Bernoulli **19**(2), 2013, 521–547
DOI: 10.3150/11-BEJ410

Least squares after model selection in high-dimensional sparse models

ALEXANDRE BELLONI[1] and VICTOR CHERNOZHUKOV[2]

[1] *100 Fuqua Drive, Durham, North Carolina 27708, USA. E-mail: abn5@duke.edu*
[2] *50 Memorial Drive, Cambridge, Massachusetts 02142, USA. E-mail: vchern@mit.edu*

In this article we study post-model selection estimators that 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 the 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 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 nonasymptotic and hold in both parametric and nonparametric models. Moreover, our analysis is not limited to the Lasso estimator acting as a 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 additional 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 those procedures as well as traditional thresholding in a wide variety of experiments.

Keywords: Lasso; OLS post-Lasso; post-model selection estimators

1. Introduction

In this work, we study post-model selection 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 [9,16] 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 authors have investigated estimation of hdsms, focusing primarily on mean regression with the ℓ_1-norm acting as a penalty function [4,6–9,12,16,22,24,26]. The results of [4,6–8,12, 16,24,26] demonstrate the fundamental result that ℓ_1-penalized least squares estimators achieve

1350-7265

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. [12,22] demonstrated a similar fundamental result on the excess forecasting error loss under both quadratic and nonquadratic 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. In addition, [3] investigated the ℓ_1-penalized quantile regression process and obtained similar results. See [4,6–8,11,13,14,17] for many other interesting developments and a detailed review of the existing literature.

In this article we derive theoretical properties of post-model selection estimators that 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 thus is hard to improve on. 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. (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 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 provide generic rate results that cover any first-step estimator for which a rate and a sparsity bound are available. We also present a generic result from 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 that we introduce here (fit-Lasso). The new thresholding scheme induces additional 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 provides 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 signal levels, all estimators perform similarly. When the signal strength is intermediate, OLS post-Lasso and OLS post-fit Lasso significantly 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 on 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.

To the best of our knowledge, this article 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

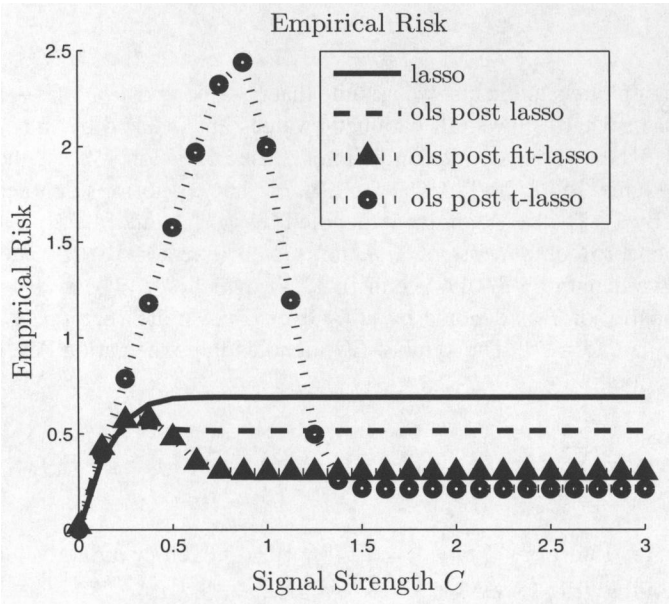

Figure 1. This figure plots the performance of the estimators listed in the text under the equicorrelated design for the covariates $x_i \sim N(0, \Sigma)$, $\Sigma_{jk} = 1/2$ if $j \neq k$. The number of regressors is $p = 500$, and the sample size is $n = 100$ with 1000 simulations for each level 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.

problem. Our analysis builds on the ideas of [3], who established the properties of postpenalized 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 by [3] in the context of median regression. Our sparsity bounds for Lasso improve on the analogous bounds of [4] and are comparable to the bounds of [26] obtained under a larger penalty level. We also rely on the maximal inequalities of [26] to provide primitive conditions for the sharp sparsity bounds to hold.

