Least Squares After Model Selection in High-dimensional Sparse Models

Submitted to the Bernoulli

Least Squares After Model Selection in High-dimensional Sparse Models

Alexandre Belloni and Victor Chernozhukov

In this paper we study post-model selection estimators which apply ordinary least squares (ols) to the model selected by first-step penalized estimators, typically lasso. It is well known that lasso can estimate the nonparametric regression function at nearly the oracle rate, and is thus hard to improve upon. We show that ols post lasso estimator performs at least as well as lasso in terms of the rate of convergence, and has the advantage of a smaller bias. Remarkably, this performance occurs even if the lasso-based model selection "fails" in the sense of missing some components of the "true" regression model. By the "true" model we mean here the best s-dimensional approximation to the nonparametric regression function chosen by the oracle. Furthermore, ols post lasso estimator can perform strictly better than lasso, in the sense of a strictly faster rate of convergence, if the lasso-based model selection correctly includes all components of the "true" model as a subset and also achieves sufficient sparsity. In the extreme case, when lasso perfectly selects the "true" model, the ols post lasso estimator becomes the oracle estimator. An important ingredient in our analysis is a new sparsity bound on the dimension of the model selected by lasso which guarantees that this dimension is at most of the same order as the dimension of the "true" model. Our rate results are non-asymptotic and hold in both parametric and nonparametric models. Moreover, our analysis is not limited to the lasso estimator acting as selector in the first step, but also applies to any other estimator, for example various forms of thresholded lasso, with good rates and good sparsity properties. Our analysis covers both traditional thresholding and a new practical, data-driven thresholding scheme that induces maximal sparsity subject to maintaining a certain goodness-of-fit. The latter scheme has theoretical guarantees similar to those of lasso or ols post lasso, but it dominates these procedures as well as traditional thresholding in a wide variety of experiments.

First arXiv version: December 2009.
Key words. lasso, ols post lasso, post-model-selection estimators.
AMS Codes. Primary 62H12, 62J99; secondary 62J07.

1. Introduction

In this work we study post-model selected estimators for linear regression in high-dimensional sparse models (hdsms). In such models, the overall number of regressors p is very large, possibly much larger than the sample size n. However, there are $s = o(n)$ regressors that capture most of the impact of all covariates on the response variable. hdsms ([9], [22]) have emerged to deal with many new applications arising in biometrics,

signal processing, machine learning, econometrics, and other areas of data analysis where high-dimensional data sets have become widely available.

Several papers have begun to investigate estimation of hdsms, primarily focusing on mean regression with the ℓ_1-norm acting as a penalty function [4, 6, 7, 8, 9, 17, 22, 28, 31, 33]. The results in [4, 6, 7, 8, 17, 22, 31, 33] demonstrated the fundamental result that ℓ_1-penalized least squares estimators achieve the rate $\sqrt{s/n}\sqrt{\log p}$, which is very close to the oracle rate $\sqrt{s/n}$ achievable when the true model is known. The works [17, 28] demonstrated a similar fundamental result on the excess forecasting error loss under both quadratic and non-quadratic loss functions. Thus the estimator can be consistent and can have excellent forecasting performance even under very rapid, nearly exponential growth of the total number of regressors p. Also, [2] investigated the ℓ_1-penalized quantile regression process, obtaining similar results. See [4, 6, 7, 8, 15, 19, 20, 24] for many other interesting developments and a detailed review of the existing literature.

In this paper we derive theoretical properties of post-model selection estimators which apply ordinary least squares (ols) to the model selected by first-step penalized estimators, typically lasso. It is well known that lasso can estimate the mean regression function at nearly the oracle rate, and hence is hard to improve upon. We show that ols post lasso can perform at least as well as lasso in terms of the rate of convergence, and has the advantage of a smaller bias. This nice performance occurs even if the lasso-based model selection "fails" in the sense of missing some components of the "true" regression model. Here by the "true" model we mean the best s-dimensional approximation to the regression function chosen by the oracle. The intuition for this result is that lasso-based model selection omits only those components with relatively small coefficients. Furthermore, ols post lasso can perform strictly better than lasso, in the sense of a strictly faster rate of convergence, if the lasso-based model correctly includes all components of the "true" model as a subset and is sufficiently sparse. Of course, in the extreme case, when lasso perfectly selects the "true" model, the ols post lasso estimator becomes the oracle estimator.

Importantly, our rate analysis is not limited to the lasso estimator in the first step, but applies to a wide variety of other first-step estimators, including, for example, thresholded lasso, the Dantzig selector, and their various modifications. We give generic rate results that cover any first-step estimator for which a rate and a sparsity bound are available. We also give a generic result on using thresholded lasso as the first-step estimator, where thresholding can be performed by a traditional thresholding scheme (t-lasso) or by a new fitness-thresholding scheme we introduce in the paper (fit-lasso). The new thresholding scheme induces maximal sparsity subject to maintaining a certain goodness-of-fit in the sample, and is completely data-driven. We show that ols post fit-lasso estimator performs at least as well as the lasso estimator, but can be strictly better under good model selection properties.

Finally, we conduct a series of computational experiments and find that the results confirm our theoretical findings. Figure 1 is a brief graphical summary of our theoretical results showing how the empirical risk of various estimators change with the signal strength C (coefficients of relevant covariates are set equal to C). For very low level of

signal, all estimators perform similarly. When the signal strength is intermediate, ols post lasso and ols post fit-lasso substantially outperform lasso and the ols post t-lasso estimators. However, we find that the ols post fit-lasso outperforms ols post lasso whenever lasso does not produce very sparse solutions which occurs if the signal strength level is not low. For large levels of signal, ols post fit-lasso and ols post t-lasso perform very well improving upon lasso and ols post lasso. Thus, the main message here is that ols post lasso and ols post fit-lasso perform at least as well as lasso and sometimes a lot better.

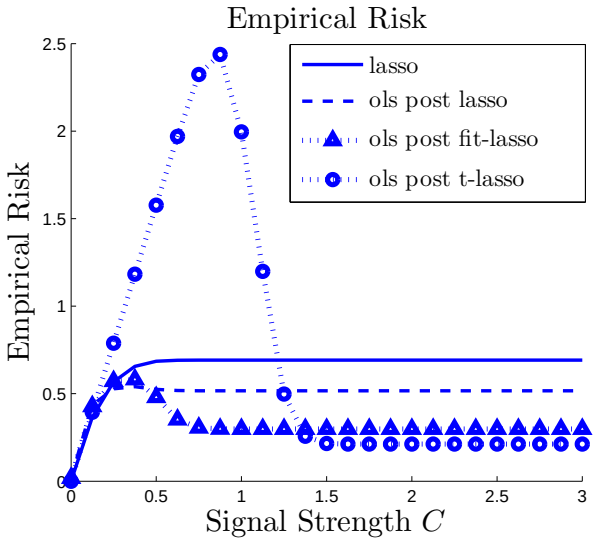

Figure 1. This figure plots the performance of the estimators listed in the text under the equi-correlated design for the covariates $x_i \sim N(0, \Sigma)$, $\Sigma_{jk} = 1/2$ if $j \neq k$. The number of regressors is $p = 500$ and the sample size is $n = 100$ with 1000 simulations for each level of signal strength C. In each simulation there are 5 relevant covariates whose coefficients are set equal to the signal strength C, and the variance of the noise is set to 1.

To the best of our knowledge, our paper is the first to establish the aforementioned rate results on ols post lasso and the proposed ols post fitness-thresholded lasso in the mean regression problem. Our analysis builds upon the ideas in [2], who established the properties of post-penalized procedures for the related, but different, problem of median regression. Our analysis also builds on the fundamental results of [4] and the other works cited above that established the properties of the first-step lasso-type estimators. An important ingredient in our analysis is a new sparsity bound on the dimension of the model selected by lasso, which guarantees that this dimension is at most of the same order as the dimension of the "true" model. This result builds on some inequalities for sparse eigenvalues and reasoning previously given in [2] in the context of median regression. Our sparsity bounds for lasso improve upon the analogous bounds in [4] and are comparable to the bounds in [33] obtained under a larger penalty level. We also rely on maximal inequalities in [33] to provide primitive conditions for the sharp sparsity

bounds to hold.

We organize the paper as follows. Section 2 reviews the model and discusses the estimators. Section 3 revisits some benchmark results of [4] for lasso, albeit allowing for a data driven choice of penalty level, develops an extension of model selection results of [19] to the nonparametric case, and derives a new sparsity bound for lasso. Section 4 presents a generic rate result on ols post-model selection estimators. Section 5 applies the generic results to the ols post lasso and the ols post thresholded lasso estimators. Appendix contains main proofs and the Supplementary Appendix contains auxiliary proofs. In the Supplementary Appendix we also present the results of our computational experiments.

Notation. In making asymptotic statements, we assume that $n \to \infty$ and $p = p_n \to \infty$, and we also allow for $s = s_n \to \infty$. In what follows, all parameter values are indexed by the sample size n, but we omit the index whenever this does not cause confusion. We use the notation $(a)_+ = \max\{a, 0\}$, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. The ℓ_2-norm is denoted by $\|\cdot\|$, the ℓ_1-norm is denoted by $\|\cdot\|_1$, the ℓ_∞-norm is denoted by $\|\cdot\|_\infty$, and the ℓ_0-norm $\|\cdot\|_0$ denotes the number of non-zero components of a vector. Given a vector $\delta \in \mathbb{R}^p$, and a set of indices $T \subset \{1, \ldots, p\}$, we denote by δ_T the vector in which $\delta_{Tj} = \delta_j$ if $j \in T$, $\delta_{Tj} = 0$ if $j \notin T$, and by $|T|$ the cardinality of T. Given a covariate vector $x_i \in \mathbb{R}^p$, we denote by $x_i[T]$ vector $\{x_{ij}, j \in T\}$. The symbol $E[\cdot]$ denotes the expectation. We also use standard empirical process notation

$$\mathbb{E}_n[f(z_\bullet)] := \sum_{i=1}^n f(z_i)/n \quad \text{and} \quad \mathbb{G}_n(f(z_\bullet)) := \sum_{i=1}^n (f(z_i) - E[f(z_i)])/\sqrt{n}.$$

We also denote the $L^2(\mathbb{P}_n)$-norm by $\|f\|_{\mathbb{P}_n,2} = (\mathbb{E}_n[f_\bullet^2])^{1/2}$. Given covariate values $x_1, \ldots, x_n$, we define the prediction norm of a vector $\delta \in \mathbb{R}^p$ as $\|\delta\|_{2,n} = \{\mathbb{E}_n[(x_i'\delta)^2]\}^{1/2}$. We use the notation $a \lesssim b$ to denote $a \leq Cb$ for some constant $C > 0$ that does not depend on n (and therefore does not depend on quantities indexed by n like p or s); and $a \lesssim_P b$ to denote $a = O_P(b)$. For an event A, we say that A wp $\to 1$ when A occurs with probability approaching one as n grows. Also we denote by $\bar{c} = (c+1)/(c-1)$ for a chosen constant $c > 1$.

2. The setting, estimators, and conditions

2.1. The setting

Condition (M). *We have data $\{(y_i, z_i), i = 1, \ldots, n\}$ such that for each n*

$$y_i = f(z_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n, \tag{2.1}$$

where y_i are the outcomes, z_i are vectors of fixed regressors, and ϵ_i are i.i.d. errors. Let $P(z_i)$ be a given p-dimensional dictionary of technical regressors with respect z_i, i.e. a p-vector of transformation of z_i, with components

$$x_i := P(z_i)$$

of the dictionary normalized so that

$$\mathbb{E}_n[x_{\bullet j}^2] = 1 \text{ for } j = 1, \ldots, p.$$

In making making asymptotic statements, we assume that $n \to \infty$ *and* $p = p_n \to \infty$*, and that all parameters of the model are implicitly indexed by* n*.*

We would like to estimate the nonparametric regression function f at the design points, namely the values $f_i = f(z_i)$ for $i = 1, \ldots, n$. In order to setup estimation and define a performance benchmark we consider the following oracle risk minimization program:

$$\min_{0 \le k \le p \wedge n} c_k^2 + \sigma^2 \frac{k}{n}, \tag{2.2}$$

where

$$c_k^2 := \min_{\|\beta\|_0 \le k} \mathbb{E}_n[(f_\bullet - x_\bullet'\beta)^2]. \tag{2.3}$$

Note that $c_k^2 + \sigma^2 k/n$ is an upper bound on the risk of the best k-sparse least squares estimator, i.e. the best estimator amongst all least squares estimators that use k out of p components of x_i to estimate f_i, for $i = 1, \ldots, n$. The oracle program (2.2) chooses the optimal value of k. Let s be the smallest integer amongst these optimal values, and let

$$\beta_0 \in \arg\min_{\|\beta\|_0 \le s} \mathbb{E}_n[(f_\bullet - x_\bullet'\beta)^2]. \tag{2.4}$$

We call β_0 the oracle target value, $T := \text{support}(\beta_0)$ the oracle model, $s := |T| = \|\beta_0\|_0$ the dimension of the oracle model, and $x_i'\beta_0$ the oracle approximation to f_i. The latter is our intermediary target, which is equal to the ultimate target f_i up to the approximation error

$$r_i := f_i - x_i'\beta_0.$$

If we knew T we could simply use $x_i[T]$ as regressors and estimate f_i, for $i = 1, \ldots, n$, using the least squares estimator, achieving the risk of at most

$$c_s^2 + \sigma^2 s/n,$$

which we call the oracle risk. Since T is not known, we shall estimate T using lasso-type methods and analyze the properties of post-model selection least squares estimators, accounting for possible model selection mistakes.

***Remark* 2.1** (The oracle program)**.** Note that if argmin is not unique in the problem (2.4), it suffices to select one of the values in the set of argmins. Supplementary Appendix provides a more detailed discussion of the oracle problem. The idea of using oracle problems such as (2.2) for benchmarking the performance follows its prior uses in the literature. For instance, see [4], Theorem 6.1, where an analogous problem appears in upper bounds on performance of lasso. □

***Remark* 2.2** (A leading special case)**.** When contrasting the performance of lasso and ols post lasso estimators in Remarks 5.1-5.2 given later, we shall mention a balanced case where

$$c_s^2 \lesssim \sigma^2 s/n \tag{2.5}$$

which says that the oracle program (2.2) is able to balance the norm of the bias squared to be not much larger than the variance term $\sigma^2 s/n$. This corresponds to the case that the approximation error bias does not dominate the estimation error of the oracle least squares estimator, so that the oracle rate of convergence simplifies to $\sqrt{s/n}$ mentioned in the introduction.

2.2. Model selectors based on lasso

Given the large number of regressors $p > n$, some regularization or covariate selection is required in order to obtain consistency. The lasso estimator [26], defined as follows, achieves both tasks by using the ℓ_1 penalization:

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \widehat{Q}(\beta) + \frac{\lambda}{n}\|\beta\|_1, \quad \text{where} \quad \widehat{Q}(\beta) = \mathbb{E}_n[(y_\bullet - x_\bullet'\beta)^2], \tag{2.6}$$

and λ is the penalty level whose choice is described below. If the solution is not unique we pick any solution with minimum support. The lasso is often used as an estimator and more often only as a model selection device, with the model selected by lasso given by:

$$\widehat{T} := \text{support}(\widehat{\beta}).$$

Moreover, we denote by $\widehat{m} := |\widehat{T} \setminus T|$ the number of components outside T selected by lasso and by $\widehat{f}_i = x_i'\widehat{\beta}, i = 1, \ldots, n$ the lasso estimate of $f_i, i = 1, \ldots, n$.

Oftentimes additional thresholding is applied to remove regressors with small estimated coefficients, defining the so called thresholded lasso estimator:

$$\widehat{\beta}(t) = (\widehat{\beta}_j 1\{|\widehat{\beta}_j| > t\}, j = 1, \ldots, p), \tag{2.7}$$

where $t \geq 0$ is the thresholding level, and the corresponding selected model is then

$$\widehat{T}(t) := \text{support}(\widehat{\beta}(t)).$$

Note that setting $t = 0$, we have $\widehat{T}(t) = \widehat{T}$, so lasso is a special case of thresholded lasso.

2.3. Post-model selection estimators

Given this all of our post-model selection estimators or ols post lasso estimators will take the form

$$\widetilde{\beta}^t = \arg\min_{\beta \in \mathbb{R}^p} \widehat{Q}(\beta) : \beta_j = 0 \text{ for each } j \in \widehat{T}^c(t). \tag{2.8}$$

That is given the model selected a threshold lasso $\widehat{T}(t)$, including the lasso's model $\widehat{T}(0)$ as a special case, the post-model selection estimator applies the ordinary least squares to the selected model.

In addition to the case of $t = 0$, we also consider the following choices for the threshold level:

$$\begin{array}{ll} \text{traditional threshold (t):} & t > \zeta = \max_{1 \le j \le p} |\widehat{\beta}_j - \beta_{0j}|, \\ \text{fitness-based threshold (fit):} & t = t_\gamma := \max_{t \ge 0} \{t : \widehat{Q}(\widetilde{\beta}^t) - \widehat{Q}(\widehat{\beta}) \le \gamma\}, \end{array} \tag{2.9}$$

where $\gamma \le 0$, and $|\gamma|$ is the gain of the in-sample fit allowed relative to lasso.

As discussed in Section 3.2, the standard thresholding method is particularly appealing in models in which oracle coefficients β_0 are well separated from zero. This scheme however may perform poorly in models with oracle coefficients not well separated from zero and in nonparametric models. Indeed, even in parametric models with many small but non-zero true coefficients, thresholding the estimates too aggressively may result in large goodness-of-fit losses, and consequently in slow rates of convergence and even inconsistency for the second-step estimators. This issue directly motivates our new goodness-of-fit based thresholding method, which sets to zero small coefficient estimates as much as possible subject to maintaining a certain goodness-of-fit level.

