbb{P})$ for all $n \geq 1$. We write $\mathbb{P}_e(\cdot) = \mathbb{P}(\,\cdot\,|\boldsymbol{X}, \varepsilon)$ and $\mathbb{E}_e[\,\cdot\,] = \mathbb{E}[\,\cdot\,|\boldsymbol{X}, \varepsilon]$ to denote the probability and expectation conditionally on $\boldsymbol{X}$ and ε.

To derive the theoretical results of the paper, it is convenient to reformulate the estimator $\hat{\lambda}_\alpha$ as follows: define $\hat{\Pi}(\gamma, e) = \max_{1 \leq j \leq p} |\hat{W}_j(\gamma, e)|$, where

$$\hat{W}(\gamma, e) = \big(\hat{W}_1(\gamma, e), \ldots, \hat{W}_p(\gamma, e)\big)^\top \quad \text{with} \quad \hat{W}_j(\gamma, e) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_{ij} \hat{\varepsilon}_{\frac{2}{\sqrt{n}}\gamma, i} e_i.$$

Moreover, let $\hat{\pi}_\alpha(\gamma)$ be the $(1-\alpha)$-quantile of $\hat{\Pi}(\gamma, e)$ conditionally on $\boldsymbol{X}$ and Y, that is, conditionally on $\boldsymbol{X}$ and ε, which is formally defined as $\hat{\pi}_\alpha(\gamma) = \inf\{q : \mathbb{P}_e(\hat{\Pi}(\gamma, e) \leq q) \geq$

$1-\alpha\}$, and set

$$\hat{\gamma}_\alpha = \inf\bigl\{\gamma > 0 : \hat{\pi}_\alpha(\gamma') \le \gamma' \text{ for all } \gamma' \ge \gamma\bigr\}.$$

The quantities $\hat{\Pi}(\gamma, e)$, $\hat{\pi}_\alpha(\gamma)$ and $\hat{\gamma}_\alpha$ are related to $\hat{Q}(\lambda, e)$, $\hat{q}_\alpha(\lambda)$ and $\hat{\lambda}_\alpha$ by the equations

$$\hat{\Pi}\Bigl(\frac{\sqrt{n}}{2}\lambda, e\Bigr) = \frac{\sqrt{n}}{2}\hat{Q}(\lambda, e), \quad \hat{\pi}_\alpha\Bigl(\frac{\sqrt{n}}{2}\lambda\Bigr) = \frac{\sqrt{n}}{2}\hat{q}_\alpha(\lambda), \quad \hat{\gamma}_\alpha = \frac{\sqrt{n}}{2}\hat{\lambda}_\alpha.$$

Hence, $\hat{\gamma}_\alpha$ is a rescaled version of the estimator $\hat{\lambda}_\alpha$. In particular, we can reformulate $\hat{\lambda}_\alpha$ in terms of $\hat{\gamma}_\alpha$ as $\hat{\lambda}_\alpha = 2\hat{\gamma}_\alpha/\sqrt{n}$.

For our proof strategy, we require some auxiliary statistics which are closely related to $\hat{\Pi}(\gamma, e)$, $\hat{\pi}_\alpha(\gamma)$ and $\hat{\gamma}_\alpha$. To start with, we define $\Pi(e) = \max_{1\le j\le p} |W_j(e)|$, where

$$W(e) = \bigl(W_1(e), \dots, W_p(e)\bigr)^\top \quad \text{with} \quad W_j(e) = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_{ij}\varepsilon_i e_i,$$

and let γ_α be the $(1-\alpha)$-quantile of $\Pi(e)$ conditionally on $\boldsymbol{X}$ and ε. Moreover, we set $\Pi^* = \max_{1\le j\le p}|W_j^*|$, where

$$W^* = (W_1^*, \dots, W_p^*)^\top \quad \text{with} \quad W_j^* = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_{ij}\varepsilon_i,$$

and let γ_α^* be the $(1-\alpha)$-quantile of Π^*. Notice that γ_α^* is a rescaled version of λ_α^*, in particular, $\gamma_\alpha^* = \sqrt{n}\lambda_\alpha^*/2$. Finally, we define $\Pi^G = \max_{1\le j\le p}|G_j|$, where $G = (G_1, \dots, G_p)^\top$ is a Gaussian random vector with the same covariance structure as W^*, that is, $\mathbb{E}[G] = \mathbb{E}[W^*] = 0$ and $\mathbb{E}[GG^\top] = \mathbb{E}[W^*(W^*)^\top]$, and we denote the $(1-\alpha)$-quantile of Π^G by γ_α^G.

In order to relate the criterion function $\hat{\Pi}(\gamma, e)$ to the term $\Pi(e)$, we make use of the simple bound

$$\hat{\Pi}(\gamma, e) \begin{cases} \le \Pi(e) + R(\gamma, e) \\ \ge \Pi(e) - R(\gamma, e), \end{cases} \tag{A.1}$$

where

$$R(\gamma, e) = \max_{1\le j\le p}\Bigl|\frac{1}{\sqrt{n}}\sum_{i=1}^n X_{ij}X_i^\top\bigl(\beta^* - \hat{\beta}_{\frac{2}{\sqrt{n}}\gamma}\bigr)e_i\Bigr|.$$

For our technical arguments, we further define the expression

$$\Delta = \max_{1\le j,k\le p}\bigl|\Sigma_{jk} - \Sigma_{jk}^*\bigr| = \max_{1\le j,k\le p}\Bigl|\frac{1}{n}\sum_{i=1}^n\bigl(X_{ij}X_{ik}\varepsilon_i^2 - \mathbb{E}[X_{ij}X_{ik}\varepsilon_i^2]\bigr)\Bigr|,$$

where $\Sigma = (\Sigma_{jk} : 1 \le j,k \le p) = \mathbb{E}_e[W(e)W(e)^\top]$ is the covariance matrix of $W(e)$ conditionally on $\boldsymbol{X}$ and ε, and $\Sigma^* = (\Sigma_{jk}^* : 1 \le j,k \le p) = \mathbb{E}[W^*(W^*)^\top]$ is the covariance matrix of W^*. We finally introduce the event

$$\mathcal{S}_\gamma = \Bigl\{\frac{1}{\sqrt{n}}\|\boldsymbol{X}^\top\varepsilon\|_\infty \le \gamma\Bigr\},$$

which relates to $\mathcal{T}_\lambda$ by the equation $\mathcal{S}_{\sqrt{n}\lambda/2} = \mathcal{T}_\lambda$, as well as the event

$$\mathcal{A}_n = \bigl\{\Delta \le B_\Delta\sqrt{\log(n \vee p)/n}\bigr\}, \tag{A.2}$$

where the constant B_Δ is defined in Lemma A.1 below.