The article is organized as follows. Section 2 reviews the model and discusses the estimators. Section 3 revisits some benchmark results of [4] for Lasso, allowing for a data-driven choice of penalty level, develops an extension of model selection results of [13] 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. The Appendix contains main proofs, and the supplemental article [2] contains auxiliary proofs, as well as the results of our computational experiments.

Notation

When making asymptotic statements, we assume that $n \to \infty$ and $p = p_n \to \infty$, and 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 omission will 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 nonzero 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 $\mathbb{R}^p$ in which $\delta_{Tj} = \delta_j$ if $j \in T$ and $\delta_{Tj} = 0$ if $j \notin T$. The cardinality of T is denoted by $|T|$. Given a covariate vector $x_i \in \mathbb{R}^p$, we let $x_i[T]$ denote the vector $\{x_{ij}, j \in T\}$. The symbol $\mathrm{E}[\cdot]$ denotes the expectation. We also use standard empirical process notation

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

We 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 $\|\delta\|_{2,n} = \{\mathbb{E}_n[(x_\bullet'\delta)^2]\}^{1/2} = (\delta'\mathbb{E}_n[x_\bullet x_\bullet']\delta)^{1/2}$. We use the notation $a \lesssim b$ to denote $a \leq Cb$ for some constant $C > 0$ that does not depend on n (and thus 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 1 as n increases. In addition, we write $\bar{c} = (c+1)/(c-1)$ for a chosen constant $c > 1$.

2. Setting, estimators, and conditions

2.1. Setting

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

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

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, that is, a p-vector of transformation of z_i, with components

$$x_i := P(z_i)$$

of the dictionary normalized so that

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

In 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 $i = 1, \ldots, n$. To set up the 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, that is, the best estimator among 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 among 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, then 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. Because T is not known, we 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), then it suffices to select one of the values in the set of argmins. Supplemental article [2] 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 previous uses in the literature (see, e.g., [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 and 5.2 given later, we 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 where 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 $\sqrt{s/n}$, as 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 needed to obtain consistency. The Lasso estimator [19], defined as follows, achieves both tasks by using the ℓ_1 penalization:

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

and λ is the penalty level, the choice of which is described later. If the solution is not unique, then we pick any solution with minimum support. The Lasso is often used as an estimator, and most often only as a model selection device, with the model selected by Lasso given by

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

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

Often, 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. The corresponding selected model is then

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

Note that, when 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 the foregoing, all of our post-model selection estimators or OLS post-Lasso estimators will take the form

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

That is, given that 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 OLS to the selected model.

Along with the case of $t = 0$, we also consider the following choices for the threshold level:

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

where $\gamma \leq 0$, 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 0. However, this scheme may perform poorly in models with oracle coefficients not well separated from 0 and in nonparametric models. Indeed, even in parametric models with many small but nonzero true coefficients, thresholding the estimates too aggressively may result in large goodness-of-fit losses and, consequently, 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 as many small coefficient estimates as possible to 0, subject to maintaining a certain goodness-of-fit level.

Depending on how we select the threshold, we consider three types of 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 Lassol and the third, by OLS applied to the model selected by fitness-thresholded Lasso.

The main purpose of this article 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 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 possible to dominate the noise, namely

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

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

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

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*

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

where $0 < \ell \leq 1$ *and* $1 \leq u$ *and* $0 \leq \tau < 1$ *are 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 that 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 supplemental article [2]. We can iterate on the last step a bounded number of times. We also can use OLS post-Lasso for this purpose.

2.5. Choices and computation of thresholding levels

Our analysis covers a wide range of possible threshold levels. Here, however, we 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, which 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 based on the results of Section 3.2 if the empirical Gram matrix is orthogonal and approximation error c_s vanishes. Thus, $\tilde{c} = 1$ is the least aggressive traditional thresholding that can be performed under conditions of Section 3.2. (Also note that $\tilde{c} = 1$ has performed better than $\tilde{c} > 1$ in our computational experiments.)