Depending on how we select the threshold, we consider the following three types of the post-model selection estimators:

$$\begin{array}{lll} \text{ols post lasso:} & \widetilde{\beta}^0 & (t = 0), \\ \text{ols post t-lasso:} & \widetilde{\beta}^t & (t > \zeta), \\ \text{ols post fit-lasso:} & \widetilde{\beta}^{t_\gamma} & (t = t_\gamma). \end{array} \tag{2.10}$$

The first estimator is defined by ols applied to the model selected by lasso, also called Gauss-lasso; the second by ols applied to the model selected by the thresholded lasso, and the third by ols applied to the model selected by fitness-thresholded lasso.

The main purpose of this paper is to derive the properties of the post-model selection estimators (2.10). If model selection works perfectly, which is possible only under rather special circumstances, then the post-model selection estimators are the oracle estimators, whose properties are well known. However, of a much more general interest is the case when model selection does not work perfectly, as occurs for many designs of interest in applications.

2.4. Choice and computation of penalty level for lasso

The key quantity in the analysis is the gradient of $\widehat{Q}$ at the true value:

$$S = 2\mathbb{E}_n[x_\bullet \epsilon_\bullet].$$

This gradient is the effective "noise" in the problem that should be dominated by the regularization. However we would like to make the bias as small as possible. This reasoning

suggests choosing the smallest penalty level λ so that to dominate the noise, namely

$$\lambda \geq cn\|S\|_\infty \text{ with probability at least } 1-\alpha, \tag{2.11}$$

where probability $1-\alpha$ needs to be close to 1 and $c > 1$. Therefore, we propose setting

$$\lambda = c' \, \widehat{\sigma} \, \Lambda(1-\alpha|X), \text{for some fixed } \; c' > c > 1, \tag{2.12}$$

where $\Lambda(1-\alpha|X)$ is the $(1-\alpha)$-quantile of $n\|S/\sigma\|_\infty$, and $\widehat{\sigma}$ is a possibly data-driven estimate of σ. Note that the quantity $\Lambda(1-\alpha|X)$ is independent of σ and can be easily approximated by simulation. We refer to this choice of λ as the data-driven choice, reflecting the dependence of the choice on the design matrix $X = [x_1, \ldots, x_n]'$ and a possibly data-driven $\widehat{\sigma}$. Note that the proposed (2.12) is sharper than $c'\widehat{\sigma}2\sqrt{2n\log(p/\alpha)}$ typically used in the literature. We impose the following conditions on $\widehat{\sigma}$.

Condition (V). *The estimated $\widehat{\sigma}$ obeys*

$$\ell \leq \widehat{\sigma}/\sigma \leq u \text{ with probability at least } 1-\tau,$$

where $0 < \ell \leq 1$ and $1 \leq u$ and $0 \leq \tau < 1$ be constants possibly dependent on n.

We can construct a $\widehat{\sigma}$ that satisfies this condition under mild assumptions as follows. First, set $\widehat{\sigma} = \widehat{\sigma}_0$, where $\widehat{\sigma}_0$ is an upper bound on σ which is possibly data-driven, for example the sample standard deviation of y_i. Second, compute the lasso estimator based on this estimate and set $\widehat{\sigma}^2 = \widehat{Q}(\widehat{\beta})$. We demonstrate that $\widehat{\sigma}$ constructed in this way satisfies Condition V and characterize quantities u and ℓ and τ in the Supplementary Appendix. We can iterate on the last step a bounded number of times. Moreover, we can similarly use ols post lasso for this purpose.

2.5. Choices and computation of thresholding levels

Our analysis will cover a wide range of possible threshold levels. Here, however, we would like to propose some basic options that give both good finite-sample and theoretical results. In the traditional thresholding method, we can set

$$t = \tilde{c}\lambda/n, \tag{2.13}$$

for some $\tilde{c} \geq 1$. This choice is theoretically motivated by Section 3.2 that presents the perfect model selection results, where under some conditions $\zeta \leq \tilde{c}\lambda/n$. This choice also leads to near-oracle performance of the resulting post-model selection estimator. Regarding the choice of $\tilde{c}$, we note that setting $\tilde{c} = 1$ and achieving $\zeta \leq \lambda/n$ is possible by the results of Section 3.2 if empirical Gram matrix is orthogonal and approximation error c_s vanishes. Thus, $\tilde{c} = 1$ is the least aggressive traditional thresholding one can perform under conditions of Section 3.2 (note also that $\tilde{c} = 1$ has performed better than $\tilde{c} > 1$ in our computational experiments).

Our fitness-based threshold t_γ requires the specification of the parameter γ. The simplest choice delivering near-oracle performance is $\gamma = 0$; this choice leads to the sparsest post-model selection estimator that has the same in-sample fit as lasso. Our preferred choice is however to set

$$\gamma = \frac{\widehat{Q}(\widetilde{\beta}^0) - \widehat{Q}(\widehat{\beta})}{2} < 0, \tag{2.14}$$

where $\widetilde{\beta}^0$ is the ols post lasso estimator. The resulting estimator is more sparse than lasso, and it also produces a better in-sample fit than lasso. This choice also results in near-oracle performance and also leads to the best performance in computational experiments. Note also that for any γ, we can compute t_γ by a binary search over $t \in \text{sort}\{|\widehat{\beta}_j|, j \in \widehat{T}\}$, where sort is the sorting operator. This is the case since the final estimator depends only on the selected support and not on the specific value of t used. Therefore, since there are at most $|\widehat{T}|$ different values of t to be tested, by using a binary search, we can compute t_γ exactly by running at most $\lceil \log_2 |\widehat{T}| \rceil$ ordinary least squares problems.

2.6. Conditions on the design

For the analysis of lasso we rely on the following restricted eigenvalue condition.

Condition (RE($\bar{c}$)). *For a given $\bar{c} \geq 0$,*

$$\kappa(\bar{c}) := \min_{\|\delta_{T^c}\|_1 \leq \bar{c}\|\delta_T\|_1, \delta \neq 0} \frac{\sqrt{s}\|\delta\|_{2,n}}{\|\delta_T\|_1} > 0.$$

This condition is a variant of the restricted eigenvalue condition introduced in [4], that is known to be quite general and plausible; see also [4] for related conditions.

For the analysis of post-model selection estimators we need the following restricted sparse eigenvalue condition.

Condition (RSE(m)). *For a given $m < n$,*

$$\widetilde{\kappa}(m)^2 := \min_{\|\delta_{T^c}\|_0 \leq m, \delta \neq 0} \frac{\|\delta\|_{2,n}^2}{\|\delta\|^2} > 0, \quad \phi(m) := \max_{\|\delta_{T^c}\|_0 \leq m, \delta \neq 0} \frac{\|\delta\|_{2,n}^2}{\|\delta\|^2} > 0.$$

Here m denotes the restriction on the number of non-zero components outside the support T. It will be convenient to define the following condition number associated with the empirical Gram matrix:

$$\mu(m) = \frac{\sqrt{\phi(m)}}{\widetilde{\kappa}(m)}. \tag{2.15}$$

The following lemma demonstrates the plausibility of conditions above for the case where the values x_i, $i = 1, \ldots, n$, have been generated as a realization of the random sample; there are also other primitive conditions. In this case we can expect that empirical restricted eigenvalue is actually bounded away from zero and (2.15) is bounded

from above with a high probability. The lemma uses the standard concept of (unrestricted) sparse eigenvalues (see, e.g. [4]) to state a primitive condition on the population Gram matrix. The lemma allows for standard arbitrary bounded dictionaries, arising in the nonparametric estimation, for example regression splines, orthogonal polynomials, and trigonometric series, see [14, 29, 32, 27]. Similar results are known to also hold for standard Gaussian regressors [33].

Lemma 1 (Plausibility of RE and RSE). *Suppose $\tilde{x}_i$, $i = 1, \ldots, n$, are i.i.d. zero-mean vectors, such that the population design matrix $\mathrm{E}[\tilde{x}_i \tilde{x}_i']$ has ones on the diagonal, and its $s \log n$-sparse eigenvalues are bounded from above by $\varphi < \infty$ and bounded from below by $\kappa^2 > 0$. Define x_i as a normalized form of $\tilde{x}_i$, namely $x_{ij} = \tilde{x}_{ij}/(\mathbb{E}_n[\tilde{x}_{\bullet j}^2])^{1/2}$. Suppose that $\tilde{x}_i \max_{1 \leq i \leq n} \|\tilde{x}_i\|_\infty \leq K_n$ a.s., and $K_n^2 s \log^2(n) \log^2(s \log n) \log(p \vee n) = o(n\kappa^4/\varphi)$. Then, for any $m + s \leq s \log n$, the empirical restricted sparse eigenvalues obey the following bounds:*

$$\phi(m) \leq 4\varphi, \quad \widetilde{\kappa}(m)^2 \geq \kappa^2/4, \quad \text{and} \quad \mu(m) \leq 4\sqrt{\varphi}/\kappa,$$

with probability approaching 1 as $n \to \infty$.

3. Results on lasso as an estimator and model selector

The properties of the post-model selection estimators will crucially depend on both the estimation and model selection properties of lasso. In this section we develop the estimation properties of lasso under the data-dependent penalty level, extending the results of [4], and develop the model selection properties of lasso for non-parametric models, generalizing the results of [19] to the nonparametric case.

3.1. Estimation Properties of lasso

The following theorem describes the main estimation properties of lasso under the data-driven choice of the penalty level.

Theorem 1 (Performance bounds for lasso under data-driven penalty). *Suppose that Conditions M and RE($\bar{c}$) hold for $\bar{c} = (c+1)/(c-1)$. If $\lambda \geq cn\|S\|_\infty$, then*

$$\|\widehat{\beta} - \beta_0\|_{2,n} \leq \left(1 + \frac{1}{c}\right) \frac{\lambda\sqrt{s}}{n\kappa(\bar{c})} + 2c_s.$$

Moreover, suppose that Condition V holds. Under the data-driven choice (2.12), for $c' \geq c/\ell$, we have $\lambda \geq cn\|S\|_\infty$ with probability at least $1 - \alpha - \tau$, so that with at least the same probability

$$\|\widehat{\beta} - \beta_0\|_{2,n} \leq (c' + c'/c) \frac{\sqrt{s}}{n\kappa(\bar{c})} \sigma u \Lambda(1-\alpha|X) + 2c_s, \text{ where } \Lambda(1-\alpha|X) \leq \sqrt{2n \log(p/\alpha)}.$$

If further RE($2\bar{c}$) holds, then

$$\|\widehat{\beta}-\beta_0\|_1 \le \left(\frac{(1+2\bar{c})\sqrt{s}}{\kappa(2\bar{c})}\|\widehat{\beta}-\beta_0\|_{2,n}\right) \vee \left(\left(1+\frac{1}{2\bar{c}}\right)\frac{2c}{c-1}\frac{n}{\lambda}c_s^2\right).$$

This theorem extends the result of [4] by allowing for data-driven penalty level and deriving the rates in ℓ_1-norm. These results may be of independent interest and are needed for subsequent results.

***Remark* 3.1.** Furthermore, a performance bound for the estimation of the regression function follows from the relation

$$\left|\|\widehat{f}-f\|_{\mathbb{P}_n,2} - \|\widehat{\beta}-\beta_0\|_{2,n}\right| \le c_s, \tag{3.16}$$

where $\widehat{f}_i = x_i'\widehat{\beta}$ is the lasso estimate of the regression function f evaluated at z_i. It is interesting to know some lower bounds on the rate which follow from Karush-Kuhn-Tucker conditions for lasso (see equation (A.25) in the appendix):

$$\|\widehat{f}-f\|_{\mathbb{P}_n,2} \ge \frac{(1-1/c)\lambda\sqrt{|\widehat{T}|}}{2n\sqrt{\phi(\widehat{m})}},$$

where $\widehat{m} = |\widehat{T}\setminus T|$. We note that a similar lower bound was first derived in [21] with $\phi(p)$ instead of $\phi(\widehat{m})$. □

The preceding theorem and discussion imply the following useful asymptotic bound on the performance of the estimators.

Corollary 1 (Asymptotic bounds on performance of lasso)**.** *Under the conditions of Theorem 1, if*

$$\phi(\widehat{m}) \lesssim 1,\ \kappa(\bar{c}) \gtrsim 1,\ \mu(\widehat{m}) \lesssim 1,\ \log(1/\alpha) \lesssim \log p,\ \alpha = o(1),\ u/\ell \lesssim 1,\ \text{and } \tau = o(1) \tag{3.17}$$

hold as n grows, we have that

$$\|\widehat{f}-f\|_{\mathbb{P}_n,2} \lesssim_P \sigma\sqrt{\frac{s\log p}{n}} + c_s.$$

Moreover, if $|\widehat{T}| \gtrsim_P s$, in particular if $T \subseteq \widehat{T}$ with probability going to 1, we have

$$\|\widehat{f}-f\|_{\mathbb{P}_n,2} \gtrsim_P \sigma\sqrt{\frac{s\log p}{n}}.$$

In Lemma 1 we established fairly general sufficient conditions for the first three relations in (3.17) to hold with high probability as n grows, when the design points $z_1,\dots,z_n$

were generated as a random sample. The remaining relations are mild conditions on the choice of α and the estimation of σ that are used in the definition of the data-driven choice (2.12) of the penalty level λ.

It follows from the corollary that provided $\kappa(\bar{c})$ is bounded away from zero, lasso with data-driven penalty estimates the regression function at a near-oracle rate. The second part of the corollary generalizes to the nonparametric case the lower bound obtained for lasso in [21]. It shows that the rate cannot be improved in general. We shall use the asymptotic rates of convergence to compare the performance of lasso and the post-model selection estimators.

3.2. Model selection properties of lasso

The main results of the paper do not require the first-step estimators like lasso to perfectly select the "true" oracle model. In fact, we are specifically interested in the most common cases where these estimators do not perfectly select the true model. For these cases, we will prove that post-model selection estimators such as ols post lasso achieve near-oracle rates like those of lasso. However, in some special cases, where perfect model selection is possible, these estimators can achieve the exact oracle rates, and thus can be even better than lasso. The purpose of this section is to describe these very special cases where perfect model selection is possible.

Theorem 2 (Some conditions for perfect model selection in nonparametric setting)**.** *Suppose that Condition M holds. (1) If the coefficients are well separated from zero, that is*

$$\min_{j\in T}|\beta_{0j}| > \zeta + t, \qquad \text{for some } t \geq \zeta := \max_{j=1,\ldots,p}|\widehat{\beta}_j - \beta_{0j}|,$$

then the true model is a subset of the selected model, $T := \mathrm{support}(\beta_0) \subseteq \widehat{T} := \mathrm{support}(\widehat{\beta})$. *Moreover,* T *can be perfectly selected by applying level* t *thresholding to* $\widehat{\beta}$*, i.e.* $T = \widehat{T}(t)$. *(2) In particular, if* $\lambda \geq cn\|S\|_\infty$*, and there is a constant* $U > 5\bar{c}$ *such that the empirical Gram matrix satisfies* $|\mathbb{E}_n[x_{\bullet j}x_{\bullet k}]| \leq 1/(Us)$ *for all* $1 \leq j < k \leq p$*, then*

$$\zeta \leq \frac{\lambda}{n}\cdot\frac{U+\bar{c}}{U-5\bar{c}} + \frac{\sigma}{\sqrt{n}}\wedge c_s + \frac{6\bar{c}}{U-5\bar{c}}\frac{c_s}{\sqrt{s}} + \frac{4\bar{c}}{U}\frac{n}{\lambda}\frac{c_s^2}{s}.$$

These results substantively generalize the parametric results of [19] on model selection by thresholded lasso. These results cover the more general nonparametric case and may be of independent interest. Note also that the conditions for perfect model selection stated require a strong assumption on the separation of coefficients of the oracle from zero, and also a near perfect orthogonality of the empirical Gram matrix. This is the sense in which the perfect model selection is a rather special, non-general phenomenon. Finally, we note that it is possible to perform perfect selection of the oracle model by lasso without applying any additional thresholding under additional technical conditions and higher penalty levels [34, 31, 5]. In the supplement we state the nonparametric extension of the parametric result due to [31].

3.3. Sparsity properties of lasso

We also derive new sharp sparsity bounds for lasso, which may be of independent interest.

We begin with a preliminary sparsity bound for lasso.

Lemma 2 (Empirical pre-sparsity for lasso)**.** *Suppose that Conditions M and RE($\bar{c}$) hold, $\lambda \geq cn\|S\|_\infty$, and let $\widehat{m} = |\widehat{T} \setminus T|$. We have for $\bar{c} = (c+1)/(c-1)$ that*

$$\sqrt{\widehat{m}} \leq \sqrt{s}\sqrt{\phi(\widehat{m})}\ 2\bar{c}/\kappa(\bar{c}) + 3(\bar{c}+1)\sqrt{\phi(\widehat{m})}\ nc_s/\lambda.$$

The lemma above states that lasso achieves the oracle sparsity up to a factor of $\phi(\widehat{m})$. Under the conditions (2.5) and $\kappa(\bar{c}) \gtrsim 1$, the lemma above immediately yields the simple upper bound on the sparsity of the form

$$\widehat{m} \lesssim_P s\phi(n), \tag{3.18}$$

as obtained for example in [4] and [22]. Unfortunately, this bound is sharp only when $\phi(n)$ is bounded. When $\phi(n)$ diverges, for example when $\phi(n) \gtrsim_P \sqrt{\log p}$ in the Gaussian design with $p \geq 2n$ by Lemma 6 of [3], the bound is not sharp. However, for this case we can construct a sharp sparsity bound by combining the preceding pre-sparsity result with the following sub-linearity property of the restricted sparse eigenvalues.

Lemma 3 (Sub-linearity of restricted sparse eigenvalues)**.** *For any integer $k \geq 0$ and constant $\ell \geq 1$ we have $\phi(\lceil \ell k \rceil) \leq \lceil \ell \rceil \phi(k)$.*

A version of this lemma for unrestricted sparse eigenvalues has been previously proven in [2]. The combination of the preceding two lemmas gives the following sparsity theorem.