A.2 Auxiliary Results

Before we prove the main results of the paper, we derive some auxiliary lemmas which are needed later on. Their proofs can be found in the Supplementary Material.

Lemma A.1 *There exist positive constants B_Δ, C_Δ and K_Δ that depend only on the model parameters Θ such that*

$$\mathbb{P}\big(\Delta > B_\Delta \sqrt{\log(n \vee p)/n}\big) \le C_\Delta n^{-K_\Delta}.$$

In particular, K_Δ can be chosen to be any positive constant with $K_\Delta < (\theta - 4)/4$, where $\theta > 4$ is defined in (C3).

Lemma A.2 *On the event $\mathcal{S}_\gamma$, it holds that*

$$\mathbb{P}_e\bigg(R(\gamma', e) > \frac{B_R (\log n)^2 \sqrt{\|\beta^*\|_1 \gamma'}}{n^{1/4}}\bigg) \le C_R\, n^{-K_R}$$

for any $\gamma' \ge \gamma$, where the constants B_R, C_R and K_R depend only on the model parameters Θ, and K_R can be chosen as large as desired by picking C_R large enough.

Lemma A.3 *For every $\alpha > 1/(n \vee p)$, it holds that*

$$\gamma_\alpha^G \le C_X C_\sigma \big[\sqrt{2\log(2p)} + \sqrt{2\log(n \vee p)}\big],$$

where the constants C_X and C_σ are defined in (C2) and (C3), respectively.

In addition to Lemmas A.1–A.3, we state some results on high-dimensional Gaussian approximations and anti-concentration bounds for Gaussian random vectors from Chernozhukov et al. (2013) and Chernozhukov et al. (2015) that are required for the proofs in the sequel. The first result is an anti-concentration bound which is taken from Chernozhukov et al. (2015)—see their Theorem 3 and Corollary 1.

Lemma A.4 *Let $(V_1, \ldots, V_p)^\top$ be a centered Gaussian random vector in $\mathbb{R}^p$. Suppose that there are constants $0 < c_3 < C_3 < \infty$ with $c_3 \le \sigma_j \le C_3$, where $\sigma_j^2 = \mathbb{E}[V_j^2]$ for $1 \le j \le p$. Then for every $\delta > 0$,*

$$\sup_{t \in \mathbb{R}} \mathbb{P}\bigg(\Big|\max_{1 \le j \le p} V_j - t\Big| \le \delta\bigg) \le C\delta\sqrt{1 \vee \log(p/\delta)},$$

where $C > 0$ depends only on c_3 and C_3.

The next two results correspond to Theorem 2 in Chernozhukov et al. (2015) and Corollary 2.1 in Chernozhukov et al. (2013), respectively.

Lemma A.5 *Let $V = (V_1, \ldots, V_p)^\top$ and $V' = (V_1', \ldots, V_p')^\top$ be centered Gaussian random vectors in $\mathbb{R}^p$ with covariance matrices $\Sigma^V = (\Sigma_{jk}^V : 1 \le j, k \le p)$ and $\Sigma^{V'} = (\Sigma_{jk}^{V'} : 1 \le j, k \le p)$, respectively, and define $\delta = \max_{1 \le j,k \le p} |\Sigma_{jk}^V - \Sigma_{jk}^{V'}|$. Suppose that there are* *constants* $0 < c_4 < C_4 < \infty$ *with* $c_4 \leq \Sigma^V_{jj} \leq C_4$ *for* $1 \leq j \leq p$. *Then there exists a constant* $C > 0$ *that depends only on* c_4 *and* C_4 *such that*

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\Big(\max_{1 \leq j \leq p} V_j \leq t \Big) - \mathbb{P}\Big(\max_{1 \leq j \leq p} V'_j \leq t \Big) \right| \\ \leq C \delta^{1/3} \big\{ 1 \vee 2 \log p \vee \log(1/\delta) \big\}^{1/3} (\log p)^{1/3}.$$

Lemma A.6 *Let* $Z_i = (Z_{i1}, \ldots, Z_{ip})^\top$ *be independent* $\mathbb{R}^p$*-valued random vectors for* $1 \leq i \leq n$ *with mean zero and the following properties:* $c_5 \leq n^{-1} \sum_{i=1}^n \mathbb{E}[Z_{ij}^2] \leq C_5$ *and* $\max_{k=1,2} \{ n^{-1} \sum_{i=1}^n \mathbb{E}[|Z_{ij}|^{2+k} / D_n^k] \} + \mathbb{E}[(\max_{1 \leq j \leq p} |Z_{ij}|/D_n)^4] \leq 4$, *where* $c_5 > 0$, $C_5 > 0$ *and* $D_n \geq 1$ *is such that* $D_n^4 (\log(pn))^7/n \leq C_6 n^{-c_6}$ *for some constants* $c_6 > 0$ *and* $C_6 > 0$. *Define*

$$W = (W_1, \ldots, W_p)^\top \quad \textit{with} \quad W_j = \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_{ij}$$

and let $V = (V_1, \ldots, V_p)^\top$ *be a Gaussian random vector with the same mean and covariance as* W, *that is,* $\mathbb{E}[V] = \mathbb{E}[W] = 0$ *and* $\mathbb{E}[VV^\top] = \mathbb{E}[WW^\top]$. *Then there exist constants* $C > 0$ *and* $K > 0$ *that depend only on* c_5, C_5, c_6 *and* C_6 *such that*

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\Big(\max_{1 \leq j \leq p} W_j \leq t \Big) - \mathbb{P}\Big(\max_{1 \leq j \leq p} V_j \leq t \Big) \right| \leq C n^{-K}.$$

The final lemma of this section concerns the quantiles of Gaussian maxima.

Lemma A.7 *Let* $(V_1, \ldots, V_p)^\top$ *be a centered Gaussian random vector in* $\mathbb{R}^p$ *which fulfills the conditions of Lemma A.4. Moreover, let* γ^V_α *be the* $(1-\alpha)$*-quantile of* $\max_{1 \leq j \leq p} V_j$, *which is formally defined as* $\gamma^V_\alpha = \inf\{ q : \mathbb{P}(\max_{1 \leq j \leq p} V_j \leq q) \geq 1 - \alpha \}$. *It holds that*

$$\mathbb{P}\Big(\max_{1 \leq j \leq p} V_j \leq \gamma^V_\alpha \Big) = 1 - \alpha$$

for every $\alpha \in (0,1)$.

Remark A.8 Note that Lemmas A.4–A.7 continue to hold for maxima of the form $\max_{1 \leq j \leq p} |V_j|$, $\max_{1 \leq j \leq p} |V'_j|$ and $\max_{1 \leq j \leq p} |W_j|$. This follows from the fact that $\max_{1 \leq j \leq p} |V_j| = \max_{1 \leq j \leq 2p} U_j$ with $U_j = V_j$ and $U_{p+j} = -V_j$ for $1 \leq j \leq p$.