Our fitness-based threshold t_γ requires 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. However, we prefer 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 sparser and produces a better in-sample fit than Lasso. This choice also results in near-oracle performance and 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 because the final estimator depends only on the selected support, not on the specific value of t used. Therefore, because there are at most $|\widehat{T}|$ different values of t to be tested, using a binary search, we can compute t_γ exactly by running at most $\lceil \log_2 |\widehat{T}| \rceil$ OLS problems.

2.6. Conditions on the design

For the analysis of Lasso, we use the following restricted eigenvalue condition on the empirical Gram matrix:

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 by [4], which is known to be quite general and plausible (see [4] for related conditions).

For the analysis of post-model selection estimators, we use the following restricted sparse eigenvalue condition on the empirical Gram matrix:

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

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

Condition $RSE(m)$ depends on T and can be viewed as an extension of the restricted isometry condition [9]. Here m denotes the restriction on the number of nonzero components outside the support T. The standard concept of (unrestricted) m-sparse eigenvalues corresponds to the restricted sparse eigenvalues when $T = \varnothing$ (see, e.g., [4]). It is 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 the foregoing conditions for the case where the values x_i, $i = 1, \ldots, n$, have been generated as a realization of the random sample; there are other primitive conditions as well. In this case, the empirical restricted sparse eigenvalues are bounded away from 0 and from above, so that (2.15) is bounded from above with high probability. The lemma assumes as a primitive condition that the sparse eigenvalues of the population Gram matrix bounded away from zero and from above. The lemma allows for many standard bounded dictionaries that arise in the nonparametric estimation, for example, regression splines, orthogonal polynomials, and trigonometric series (see [10,20,21,25]). Similar results are known to hold for standard Gaussian regressors as well [26].

Lemma 1 (Plausibility of *RE* and *RSE*). *Suppose that* $\tilde{x}_i$, $i = 1, \ldots, n$, *are i.i.d. mean-zero vectors, such that the population Gram matrix* $\mathrm{E}[\tilde{x}\tilde{x}']$ *has all of the diagonal elements equal to* 1, *and*

$$0 < \kappa^2 \leq \min_{1 \leq \|\delta\|_0 \leq s \log n} \frac{\delta' \mathrm{E}[\tilde{x}\tilde{x}']\delta}{\|\delta\|} \leq \max_{1 \leq \|\delta\|_0 \leq s \log n} \frac{\delta' \mathrm{E}[\tilde{x}\tilde{x}']\delta}{\|\delta\|} \leq \varphi < \infty.$$

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

$$\max_{1\le i\le n} \|\tilde{x}_i\|_\infty \le K_n \quad a.s. \quad \text{and} \quad K_n^2 s \log^2(n) \log^2(s \log n) \log(p \vee n) = o(n\kappa^4/\varphi).$$

Then, for any $m + s \le s \log n$, the restricted sparse eigenvalues of the empirical Gram matrix obey the following bounds:

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

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 depend crucially 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 also develop the model selection properties of Lasso for nonparametric models, generalizing the results of [13] 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 $\lambda \ge cn\|S\|_\infty$, then*

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

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

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

If in addition 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 a data-driven penalty level and deriving the rates in ℓ_1-norm. These results may be of independent interest and are necessary for the subsequent results.

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

$$\left| \|\widehat{f} - f\|_{\mathbb{P}_n,2} - \|\widehat{\beta} - \beta_0\|_{2,n} \right| \leq c_s, \tag{3.1}$$

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.1) in the Appendix):

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

where $\widehat{m} = |\widehat{T} \setminus T|$. We note that a similar lower bound was first derived by [15] 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*

$$\begin{gathered} \phi(\widehat{m}) \lesssim 1, \qquad \kappa(\bar{c}) \gtrsim 1, \qquad \mu(\widehat{m}) \lesssim 1, \qquad \log(1/\alpha) \lesssim \log p, \\ \alpha = o(1), \qquad u/\ell \lesssim 1 \quad \textit{and} \quad \tau = o(1) \end{gathered} \tag{3.2}$$

hold as n grows, then we have

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

Moreover, if $|\widehat{T}| \gtrsim_P s$ *– in particular, if* $T \subseteq \widehat{T}$ *with probability going to* 1 *– then 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 (3.2) to hold with high probability as n grows, when the design points $z_1, \ldots, z_n$ are 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 as long as $\kappa(\bar{c})$ is bounded away from 0, 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 by [15]. It shows that the rate cannot be improved in general. We 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

Our main results do not require that the first-step estimators like Lasso 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 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. In this section we describe these very special cases in which perfect model selection is possible.