Theorem 3 (Sparsity bound for lasso under data-driven penalty)**.** *Suppose that Conditions M and RE($\bar{c}$) hold, and let $\widehat{m} := |\widehat{T} \setminus T|$. The event $\lambda \geq cn\|S\|_\infty$ implies that*

$$\widehat{m} \leq s \cdot \left[\min_{m \in \mathcal{M}} \phi(m \wedge n) \right] \cdot L_n,$$

where $\mathcal{M} = \{m \in \mathbb{N} : m > s\phi(m \wedge n) \cdot 2L_n\}$ and $L_n = [2\bar{c}/\kappa(\bar{c}) + 3(\bar{c}+1)nc_s/(\lambda\sqrt{s})]^2$.

The main implication of Theorem 3 is that under (2.5), if $\min_{m \in \mathcal{M}} \phi(m \wedge n) \lesssim 1$ and $\lambda \geq cn\|S\|_\infty$ hold with high probability, which is valid by Lemma 1 for important designs and by the choice of penalty level (2.12), then with high probability

$$\widehat{m} \lesssim s. \tag{3.19}$$

Consequently, for these designs and penalty level, lasso's sparsity is of the same order as the oracle sparsity, namely $\widehat{s} := |\widehat{T}| \leq s + \widehat{m} \lesssim s$ with high probability. The reason for this is that $\min_{m \in \mathcal{M}} \phi(m) \ll \phi(n)$ for these designs, which allows us to sharpen the previous sparsity bound (3.18) considered in [4] and [22]. Also, our new bound is comparable to the bounds in [33] in terms of order of sharpness, but it requires a smaller penalty level λ which also does not depend on the unknown sparse eigenvalues as in [33].

4. Performance of post-model selection estimators with a generic model selector

Next, we present a general result on the performance of a post-model selection estimator with a generic model selector.

Theorem 4 (Performance of post-model selection estimator with a generic model selector)**.** *Suppose Condition M holds and let $\widehat{\beta}$ be any first-step estimator acting as the model selector and denote by $\widehat{T} := \text{support}(\widehat{\beta})$ the model it selects, such that $|\widehat{T}| \leq n$. Let $\widetilde{\beta}$ be the post-model selection estimator defined by*

$$\widetilde{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \widehat{Q}(\beta) : \beta_j = 0, \text{ for each } j \in \widehat{T}^c. \tag{4.20}$$

Let $B_n := \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0)$ and $C_n := \widehat{Q}(\beta_{0\widehat{T}}) - \widehat{Q}(\beta_0)$ and $\widehat{m} = |\widehat{T} \setminus T|$ be the number of wrong regressors selected. Then, if condition RSE($\widehat{m}$) holds, for any $\varepsilon > 0$, there is a constant K_ε independent of n such that with probability at least $1 - \varepsilon$, for $\widetilde{f}_i = x_i'\widetilde{\beta}$ we have

$$\|\widetilde{f} - f\|_{\mathbb{P}_n,2} \leq K_\varepsilon \sigma \sqrt{\frac{\widehat{m} \log p + (\widehat{m} + s) \log(e\mu(\widehat{m}))}{n}} + 3c_s + \sqrt{(B_n)_+ \wedge (C_n)_+}.$$

Furthermore, for any $\varepsilon > 0$, there is a constant K_ε independent of n such that with probability at least $1 - \varepsilon$,

$$B_n \leq \|\widehat{\beta} - \beta_0\|_{2,n}^2 + \left[K_\varepsilon \sigma \sqrt{\frac{\widehat{m} \log p + (\widehat{m} + s) \log(e\mu(\widehat{m}))}{n}} + 2c_s \right] \|\widehat{\beta} - \beta_0\|_{2,n}$$

$$C_n \leq 1\{T \not\subseteq \widehat{T}\} \left(\|\beta_{0\widehat{T}^c}\|_{2,n}^2 + \left[K_\varepsilon \sigma \sqrt{\frac{\log \binom{s}{\widehat{k}} + \widehat{k} \log(e\mu(0))}{n}} + 2c_s \right] \|\beta_{0\widehat{T}^c}\|_{2,n} \right).$$

Three implications of Theorem 4 are worth noting. First, the bounds on the prediction norm stated in Theorem 4 apply to the ols estimator on the components selected by any first-step estimator $\widehat{\beta}$, provided we can bound both the rate of convergence $\|\widehat{\beta} - \beta_0\|_{2,n}$ of the first-step estimator and $\widehat{m}$, the number of wrong regressors selected by the model selector. Second, note that if the selected model contains the true model, $T \subseteq \widehat{T}$, then we have $(B_n)_+ \wedge (C_n)_+ = C_n = 0$, and B_n does not affect the rate at all, and the performance of the second-step estimator is determined by the sparsity $\widehat{m}$ of the first-step estimator, which controls the magnitude of the empirical errors. Otherwise, if the selected model fails to contain the true model, that is, $T \not\subseteq \widehat{T}$, the performance of the second-step estimator is determined by both the sparsity $\widehat{m}$ and the minimum between B_n and C_n. The quantity B_n measures the in-sample loss-of-fit induced by the first-step

estimator relative to the "true" parameter value β_0, and C_n measures the in-sample loss-of-fit induced by truncating the "true" parameter β_0 outside the selected model $\widehat{T}$.

The proof of Theorem 4 relies on the sparsity-based control of the empirical error provided by the following lemma.

Lemma 4 (Sparsity-based control of empirical error)**.** *Suppose Condition M holds. (1) For any $\varepsilon > 0$, there is a constant K_ε independent of n such that with probability at least $1 - \varepsilon$,*

$$|\widehat{Q}(\beta_0 + \delta) - \widehat{Q}(\beta_0) - \|\delta\|_{2,n}^2| \leq K_\varepsilon \sigma \sqrt{\frac{m \log p + (m+s)\log(e\mu(m))}{n}} \|\delta\|_{2,n} + 2c_s \|\delta\|_{2,n},$$

uniformly for all $\delta \in \mathbb{R}^p$ such that $\|\delta_{T^c}\|_0 \leq m$, and uniformly over $m \leq n$.

(2) Furthermore, with at least the same probability,

$$|\widehat{Q}(\beta_{0\widetilde{T}}) - \widehat{Q}(\beta_0) - \|\beta_{0\widetilde{T}^c}\|_{2,n}^2| \leq K_\varepsilon \sigma \sqrt{\frac{\log \binom{s}{k} + k \log(e\mu(0))}{n}} \|\beta_{0\widetilde{T}^c}\|_{2,n} + 2c_s \|\beta_{0\widetilde{T}^c}\|_{2,n},$$

uniformly for all $\widetilde{T} \subset T$ such that $|T \setminus \widetilde{T}| = k$, and uniformly over $k \leq s$.

The proof of the lemma in turn relies on the following maximal inequality, whose proof involves the use of Samorodnitsky-Talagrand's type inequality.

Lemma 5 (Maximal inequality for a collection of empirical processes)**.** *Let $\epsilon_i \sim N(0, \sigma^2)$ be independent for $i = 1, \ldots, n$, and for $m = 1, \ldots, n$ define*

$$e_n(m, \eta) := \sigma 2\sqrt{2} \left(\sqrt{\log \binom{p}{m}} + \sqrt{(m+s)\log\left(D\mu(m)\right)} + \sqrt{(m+s)\log(1/\eta)} \right)$$

for any $\eta \in (0,1)$ and some universal constant D. Then

$$\sup_{\|\delta_{T^c}\|_0 \leq m, \|\delta\|_{2,n} > 0} \left| \mathbb{G}_n \left(\frac{\epsilon_i x_i' \delta}{\|\delta\|_{2,n}} \right) \right| \leq e_n(m, \eta), \text{ for all } m \leq n,$$

with probability at least $1 - \eta e^{-s}/(1 - 1/e)$.

5. Performance of least squares after lasso-based model selection

In this section we specialize our results on post-model selection estimators to the case of lasso being the first-step estimator. The previous generic results allow us to use sparsity bounds and rate of convergence of lasso to derive the rate of convergence of post-model selection estimators in the parametric and nonparametric models.

5.1. Performance of ols post lasso

Here we show that the ols post lasso estimator enjoys good theoretical performance despite (generally) imperfect selection of the model by lasso.

Theorem 5 (Performance of ols post lasso)**.** *Suppose Conditions M , RE($\bar{c}$), and RSE($\widehat{m}$) hold where $\bar{c} = (c+1)/(c-1)$ and $\widehat{m} = |\widehat{T} \setminus T|$. If $\lambda \geq cn\|S\|_\infty$ occurs with probability at least $1-\alpha$, then for any $\varepsilon > 0$ there is a constant K_ε independent of n such that with probability at least $1-\alpha-\varepsilon$, for $\widetilde{f}_i = x_i'\widetilde{\beta}$ we have*

$$\|\widetilde{f} - f\|_{\mathbb{P}_n,2} \quad \leq K_\varepsilon \sigma \sqrt{\tfrac{\widehat{m}\log p + (\widehat{m}+s)\log(e\mu(\widehat{m}))}{n}} + 3c_s + 1\{T \not\subseteq \widehat{T}\}\sqrt{\tfrac{\lambda\sqrt{s}}{n\kappa(1)}\left(\tfrac{(1+c)\lambda\sqrt{s}}{cn\kappa(1)} + 2c_s\right)}.$$

In particular, under Condition V and the data-driven choice of λ specified in (2.12) with $\log(1/\alpha) \lesssim \log p$, $u/\ell \lesssim 1$, for any $\varepsilon > 0$ there is a constant $K'_{\varepsilon,\alpha}$ such that

$$\begin{aligned} \|\widetilde{f} - f\|_{\mathbb{P}_n,2} \quad &\leq 3c_s + K'_{\varepsilon,\alpha}\sigma\left[\sqrt{\tfrac{\widehat{m}\log(pe\mu(\widehat{m}))}{n}} + \sqrt{\tfrac{s\log(e\mu(\widehat{m}))}{n}}\right] + \\ &\quad + 1\{T \not\subseteq \widehat{T}\}\left[K'_{\varepsilon,\alpha}\sigma\sqrt{\tfrac{s\log p}{n}}\tfrac{1}{\kappa(1)} + c_s\right] \end{aligned} \tag{5.21}$$

with probability at least $1-\alpha-\varepsilon-\tau$.

This theorem provides a performance bound for ols post lasso as a function of 1) lasso's sparsity characterized by $\widehat{m}$, 2) lasso's rate of convergence, and 3) lasso's model selection ability. For common designs this bound implies that ols post lasso performs at least as well as lasso, but it can be strictly better in some cases, and has smaller regularization bias. We provide further theoretical comparisons in what follows, and computational examples supporting these comparisons appear in Supplementary Appendix. It is also worth repeating here that performance bounds in other norms of interest immediately follow by the triangle inequality and by definition of $\widetilde{\kappa}$ as discussed in Remark 3.1.

The following corollary summarizes the performance of ols post lasso under commonly used designs.

Corollary 2 (Asymptotic performance of ols post lasso)**.** *Under the conditions of Theorem 5, (2.5) and (3.17), as n grows, we have that*

$$\|\widetilde{f} - f\|_{\mathbb{P}_n,2} \quad \lesssim_P \quad \begin{cases} \sigma\sqrt{\frac{s\log p}{n}} + c_s, & \text{in general,} \\ \sigma\sqrt{\frac{o(s)\log p}{n}} + \sigma\sqrt{\frac{s}{n}} + c_s, & \text{if } \widehat{m} = o_P(s) \text{ and } T \subseteq \widehat{T} \text{ wp} \to 1, \\ \sigma\sqrt{s/n} + c_s, & \text{if } T = \widehat{T} \text{ wp} \to 1. \end{cases}$$

***Remark* 5.1** (Comparison of the performance of ols post lasso vs lasso)**.** We now compare the upper bounds on the rates of convergence of lasso and ols post lasso under conditions of the corollary. In general, the rates coincide. Notably, this occurs despite the fact that lasso may in general fail to correctly select the oracle model T as a subset,

that is $T \not\subseteq \widehat{T}$. However, if the oracle model has well-separated coefficients and condition and the approximation error does not dominated the estimation error – then ols post lasso rate improves upon lasso's rate. Specifically, this occurs if condition (2.5) holds and $\widehat{m} = o_P(s)$ and $T \subseteq \widehat{T}$ wp $\to 1$, as under conditions of Theorem 2 Part 1 or in the case of perfect model selection, when $T = \widehat{T}$ wp $\to 1$, as under conditions of [31]. Under such cases, we know from Corollary 1, that the rates found for lasso are sharp, and they cannot be faster than $\sigma\sqrt{s \log p/n}$. Thus the improvement in the rate of convergence of ols post lasso over lasso in such cases is strict.

5.2. Performance of ols post fit-lasso

In what follows we provide performance bounds for ols post fit-lasso $\widetilde{\beta}$ defined in equation (4.20) with threshold (2.9) for the case where the first-step estimator $\widehat{\beta}$ is lasso. We let $\widetilde{T}$ denote the model selected.

Theorem 6 (Performance of ols post fit-lasso)**.** *Suppose Conditions M , RE($\bar{c}$), and RSE($\widetilde{m}$) hold where $\bar{c} = (c+1)/(c-1)$ and $\widetilde{m} = |\widetilde{T} \setminus T|$. If $\lambda \geq cn\|S\|_\infty$ occurs with probability at least $1-\alpha$, then for any $\varepsilon > 0$ there is a constant K_ε independent of n such that with probability at least $1-\alpha-\varepsilon$, for $\widetilde{f}_i = x_i'\widetilde{\beta}$ we have*

$$\|\widetilde{f} - f\|_{\mathbb{P}_n,2} \leq K_\varepsilon \sigma \sqrt{\tfrac{\widetilde{m}\log p + (\widetilde{m}+s)\log(e\mu(\widetilde{m}))}{n}} + 3c_s + 1\{T \not\subseteq \widetilde{T}\}\sqrt{\tfrac{\lambda\sqrt{s}}{n\kappa(1)}\left(\tfrac{(1+c)\lambda\sqrt{s}}{cn\kappa(1)} + 2c_s\right)}.$$

Under Condition V and the data-driven choice of λ specified in (2.12) with $\log(1/\alpha) \lesssim \log p$, $u/\ell \lesssim 1$, for any $\varepsilon > 0$ there is a constant $K'_{\varepsilon,\alpha}$ such that

$$\begin{aligned} \|\widetilde{f} - f\|_{\mathbb{P}_n,2} \quad &\leq 3c_s + K'_{\varepsilon,\alpha}\sigma\left[\sqrt{\tfrac{\widetilde{m}\log(pe\mu(\widetilde{m}))}{n}} + \sqrt{\tfrac{s\log(e\mu(\widetilde{m}))}{n}}\right] + \\ &\quad + 1\{T \not\subseteq \widetilde{T}\}\left[K'_{\varepsilon,\alpha}\sigma\sqrt{\tfrac{s\log p}{n}}\tfrac{1}{\kappa(1)} + \ c_s\right] \end{aligned} \tag{5.22}$$

with probability at least $1-\alpha-\varepsilon-\tau$.

This theorem provides a performance bound for ols post fit- lasso as a function of 1) its sparsity characterized by $\widetilde{m}$, 2) lasso's rate of convergence, and 3) the model selection ability of the thresholding scheme. Generally, this bound is as good as the bound for ols post lasso, since the ols post fitness-thresholded lasso thresholds as much as possible subject to maintaining certain goodness-of-fit. It is also appealing that this estimator determines the thresholding level in a completely data-driven fashion. Moreover, by construction the estimated model is sparser than ols post lasso's model, which leads to an improved performance of ols post fitness-thresholded lasso over ols post lasso in some cases. We provide further theoretical comparisons below and computational examples in the Supplementary Appendix.

The following corollary summarizes the performance of ols post fit-lasso under commonly used designs.

Corollary 3 (Asymptotic performance of ols post fit-lasso)**.** *Under the conditions of Theorem 6, if conditions in (2.5) and (3.17) hold, as n grows, we have that the ols post fitness-thresholded lasso satisfies*

$$\|\widetilde{f}-f\|_{\mathbb{P}_n,2} \lesssim_P \begin{cases} \sigma\sqrt{\frac{s\log p}{n}}+c_s, & \text{in general,} \\ \sigma\sqrt{\frac{o(s)\log p}{n}}+\sigma\sqrt{\frac{s}{n}}+c_s, & \text{if } \widetilde{m}=o_P(s) \text{ and } T\subseteq\widetilde{T} \text{ wp} \to 1, \\ \sigma\sqrt{\frac{s}{n}}+c_s, & \text{if } T=\widetilde{T} \text{ wp} \to 1. \end{cases}$$

***Remark* 5.2** (Comparison of the performance of ols post fit-lasso vs lasso and ols post lasso)**.** Under the conditions of the corollary, the ols post fitness-thresholded lasso matches the near oracle rate of convergence of lasso and ols post lasso: $\sigma\sqrt{s\log p/n}+c_s$. If $\widetilde{m}=o_P(s)$ and $T\subseteq\widetilde{T}$ wp $\to 1$ and (2.5) hold, then ols post fit-lasso strictly improves upon lasso's rate. That is, if the oracle models has coefficients well-separated from zero and the approximation error is not dominant, the improvement is strict. An interesting question is whether ols post fit-lasso can outperform ols post lasso in terms of the rates. We cannot rank these estimators in terms of rates in general. However, this necessarily occurs when the lasso does not achieve the sufficient sparsity while the model selection works well, namely when $\widetilde{m}=o_P(\widehat{m})$ and $T\subseteq\widetilde{T}$ wp $\to 1$. Lastly, under conditions ensuring perfect model selection, namely condition of Theorem 2 holding for $t=t_\gamma$, ols post fit-lasso achieves the oracle performance, $\sigma\sqrt{s/n}+c_s$. □

5.3. Performance of the ols post thresholded lasso

Next we consider the traditional thresholding scheme which truncates to zero all components below a set threshold t. This is arguably the most used thresholding scheme in the literature. To state the result, recall that $\widehat{\beta}_{tj}=\widehat{\beta}_j 1\{|\widehat{\beta}_j|>t\}$, $\widetilde{m}:=|\widetilde{T}\setminus T|$, $m_t:=|\widehat{T}\setminus\widetilde{T}|$ and $\gamma_t:=\|\widehat{\beta}_t-\widehat{\beta}\|_{2,n}$ where $\widehat{\beta}$ is the lasso estimator.