A.3 Proof of Theorem 1

The proof proceeds in several steps. To start with, we formally relate the quantiles γ^*_α, γ_α and γ^G_α to each other.

Proposition A.9 *There exist positive constants* C *and* K *that depend only on the model parameters* Θ *such that*

$$\gamma^*_{\alpha + \kappa_n} \leq \gamma^G_\alpha \leq \gamma^*_{\alpha - \kappa_n}$$
$$\gamma^G_{\alpha + \kappa_n} \leq \gamma^*_\alpha \leq \gamma^G_{\alpha - \kappa_n}$$

for any $\alpha \in (\kappa_n, 1 - \kappa_n)$ *with* $\kappa_n = C n^{-K}$.

Proof From (C2) and (C3), it immediately follows that $0 < c_5 \le n^{-1} \sum_{i=1}^n \mathbb{E}[(X_{ij}\varepsilon_i)^2] \le C_5 < \infty$ and $\max_{k=1,2}\{n^{-1} \sum_{i=1}^n \mathbb{E}[|X_{ij}\varepsilon_i|^{2+k}/D^k]\} + \mathbb{E}[(\max_{1\le j \le p} |X_{ij}\varepsilon_i|/D)^4] \le 4$ for some appropriately chosen constants c_5, C_5 and D that depend only on the parameters Θ. Since D does not depend on n, it further holds that $D^4(\log(pn))^7/n \le C_6 n^{-c_6}$, where c_6 can be chosen to be any positive constant strictly smaller than 1 provided that C_6 is picked sufficiently large. Hence, we can apply Lemma A.6 to the terms $\Pi^* = \max_{1\le j\le p} |W_j^*|$ and $\Pi^G = \max_{1 \le j \le p} |G_j|$ to obtain the following: there exist constants $C > 0$ and $K > 0$ depending only on c_5, C_5, c_6 and C_6 such that

$$\sup_{t\in\mathbb{R}} \big|\mathbb{P}\big(\Pi^* \le t\big) - \mathbb{P}\big(\Pi^G \le t\big)\big| < Cn^{-K}. \tag{A.3}$$

By Lemma A.7, it holds that $\mathbb{P}(\Pi^G \le \gamma_\alpha^G) = 1 - \alpha$. Using this identity together with (A.3) yields that

$$1 - (\alpha + Cn^{-K}) < \mathbb{P}\big(\Pi^* \le \gamma_\alpha^G\big) < 1 - (\alpha - Cn^{-K}),$$

which in turn implies that $\gamma^*_{\alpha + Cn^{-K}} \le \gamma^G_\alpha \le \gamma^*_{\alpha - Cn^{-K}}$ for any $\alpha \in (Cn^{-K}, 1 - Cn^{-K})$. This is the first statement of Proposition A.9. The second statement is an immediate consequence thereof. ■

Proposition A.10 *There exist positive constants C and K that depend only on the model parameters Θ such that on the event $\mathcal{A}_n$,*

$$\begin{aligned} \gamma^*_{\alpha+\xi_n} &\le \gamma_\alpha \le \gamma^*_{\alpha-\xi_n} \\ \gamma_{\alpha+\xi_n} &\le \gamma^*_\alpha \le \gamma_{\alpha-\xi_n} \end{aligned}$$

for any $\alpha \in (\xi_n, 1-\xi_n)$ with $\xi_n = Cn^{-K}$.

Proof Conditionally on $\boldsymbol{X}$ and ε, $W(e)$ is a Gaussian random vector with the covariance matrix $\Sigma = (\Sigma_{jk} : 1 \le j,k \le p)$, where $\Sigma_{jk} = n^{-1}\sum_{i=1}^n X_{ij}X_{ik}\varepsilon_i^2$. Moreover, by definition, G is a Gaussian random vector with the covariance matrix $\Sigma^* = (\Sigma^*_{jk} : 1 \le j,k \le p)$, where $\Sigma^*_{jk} = n^{-1}\sum_{i=1}^n \mathbb{E}[X_{ij}X_{ik}\varepsilon_i^2]$. It is straightforward to verify that under (C2) and (C3), $c_4 \le \Sigma^*_{jj} \le C_4$ with some constants $0 < c_4 \le C_4 < \infty$ that depend only on the parameters Θ. Hence, by Lemma A.5, the distribution of $\Pi^G = \max_{1\le j\le p} |G_j|$ is close to the conditional distribution of $\Pi(e) = \max_{1 \le j \le p} |W_j(e)|$ in the following sense:

$$\sup_{t\in\mathbb{R}} \big|\mathbb{P}_e\big(\Pi(e) \le t\big) - \mathbb{P}\big(\Pi^G \le t\big)\big| \le \pi(\Delta), \tag{A.4}$$

where $\pi(\Delta) = C\Delta^{1/3}\{1 \vee 2\log(2p) \vee \log(1/\Delta)\}^{1/3}\{\log(2p)\}^{1/3}$ with C depending only on c_4 and C_4. Notice that the logarithm in the expression $\pi(\Delta)$ takes the argument $2p$ rather than p as in the formulation of Lemma A.5. This is due to the fact that $\Pi(e)$ and Π^G are maxima over absolute values as discussed in Remark A.8. From (A.4), it immediately follows that on the event $\mathcal{A}_n$,

$$\sup_{t\in\mathbb{R}} \big|\mathbb{P}_e\big(\Pi(e) \le t\big) - \mathbb{P}\big(\Pi^G \le t\big)\big| < \pi_n, \tag{A.5}$$

where we let π_n be such that $\pi(B_\Delta\sqrt{\log(n \vee p)/n}) < \pi_n \le Cn^{-K}$ with some positive constants C and K. With the help of (A.5) and analogous arguments as in the proof of Proposition A.9, we can infer that on the event $\mathcal{A}_n$,

$$\begin{aligned}\gamma_{\alpha+\pi_n} &\le \gamma_\alpha^G \le \gamma_{\alpha-\pi_n}\\ \gamma_{\alpha+\pi_n}^G &\le \gamma_\alpha \le \gamma_{\alpha-\pi_n}^G.\end{aligned}$$