Theorem 2 (Some conditions for perfect model selection in nonparametric settings). *Suppose that Condition* M *holds.*

(1) *If the coefficients are well separated from* 0, *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, $T := \text{support}(\beta_0) \subseteq \widehat{T} := \text{support}(\widehat{\beta})$. *Moreover,* T *can be perfectly selected by applying level* t *thresholding to* $\widehat{\beta}$, *that is,* $T = \widehat{T}(t)$.

(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 [13] on model selection by thresholded Lasso. These results cover the more general nonparametric case and may be of independent interest. Also note that the stated conditions for perfect model selection require a strong assumption on the separation of coefficients of the oracle from 0, along with near-perfect orthogonality of the empirical Gram matrix. This is the sense in which the perfect model selection is a rather special, nongeneral 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 [5,24,27]. In the supplement, we state the nonparametric extension of the parametric result due to [24].

3.3. Sparsity properties of Lasso

Here we 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 presparsity for Lasso). *Suppose that Conditions* M *and* $RE(\bar{c})$ *hold and that* $\lambda \geq cn\|S\|_\infty$, *and let* $\widehat{m} = |\widehat{T} \setminus T|$. *For* $\bar{c} = (c+1)/(c-1)$, *we have 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 foregoing lemma 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 immediately yields the simple upper bound on the sparsity of the form

$$\widehat{m} \lesssim_P s\phi(n), \tag{3.3}$$

as obtained for examples of [4] and [16]. 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 [1] – the bound is not sharp. However, for this case we can construct a sharp sparsity bound by combining the preceding presparsity result with the following sublinearity property of the restricted sparse eigenvalues.

Lemma 3 (Sublinearity 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 proven by [3]. 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,

$$\widehat{m} \lesssim s. \tag{3.4}$$

Consequently, for these designs and penalty levels,the sparsity of Lasso is of the same order as that of the oracle, namely $\widehat{s} := |\widehat{T}| \leq s + \widehat{m} \lesssim s$, with high probability. This is because $\min_{m \in \mathcal{M}} \phi(m) \ll \phi(n)$ for these designs, which allows us to sharpen the previous sparsity bound (3.3) considered by [4] and [16]. Moreover, our new bound is comparable to the bounds of [26] 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 [26]).

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

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

$$\widetilde{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \widehat{Q}(\beta) : \beta_j = 0 \qquad \text{for each } j \in \widehat{T}^c. \tag{4.1}$$

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 incorrect 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*

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

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

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

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

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 that we can bound both $\|\widehat{\beta} - \beta_0\|_{2,n}$, the rate of convergence of the first-step estimator, and $\widehat{m}$, the number of incorrect regressors selected by the model selector. Second, note that if the selected model contains the true model, $T \subseteq \widehat{T}$, then we have $(B_n)_+ \wedge (C_n)_+ = C_n = 0$. In that case, B_n has no affect on the rate, 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 (i.e., $T \not\subseteq \widehat{T}$), then 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 that Condition* M *holds.*

(1) *For any* $\varepsilon > 0$, *there is a constant* K_ε *independent of* n *such that with probability at least* $1 - \varepsilon$,

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

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,*

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

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

The proof of this lemma in turn relies on the following maximal inequality, the proof of which involves the use of a Samorodnitsky–Talagrand type of 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*

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

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

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

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 apply our results on post-model selection estimators to the case where Lasso is the first-step estimator. Our previous generic results allow us to use the 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 has good theoretical performance despite (generally) imperfect selection of the model by Lasso.