Theorem 7 (Performance of ols post t-lasso)**.** *Suppose Conditions M , RE($\bar{c}$), and RSE($\widetilde{m}$) hold where $\bar{c}=(c+1)/(c-1)$ and $\widetilde{m}=|\widetilde{T}\setminus T|$. If $\lambda\geq cn\|S\|_\infty$ occurs with probability at least $1-\alpha$, then for any $\varepsilon>0$ there is a constant K_ε independent of n such that with probability at least $1-\alpha-\varepsilon$, for $\widetilde{f}_i=x_i'\widetilde{\beta}$ we have*

$$\begin{aligned}\|\widetilde{f}-f\|_{\mathbb{P}_n,2} \quad &\leq K_\varepsilon\sigma\sqrt{\tfrac{\widetilde{m}\log p+(\widetilde{m}+s)\log(e\mu(\widetilde{m}))}{n}}+3c_s+1\{T\not\subseteq\widetilde{T}\}\left(\gamma_t+\tfrac{1+c}{c}\tfrac{\lambda\sqrt{s}}{n\kappa(\bar{c})}+2c_s\right)+ \\ &+1\{T\not\subseteq\widetilde{T}\}\sqrt{\left[K_\varepsilon\sigma\sqrt{\tfrac{\widetilde{m}\log p+(\widetilde{m}+s)\log(e\mu(\widetilde{m}))}{n}}+2c_s\right]\left(\gamma_t+\tfrac{1+c}{c}\tfrac{\lambda\sqrt{s}}{n\kappa(\bar{c})}+2c_s\right)}\end{aligned}$$

where $\gamma_t\leq t\sqrt{\phi(m_t)m_t}$. Under Condition V and the data-driven choice of λ specified in (2.12) for $\log(1/\alpha)\lesssim\log p$, $u/\ell\lesssim 1$, for any $\varepsilon>0$ there is a constant $K'_{\varepsilon,\alpha}$ such that

with probability at least $1-\alpha-\varepsilon-\tau$

$$\begin{aligned}\|\widetilde{f}-f\|_{\mathbb{P}_n,2} \quad &\leq \quad 3c_s + K'_{\varepsilon,\alpha}\left[\sigma\sqrt{\frac{\widetilde{m}\log(pe\mu(\widetilde{m}))}{n}}+\sigma\sqrt{\frac{s\log(e\mu(\widetilde{m}))}{n}}\right]+ \\ &\quad +1\{T\not\subseteq\widetilde{T}\}\left[\gamma_t + K'_{\varepsilon,\alpha}\sigma\sqrt{\tfrac{s\log p}{n}}\tfrac{1}{\kappa(\bar{c})}+4c_s\right].\end{aligned}$$

This theorem provides a performance bound for ols post thresholded lasso as a function of 1) its sparsity characterized by $\widetilde{m}$ and improvements in sparsity over lasso characterized by m_t, 2) lasso's rate of convergence, 3) the thresholding level t and resulting goodness-of-fit loss γ_t relative to lasso induced by thresholding, and 4) model selection ability of the thresholding scheme. Generally, this bound may be worse than the bound for lasso, and this arises because the ols post thresholded lasso may potentially use too much thresholding resulting in large goodness-of-fit losses γ_t. We provide further theoretical comparisons below and computational examples in Section D of the Supplementary Appendix.

***Remark* 5.3** (Comparison of the performance of ols post thresholded lasso vs lasso and ols post lasso)**.** In this discussion we also assume conditions in (2.5) and (3.17) made in the previous formal comparisons. Under these conditions, ols post thresholded lasso obeys the bound:

$$\|\widetilde{f}-f\|_{\mathbb{P}_n,2} \lesssim_P \sigma\sqrt{\frac{\widetilde{m}\log p}{n}}+\sigma\sqrt{\frac{s}{n}}+c_s+1\{T\not\subseteq\widetilde{T}\}\left(\gamma_t\vee\sigma\sqrt{\frac{s\log p}{n}}\right). \tag{5.23}$$

In this case we have $\widetilde{m}\vee m_t \leq s+\widehat{m}\lesssim_P s$ by Theorem 3, and, in general, the rate above cannot improve upon lasso's rate of convergence given in Lemma 1.

As expected, the choice of t, which controls γ_t via the bound $\gamma_t \leq t\sqrt{\phi(m_t)m_t}$, can have a large impact on the performance bounds: If

$$t\lesssim\sigma\sqrt{\tfrac{\log p}{n}} \quad \text{then} \quad \|\widetilde{f}-f\|_{\mathbb{P}_n,2}\lesssim_P \sigma\sqrt{\tfrac{s\log p}{n}}+c_s. \tag{5.24}$$

The choice (5.24), suggested by [19] and Theorem 3, is theoretically sound, since it guarantees that ols post thresholded lasso achieves the near-oracle rate of lasso. Note that to implement the choice (5.24) in practice we suggest to set $t=\lambda/n$, since the separation from zero of the coefficients is unknown in practice. Note that using a much larger t can lead to inferior rates of convergence.

Furthermore, there is a special class of models – a neighborhood of parametric models with well-separated coefficients – for which improvements upon the rate of convergence of lasso is possible. Specifically, if $\widetilde{m}=o_P(s)$ and $T\subseteq\widetilde{T}$ wp $\to 1$ then ols post thresholded lasso strictly improves upon lasso's rate. Furthermore, if $\widetilde{m}=o_P(\widehat{m})$ and $T\subseteq\widetilde{T}$ wp $\to 1$, ols post thresholded lasso also outperforms ols post lasso:

$$\|\widetilde{f}-f\|_{\mathbb{P}_n,2}\lesssim_P \quad \sigma\sqrt{\frac{o(\widehat{m})\log p}{n}}+\sigma\sqrt{\frac{s}{n}}+c_s.$$

Lastly, under the conditions of Theorem 2 holding for the given t, ols post thresholded lasso achieves the oracle performance, $\|\widetilde{f} - f\|_{\mathbb{P}_n,2} \lesssim_P \ \sigma\sqrt{s/n} + c_s$. □

Appendix A: Proofs

A.1. Proofs for Section 3

Proof of Theorem 1. The bound in $\|\cdot\|_{2,n}$ norm follows by the same steps as in [4], so we omit the derivation to the supplement.

Under the data-driven choice (2.12) of λ and Condition V, we have $c'\widehat{\sigma} \geq c\sigma$ with probability at least $1-\tau$ since $c' \geq c/\ell$. Moreover, with the same probability we also have $\lambda \leq c'u\sigma\Lambda(1-\alpha|X)$. The result follows by invoking the $\|\cdot\|_{2,n}$ bound.

The bound in $\|\cdot\|_1$ is proven as follows. First, assume $\|\delta_{T^c}\|_1 \leq 2\bar{c}\|\delta_T\|_1$. In this case, by definition of the restricted eigenvalue, we have $\|\delta\|_1 \leq (1+2\bar{c})\|\delta_T\|_1 \leq (1+2\bar{c})\sqrt{s}\|\delta\|_{2,n}/\kappa(2\bar{c})$ and the result follows by applying the first bound to $\|\delta\|_{2,n}$ since $\bar{c} > 1$. On the other hand, consider the case that $\|\delta_{T^c}\|_1 > 2\bar{c}\|\delta_T\|_1$. The relation

$$-\frac{\lambda}{cn}(\|\delta_T\|_1 + \|\delta_{T^c}\|_1) + \|\delta\|_{2,n}^2 - 2c_s\|\delta\|_{2,n} \leq \frac{\lambda}{n}(\|\delta_T\|_1 - \|\delta_{T^c}\|_1),$$

which is established in (B.35) in the supplementary appendix, implies that $\|\delta\|_{2,n} \leq 2c_s$ and also

$$\|\delta_{T^c}\|_1 \leq \bar{c}\|\delta_T\|_1 + \frac{c}{c-1}\frac{n}{\lambda}\|\delta\|_{2,n}(2c_s - \|\delta\|_{2,n}) \leq \|\delta_T\|_1 + \frac{c}{c-1}\frac{n}{\lambda}c_s^2 \leq \frac{1}{2}\|\delta_{T^c}\|_1 + \frac{c}{c-1}\frac{n}{\lambda}c_s^2.$$

Thus,

$$\|\delta\|_1 \leq \left(1 + \frac{1}{2\bar{c}}\right)\|\delta_{T^c}\|_1 \leq \left(1 + \frac{1}{2\bar{c}}\right)\frac{2c}{c-1}\frac{n}{\lambda}c_s^2.$$

The result follows by taking the maximum of the bounds on each case and invoking the bound on $\|\delta\|_{2,n}$.

□

Proof of Theorem 2. Part (1) follows immediately from the assumptions.

To show part(2), let $\delta := \widehat{\beta} - \beta_0$, and proceed in two steps.

Step 1. By the first order optimality conditions of $\widehat{\beta}$ and the assumption on λ

$$\begin{aligned}\|\mathbb{E}_n[x_\bullet x_\bullet'\delta]\|_\infty &\leq \|\mathbb{E}_n[x_\bullet(y_\bullet - x_\bullet'\widehat{\beta})]\|_\infty + \|S/2\|_\infty + \|\mathbb{E}_n[x_\bullet r_\bullet]\|_\infty \\ &\leq \tfrac{\lambda}{2n} + \tfrac{\lambda}{2cn} + \min\left\{\tfrac{\sigma}{\sqrt{n}}, c_s\right\}\end{aligned}$$

since $\|\mathbb{E}_n[x_\bullet r_\bullet]\|_\infty \leq \min\left\{\frac{\sigma}{\sqrt{n}}, c_s\right\}$ by Step 2 below.

Next let e_j denote the jth-canonical direction. Thus, for every $j = 1, \ldots, p$ we have

$$\begin{aligned}|\mathbb{E}_n[e_j'x_\bullet x_\bullet'\delta] - \delta_j| = |\mathbb{E}_n[e_j'(x_\bullet x_\bullet' - I)\delta]| &\leq \max_{1\leq j,k\leq p}|(\mathbb{E}_n[x_\bullet x_\bullet' - I])_{jk}|\ \|\delta\|_1 \\ &\leq \|\delta\|_1/[Us].\end{aligned}$$

Then, combining the two bounds above and using the triangle inequality we have

$$\|\delta\|_\infty \leq \|\mathbb{E}_n[x_\bullet x_\bullet'\delta]\|_\infty + \|\mathbb{E}_n[x_\bullet x_\bullet'\delta] - \delta\|_\infty \leq \left(1+\frac{1}{c}\right)\frac{\lambda}{2n} + \min\left\{\frac{\sigma}{\sqrt{n}}, c_s\right\} + \frac{\|\delta\|_1}{Us}.$$

The result follows by Theorem 1 to bound $\|\delta\|_1$ and the arguments in [4] and [19] to show that the bound on the correlations imply that for any $C > 0$

$$\kappa(C) \geq \sqrt{1 - s(1+2C)\|\mathbb{E}_n[x_\bullet x_\bullet' - I]\|_\infty}$$

so that $\kappa(\bar{c}) \geq \sqrt{1-[(1+2\bar{c})/U]}$ and $\kappa(2\bar{c}) \geq \sqrt{1-[(1+4\bar{c})/U]}$ under this particular design.

Step 2. In this step we show that $\|\mathbb{E}_n[x_\bullet r_\bullet]\|_\infty \leq \min\left\{\frac{\sigma}{\sqrt{n}}, c_s\right\}$. First note that for every $j = 1, \ldots, p$, we have $|\mathbb{E}_n[x_{\bullet j} r_\bullet]| \leq \sqrt{\mathbb{E}_n[x_{\bullet j}^2]\mathbb{E}_n[r_\bullet^2]} = c_s$. Next, by definition of β_0 in (2.2), for $j \in T$ we have $\mathbb{E}_n[x_{\bullet j}(f_\bullet - x_\bullet'\beta_0)] = \mathbb{E}_n[x_{\bullet j} r_\bullet] = 0$ since β_0 is a minimizer over the support of β_0. For $j \in T^c$ we have that for any $t \in \mathbb{R}$

$$\mathbb{E}_n[(f_\bullet - x_\bullet'\beta_0)^2] + \sigma^2\frac{s}{n} \leq \mathbb{E}_n[(f_\bullet - x_\bullet'\beta_0 - t x_{\bullet j})^2] + \sigma^2\frac{s+1}{n}.$$

Therefore, for any $t \in \mathbb{R}$ we have

$$-\sigma^2/n \leq \mathbb{E}_n[(f_\bullet - x_\bullet'\beta_0 - t x_{\bullet j})^2] - \mathbb{E}_n[(f_\bullet - x_\bullet'\beta_0)^2] = -2t\mathbb{E}_n[x_{\bullet j}(f_\bullet - x_\bullet'\beta_0)] + t^2\mathbb{E}_n[x_{\bullet j}^2].$$

Taking the minimum over t in the right hand side at $t^* = \mathbb{E}_n[x_{\bullet j}(f_\bullet - x_\bullet'\beta_0)]$ we obtain $-\sigma^2/n \leq -(\mathbb{E}_n[x_{\bullet j}(f_\bullet - x_\bullet'\beta_0)])^2$ or equivalently, $|\mathbb{E}_n[x_{\bullet j}(f_\bullet - x_\bullet'\beta_0)]| \leq \sigma/\sqrt{n}$. □

Proof of Lemma 2. Let $\widehat{T} = \text{support}(\widehat{\beta})$, and $\widehat{m} = |\widehat{T}\backslash T|$. We have from the optimality conditions that $|2\mathbb{E}_n[x_{\bullet j}(y_\bullet - x_\bullet'\widehat{\beta})]| = \lambda/n$ for all $j \in \widehat{T}$. Therefore we have for $R = (r_1, \ldots, r_n)'$

$$\begin{aligned}
\sqrt{|\widehat{T}|}\lambda &\leq 2\|(X'(Y - X\widehat{\beta}))_{\widehat{T}}\| \\
&\leq 2\|(X'(Y - R - X\beta_0))_{\widehat{T}}\| + 2\|(X'(R + X\beta_0 - X\widehat{\beta}))_{\widehat{T}}\| \\
&\leq \sqrt{|\widehat{T}|}\cdot n\|S\|_\infty + 2n\sqrt{\phi(\widehat{m})}(\mathbb{E}_n[(x_\bullet'\widehat{\beta} - f_\bullet)^2])^{1/2},
\end{aligned}$$

where we used the definition of $\phi(\widehat{m})$ and the Holder inequality. Since $\lambda/c \geq n\|S\|_\infty$ we have

$$(1-1/c)\sqrt{|\widehat{T}|}\lambda \leq 2n\sqrt{\phi(\widehat{m})}(\mathbb{E}_n[(x_\bullet'\widehat{\beta} - f_\bullet)^2])^{1/2}. \tag{A.25}$$

Moreover, since $\widehat{m} \leq |\widehat{T}|$, and by Theorem 1 and Remark 3.1, $(\mathbb{E}_n[(x_\bullet'\widehat{\beta} - f_\bullet)^2])^{1/2} \leq \|\widehat{\beta} - \beta_0\|_{2,n} + c_s \leq \left(1+\frac{1}{c}\right)\frac{\lambda\sqrt{s}}{n\kappa(\bar{c})} + 3c_s$ we have

$$(1-1/c)\sqrt{\widehat{m}} \leq 2\sqrt{\phi(\widehat{m})}(1+1/c)\sqrt{s}/\kappa(\bar{c}) + 6\sqrt{\phi(\widehat{m})}\, nc_s/\lambda.$$

The result follows by noting that $(1-1/c) = 2/(\bar{c}+1)$ by definition of $\bar{c}$. □

Proof of Theorem 3. In the event $\lambda \geq c \cdot n\|S\|_\infty$, by Lemma 2 $\sqrt{\widehat{m}} \leq \sqrt{\phi(\widehat{m})} \cdot 2\bar{c}\sqrt{s}/\kappa(\bar{c}) + 3(\bar{c}+1)\sqrt{\phi(\widehat{m})} \cdot nc_s/\lambda$, which, by letting $L_n = \left(\frac{2\bar{c}}{\kappa(\bar{c})} + 3(\bar{c}+1)\frac{nc_s}{\lambda\sqrt{s}}\right)^2$, can be rewritten as

$$\widehat{m} \leq s \cdot \phi(\widehat{m}) L_n. \tag{A.26}$$

Note that $\widehat{m} \leq n$ by optimality conditions. Consider any $M \in \mathcal{M}$, and suppose $\widehat{m} > M$. Therefore by Lemma 3 on sublinearity of sparse eigenvalues

$$\widehat{m} \leq s \cdot \left\lceil \frac{\widehat{m}}{M} \right\rceil \phi(M) L_n.$$

Thus, since $\lceil k \rceil < 2k$ for any $k \geq 1$ we have $M < s \cdot 2\phi(M)L_n$ which violates the condition of $M \in \mathcal{M}$. Therefore, we must have $\widehat{m} \leq M$. In turn, applying (A.26) once more with $\widehat{m} \leq (M \wedge n)$ we obtain $\widehat{m} \leq s \cdot \phi(M \wedge n)L_n$. The result follows by minimizing the bound over $M \in \mathcal{M}$. □

A.2. Proofs for Section 4

Proof of Theorem 4. Let $\widetilde{\delta} := \widetilde{\beta} - \beta_0$. By definition of the second-step estimator, it follows that $\widehat{Q}(\widetilde{\beta}) \leq \widehat{Q}(\widehat{\beta})$ and $\widehat{Q}(\widetilde{\beta}) \leq \widehat{Q}(\beta_{0\widehat{T}})$. Thus,

$$\widehat{Q}(\widetilde{\beta}) - \widehat{Q}(\beta_0) \leq \left(\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0)\right) \wedge \left(\widehat{Q}(\beta_{0\widehat{T}}) - \widehat{Q}(\beta_0)\right) \leq B_n \wedge C_n.$$

By Lemma 4 part (1), for any $\varepsilon > 0$ there exists a constant K_ε such that with probability at least $1 - \varepsilon$: $|\widehat{Q}(\widetilde{\beta}) - \widehat{Q}(\beta_0) - \|\widetilde{\delta}\|_{2,n}^2| \leq A_{\varepsilon,n}\|\widetilde{\delta}\|_{2,n} + 2c_s\|\widetilde{\delta}\|_{2,n}$ where

$$A_{\varepsilon,n} := K_\varepsilon \sigma \sqrt{(\widehat{m}\log p + (\widehat{m}+s)\log(e\mu(\widehat{m})))/n}.$$

Combining these relations we obtain the inequality $\|\widetilde{\delta}\|_{2,n}^2 - A_{\varepsilon,n}\|\widetilde{\delta}\|_{2,n} - 2c_s\|\widetilde{\delta}\|_{2,n} \leq B_n \wedge C_n$, solving which we obtain the stated inequality: $\|\widetilde{\delta}\|_{2,n} \leq A_{\varepsilon,n} + 2c_s + \sqrt{(B_n)_+ \wedge (C_n)_+}$. Finally, the bound on B_n follows from Lemma 4 result (1). The bound on C_n follows from Lemma 4 result (2). □

Proof of Lemma 4. Part (1) follows from the relation

$$|\widehat{Q}(\beta_0 + \delta) - \widehat{Q}(\beta_0) - \|\delta\|_{2,n}^2| = |2\mathbb{E}_n[\epsilon_\bullet x_\bullet'\delta] + 2\mathbb{E}_n[r_\bullet x_\bullet'\delta]|,$$

then bounding $|2\mathbb{E}_n[r_\bullet x_\bullet'\delta]|$ by $2c_s\|\delta\|_{2,n}$ using the Cauchy-Schwarz inequality, applying Lemma 5 on sparse control of noise to $|2\mathbb{E}_n[\epsilon_\bullet x_\bullet'\delta]|$ where we bound $\binom{p}{m}$ by p^m and set $K_\varepsilon = 6\sqrt{2}\log^{1/2}\max\{e, D, 1/(e^s\varepsilon[1-1/e])\}$. Part (2) also follows from Lemma 5 but applying it with $s = 0$, $p = s$ (since only the components in T are modified), $m = k$, and noting that we can take $\mu(m)$ with $m = 0$. □

Proof of Lemma 5. We divide the proof into steps.