Combining this with the statement of Proposition A.9, we finally get that on $\mathcal{A}_n$,

$$\begin{aligned}\gamma_{\alpha+\xi_n}^* &\le \gamma_\alpha \le \gamma_{\alpha-\xi_n}^*\\ \gamma_{\alpha+\xi_n} &\le \gamma_\alpha^* \le \gamma_{\alpha-\xi_n}\end{aligned}$$

with $\xi_n = \kappa_n + \pi_n$, which completes the proof. ■

We now turn to the main part of the proof of Theorem 1. To make the notation more compact, we introduce the shorthands $\rho_n = B_R(\log n)^2\{C_\beta n^{1/2-\delta_\beta}\}^{1/2}/n^{1/4}$ and $\psi_n = C_X C_\sigma[\sqrt{2\log(2p)} + \sqrt{2\log(n \vee p)}]$. Moreover, we let $\{\nu_n\}$ be any null sequence with

$$\nu_n > \xi_n + C_R n^{-K_R} + C_7 \max\left\{ g_n\sqrt{1 \vee \log\left(\frac{2p}{g_n}\right)}, h_n\sqrt{1 \vee \log\left(\frac{2p}{h_n}\right)}\right\}, \tag{A.6}$$

where $g_n = \rho_n(1+\psi_n)$, $h_n = \{\rho_n + \rho_n(1+\rho_n)\}\psi_n + \rho_n$, and C_7 is a positive constant that depends only on the parameters Θ and that is specified below.

Our aim is to prove that on the event $\mathcal{T}_{\lambda^*_{\alpha+\nu_n}} \cap \mathcal{A}_n$, $\lambda^*_{\alpha+\nu_n} \le \hat\lambda_\alpha \le \lambda^*_{\alpha-\nu_n}$. This is equivalent to the following statement: on the event $\mathcal{S}_{\gamma^*_{\alpha+\nu_n}} \cap \mathcal{A}_n$, it holds that $\gamma^*_{\alpha+\nu_n} \le \hat\gamma_\alpha \le \gamma^*_{\alpha-\nu_n}$. By definition of $\hat\gamma_\alpha$,

$$\hat\pi_\alpha(\gamma) \le \gamma \text{ for all } \gamma \ge \gamma^*_{\alpha-\nu_n} \quad\Longrightarrow\quad \hat\gamma_\alpha \le \gamma^*_{\alpha-\nu_n} \tag{A.7}$$

$$\hat\pi_\alpha(\gamma) > \gamma \text{ for some } \gamma > \gamma^*_{\alpha+\nu_n} \quad\Longrightarrow\quad \hat\gamma_\alpha > \gamma^*_{\alpha+\nu_n}, \tag{A.8}$$

and by definition of $\hat\pi_\alpha(\gamma)$,

$$\mathbb{P}_e(\hat\Pi(\gamma,e) \le \gamma) > 1-\alpha \quad\Longrightarrow\quad \hat\pi_\alpha(\gamma) \le \gamma \tag{A.9}$$

$$\mathbb{P}_e(\hat\Pi(\gamma,e) \le \gamma) < 1-\alpha \quad\Longrightarrow\quad \hat\pi_\alpha(\gamma) > \gamma. \tag{A.10}$$

Hence, it suffices to prove the following two statements:

(I) On the event $\mathcal{S}_{\gamma^*_{\alpha+\nu_n}} \cap \mathcal{A}_n$, $\mathbb{P}_e(\hat\Pi(\gamma,e) \le \gamma) > 1-\alpha$ for all $\gamma \ge \gamma^*_{\alpha-\nu_n}$.

(II) On the event $\mathcal{S}_{\gamma^*_{\alpha+\nu_n}} \cap \mathcal{A}_n$, $\mathbb{P}_e(\hat\Pi(\gamma,e) \le \gamma) < 1-\alpha$ for some $\gamma > \gamma^*_{\alpha+\nu_n}$.

Proof of (I) Suppose we are on the event $\mathcal{S}_{\gamma^*_{\alpha+\nu_n}} \cap \mathcal{A}_n$ and let $\gamma \ge \gamma^*_{\alpha-\nu_n}$. Using the simple bound (A.1), we obtain that

$$\begin{aligned}\mathbb{P}_e\big(\hat\Pi(\gamma,e) \le \gamma\big) &\ge \mathbb{P}_e\big(\Pi(e) + R(\gamma,e) \le \gamma\big)\\ &\ge \mathbb{P}_e\big(\Pi(e) + R(\gamma,e) \le \gamma, R(\gamma,e) \le \rho_n\sqrt{\gamma}\big)\\ &\ge \mathbb{P}_e\big(\Pi(e) + \rho_n\sqrt{\gamma} \le \gamma, R(\gamma,e) \le \rho_n\sqrt{\gamma}\big)\\ &\ge \mathbb{P}_e\big(\Pi(e) \le \gamma - \rho_n\sqrt{\gamma}\big) - \mathbb{P}_e\big(R(\gamma,e) > \rho_n\sqrt{\gamma}\big)\\ &\ge \mathbb{P}_e\big(\Pi(e) \le \gamma - \rho_n\sqrt{\gamma}\big) - C_R n^{-K_R},\end{aligned} \tag{A.11}$$

where the last inequality is by Lemma A.2. Since $\gamma-\rho_n\sqrt{\gamma}\geq\gamma-\rho_n(1+\gamma)=(1-\rho_n)\gamma-\rho_n$ and $\gamma\geq\gamma^*_{\alpha-\nu_n}$, we further get that

$$\begin{aligned}\mathbb{P}_e\big(\Pi(e)\leq\gamma-\rho_n\sqrt{\gamma}\big)&\geq\mathbb{P}_e\big(\Pi(e)\leq(1-\rho_n)\gamma-\rho_n\big)\\&\geq\mathbb{P}_e\big(\Pi(e)\leq(1-\rho_n)\gamma^*_{\alpha-\nu_n}-\rho_n\big)\\&=\mathbb{P}_e\big(\Pi(e)\leq\gamma^*_{\alpha-\nu_n}-\rho_n(1+\gamma^*_{\alpha-\nu_n})\big).\end{aligned}$$

Moreover, since $\gamma^*_{\alpha-\nu_n}\geq\gamma_{\alpha+\xi_n-\nu_n}$ on the event $\mathcal{A}_n$ by Proposition A.10 and $\gamma^*_{\alpha-\nu_n}\leq\gamma^G_{\alpha-\kappa_n-\nu_n}\leq\psi_n$ by Proposition A.9 and Lemma A.3, it follows that

$$\begin{aligned}\mathbb{P}_e\big(\Pi(e)\leq\gamma-\rho_n\sqrt{\gamma}\big)&\geq\mathbb{P}_e\big(\Pi(e)\leq\gamma_{\alpha+\xi_n-\nu_n}-\rho_n(1+\psi_n)\big)\\&=\mathbb{P}_e\big(\Pi(e)\leq\gamma_{\alpha+\xi_n-\nu_n}\big)\\&\quad-\mathbb{P}_e\big(\gamma_{\alpha+\xi_n-\nu_n}-\rho_n(1+\psi_n)<\Pi(e)\leq\gamma_{\alpha+\xi_n-\nu_n}\big).\end{aligned}\tag{A.12}$$