Step 0. Note that we can restrict the supremum over $\|\delta\| = 1$ since the function is homogenous of degree zero.

Step 1. For each non-negative integer $m \leq n$, and each set $\widetilde{T} \subset \{1, \ldots, p\}$, with $|\widetilde{T} \setminus T| \leq m$, define the class of functions

$$\mathcal{G}_{\widetilde{T}} = \{\epsilon_i x_i'\delta/\|\delta\|_{2,n} : \text{support}(\delta) \subseteq \widetilde{T}, \|\delta\| = 1\}. \tag{A.27}$$

Also define $\mathcal{F}_m = \{\mathcal{G}_{\widetilde{T}} : \widetilde{T} \subset \{1, \ldots, p\} : |\widetilde{T} \setminus T| \leq m\}$. It follows that

$$P\left(\sup_{f\in\mathcal{F}_m} |\mathbb{G}_n(f)| \geq e_n(m,\eta)\right) \leq \binom{p}{m} \max_{|\widetilde{T}\setminus T|\leq m} P\left(\sup_{f\in\mathcal{G}_{\widetilde{T}}} |\mathbb{G}_n(f)| \geq e_n(m,\eta)\right). \tag{A.28}$$

We apply Samorodnitsky-Talagrand's inequality (Proposition A.2.7 in van der Vaart and Wellner [30]) to bound the right hand side of (A.28). Let

$$\rho(f,g) := \sqrt{E[\mathbb{G}_n(f) - \mathbb{G}_n(g)]^2} = \sqrt{E\mathbb{E}_n[(f-g)^2]}$$

for $f, g \in \mathcal{G}_{\widetilde{T}}$; by Step 2 below, the covering number of $\mathcal{G}_{\widetilde{T}}$ with respect to ρ obeys

$$N(\varepsilon, \mathcal{G}_{\widetilde{T}}, \rho) \leq (6\sigma\mu(m)/\varepsilon)^{m+s}, \text{ for each } 0 < \varepsilon \leq \sigma, \tag{A.29}$$

and $\sigma^2(\mathcal{G}_{\widetilde{T}}) := \max_{f\in\mathcal{G}_{\widetilde{T}}} E[\mathbb{G}_n(f)]^2 = \sigma^2$. Then, by Samorodnitsky-Talagrand's inequality

$$P\left(\sup_{f\in\mathcal{G}_{\widetilde{T}}} |\mathbb{G}_n(f)| \geq e_n(m,\eta)\right) \leq \left(\frac{D\sigma\mu(m)e_n(m,\eta)}{\sqrt{m+s}\sigma^2}\right)^{m+s} \bar{\Phi}(e_n(m,\eta)/\sigma) \tag{A.30}$$

for some universal constant $D \geq 1$, where $\bar{\Phi} = 1 - \Phi$ and Φ is the cumulative probability distribution function for a standardized Gaussian random variable. For $e_n(m,\eta)$ defined in the statement of the theorem, it follows that $P\left(\sup_{f\in\mathcal{G}_{\widetilde{T}}} |\mathbb{G}_n(f)| \geq e_n(m,\eta)\right) \leq \eta e^{-m-s}/\binom{p}{m}$ by simple substitution into (A.30). Then,

$$\begin{aligned} P\left(\sup_{f\in\mathcal{F}_m} |\mathbb{G}_n(f)| > e_n(m,\eta), \exists m \leq n\right) &\leq \sum_{m=0}^{n} P\left(\sup_{f\in\mathcal{F}_m} |\mathbb{G}_n(f)| > e_n(m,\eta)\right) \\ &\leq \sum_{m=0}^{n} \eta e^{-m-s} \leq \eta e^{-s}/(1-1/e), \end{aligned}$$

which proves the claim.

Step 2. This step establishes (A.29). For $t \in \mathbb{R}^p$ and $\widetilde{t} \in \mathbb{R}^p$, consider any two functions

$$\epsilon_i \frac{(x_i't)}{\|t\|_{2,n}} \text{ and } \epsilon_i \frac{(x_i'\widetilde{t})}{\|\widetilde{t}\|_{2,n}} \text{ in } \mathcal{G}_{\widetilde{T}}, \text{ for a given } \widetilde{T} \subset \{1, \ldots, p\} : |\widetilde{T} \setminus T| \leq m.$$

We have that

$$\sqrt{E\mathbb{E}_n\left[\epsilon_\bullet^2\left(\frac{(x_\bullet' t)}{\|t\|_{2,n}}-\frac{(x_\bullet'\widetilde{t})}{\|\widetilde{t}\|_{2,n}}\right)^2\right]}\leq\sqrt{E\mathbb{E}_n\left[\epsilon_\bullet^2\frac{(x_\bullet'(t-\widetilde{t}))^2}{\|t\|_{2,n}^2}\right]}+\sqrt{E\mathbb{E}_n\left[\epsilon_\bullet^2\left(\frac{(x_\bullet'\widetilde{t})}{\|t\|_{2,n}}-\frac{(x_\bullet'\widetilde{t})}{\|\widetilde{t}\|_{2,n}}\right)^2\right]}.$$

By definition of $\mathcal{G}_{\widetilde{T}}$ in (A.27), support$(t)\subseteq\widetilde{T}$ and support$(\widetilde{t})\subseteq\widetilde{T}$, so that support$(t-\widetilde{t})\subseteq\widetilde{T}$, $|\widetilde{T}\setminus T|\leq m$, and $\|t\|=1$ by (A.27). Hence by definition RSE(m),

$$E\mathbb{E}_n\left[\epsilon_\bullet^2\frac{(x_\bullet'(t-\widetilde{t}))^2}{\|t\|_{2,n}^2}\right]\leq\sigma^2\phi(m)\|t-\widetilde{t}\|^2/\widetilde{\kappa}(m)^2,\text{ and}$$

$$E\mathbb{E}_n\left[\epsilon_\bullet^2\left(\frac{(x_\bullet'\widetilde{t})}{\|t\|_{2,n}}-\frac{(x_\bullet'\widetilde{t})}{\|\widetilde{t}\|_{2,n}}\right)^2\right]=E\mathbb{E}_n\left[\epsilon_\bullet^2\frac{(x_\bullet'\widetilde{t})^2}{\|\widetilde{t}\|_{2,n}^2}\left(\frac{\|\widetilde{t}\|_{2,n}-\|t\|_{2,n}}{\|t\|_{2,n}}\right)^2\right]$$

$$=\sigma^2\left(\frac{\|\widetilde{t}\|_{2,n}-\|t\|_{2,n}}{\|t\|_{2,n}}\right)^2\leq\sigma^2\|\widetilde{t}-t\|_{2,n}^2/\|t\|_{2,n}^2\leq\sigma^2\phi(m)\|\widetilde{t}-t\|^2/\widetilde{\kappa}(m)^2,$$

so that

$$\sqrt{E\mathbb{E}_n\left[\epsilon_\bullet^2\left(\frac{(x_\bullet' t)}{\|t\|_{2,n}}-\frac{(x_\bullet'\widetilde{t})}{\|\widetilde{t}\|_{2,n}}\right)^2\right]}\leq 2\sigma\|t-\widetilde{t}\|\sqrt{\phi(m)}/\widetilde{\kappa}(m)=2\sigma\mu(m)\|t-\widetilde{t}\|.$$

Then the bound (A.29) follows from the bound in [30] page 94, $N(\varepsilon,\mathcal{G}_{\widetilde{T}},\rho)\leq N(\varepsilon/R,B(0,1),\|\cdot\|)\leq(3R/\varepsilon)^{m+s}$ with $R=2\sigma\mu(m)$ for any $\varepsilon\leq\sigma$. □

A.3. Proofs for Section 5

Proof of Theorem 5. First note that if $T\subseteq\widehat{T}$ we have $C_n=0$ so that $B_n\wedge C_n\leq 1\{T\not\subseteq\widehat{T}\}B_n$.

Next we bound B_n. Note that by the optimality of $\widehat{\beta}$ in the lasso problem, and letting $\widehat{\delta}=\widehat{\beta}-\beta_0$,

$$B_n:=\widehat{Q}(\widehat{\beta})-\widehat{Q}(\beta_0)\quad\leq\tfrac{\lambda}{n}(\|\beta_0\|_1-\|\widehat{\beta}\|_1)\leq\tfrac{\lambda}{n}(\|\widehat{\delta}_T\|_1-\|\widehat{\delta}_{T^c}\|_1).\tag{A.31}$$

If $\|\widehat{\delta}_{T^c}\|_1>\|\widehat{\delta}_T\|_1$, we have $\widehat{Q}(\widehat{\beta})-\widehat{Q}(\beta_0)\leq 0$. Otherwise, if $\|\widehat{\delta}_{T^c}\|_1\leq\|\widehat{\delta}_T\|_1$, by RE(1) we have

$$B_n:=\widehat{Q}(\widehat{\beta})-\widehat{Q}(\beta_0)\leq\tfrac{\lambda}{n}\|\widehat{\delta}_T\|_1\leq\tfrac{\lambda}{n}\tfrac{\sqrt{s}\|\widehat{\delta}\|_{2,n}}{\kappa(1)}.\tag{A.32}$$

The result follows by applying Theorem 1 to bound $\|\widehat{\delta}\|_{2,n}$, under the condition that RE(1) holds, and Theorem 4.

The second claim follows from the first by using $\lambda\lesssim\sqrt{n\log p}$ under Condition V, the specified conditions on the penalty level. The final bound follows by applying the relation that for any nonnegative numbers a,b, we have $\sqrt{ab}\leq(a+b)/2$. □

Acknowledgements

We thank Don Andrews, Whitney Newey, and Alexandre Tsybakov as well as participants of the Cowles Foundation Lecture at the 2009 Summer Econometric Society meeting and the joint Harvard-MIT seminar for useful comments. We thank Denis Chetverikov, Brigham Fradsen, Joonhwan Lee, two refereers and the associate editor for numerous suggestions that helped improve the paper. We thank Kengo Kato for pointing to use the usefulness of [25] for bounding empirical sparse eigenvalues. We gratefully acknowledge the financial support from the National Science Foundation.

References

[1] D. Achlioptas (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins, Journal of Computer and System Sciences, 66, 671-687.

[2] A. Belloni and V. Chernozhukov (2011). ℓ_1-penalized quantile regression for high dimensional sparse models, Ann. Statist. Volume 39, Number 1, 82-130.

[3] A. Belloni and V. Chernozhukov (2011). Supplementary material: Supplement to "ℓ_1-penalized quantile regression in high-dimensional sparse models." Digital Object Identifier: doi:10.1214/10-AOS827SUPP.

[4] P. J. Bickel, Y. Ritov and A. B. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector, Ann. Statist. Volume 37, Number 4 (2009), 1705-1732.

[5] F. Bunea (2008). Consistent selection via the Lasso for high-dimensional approximating models. In: IMS Lecture Notes Monograph Series, vol.123, 123-137.

[6] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp (2006). Aggregation and sparsity via ℓ_1-penalized least squares, in Proceedings of 19th Annual Conference on Learning Theory (COLT 2006) (G. Lugosi and H. U. Simon, eds.). Lecture Notes in Artificial Intelligence 4005 379-391. Springer, Berlin.

[7] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp (2007). Aggregation for Gaussian regression, The Annals of Statistics, Vol. 35, No. 4, 1674-1697.

[8] F. Bunea, A. Tsybakov, and M. H. Wegkamp (2007). Sparsity oracle inequalities for the Lasso, Electronic Journal of Statistics, Vol. 1, 169-194.

[9] E. Candès and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. Ann. Statist. Volume 35, Number 6, 2313–2351.

[10] D. L. Donoho (2006). For most large underdetermined systems of linear equations, the minimal l1-norm solution is also the sparsest solution. Comm. Pure Appl. Math. 59 797–829.

[11] D. L. Donoho, M. Elad and V. N. Temlyakov (2006). Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Trans. Inform. 52, 6–18.

[12] D. L. Donoho and J. M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage, Biometrika 1994 81(3):425-455.

[13] R. Dudley (2000). Uniform cental limit theorems, Cambridge Studies in advanced mathematics.

[14] S. Efromovich (1999). Nonparametric curve estimation: methods, theory and applications, Springer.

[15] J. Fan and J. Lv. (2008). Sure independence screening for ultra-high dimensional feature space, Journal of the Royal Statistical Society. Series B, vol. 70 (5), pp. 849–911.

[16] O. Guédon and M. Rudelson (2007). L_p-moments of random vectors via majorizing measures, Advances in Mathematics, Volume 208, Issue 2, Pages 798-823.
[17] V. Koltchinskii (2009). Sparsity in penalized empirical risk minimization, Ann. Inst. H. Poincar Probab. Statist. Volume 45, Number 1, 7-57.
[18] M. Ledoux and M. Talagrand (1991). Probability in Banach Spaces (Isoperimetry and processes). Ergebnisse der Mathematik und ihrer Grenzgebiete, Springer-Verlag.
[19] K. Lounici (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators, Electron. J. Statist. Volume 2, 90-102.
[20] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer (2009). Taking advantage of sparsity in multi-task learning, in Proceedings of COLT-2009.
[21] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer (2010). Oracle inequalities and optimal inference under group sparsity, accepted at the Annals of Statistics.
[22] N. Meinshausen and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data. Annals of Statistics, vol. 37(1), 2246–2270.
[23] P. Rigollet and A. B. Tsybakov (2011). Exponential Screening and optimal rates of sparse estimation, Annals of Statistics, Volume 39, Number 2, 731-771.
[24] M. Rosenbaum and A. B. Tsybakov (2010). Sparse recovery under matrix uncertainty, Ann. Statist. Volume 38, Number 5, 2620-2651.
[25] Mark Rudelson and Roman Vershynin (2008). On sparse reconstruction from Fourier and Gaussian measurements, Communications on Pure and Applied Mathematics Volume 61, Issue 8, pages 10251045.
[26] R. Tibshirani (1996). Regression shrinkage and selection via the Lasso. J. Roy. Statist. Soc. Ser. B 58 267-288.
[27] A. Tsybakov (2008). Introduction to nonparametric estimation, Springer.
[28] S. A. van de Geer (2008). High-dimensional generalized linear models and the lasso, Annals of Statistics, Vol. 36, No. 2, 614–645.
[29] S. van de Geer (2000). Empirical Processes in M-Estimation, Cambridge University Press.
[30] A. W. van der Vaart and J. A. Wellner (1996). Weak convergence and empirical processes, Springer Series in Statistics.
[31] M. Wainwright (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (Lasso) , IEEE Transactions on Information Theory, 55:2183–2202, May.
[32] L. Wasserman (2005). All of Nonparametric Statistics, Springer.
[33] C.-H. Zhang and J. Huang (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. Ann. Statist. Volume 36, Number 4, 1567–1594.
[34] P. Zhao and B. Yu (2006). On model selection consistency of Lasso. J. Machine Learning Research, 7 (nov), 2541-2567.