On the event $\mathcal{A}_n$, we have that $c_X^2c_\sigma^2-B_\Delta\sqrt{\log(n\vee p)/n}\leq\mathbb{E}_e[W_j^2(e)]\leq C_X^2C_\sigma^2+B_\Delta\sqrt{\log(n\vee p)/n}$. Hence, we can apply Lemma A.4 to get that

$$\begin{aligned}&\mathbb{P}_e\big(\gamma_{\alpha+\xi_n-\nu_n}-\rho_n(1+\psi_n)<\Pi(e)\leq\gamma_{\alpha+\xi_n-\nu_n}\big)\\&\quad\leq\sup_{t\in\mathbb{R}}\mathbb{P}_e\big(|\Pi(e)-t|\leq\rho_n(1+\psi_n)\big)\\&\quad\leq C_7\rho_n(1+\psi_n)\sqrt{1\vee\log\Big(\frac{2p}{\rho_n(1+\psi_n)}\Big)}\end{aligned}\tag{A.13}$$

with C_7 depending only on Θ. By Lemma A.7, it further holds that $\mathbb{P}_e(\Pi(e)\leq\gamma_{\alpha+\xi_n-\nu_n})=1-(\alpha+\xi_n-\nu_n)$. Plugging this identity and (A.13) into (A.12) yields that

$$\mathbb{P}_e\big(\Pi(e)\leq\gamma-\rho_n\sqrt{\gamma}\big)\geq1-\alpha+\nu_n-\xi_n-C_7g_n\sqrt{1\vee\log\Big(\frac{2p}{g_n}\Big)}$$

with $g_n=\rho_n(1+\psi_n)$. Inserting this into (A.11), we finally arrive at

$$\mathbb{P}_e\big(\hat{\Pi}(\gamma,e)\leq\gamma\big)\geq1-\alpha+\nu_n-\xi_n-C_Rn^{-K_R}-C_7g_n\sqrt{1\vee\log\Big(\frac{2p}{g_n}\Big)}>1-\alpha,$$

where the last inequality follows from the definition of ν_n in (A.6). ■

Proof of (II) Suppose we are on the event $\mathcal{S}_{\gamma^*_{\alpha+\nu_n}}\cap\mathcal{A}_n$ and let $\gamma=(1+\phi_n)\gamma^*_{\alpha+\nu_n}$, where $\{\phi_n\}$ is a null sequence of positive numbers with $\phi_n\leq Cn^{-K}$ for some constants C and K. For convenience, we set $\phi_n=\rho_n$, but we could also work with any other choice of ϕ_n that satisfies the conditions mentioned in the previous sentence. With the bound (A.1), we get that

$$\begin{aligned}\mathbb{P}_e\big(\hat{\Pi}(\gamma,e)>\gamma\big)&\geq\mathbb{P}_e\big(\Pi(e)-R(\gamma,e)>\gamma\big)\\&\geq\mathbb{P}_e\big(\Pi(e)-R(\gamma,e)>\gamma,R(\gamma,e)\leq\rho_n\sqrt{\gamma}\big)\\&\geq\mathbb{P}_e\big(\Pi(e)-\rho_n\sqrt{\gamma}>\gamma,R(\gamma,e)\leq\rho_n\sqrt{\gamma}\big)\\&\geq\mathbb{P}_e\big(\Pi(e)>\gamma+\rho_n\sqrt{\gamma}\big)-\mathbb{P}_e\big(R(\gamma,e)>\rho_n\sqrt{\gamma}\big)\\&\geq\mathbb{P}_e\big(\Pi(e)>\gamma+\rho_n\sqrt{\gamma}\big)-C_Rn^{-K_R},\end{aligned}\tag{A.14}$$

where the final inequality is a direct consequence of Lemma A.2. Analogous arguments as in the proof of (I) yield that

$$\begin{aligned}
&\mathbb{P}_e\big(\Pi(e) > \gamma + \rho_n\sqrt{\gamma}\big)\\
&\geq \mathbb{P}_e\big(\Pi(e) > \gamma + \rho_n(1+\gamma)\big)\\
&= \mathbb{P}_e\big(\Pi(e) > (1+\phi_n)\gamma^*_{\alpha+\nu_n} + \rho_n(1+(1+\phi_n)\gamma^*_{\alpha+\nu_n})\big)\\
&= \mathbb{P}_e\big(\Pi(e) > \gamma^*_{\alpha+\nu_n} + \{\phi_n + \rho_n(1+\phi_n)\}\gamma^*_{\alpha+\nu_n} + \rho_n\big)\\
&\geq \mathbb{P}_e\big(\Pi(e) > \gamma_{\alpha-\xi_n+\nu_n} + h_n\big)\\
&= \mathbb{P}_e\big(\Pi(e) > \gamma_{\alpha-\xi_n+\nu_n}\big) - \mathbb{P}_e\big(\gamma_{\alpha-\xi_n+\nu_n} < \Pi(e) \leq \gamma_{\alpha-\xi_n+\nu_n} + h_n\big)\\
&\geq \alpha + \nu_n - \xi_n - C_7 h_n\sqrt{1 \vee \log(2p/h_n)},
\end{aligned}$$

where $h_n = \{\phi_n + \rho_n(1+\phi_n)\}\psi_n + \rho_n = \{\rho_n + \rho_n(1+\rho_n)\}\psi_n + \rho_n$ under the assumption that $\phi_n = \rho_n$. Inserting this into (A.14), we arrive that

$$\mathbb{P}_e\big(\hat{\Pi}(\gamma, e) > \gamma\big) \geq \alpha + \nu_n - \xi_n - C_R n^{-K_R} - C_7 h_n\sqrt{1 \vee \log\Big(\frac{2p}{h_n}\Big)} > \alpha,$$

which is equivalent to the statement that $\mathbb{P}_e\big(\hat{\Pi}(\gamma, e) \leq \gamma\big) < 1 - \alpha$. ■

A.4 Proof of Proposition 3

From (2), it follows that $\|\boldsymbol{X}(\beta^* - \hat{\beta}_\lambda)\|_2^2/n \leq 2\lambda\|\beta^*\|_1$ for every $\lambda \geq \lambda^*_{\alpha+\nu_n}$ on the event $\mathcal{T}_{\lambda^*_{\alpha+\nu_n}}$. Moreover, by Theorem 1, $\lambda^*_{\alpha+\nu_n} \leq \hat{\lambda}_\alpha \leq \lambda^*_{\alpha-\nu_n}$ on the event $\mathcal{T}_{\lambda^*_{\alpha+\nu_n}} \cap \mathcal{A}_n$. Hence, we can infer that

$$\frac{1}{n}\|\boldsymbol{X}(\beta^* - \hat{\beta}_{\hat{\lambda}_\alpha})\|_2^2 \leq 2\hat{\lambda}_\alpha\|\beta^*\|_1 \leq 2\lambda^*_{\alpha-\nu_n}\|\beta^*\|_1$$

on the event $\mathcal{T}_{\lambda^*_{\alpha+\nu_n}} \cap \mathcal{A}_n$, which occurs with probability $\mathbb{P}(\mathcal{T}_{\lambda^*_{\alpha+\nu_n}} \cap \mathcal{A}_n) \geq 1 - \alpha - \nu_n - C_1 n^{-K_1}$.

A.5 Proof of Proposition 5

From Lemma 4, we know that on the event $\mathcal{T}_{\lambda^*_{\alpha+\nu_n}} \cap \mathcal{B}_n$, $\|\hat{\beta}_\lambda - \beta^*\|_\infty \leq \kappa\lambda$ for every $\lambda \geq (1+\delta)\lambda^*_{\alpha+\nu_n}$. Moreover, on the event $\mathcal{T}_{\lambda^*_{\alpha+\nu_n}} \cap \mathcal{A}_n$, it holds that $\lambda^*_{\alpha+\nu_n} \leq \hat{\lambda}_\alpha \leq \lambda^*_{\alpha-\nu_n}$ by Theorem 1. Hence, we can infer that

$$\|\hat{\beta}_{(1+\delta)\hat{\lambda}_\alpha} - \beta^*\|_\infty \leq (1+\delta)\kappa\hat{\lambda}_\alpha \leq (1+\delta)\kappa\lambda^*_{\alpha-\nu_n}$$

on the event $\mathcal{T}_{\lambda^*_{\alpha+\nu_n}} \cap \mathcal{A}_n \cap \mathcal{B}_n$, which occurs with probability $\mathbb{P}(\mathcal{T}_{\lambda^*_{\alpha+\nu_n}} \cap \mathcal{A}_n \cap \mathcal{B}_n) \geq 1 - \alpha - \mathbb{P}(\mathcal{B}_n^\complement) - \nu_n - C_1 n^{-K_1}$.

A.6 Proof of Proposition 6

For the proof, we reformulate the test of H_0 as follows: slightly abusing notation, we redefine the test statistic as $T = \|\boldsymbol{X}^\top Y\|_\infty/\sqrt{n}$. As above, we further let $\gamma^*_\alpha = \sqrt{n}\lambda^*_\alpha/2$ be the $(1-\alpha)$-quantile of $\|\boldsymbol{X}^\top\varepsilon\|_\infty/\sqrt{n}$ and define the estimator $\hat{\gamma}_\alpha = \sqrt{n}\hat{\lambda}_\alpha/2$. Our test of H_0 can now be expressed as follows: reject H_0 at the significance level α if $T > \hat{\gamma}_\alpha$.

We first prove that $\mathbb{P}(T \leq \hat{\gamma}_\alpha) \geq 1 - \alpha + o(1)$ under H_0. With the help of Theorem 1, we obtain that under H_0,

$$\begin{aligned}\mathbb{P}(T \leq \hat{\gamma}_\alpha) &\geq \mathbb{P}(T \leq \hat{\gamma}_\alpha, \mathcal{T}_{\lambda^*_{\alpha+\nu_n}} \cap \mathcal{A}_n) \geq \mathbb{P}(T \leq \gamma^*_{\alpha+\nu_n}, \mathcal{T}_{\lambda^*_{\alpha+\nu_n}} \cap \mathcal{A}_n) \\ &= \mathbb{P}(\mathcal{T}_{\lambda^*_{\alpha+\nu_n}} \cap \mathcal{A}_n) \geq 1 - \alpha - \nu_n - C_1 n^{-K_1},\end{aligned}$$

where the equality in the last line follows from the fact that the two events $\mathcal{T}_{\lambda^*_{\alpha+\nu_n}}$ and $\{T \leq \gamma^*_{\alpha+\nu_n}\}$ are identical. As a result, we get that $\mathbb{P}(T \leq \hat{\gamma}_\alpha) \geq 1 - \alpha + o(1)$ under H_0.

We next prove that $\mathbb{P}(T > \hat{\gamma}_\alpha) = 1 - o(1)$ under any alternative $H_1 : \beta^* \neq 0$ that fulfills the conditions of Proposition 6. Suppose we are on such an alternative and let $\{\alpha_n\}$ be a null sequence with $2\nu_n + (n \vee p)^{-1} < \alpha_n < \alpha$. It holds that

$$\begin{aligned}\mathbb{P}(T > \hat{\gamma}_\alpha) \geq \mathbb{P}(T > \hat{\gamma}_{\alpha_n}) &= \mathbb{P}\Big(\frac{1}{\sqrt{n}}\|\boldsymbol{X}^\top\boldsymbol{X}\beta^* + \boldsymbol{X}^\top\varepsilon\|_\infty > \hat{\gamma}_{\alpha_n}\Big) \\ &\geq \mathbb{P}\Big(\frac{\|\boldsymbol{X}^\top\boldsymbol{X}\beta^*\|_\infty}{\sqrt{n}} > \hat{\gamma}_{\alpha_n} + \frac{\|\boldsymbol{X}^\top\varepsilon\|_\infty}{\sqrt{n}}\Big), \qquad \text{(A.15)}\end{aligned}$$

where the last line is due to the triangle inequality. Applying Theorem 1, Proposition A.9 and Lemma A.3, the term $\hat{\gamma}_{\alpha_n}$ can be bounded by

$$\hat{\gamma}_{\alpha_n} \leq \gamma^*_{\alpha_n - \nu_n} \leq \gamma^G_{\alpha_n - \nu_n - \kappa_n} \leq \psi_n \qquad \text{(A.16)}$$

on the event $\mathcal{T}_{\lambda^*_{\alpha_n+\nu_n}} \cap \mathcal{A}_n$, where $\psi_n = C_X C_\sigma[\sqrt{2\log(2p)} + \sqrt{2\log(n \vee p)}]$. Moreover, for the term $\Pi^* = \|\boldsymbol{X}^\top\varepsilon\|_\infty/\sqrt{n}$, we have that

$$\begin{aligned}\mathbb{P}\Big(\frac{\|\boldsymbol{X}^\top\varepsilon\|_\infty}{\sqrt{n}} \leq \psi_n\Big) = \mathbb{P}(\Pi^* \leq \psi_n) &= \mathbb{P}(\Pi^G \leq \psi_n) + \big[\mathbb{P}(\Pi^* \leq \psi_n) - \mathbb{P}(\Pi^G \leq \psi_n)\big] \\ &\geq \mathbb{P}(\Pi^G \leq \psi_n) - Cn^{-K} \\ &\geq \mathbb{P}(\Pi^G \leq \gamma^G_{2/(n\vee p)}) - Cn^{-K} \\ &= 1 - \frac{2}{n \vee p} - Cn^{-K} = 1 - o(1) \qquad \text{(A.17)}\end{aligned}$$