Supplementary Appendix

Appendix A: Additional Results and Comments

A.1. On the Oracle Problem

Let us now briefly explain what is behind problem (2.2). Under some mild assumptions, this problem directly arises as the (infeasible) oracle risk minimization problem. Indeed, consider a least squares estimator $\widehat{\beta}_{\widetilde{T}}$, which is obtained by using a model $\widetilde{T}$, i.e. by regressing y_i on regressors $x_i[\widetilde{T}]$, where $x_i[\widetilde{T}] = \{x_{ij}, j \in \widetilde{T}\}$. This estimator takes value $\widehat{\beta}_{\widetilde{T}} = \mathbb{E}_n[x_\bullet[\widetilde{T}]x_\bullet[\widetilde{T}]']^- \mathbb{E}_n[x_\bullet[\widetilde{T}]y_\bullet]$. The expected risk of this estimator $\mathbb{E}_n \mathrm{E}[f_\bullet - x_\bullet'\widehat{\beta}_{\widetilde{T}}]^2$ is equal to

$$\min_{\beta \in \mathbb{R}^{|\widetilde{T}|}} \mathbb{E}_n[(f_\bullet - x_\bullet[\widetilde{T}]'\beta)^2] + \sigma^2 \frac{k}{n},$$

where $k = \mathrm{rank}(\mathbb{E}_n[x_\bullet[\widetilde{T}]x_\bullet[\widetilde{T}]'])$. The oracle knows the risk of each of the models $\widetilde{T}$ and can minimize this risk

$$\min_{\widetilde{T}} \min_{\beta \in \mathbb{R}^{|\widetilde{T}|}} \mathbb{E}_n[(f_\bullet - x_\bullet[\widetilde{T}]'\beta)^2] + \sigma^2 \frac{k}{n},$$

by choosing the best model or the oracle model T. This problem is in fact equivalent to (2.2), provided that $\mathrm{rank}\,(\mathbb{E}_n[x_\bullet[T]x_\bullet[T]']) = \|\beta_0\|_0$, i.e. full rank. Thus, in this case any value β_0 solving (2.2) is the expected value of the oracle least squares estimator $\widehat{\beta}_T = \mathbb{E}_n[x_\bullet[T]x_\bullet[T]']^{-1}\mathbb{E}_n[x_\bullet[T]y_\bullet]$, i.e. $\beta_0 = \mathbb{E}_n[x_\bullet[T]x_\bullet[T]']^{-1}\mathbb{E}_n[x_\bullet[T]f_\bullet]$. This value is our target or "true" parameter value and the oracle model T is the target or "true" model. Note that when $c_s = 0$ we have that $f_i = x_i'\beta_0$, which gives us the special parametric case.

A.2. Estimation of σ – finite-sample analysis

Consider the following algorithm to estimate σ.

Algorithm (Estimation of σ using lasso iterations) Set $\widehat{\sigma}_0 = \sqrt{\mathrm{Var}_n[y_\bullet]}$.

(1) Compute the lasso estimator $\widehat{\beta}$ based on $\lambda = c'\widehat{\sigma}_0\Lambda(1-\alpha|X)$;

(2) Set $\widehat{\sigma} = \sqrt{\widehat{Q}(\widehat{\beta})}$.

The following lemmas establish the finite sample bounds on ℓ, u, and τ that appear in Condition V associated with using $\widehat{\sigma}_0$ and $\sqrt{\widehat{Q}(\widehat{\beta})}$ as an estimator for σ.

Lemma 6. *Assume that for some $k > 4$ we have $\mathrm{E}[|y_i|^k] < C$ uniformly in n. There is a constant K such that for any positive numbers v and r we have with probability at least $1 - \frac{KC}{n^{k/4}v^{k/2}} - \frac{KC}{n^{k/2}r^k}$*

$$|\widehat{\sigma}_0^2 - \sigma_0^2| \leq v + r(r + 2C^{1/k})$$

where $\sigma_0 = \sqrt{Var[y_\bullet]}$.

Proof. We have that $\widehat{\sigma}_0^2 - \sigma_0^2 = \mathbb{E}_n[y_\bullet^2 - \mathrm{E}[y_\bullet^2]] - (\mathbb{E}_n[y_\bullet])^2 + (\mathrm{E}\mathbb{E}_n[y_\bullet])^2$.

Next note that by Markov inequality and Rosenthal inequality, for some constant $A(r/2)$ we have

$$\begin{aligned} P(|\mathbb{E}_n[y_\bullet^2 - \mathrm{E}[y_\bullet^2]]| > v) &\leq \frac{\mathrm{E}\left|\sum_{i=1}^n y_i^2 - \mathrm{E}[y_i^2]\right|^{k/2}}{n^{k/2}v^{k/2}} \leq \frac{A(r/2)\max\{\sum_{i=1}^n \mathrm{E}|y_i|^k,\ (\sum_{i=1}^n \mathrm{E}|y_i|^4)^{k/4}\}}{n^{k/2}v^{k/2}} \\ &\leq \frac{A(k/2)\max\{nC,\ Cn^{k/4}\}}{n^{k/2}v^{k/2}} \leq \frac{A(k/2)C}{n^{k/4}v^{k/2}}. \end{aligned}$$

Next note that $(\mathbb{E}_n[y_\bullet])^2 - (\mathrm{E}\mathbb{E}_n[y_\bullet])^2 = (\mathbb{E}_n[y_\bullet + \mathrm{E}[y_\bullet]])(\mathbb{E}_n[y_\bullet - \mathrm{E}[y_\bullet]])$. Similarly, by Markov inequality and Rosenthal inequality, for some constant $A(r)$, we have $P(|\mathbb{E}_n[y_\bullet - \mathrm{E}[y_\bullet]]| > r) \leq \frac{A(k)C}{n^{k/2}r^k}$. Thus,

$$P(|(\mathbb{E}_n[y_\bullet])^2 - (\mathrm{E}\mathbb{E}_n[y_\bullet])^2| > r(r + 2C^{1/k})) \leq \frac{A(k)C}{n^{k/2}r^k}.$$

The result follows by choosing $K \geq A(k) \vee A(k/2)$. □

Lemma 7. *Suppose that Condition M holds and that $\lambda \geq cn\|S\|_\infty$ with probability at least $1 - \alpha$. Then, for any $\varepsilon, \gamma \in (0, 1)$ we have*

$$\frac{\widehat{Q}(\widehat{\beta})}{\sigma^2} \leq 1 + \frac{2\lambda^2 s}{\sigma^2 n^2 \kappa(1)^2} + \frac{2c_s\lambda\sqrt{s}}{\sigma^2 n\kappa(1)} + \frac{c_s^2}{\sigma^2} + \frac{2c_s}{\sigma\sqrt{n}}\sqrt{2\log 1/\gamma} + \varepsilon,$$

$$\frac{\widehat{Q}(\widehat{\beta})}{\sigma^2} \geq 1 - \frac{c_s^2}{\sigma^2} - \frac{(2 + 4\bar{c})}{c\sigma^2}\left[\frac{\lambda^2 s}{n^2\kappa(2\bar{c})\kappa(\bar{c})} + \frac{c_s\lambda\sqrt{s}}{n\kappa(2\bar{c})} + c_s^2\right] - \frac{2c_s}{\sigma\sqrt{n}}\sqrt{2\log 1/\gamma} - \varepsilon$$

with probability $1 - \alpha - 2\exp(-n\varepsilon^2/12) - \gamma$.

Proof. We start by

$$\frac{\widehat{Q}(\widehat{\beta})}{\sigma^2} = \frac{\widehat{Q}(\widehat{\beta}) - \mathbb{E}_n[\epsilon_\bullet^2]}{\sigma^2} + \frac{\mathbb{E}_n[\epsilon_\bullet^2]}{\sigma^2}.$$

To control the second term we invoke tail-bounds for the chi-square distribution, see for instance Lemma 4.1 in [1]. Indeed, for any $\varepsilon > 0$ we have

$$P(\mathbb{E}_n[\epsilon_\bullet^2] \leq \sigma^2(1 - \varepsilon)) \leq \exp\left(-\frac{n\varepsilon^2}{2} \cdot \left(\frac{1}{2} - \frac{\varepsilon}{3}\right)\right) \quad \text{and}$$

$$P(\mathbb{E}_n[\epsilon_\bullet^2] \geq \sigma^2(1 + \varepsilon)) \leq \exp\left(-\frac{n\varepsilon^2}{2} \cdot \left(\frac{1}{2} - \frac{\varepsilon}{3}\right)\right).$$

To bound the first term, we have

$$\widehat{Q}(\widehat{\beta}) - \mathbb{E}_n[\epsilon_\bullet^2] = \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) + \mathbb{E}_n[r_\bullet^2] + 2\mathbb{E}_n[\epsilon_\bullet r_\bullet].$$

where $\mathbb{E}_n[r_\bullet^2] = c_s^2$ and since $2\mathbb{E}_n[\epsilon_\bullet r_\bullet] \sim N(0, 4\sigma^2 c_s^2/n)$ it follows that $2\mathbb{E}_n[\epsilon_\bullet r_\bullet] \leq \sigma(2c_s/\sqrt{n})(\sqrt{2\log 1/\gamma})$ with probability $1-\gamma$.

Finally, we bound the term $\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0)$ from above and below. To bound above, we use the optimality of $\widehat{\beta}$, so that $\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) \leq \frac{\lambda}{n}(\|\delta_T\|_1 - \|\delta_{T^c}\|_1)$. If $\|\delta_T\|_1 \leq \|\delta_{T^c}\|_1$ we have $\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) \leq 0$. Thus we can assume $\|\delta_{T^c}\|_1 \leq \|\delta_T\|_1$. Then, with probability at least $1-\alpha$ we have $\lambda \geq cn\|S\|_\infty$ and by the definition of RE(1) and Theorem 1 we have

$$\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) \leq \frac{\lambda\sqrt{s}}{n\kappa(1)}\|\delta\|_{2,n} \leq \left(1 + \frac{1}{c}\right)\frac{\lambda^2 s}{n^2\kappa(1)^2} + \frac{2c_s\lambda\sqrt{s}}{n\kappa(1)}.$$

To bound from below note that by convexity

$$\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) \geq \|\delta\|_{2,n}^2 - \|S\|_\infty\|\delta\|_1 - 2c_s\|\delta\|_{2,n}$$

It follows that $\|\delta\|_{2,n}^2 - 2c_s\|\delta\|_{2,n} \geq -c_s^2$. Next, we invoke the ℓ_1-norm bound in Theorem 1 so that

$$\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) \geq -c_s^2 - \left[\frac{\lambda(2+4\bar{c})}{cn}\frac{\sqrt{s}}{\kappa(2\bar{c})}\left(\frac{\lambda\sqrt{s}}{n\kappa(\bar{c})} + c_s\right)\right] \vee \left[\frac{(2+4\bar{c})c_s^2}{c}\right].$$

The result follows by simplifying the expression above. □

The result below verifies Condition V relying on Lemmas 6 and 7.

Theorem 8. *Assume that Condition M hold and for some $k > 4$ we have $\mathrm{E}[|y_i|^k] < C$ uniformly in n. Then, for any $\varepsilon, \gamma \in (0,1)$ we have that Condition V holds with*

$$\tau = 1 - \alpha - \frac{K2^{k/2}}{n^{k/4}}(C/\sigma_0^k)(1-c/c')^{-k/2} - \frac{K6^k}{n^{k/2}}(C^2/\sigma_0^{2k})\cdot(1-c/c')^{-k} - 2\exp(-n\varepsilon^2/12) - \gamma,$$

$$u \leq 1 + \frac{2s\left(3c'\sigma_0\Lambda(1-\alpha|X)\right)^2}{\sigma^2 n^2\kappa(1)^2} + \left(3c'\sigma_0\Lambda(1-\alpha|X)\right)\frac{2c_s\sqrt{s}}{\sigma^2 n\kappa(1)} + \frac{c_s^2}{\sigma^2} + \frac{2c_s}{\sigma\sqrt{n}}\sqrt{2\log 1/\gamma} + \varepsilon,$$

$$\ell \geq 1 - \frac{c_s^2}{\sigma^2} - \frac{(2+4\bar{c})}{c\sigma^2}\left[\frac{\left(3c'\sigma_0\Lambda(1-\alpha|X)\right)^2 s}{n^2\kappa(2\bar{c})\kappa(\bar{c})} + \frac{c_s\left(3c'\sigma_0\Lambda(1-\alpha|X)\right)\sqrt{s}}{n\kappa(2\bar{c})} + c_s^2\right] - \frac{2c_s}{\sigma\sqrt{n}}\sqrt{2\log 1/\gamma} - \varepsilon.$$

Proof. By Lemma 6 with $v = \sigma_0^2\cdot(1-c/c')/2$ and $r = (\sigma_0^2/C^{1/k})\cdot(1-c/c')/6$, with probability at least $1 - \frac{K2^{k/2}}{n^{k/4}}(C/\sigma_0^k)(1-c/c')^{-k/2} - \frac{K6^k}{n^{k/2}}(C^2/\sigma_0^{2k})\cdot(1-c/c')^{-k}$, we have $|\widehat{\sigma}_0^2 - \sigma_0^2| \leq \sigma_0^2(1-c/c')$ so that

$$c/c' \leq \frac{\widehat{\sigma}_0^2}{\sigma_0^2} \leq 2 + c/c' \leq 3.$$

Since $\sigma \leq \sigma_0$, for $\lambda = c'\cdot\widehat{\sigma}_0\cdot\Lambda(1-\alpha|X)$, we have $\lambda \geq cn\|S\|_\infty$. with probability at least $1 - \alpha - \frac{K2^{k/2}}{n^{k/4}}(C/\sigma_0^k)(1-c/c')^{-k/2} - \frac{K6^k}{n^{k/2}}(C^2/\sigma_0^{2k})\cdot(1-c/c')^{-k}$.

Thus, by Lemma 7, we have that Condition V holds with the stated bounds. □

Under the typical design conditions

$$\kappa(2\bar{c}) \gtrsim 1,\ \alpha = o(1),\ \text{and}\ s\log(p/\alpha) = o(n), \tag{A.33}$$

the bounds stated in Theorem 8 establish that $\ell \to 1$, $u \to 1$ and $\tau \to 0$ asymptotically. In finite samples, the following lemma ensures that $\ell > 0$.

Lemma 8. *We have that $\widehat{\sigma}_0 > 0$ and $\widehat{\sigma} = \sqrt{\widehat{Q}(\widehat{\beta})} > 0$ with probability 1.*

Proof. First note that $\widehat{\sigma}_0 = \sqrt{\mathrm{Var}_n[y_\bullet]} = 0$ only if $y_i = \bar{y}$ for every $i = 1, \ldots, n$. That is, $\epsilon_i = \mathbb{E}_n[x_\bullet'\beta_0 + \epsilon_\bullet] - x_i'\beta_0$ which is a zero measure event.

Next note that $\widehat{\sigma} = \sqrt{\widehat{Q}(\widehat{\beta})} = 0$ only if $y_i = x_i'\widehat{\beta}$ for every $i = 1, \ldots, n$. By the optimality conditions we have $0 \in \nabla\widehat{Q}(\widehat{\beta}) + \frac{\lambda}{n}\partial\|\cdot\|_1(\widehat{\beta})$. Since $\nabla\widehat{Q}(\widehat{\beta}) = 0$, we have $0 \in \partial\|\cdot\|_1(\widehat{\beta})$ which implies that $\widehat{\beta} = 0$. In turn $y_i = x_i'\widehat{\beta} = 0$ for every $i = 1, \ldots, n$ which is a zero measure event since $y_i = x_i'\beta_0 + \epsilon_i$. □

A.3. Perfect Model Selection

The following result on perfect model selection also requires strong assumptions on separation of coefficients and the empirical Gram matrix. Recall that for a scalar v, $\mathrm{sign}(v) = v/|v|$ if $|v| > 0$, and 0 otherwise. If v is a vector, we apply the definition componentwise. Also, given a vector $x \in \mathbb{R}^p$ and a set $T \subset \{1, \ldots, p\}$, let us denote $x_i[T] := \{x_{ij}, j \in T\}$.

Lemma 9 (Cases with Perfect Model Selection by lasso)**.** *Suppose Condition M holds. We have perfect model selection for lasso, $\widehat{T} = T$, if and only if*

$$\begin{aligned} &\Big\| \mathbb{E}_n\left[x_\bullet[T^c]x_\bullet[T]'\right] \mathbb{E}_n\left[x_\bullet[T]x_\bullet[T]'\right]^{-1} \Big\{ \mathbb{E}_n[x_\bullet[T]u_\bullet] \\ &\qquad - \tfrac{\lambda}{2n}\mathrm{sign}(\beta_0[T]) \Big\} - \mathbb{E}_n[x_\bullet[T^c]u_\bullet] \Big\|_\infty \leq \tfrac{\lambda}{2n}, \\ &\min_{j\in T} \left| \beta_{0j} + \left(\mathbb{E}_n\left[x_\bullet[T]x_\bullet[T]'\right]^{-1} \left\{ \mathbb{E}_n[x_\bullet[T]u_\bullet] - \tfrac{\lambda}{2n}\mathrm{sign}(\beta_0[T]) \right\} \right)_j \right| > 0. \end{aligned}$$

The result follows immediately from the first order optimality conditions, see [31]. The paper [34] provide further primitive sufficient conditions for perfect model selection for the parametric case in which $u_i = \varepsilon_i$, and [5] provide some conditions for the nonparametric case. The conditions above might typically require a slightly larger choice of λ than (2.12), and larger separation from zero of the minimal non-zero coefficient $\min_{j\in T}|\beta_{0j}|$.

Appendix B: Omitted Proofs

B.1. Section 2: Proof of Lemma 1

Proof of Lemma 1. We can assume that $m+s \geq 1$. Let $\widehat{\sigma}_j^2 = \mathbb{E}_n[\tilde{x}_{\bullet j}^2]$ for $j = 1,\ldots,p$. Moreover, let $c_*(m)$ and $c^*(m)$ denote the minimum and maximum m-sparse eigenvalues associated with $\mathbb{E}_n[\tilde{x}_\bullet \tilde{x}_\bullet']$ (unnormalized covariates). It follows that $\phi(m) \leq \max_{1\leq j\leq p} \widehat{\sigma}_j^2 c^*(m+s)$ and $\widetilde{\kappa}(m)^2 \geq \min_{1\leq j\leq p} \widehat{\sigma}_j^2 c_*(m+s)$. These relations shows that for bounding $c^*(m+s)$ and $c_*(m+s)$ it suffices to bound $\phi(m)$, $\widetilde{\kappa}(m)$, and deviations of $\widehat{\sigma}_j$'s away from 1.