with some positive constants C and K, where the first inequality follows from (A.3) and the second one from Lemma A.3. Using (A.16) and (A.17) in the right-hand side of equation (A.15), we can infer that

$$\begin{aligned}&\mathbb{P}\Big(\frac{\|\boldsymbol{X}^\top\boldsymbol{X}\beta^*\|_\infty}{\sqrt{n}} > \hat{\gamma}_{\alpha_n} + \frac{\|\boldsymbol{X}^\top\varepsilon\|_\infty}{\sqrt{n}}\Big) \\ &\geq \mathbb{P}\Big(\frac{\|\boldsymbol{X}^\top\boldsymbol{X}\beta^*\|_\infty}{\sqrt{n}} > \hat{\gamma}_{\alpha_n} + \frac{\|\boldsymbol{X}^\top\varepsilon\|_\infty}{\sqrt{n}},\ \frac{\|\boldsymbol{X}^\top\varepsilon\|_\infty}{\sqrt{n}} \leq \psi_n,\ \mathcal{T}_{\lambda^*_{\alpha_n+\nu_n}} \cap \mathcal{A}_n\Big) \\ &\geq \mathbb{P}\Big(\frac{\|\boldsymbol{X}^\top\boldsymbol{X}\beta^*\|_\infty}{\sqrt{n}} > 2\psi_n,\ \frac{\|\boldsymbol{X}^\top\varepsilon\|_\infty}{\sqrt{n}} \leq \psi_n,\ \mathcal{T}_{\lambda^*_{\alpha_n+\nu_n}} \cap \mathcal{A}_n\Big) \\ &= \mathbb{P}\Big(\frac{\|\boldsymbol{X}^\top\boldsymbol{X}\beta^*\|_\infty}{\sqrt{n}} > 2\psi_n\Big) - o(1) = 1 - o(1), \qquad \text{(A.18)}\end{aligned}$$

the last equality following from the assumption that $\mathbb{P}(\|\boldsymbol{X}^\top\boldsymbol{X}\beta^*\|_\infty/n \geq c\sqrt{\log(n \vee p)/n})$ → 1 for every $c > 0$. Combining (A.18) with (A.15) yields that $\mathbb{P}(T > \hat{\gamma}_\alpha) = 1 - o(1)$ under the alternative, which completes the proof.

A.7 Proof of Proposition 7

Similarly as in the proof of Proposition 6, we reformulate the test of $H_{0,B} : \beta^*_B = 0$. Slightly abusing notation, we redefine the test statistic as

$$T_B = \frac{\|(\mathcal{P}\boldsymbol{X}_B)^\top \mathcal{P}Y\|_\infty}{\sqrt{n}}.$$

Moreover, we let $\gamma^*_{\alpha,B} = \sqrt{n}\lambda^*_{\alpha,B}/2$ be the $(1-\alpha)$-quantile of $\|(\mathcal{P}\boldsymbol{X}_B)^\top u\|_\infty/\sqrt{n}$ and set $\hat{\gamma}_{\alpha,B} = \sqrt{n}\hat{\lambda}_{\alpha,B}/2$. Our test of $H_{0,B}$ can now be formulated as follows: reject $H_{0,B}$ at the significance level α if $T_B > \hat{\gamma}_{\alpha,B}$. This test has the same structure as the test of the simpler hypothesis $H_0 : \beta^* = 0$. The only difference is that it is based on the transformed model $\mathcal{P}Y = \mathcal{P}\boldsymbol{X}_B\beta^*_B + u$ rather than on the original model $Y = \boldsymbol{X}\beta^* + \varepsilon$. Even though a minor detail at first sight, this change of model brings about some technical complications. The issue is that the entries of the noise vector u are in general not independent, whereas those of ε are. Similarly, the rows of the design matrix $\mathcal{P}\boldsymbol{X}_B$ are in general not independent in contrast to those of $\boldsymbol{X}$. As a consequence, the central result of our theory, Theorem 1, cannot be applied to the estimator $\hat{\gamma}_{\alpha,B}$ directly. To adapt it to the present situation, we define the event

$$\mathcal{S}'_\gamma = \Big\{\frac{1}{\sqrt{n}}\|(\mathcal{P}\boldsymbol{X}_B)^\top u\|_\infty \le \gamma\Big\}$$

and let C'_1, K'_1, C'_2 and K'_2 be positive constants that depend only on the model parameters $\Theta' = \Theta \cup \{c_\vartheta, |A|, \|\Psi_A^{-1}\|_2\}$. With this notation at hand, we can prove the following.

Proposition A.11 *There exist an event $\mathcal{A}'_n$ with $\mathbb{P}(\mathcal{A}'_n) \ge 1 - C'_1 n^{-K'_1}$ for some positive constants C'_1 and K'_1 and a sequence of real numbers ν'_n with $0 < \nu'_n \le C'_2 n^{-K'_2}$ for some positive constants C'_2 and K'_2 such that the following holds: on the event $\mathcal{S}'_{\gamma^*_{\alpha+\nu'_n,B}} \cap \mathcal{A}'_n$,*

$$\gamma^*_{\alpha+\nu'_n,B} \le \hat{\gamma}_{\alpha,B} \le \gamma^*_{\alpha-\nu'_n,B}$$

for any $\alpha \in (a_n, 1-a_n)$ with $a_n = 2\nu'_n + (n \vee p)^{-1}$.

The overall strategy to prove Proposition A.11 is the same as the one for Theorem 1. There are some complications, however, that stem from the fact that the entries of u and the rows of $\mathcal{P}\boldsymbol{X}_B$ are not independent. We provide the proof of Proposition A.11 in the Supplementary Material, where we highlight the main differences to the proof of Theorem 1.

With Proposition A.11 in place, the proof of Proposition 7 proceeds analogously to the one of Proposition 6. For this reason, we only give a brief summary of the main steps. First suppose that the null hypothesis $H_{0,B}$ holds true. With the help of Proposition A.11, we get that

$$\begin{aligned}\mathbb{P}(T_B \le \hat{\gamma}_{\alpha,B}) &\ge \mathbb{P}\big(T_B \le \hat{\gamma}_{\alpha,B}, \mathcal{S}'_{\gamma^*_{\alpha+\nu'_n,B}} \cap \mathcal{A}'_n\big) \\ &\ge \mathbb{P}\big(T_B \le \gamma^*_{\alpha+\nu'_n,B}, \mathcal{S}'_{\gamma^*_{\alpha+\nu'_n,B}} \cap \mathcal{A}'_n\big) \\ &= \mathbb{P}\big(\mathcal{S}'_{\gamma^*_{\alpha+\nu'_n,B}} \cap \mathcal{A}'_n\big) \ge 1 - \alpha - \nu'_n - C'_1 n^{-K'_1},\end{aligned}$$

which implies that $\mathbb{P}(T_B \le \hat{\gamma}_{\alpha,B}) \ge 1 - \alpha + o(1)$ under $H_{0,B}$.

Next suppose we are on an alternative $H_{1,B} : \beta^*_B \neq 0$ that satisfies the conditions of Proposition 7 and let $\{\alpha_n\}$ be a null sequence with $2\nu'_n + (n \vee p)^{-1} < \alpha_n < \alpha$. Similarly as in the proof of Proposition 6, we can establish the bound

$$\mathbb{P}(T_B > \hat{\gamma}_{\alpha,B}) \ge \mathbb{P}\Big(\frac{\|(\mathcal{P}\boldsymbol{X}_B)^\top \mathcal{P}\boldsymbol{X}_B \beta^*_B\|_\infty}{\sqrt{n}} > \hat{\gamma}_{\alpha_n,B} + \frac{\|(\mathcal{P}\boldsymbol{X}_B)^\top u\|_\infty}{\sqrt{n}} \Big) \tag{A.19}$$

and verify the following: (i) $\hat{\gamma}_{\alpha_n,B} \le \psi'_n$ on the event $\mathcal{S}'_{\gamma^*_{\alpha_n+\nu'_n,B}} \cap \mathcal{A}'_n$, where $\psi'_n = C\ [\sqrt{2\log(2p)} + \sqrt{2\log(n \vee p)}]$ with some sufficiently large constant C that depends only on Θ', and (ii) $\mathbb{P}(\|(\mathcal{P}\boldsymbol{X}_B)^\top u\|_\infty / \sqrt{n} \le \psi'_n) = 1 - o(1)$. Applying (i) and (ii) to the right-hand side of (A.19) yields that

$$\begin{aligned} &\mathbb{P}\Big(\frac{\|(\mathcal{P}\boldsymbol{X}_B)^\top \mathcal{P}\boldsymbol{X}_B \beta^*_B\|_\infty}{\sqrt{n}} > \hat{\gamma}_{\alpha_n,B} + \frac{\|(\mathcal{P}\boldsymbol{X}_B)^\top u\|_\infty}{\sqrt{n}} \Big) \\ &\qquad \ge \mathbb{P}\Big(\frac{\|(\mathcal{P}\boldsymbol{X}_B)^\top \mathcal{P}\boldsymbol{X}_B \beta^*_B\|_\infty}{\sqrt{n}} > 2\psi'_n \Big) - o(1) = 1 - o(1), \end{aligned}$$

which in turn implies that $\mathbb{P}(T_B > \hat{\gamma}_{\alpha,B}) = 1 - o(1)$ under the alternative.

References

A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19:521–547, 2013.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98:791–806, 2011.

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, 2013.

P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, 2011.

T. Cai and Z. Guo. Accuracy assessment for high-dimensional linear regression. *Ann. Statist.*, 46:1807–1836, 2018.

V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.*, 41: 2786–2819, 2013.

V. Chernozhukov, D. Chetverikov, and K. Kato. Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probab. Theory and Related Fields*, 162:47–70, 2015.

D. Chetverikov, Z. Liao, and V. Chernozhukov. On cross-validated lasso in high dimensions. *Ann. Statist.*, 49:1300–1317, 2021.

M. Chichignoud, J. Lederer, and M. Wainwright. A practical scheme and fast algorithm to tune the lasso with optimality guarantees. *J. Mach. Learn. Res.*, 17:1–20, 2016.

A. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the lasso. *Bernoulli*, 23:552–581, 2017.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22, 2010.

C. Giraud. *Introduction to High-Dimensional Statistics*. CRC Press, 2014.

D. Gold, J. Lederer, and J. Tao. Inference for high-dimensional instrumental variables regression. *Journal of Econometrics*, 217:79–111, 2020.

T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: the Lasso and Generalizations*. CRC press, 2015.

M. Hebiri and J. Lederer. How correlations influence lasso prediction. *IEEE Trans. Inform. Theory*, 59:1846–1854, 2012.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15:2869–2909, 2014.

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Ann. Statist.*, 28:1356–1378, 2000.

V. Koltchinskii, K. Lounici, and A. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.*, 39:2302–2329, 2011.

J. Lederer. *Fundamentals of High-Dimensional Statistics—with Exercises and R Labs*. Springer, 2021.

J. Lederer, L. Yu, and I. Gaynanova. Oracle inequalities for high-dimensional prediction. *Bernoulli*, 25:1225–1255, 2019.

J. Lee, D. Sun, Y. Sun, and J. Taylor. Exact post-selection inference, with application to the lasso. *Ann. Statist.*, 44:907–927, 2016.

H. Leeb and B. Pötscher. Model selection and inference: facts and fiction. *Econometric Theory*, 21:21–59, 2005.

R. Lockhart, J. Taylor, R. Tibshirani, and R. Tibshirani. A significance test for the lasso. *Ann. Statist.*, 42:413–468, 2014.

A. Maurer, M. Pontil, and G. Lugosi. Structured sparsity and generalization. *J. Mach. Learn. Res.*, 13:671–690, 2012.

C. Micchelli, J. Morales, and M. Pontil. Regularizers for structured sparsity. *Adv. Comput. Math.*, 38:455–489, 2013.

R. Nickl and S. van de Geer. Confidence sets in sparse regression. *Ann. Statist.*, 41:2852–2876, 2013.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B*, 58: 267–288, 1996.

R. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. Exact post-selection inference for sequential regression procedures. *J. Amer. Statist. Assoc.*, 111:600–620, 2016.

S. van de Geer. *Estimation and Testing under Sparsity.* Springer, 2016.

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.

S. van de Geer and J. Lederer. The lasso, correlated design, and improved oracle inequalities. In *From Probability to Statistics and Back: High-Dimensional Models and Processes—A Festschrift in Honor of Jon A. Wellner*, pages 303–316. Institute of Mathematical Statistics, 2013.

S. van de Geer and A. Muro. On higher order isotropy conditions and lower bounds for sparse quadratic forms. *Electron. J. Stat.*, 8:3031–3061, 2014.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42:1166–1202, 2014.

C.-H. Zhang and S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B*, 76:217–242, 2014.

P. Zhao and B. Yu. On model selection consistency of lasso. *J. Mach. Learn. Res.*, 7: 2541–2563, 2006.