Note that $P(\max_{1\leq j\leq p} |\widehat{\sigma}_j - 1| \leq 1/4) \to 1$ as n grows, since

$$\begin{array}{ll} P(\max_{1\leq j\leq p} |\widehat{\sigma}_j - 1| > 1/4) & \leq p \max_{1\leq j\leq p} P(|\widehat{\sigma}_j^2 - 1| > 1/4) \\ & \leq p \max_{1\leq j\leq p} P(|\sum_{i=1}^n (\tilde{x}_{ij}^2 - 1)| > n/4) \\ & \leq 2p \exp(-n^2/[32nK_n^2 + 8K_n^2 n/3]) \to 0 \end{array}$$

by Bernstein's inequality (Lemma 2.2.9 in [30]), $\mathrm{Var}(\tilde{x}_{ij}^2) \leq K_n^2$, and the side condition $K_n^2 \log p = o(n)$.

Under $s\log(n)\log^2(s\log n) \leq n[\kappa/\varphi^{1/2}][\epsilon/K_n]^2/[(\log p)(\log n)]$ for some $\epsilon > 0$ small enough, the bound on $\phi(m)$ and $\widetilde{\kappa}(m)^2$ follows from the application of (a simple extension of) results of Rudelson and Vershynin [25], namely Corollary 4 in Appendix C. □

B.2. Section 3: Proofs of Theorem 1 and Proof of Lemma 3

Proof of $\|\cdot\|_{2,n}$ bound in Theorem 1. Similar to [4], we make the use of the following relation: for $\delta = \widehat{\beta} - \beta_0$, if $\lambda \geq cn\|S\|_\infty$

$$\begin{array}{rcl} \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) - \|\delta\|_{2,n}^2 & = & 2\mathbb{E}_n[\epsilon_\bullet x_\bullet'\delta] + 2\mathbb{E}_n[r_\bullet x_\bullet'\delta] \geq -\|S\|_\infty\|\delta\|_1 - 2c_s\|\delta\|_{2,n} \\ & \geq & -\dfrac{\lambda}{cn}(\|\delta_T\|_1 + \|\delta_{T^c}\|_1) - 2c_s\|\delta\|_{2,n} \end{array} \tag{B.34}$$

By definition of $\widehat{\beta}$, $\widehat{Q}(\widehat{\beta}) - Q(\beta_0) \leq \frac{\lambda}{n}\|\beta_0\|_1 - \frac{\lambda}{n}\|\widehat{\beta}\|_1$, which implies that

$$-\frac{\lambda}{cn}(\|\delta_T\|_1 + \|\delta_{T^c}\|_1) + \|\delta\|_{2,n}^2 - 2c_s\|\delta\|_{2,n} \leq \frac{\lambda}{n}(\|\delta_T\|_1 - \|\delta_{T^c}\|_1) \tag{B.35}$$

If $\|\delta\|_{2,n}^2 - 2c_s\|\delta\|_{2,n} < 0$, then we have established the bound in the statement of the theorem. On the other hand, if $\|\delta\|_{2,n}^2 - 2c_s\|\delta\|_{2,n} \geq 0$ we get for $\bar{c} = (c+1)/(c-1)$

$$\|\delta_{T^c}\|_1 \leq \bar{c} \cdot \|\delta_T\|_1, \tag{B.36}$$

and therefore δ belongs to the restricted set in condition RE($\bar{c}$). From (B.35) and using RE($\bar{c}$) we get

$$\|\delta\|_{2,n}^2 - 2c_s\|\delta\|_{2,n} \leq \left(1 + \frac{1}{c}\right)\frac{\lambda}{n}\|\delta_T\|_1 \leq \left(1 + \frac{1}{c}\right)\frac{\sqrt{s}\lambda}{n}\frac{\|\delta\|_{2,n}}{\kappa(\bar{c})}$$

which gives the result on the prediction norm. □

Proof of Lemma 3. Let $W := \mathbb{E}_n[x_\bullet x_\bullet']$ and $\bar\alpha \in \mathbb{R}^p$ be such that $\phi(\lceil \ell k \rceil) = \bar\alpha' W \bar\alpha$ and $\|\bar\alpha\| = 1$. We can decompose

$$\bar\alpha = \sum_{i=1}^{\lceil \ell \rceil} \alpha_i, \text{ with } \sum_{i=1}^{\lceil \ell \rceil} \|\alpha_{iT^c}\|_0 = \|\bar\alpha_{T^c}\|_0 \text{ and } \alpha_{iT} = \bar\alpha_T / \lceil \ell \rceil ,$$

where we can choose α_i's such that $\|\alpha_{iT^c}\|_0 \le k$ for each $i = 1, \ldots, \lceil \ell \rceil$, since $\lceil \ell \rceil k \ge \lceil \ell k \rceil$. Note that the vectors α_i's have no overlapping support outside T. Since W is positive semi-definite, $\alpha_i' W \alpha_i + \alpha_j' W \alpha_j \ge 2 |\alpha_i' W \alpha_j|$ for any pair (i, j). Therefore

$$\begin{aligned}
&\phi(\lceil \ell k \rceil) = \bar\alpha' W \bar\alpha \;=\; \textstyle\sum_{i=1}^{\lceil \ell \rceil} \sum_{j=1}^{\lceil \ell \rceil} \alpha_i' W \alpha_j \\
&\le \; \textstyle\sum_{i=1}^{\lceil \ell \rceil} \sum_{j=1}^{\lceil \ell \rceil} \frac{\alpha_i' W \alpha_i + \alpha_j' W \alpha_j}{2} = \lceil \ell \rceil \sum_{i=1}^{\lceil \ell \rceil} \alpha_i' W \alpha_i \\
&\le \lceil \ell \rceil \textstyle\sum_{i=1}^{\lceil \ell \rceil} \|\alpha_i\|^2 \phi(\|\alpha_{iT^c}\|_0) \le \lceil \ell \rceil \max_{i=1,\ldots,\lceil \ell \rceil} \phi(\|\alpha_{iT^c}\|_0) \le \lceil \ell \rceil \phi(k),
\end{aligned}$$

where we used that

$$\sum_{i=1}^{\lceil \ell \rceil} \|\alpha_i\|^2 = \sum_{i=1}^{\lceil \ell \rceil} (\|\alpha_{iT}\|^2 + \|\alpha_{iT^c}\|^2) = \frac{\|\bar\alpha_T\|^2}{\lceil \ell \rceil} + \sum_{i=1}^{\lceil \ell \rceil} \|\alpha_{iT^c}\|^2 \le \|\bar\alpha\|^2 = 1.$$

□

B.3. Section 4: Relation after (A.30) in Proof of Lemma 5

Proof of Lemma 5: Relation after (A.30). First note that $\bar\Phi(t) \le \exp(-t^2/2)$ for $t \ge 1$. Then,

$$\begin{aligned}
I \;&:= \left(\tfrac{D\sigma\mu(m) e_n(m,\eta)}{\sqrt{m + s\sigma^2}} \right)^{m+s} \bar\Phi(e_n(m,\eta)/\sigma) \\
&\le \exp\left(-\tfrac{e_n^2(m,\eta)}{2\sigma^2} + (m+s) \log\left[\tfrac{e_n(m,\eta)}{\sqrt{m+s}\sigma} \right] + (m+s)\log(D\sigma\mu(m)) \right) \\
&= \exp\left(-\tfrac{(m+s)}{2} \left[\tfrac{e_n(m,\eta)}{\sqrt{m+s}\sigma} \right]^2 + (m+s) \log\left[\tfrac{e_n(m,\eta)}{\sqrt{m+s}\sigma} \right] + (m+s)\log(D\sigma\mu(m)) \right)
\end{aligned}$$

Next note that $\log x \le x^2/4$ if $x \ge 2\sqrt{2}$. Note that $e_n(m,\eta)/[\sqrt{m+s}\sigma] \ge 2\sqrt{2}$ since $\mu(m) \ge 1$ and we can take $D \ge e$. Thus, the expression above is bounded by

$$\begin{aligned}
I \;&\le \exp\left(-\frac{(m+s)}{4} \left[\frac{e_n(m,\eta)}{\sqrt{m+s}\sigma} \right]^2 + (m+s)\log(D\sigma\mu(m)) \right) \\
&= \exp\left(-\frac{e_n^2(m,\eta)}{4\sigma^2} + (m+s)\log(D\sigma\mu(m)) \right) \\
&\le \exp\left(-\log\binom{p}{m} - (m+s)\log(1/\eta) \right).
\end{aligned}$$

□

B.4. Section 5: Proofs of Theorem 6 and 7

In this Section we provide the proof for Theorems 6 and 7. We begin with Theorem 6 which threshold level is set based on the fit of the second step estimator relative to the fit of the original estimator, in this case lasso.

Proof of Theorem 6. Let $\widetilde{B}_n := \widehat{Q}(\widetilde{\beta}) - \widehat{Q}(\beta_0)$ and $\widetilde{C}_n := \widehat{Q}(\beta_{0\widetilde{T}}) - \widehat{Q}(\beta_0)$. It follows by definition of the estimator that $\widetilde{B}_n \leq \gamma + \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0)$. Thus, by Theorem 4, for any $\varepsilon > 0$, there is a constant K_ε independent of n such that with probability at least $1 - \varepsilon$ we have

$$\begin{aligned}
&\|\widetilde{\beta} - \beta_0\|_{2,n} \leq K_\varepsilon \sigma \sqrt{\frac{\widetilde{m} \log p + (\widetilde{m} + s) \log(e\mu(\widetilde{m}))}{n}} + 2c_s + \sqrt{(\widetilde{B}_n)_+ \wedge (\widetilde{C}_n)_+}, \\
&(\widetilde{B}_n)_+ \leq \gamma + \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0), \\
&(\widetilde{B}_n)_+ \wedge (\widetilde{C}_n)_+ \leq 1\{T \not\subseteq \widetilde{T}\}(\widetilde{B}_n)_+,
\end{aligned}$$

since $\widetilde{C}_n = 0$ if $T \subseteq \widetilde{T}$.

We bound $B_n = \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0)$ as in Theorem 5, namely,

$$B_n \leq \frac{\lambda\sqrt{s}}{n\kappa(1)} \|\widehat{\beta} - \beta_0\|_{2,n} \leq \left(1 + \frac{1}{c}\right) \frac{\lambda^2 s}{n^2 \kappa(1)^2} + \frac{2c_s \lambda \sqrt{s}}{n\kappa(1)}.$$

The second claim follows from the first by using $\lambda \lesssim \sqrt{n \log p}$ under Condition V, the specified conditions on the penalty level. The final bound follows by applying the relation that for any nonnegative numbers a, b, we have $\sqrt{ab} \leq (a + b)/2$. □

The traditional thresholding scheme which truncates to zero all components below a set threshold t. This is arguably the most used thresholding scheme in the literature. Recall that $\widehat{\beta}_{tj} = \widehat{\beta}_j 1\{|\widehat{\beta}_j| > t\}$, $\widetilde{m} := |\widetilde{T} \setminus T|$, $m_t := |\widehat{T} \setminus \widetilde{T}|$ and $\gamma_t := \|\widehat{\beta}_t - \widehat{\beta}\|_{2,n}$ where $\widehat{\beta}$ is the lasso estimator.

Proof of Theorem 7. Let $\widetilde{B}_n := \widehat{Q}(\widehat{\beta}^t) - \widehat{Q}(\beta_0)$ and $\widetilde{C}_n := \widehat{Q}(\beta_{0\widetilde{T}}) - \widehat{Q}(\beta_0)$.

By Theorem 4 and Lemma 4, for any $\varepsilon > 0$, there is a constant K_ε independent of n such that with probability at least $1 - \varepsilon$ we have

$$\begin{aligned}
&\|\widetilde{\beta} - \beta_0\|_{2,n} \leq K_\varepsilon \sigma \sqrt{\frac{\widetilde{m} \log p + (\widetilde{m} + s) \log(e\mu(\widetilde{m}))}{n}} + 2c_s + \sqrt{(\widetilde{B}_n)_+ \wedge (\widetilde{C}_n)_+}, \\
&(\widetilde{B}_n)_+ \leq \|\widehat{\beta}^t - \beta_0\|_{2,n}^2 + \left[K_\varepsilon \sigma \sqrt{\frac{\widetilde{m} \log p + (\widetilde{m} + s) \log(e\mu(\widetilde{m}))}{n}} + 2c_s \right] \|\widehat{\beta}^t - \beta_0\|_{2,n}, \\
&(\widetilde{B}_n)_+ \wedge (\widetilde{C}_n)_+ \leq 1\{T \not\subseteq \widetilde{T}\}(\widetilde{B}_n)_+,
\end{aligned}$$

since $\widetilde{C}_n = 0$ if $T \subset \widetilde{T}$.

Next note that by definition of γ_t, we have $\|\widehat{\beta}^t - \beta_0\|_{2,n} \leq \gamma_t + \|\widehat{\beta} - \beta_0\|_{2,n}$. The result follows by applying Theorem 1 to bound $\|\widehat{\beta} - \beta_0\|_{2,n}$.

The second claim follows from the first by using $\lambda \lesssim \sqrt{n \log p}$ under Condition V, the specified conditions on the penalty level, and the relation that for any nonnegative numbers a, b, we have $\sqrt{ab} \leq (a+b)/2$. □

Appendix C: Uniform Control of Sparse Eigenvalues

In this section we provide a simple extension of the sparse law of large numbers for matrices derived in [25] to the case where the population matrices are non-isotropic.

Lemma 10 (Essentially in [25] Lemma 3.8)**.** *Let $x_1, \ldots, x_n$, be vectors in $\mathbb{R}^p$ with uniformly bounded entries, $\|x_i\|_\infty \leq K$ for all $i = 1, \ldots, n$. Then, for independent Rademacher random variables ε_i, $i = 1, \ldots, n$, we have*

$$\mathrm{E}\left[\sup_{\|\alpha\|_0 \leq k, \|\alpha\|=1} \left|\sum_{i=1}^n \varepsilon_i (x_i'\alpha)^2\right|\right] \leq \left(CK\sqrt{k}\log(k)\sqrt{\log(p \vee n)}\sqrt{\log n}\right) \sup_{\|\alpha\|_0 \leq k, \|\alpha\|=1} \left(\sum_{i=1}^n (x_i'\alpha)^2\right)^{1/2}$$

where C is a universal constant.

Proof. The proof follows from Rudelson and Vershynin [25] Lemma 3.8 setting $A = K/\sqrt{k}$ instead of $A = 1/\sqrt{k}$ so that the constant $C(K)$ can be taken $C \cdot K$. □

Lemma 11 (Essentially in [25] Theorem 3.6)**.** *Let x_i, $i = 1, \ldots, n$, be i.i.d. random vectors in $\mathbb{R}^p$ with uniformly bounded entries, $\|x_i\|_\infty \leq K$ a.s. for all $i = 1, \ldots, n$. Let $\delta_n := 2\left(CK\sqrt{k}\log(k)\sqrt{\log(p \vee n)}\sqrt{\log n}\right)/\sqrt{n}$, where C is the universal constant in Lemma 10. Then,*

$$\mathrm{E}\left[\sup_{\|\alpha\|_0 \leq k, \|\alpha\|=1} \left|\mathbb{E}_n\left[(\alpha' x_i)^2 - \mathrm{E}[(\alpha' x_i)^2]\right]\right|\right] \leq \delta_n^2 + \delta_n \sup_{\|\alpha\|_0 \leq k, \|\alpha\|=1} \sqrt{\mathrm{E}[(\alpha' x_i)^2]}.$$

Proof. Let

$$V_k = \sup_{\|\alpha\|_0 \leq k, \|\alpha\|=1} \left|\mathbb{E}_n\left[(\alpha' x_i)^2 - \mathrm{E}[(\alpha' x_i)^2]\right]\right|.$$

Then, by a standard symmetrization argument (see Guédon and Rudelson [16], page 804)

$$n\mathrm{E}[V_k] \quad \leq 2\mathrm{E}_x\mathrm{E}_\varepsilon\left[\sup_{\|\alpha\|_0 \leq k, \|\alpha\|=1} \left|\textstyle\sum_{i=1}^n \varepsilon_i (\alpha' x_i)^2\right|\right].$$

Letting

$$\phi(k) = \sup_{\|\alpha\|_0 \leq k, \|\alpha\| \leq 1} \mathbb{E}_n[(\alpha' x_i)^2] \quad \text{and} \quad \varphi(k) = \sup_{\|\alpha\|_0 \leq k, \|\alpha\|=1} \mathrm{E}[(\alpha' x_i)^2],$$

we have $\phi(k) \leq \varphi(k) + V_k$ and by Lemma 10

$$\begin{aligned} n\mathrm{E}[V_k] \quad &\leq 2\left(CK\sqrt{k}\log(k)\sqrt{\log(p \vee n)}\sqrt{\log n}\right)\sqrt{n}\mathrm{E}_X\left[\sqrt{\phi(k)}\right] \\ &\leq 2\left(CK\sqrt{k}\log(k)\sqrt{\log(p \vee n)}\sqrt{\log n}\right)\sqrt{n}\sqrt{\varphi(k) + \mathrm{E}[V_k]}. \end{aligned}$$

The result follows by noting that for positive numbers v, A, B, $v \leq A(v+B)^{1/2}$ implies $v \leq A^2 + A\sqrt{B}$. □

Corollary 4. *Suppose x_i, $i = 1, \ldots, n$, are i.i.d. vectors, such that the population design matrix $\mathrm{E}[x_i x_i']$ has its k-sparse eigenvalues bounded from above by $\varphi < \infty$ and bounded from below by $\kappa^2 > 0$. If x_i are arbitrary with $\max_{1 \leq i \leq n} \|x_i\|_\infty \leq K_n$ a.s., and the condition $K_n^2 k \log^2(k) \log(n) \log(p \vee n) = o(n\kappa^4/\varphi)$ holds,*

$$P\left(\sup_{\|\alpha\|_0 \leq k, \|\alpha\|=1} \mathbb{E}_n[(\alpha' x_i)^2] \leq 2\varphi, \inf_{\|\alpha\|_0 \leq k, \|\alpha\|=1} \mathbb{E}_n[(\alpha' x_i)^2] \geq \kappa^2/2 \right) = 1 - o(1).$$

Proof. Let $V_k = \sup_{\|\alpha\|_0 \leq k, \|\alpha\|=1} \left| \mathbb{E}_n \left[(\alpha' x_i)^2 - \mathrm{E}[(\alpha' x_i)^2] \right] \right|$. It suffices to prove that $P(V_k > \kappa^2/2) = o(1)$. Indeed,

$$\sup_{\|\alpha\|_0 \leq k, \|\alpha\|=1} \mathbb{E}_n[(\alpha' x_i)^2] \leq V_k + \varphi \quad \text{and} \quad \inf_{\|\alpha\|_0 \leq k, \|\alpha\|=1} \mathbb{E}_n[(\alpha' x_i)^2] > \kappa^2 - V_k.$$

By Markov inequality, $P(V_k > \kappa^2/2) \leq 2E[V_k]/\kappa^2$ and the result follows provided that $E[V_k] = o(\kappa^2)$.

For $\delta_n := 2\left(CK_n \sqrt{k} \log(k) \sqrt{\log(p \vee n)} \sqrt{\log n} \right) / \sqrt{n}$, by Lemma 11, we have $E[V_k] \leq \delta_n^2 + \delta_n \sqrt{\varphi} = o(\kappa^2)$ by the growth condition in the statement. □

Appendix D: Empirical Performance Relative to lasso

In this section we assess the finite sample performance of the following estimators: 1) lasso, which is our benchmark, 2) ols post lasso, 3) ols post fit-lasso, and 4) ols post t-lasso with the threshold $t = \lambda/n$. We consider a "parametric" and a "nonparametric" model of the form:

$$y_i = f_i + \epsilon_i, \quad f_i = z_i' \theta_0, \quad \epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n,$$

where in the "parametric" model

$$\theta_0 = C \cdot [1, 1, 1, 1, 1, 0, 0, \ldots, 0]', \tag{D.37}$$

and in the "nonparametric" model

$$\theta_0 = C \cdot [1, 1/2, 1/3, \ldots, 1/p]'. \tag{D.38}$$

The reason the latter model is called "nonparametric" is because in that model the function $f(z) = \sum_{j=1}^p z_j \theta_{0j}$ is numerically indistinguishable from the function $g(z) = \sum_{j=1}^\infty z_j \gamma_{j0}$, characterized by the infinite-dimensional parameter γ_j with true values $\gamma_{j0} = 1/j$.

The parameter C determines the size of the coefficients, representing the "strength of the signal", and we vary C between 0 and 2. The number of regressors is $p = 500$, the sample size is $n = 100$, the variance of the noise is $\sigma^2 = 1$, and we used 1000 simulations for each design. We generate regressors from the normal law $z_i \sim N(0, \Sigma)$, and consider three designs of the covariance matrix Σ: a) the isotropic design with $\Sigma_{jk} = 0$ for $j \neq k$, b) the Toeplitz design with $\Sigma_{jk} = (1/2)^{|j-k|}$, and c) the equi-correlated design with $\Sigma_{jk} = 1/2$ for $j \neq k$; in all designs $\Sigma_{jj} = 1$. Thus our parametric model is very sparse and offers a rather favorable setting for applying lasso-type methods, while our nonparametric model is non-sparse and much less favorable.

We present the results of computational experiments for each design a)-c) in Figures 2-4. The left column of each figure reports the results for the parametric model, and the right column of each figure reports the results for the nonparametric model. For each model the figures plot the following as a function of the signal strength for each estimator $\widetilde{\beta}$:

- in the top panel, the number of regressors selected, $E[|\widetilde{T}|]$,
- in the middle panel, the norm of the bias, namely $\|E[\widetilde{\beta} - \theta_0]\|$,
- in the bottom panel, the average empirical risk, namely $E[\mathbb{E}_n[f_i - z_i'\widetilde{\beta}]^2]$.

We will focus the discussion on the isotropic design, and only highlight differences for other designs.

Figure 2, left panel, shows the results for the parametric model with the isotropic design. We see from the bottom panel that, for a wide range of signal strength C, both ols post lasso and ols post fit-lasso significantly outperform both lasso and ols post t-lasso in terms of empirical risk. The middle panel shows that the first two estimators' superior performance stems from their much smaller bias. We see from the top panel that lasso achieves good sparsity, ensuring that ols post lasso performs well, but ols post fit-lasso achieves even better sparsity. Under very high signal strength, ols post fit-lasso achieves the performance of the oracle estimator; ols post t-lasso also achieves this performance; ols post lasso nearly matches it; while lasso does not match this performance. Interestingly, the ols post t-lasso performs very poorly for intermediate ranges of signal.

Figure 2, right panel, shows the results for the nonparametric model with the isotropic design. We see from the bottom panel that, as in the parametric model, both ols post lasso and ols post fit-lasso significantly outperform both lasso and ols post fit-lasso in terms of empirical risk. As in the parametric model, the middle panel shows that the first two estimators are able to outperform the last two because they have a much smaller bias. We also see from the top panel that, as in the parametric model, lasso achieves good sparsity, while ols post fit-lasso achieves excellent sparsity. In contrast to the parametric model, in the nonparametric setting the ols post t-lasso performs poorly in terms of empirical risk for almost all signals, except for very weak signals. Also in contrast to the parametric model, no estimator achieves the exact oracle performance, although lasso, and especially ols post lasso and ols post fit-lasso perform nearly as well, as we would expect from the theoretical results.

Figure 3 shows the results for the parametric and nonparametric model with the

Toeplitz design. This design deviates only moderately from the isotropic design, and we see that all of the previous findings continue to hold. Figure 4 shows the results under the equi-correlated design. This design strongly deviates from the isotropic design, but we still see that the previous findings continue to hold with only a few differences. Specifically, we see from the top panels that in this case lasso no longer selects very sparse models, while ols post fit-lasso continues to perform well and selects very sparse models. Consequently, in the case of the parametric model, ols post fit-lasso substantially outperforms ols post lasso in terms of empirical risk, as the bottom-left panel shows. In contrast, we see from the bottom right panel that in the nonparametric model, ols post fit-lasso performs equally as well as ols post lasso in terms of empirical risk, despite the fact that it uses a much sparser model for estimation.

The findings above confirm our theoretical results on post-model selection estimators in parametric and nonparametric models. Indeed, we see that ols post fit-lasso and ols post lasso are at least as good as lasso, and often perform considerably better since they remove penalization bias. ols post fit-lasso outperforms ols post lasso whenever lasso does not produce excellent sparsity. Moreover, when the signal is strong and the model is parametric and sparse (or very close to being such), the lasso-based model selection permits the selection of oracle or near-oracle model. That allows for post-model selection estimators to achieve improvements in empirical risk over lasso. Of particular note is the excellent performance of ols post fit-lasso, which uses data-driven threshold to select a sparse model. This performance is fully consistent with our theoretical results. Finally, traditional thresholding performs poorly for intermediate ranges of signal. In particular, it exhibits very large biases leading to large goodness-of-fit losses.

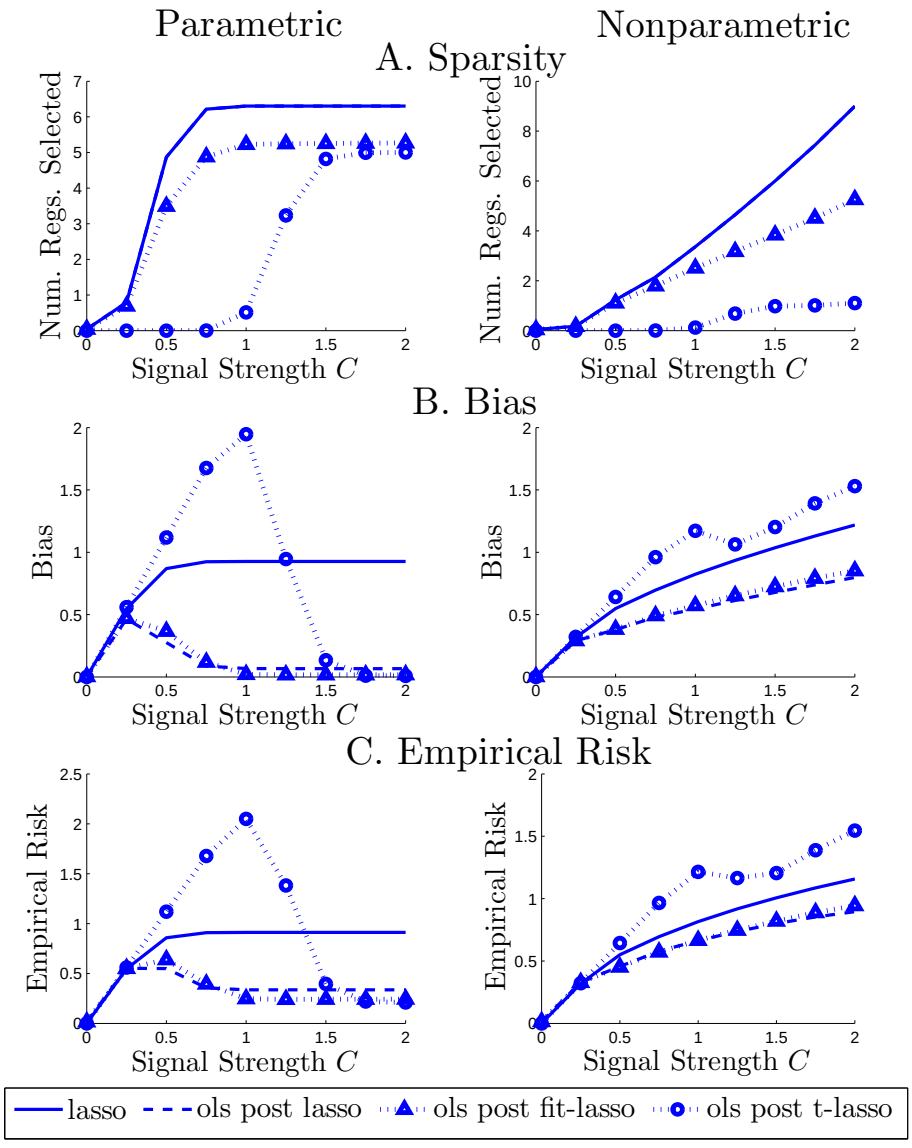

Figure 2. This figure plots the performance of the estimators listed in the text under the isotropic design for the covariates, $\Sigma_{jk} = 0$ if $j \neq k$. The left column corresponds to the parametric case and the right column corresponds to the nonparametric case described in the text. The number of regressors is $p = 500$ and the sample size is $n = 100$ with 1000 simulations for each value of C.

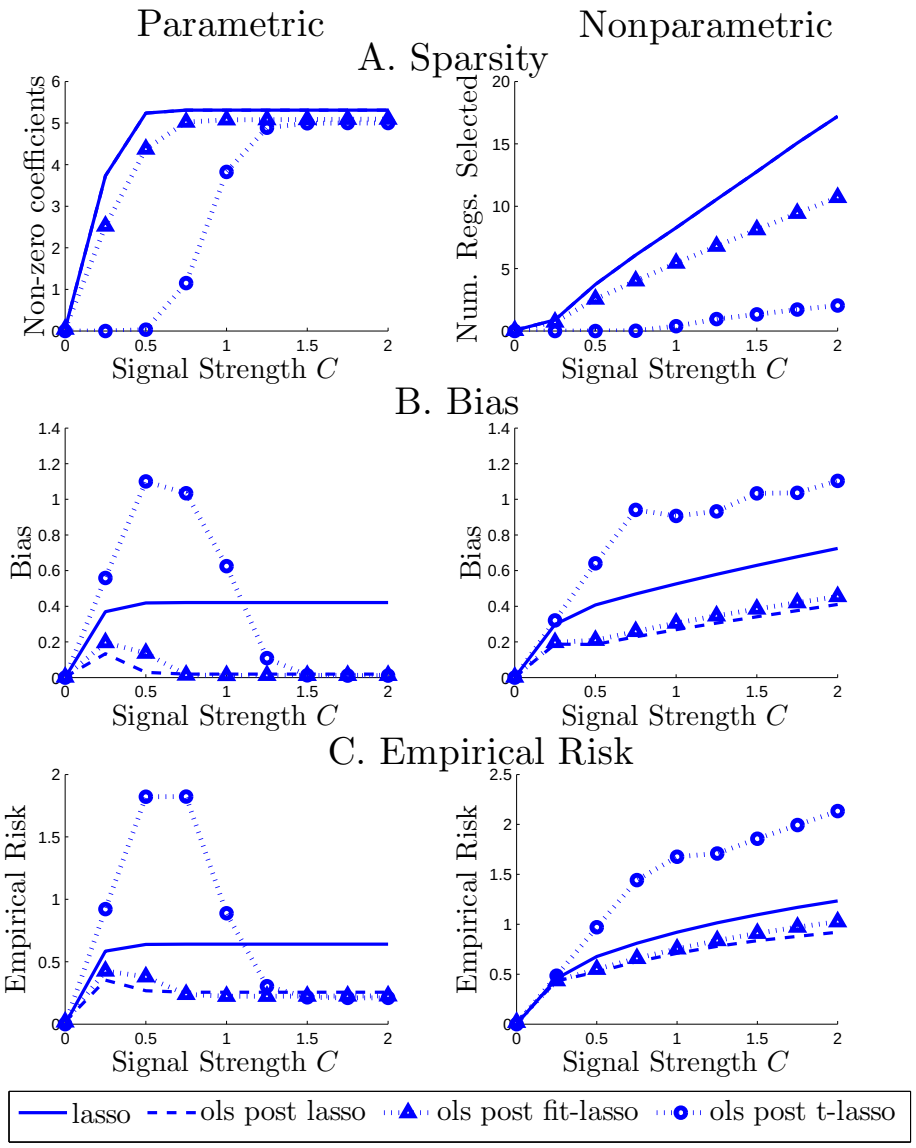

Figure 3. This figure plots the performance of the estimators listed in the text under the Toeplitz design for the covariates, $\Sigma_{jk} = \rho^{|j-k|}$ if $j \neq k$. The left column corresponds to the parametric case and the right column corresponds to the nonparametric case described in the text. The number of regressors is $p = 500$ and the sample size is $n = 100$ with 1000 simulations for each value of C.

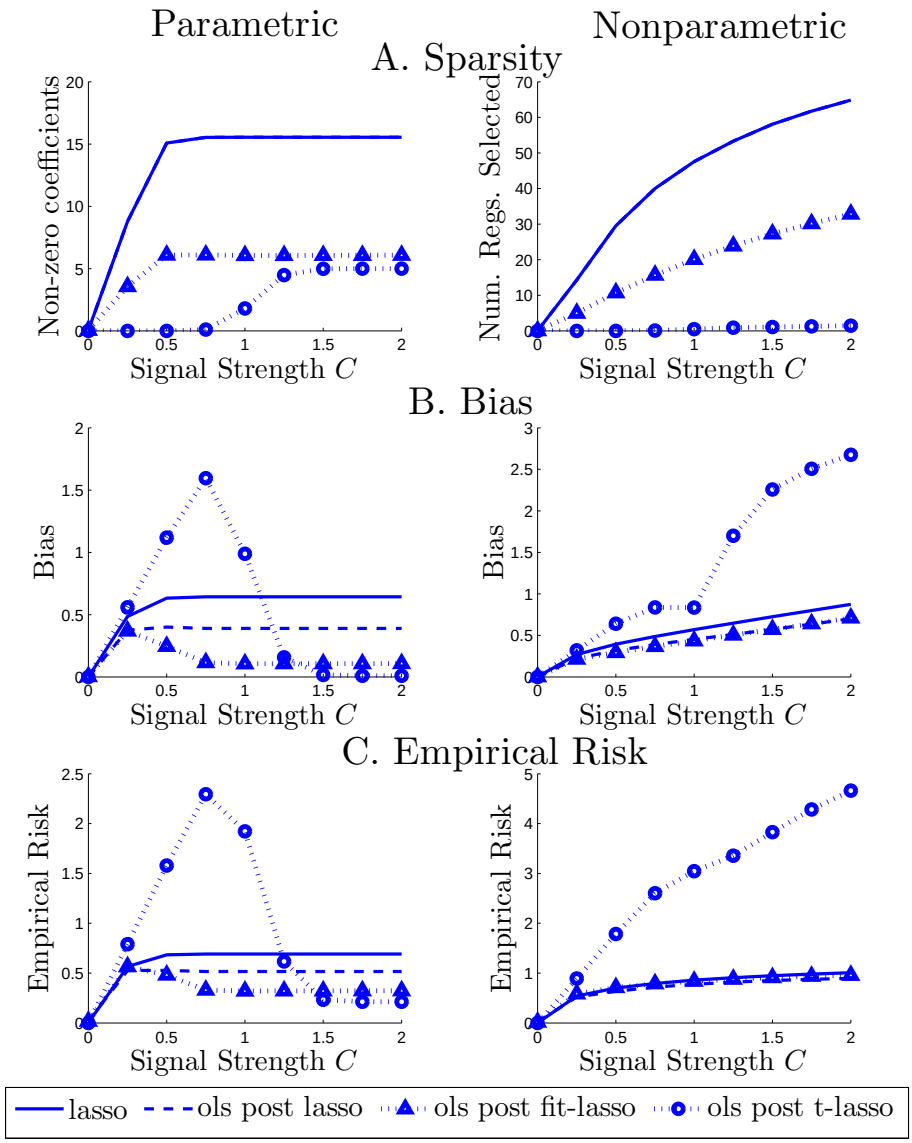

Figure 4. This figure plots the performance of the estimators listed in the text under the equi-correlated design for the covariates, $\Sigma_{jk} = \rho$ if $j \neq k$. The left column corresponds to the parametric case and the right column corresponds to the nonparametric case described in the text. The number of regressors is $p = 500$ and the sample size is $n = 100$ with 1000 simulations for each value of C.