Theorem 5 (Performance of OLS post-Lasso). *Suppose that 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*

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

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*

$$\begin{aligned}\|\widetilde{f} - f\|_{\mathbb{P}_n,2} \leq 3c_s + K'_{\varepsilon,\alpha}\sigma\left[\sqrt{\frac{\widehat{m}\log(pe\mu(\widehat{m}))}{n}} + \sqrt{\frac{s\log(e\mu(\widehat{m}))}{n}}\right] \\ + 1\{T \not\subseteq \widehat{T}\}\left[K'_{\varepsilon,\alpha}\sigma\sqrt{\frac{s\log p}{n}}\frac{1}{\kappa(1)} + c_s\right]\end{aligned} \tag{5.1}$$

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

This theorem provides a performance bound for OLS post-Lasso as a function of Lasso's sparsity (characterized by $\widehat{m}$), rate of convergence, and model selection ability. For common designs, this bound implies that OLS post-Lasso performs at least as well as Lasso and can be strictly better in some cases, and has a smaller regularization bias. We provide further theoretical comparisons in what follows, and give computational examples supporting these comparisons in the supplemental article [2]. It is also worth repeating here that performance bounds in other norms of interest follow immediately by the triangle inequality and by the 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* (3.2), *as n grows, we have that*

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

Remark 5.1 (Comparison of the performance of OLS post-Lasso and 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. Of note, this occurs despite the fact that Lasso

generally may 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 conditions and the approximation error does not dominate the estimation error, then the OLS post-Lasso rate improves on the rate of Lasso. Specifically, this occurs if condition (2.5) holds and $\widehat{m} = o_P(s)$ and $T \subseteq \widehat{T}$ wp $\to 1$, as under the conditions of Theorem 2 Part 1 or, in the case of perfect model selection, when $T = \widehat{T}$ wp $\to 1$, as under the conditions specified by [24]. In such cases, we know from Corollary 1 that the rates for Lasso are sharp and cannot be faster than $\sigma\sqrt{s \log p/n}$. Thus the faster rate of convergence of OLS post-Lasso over Lasso is strict in such cases.

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 (4.1) 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 that 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*

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

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*

$$\|\widetilde{f} - f\|_{\mathbb{P}_n,2} \leq 3c_s + K'_{\varepsilon,\alpha}\sigma\left[\sqrt{\frac{\widetilde{m}\log(pe\mu(\widetilde{m}))}{n}} + \sqrt{\frac{s\log(e\mu(\widetilde{m}))}{n}}\right] + 1\{T \not\subseteq \widetilde{T}\}\left[K'_{\varepsilon,\alpha}\sigma\sqrt{\frac{s\log p}{n}}\frac{1}{\kappa(1)} + c_s\right], \tag{5.2}$$

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

This theorem provides a performance bound for OLS post-fit Lasso as a function of its sparsity (characterized by $\widetilde{m}$), Lasso's rate of convergence, and the model selection ability of the thresholding scheme. Generally, this bound is as good as the bound for OLS post-Lasso, because the OLS post-fitness-thresholded Lasso thresholds as much as possible subject to maintaining a certain goodness of fit. Another appealing feature is that this estimator determines the thresholding level in a completely data-driven fashion. Moreover, by construction, the estimated model is

sparser than the OLS post-Lasso 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 supplemental article [2].

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* (3.2) *hold, then as n grows, the OLS post-fitness-thresholded Lasso satisfies*

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

Remark 5.2 (Comparison of the performance of OLS post-fit Lasso, 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 on Lasso's rate. That is, if the oracle model has coefficients well separated from 0 and the approximation error is not dominant, then 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 but the model selection works well, namely when $\widetilde{m} = o_P(\widehat{m})$ and $T \subseteq \widetilde{T}$ wp $\to 1$. Finally, under conditions ensuring perfect model selection – namely, the 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

We next consider the traditional thresholding scheme, which truncates to 0 all components below a set threshold, t. This is arguably the most widely 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 that 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*

$$\|\widetilde{f}-f\|_{\mathbb{P}_n,2} \le K_\varepsilon\sigma\sqrt{\frac{\widetilde{m}\log p+(\widetilde{m}+s)\log(\mathrm{e}\mu(\widetilde{m}))}{n}}+3c_s$$
$$+1\{T\not\subseteq\widetilde{T}\}\left(\gamma_t+\frac{1+c}{c}\frac{\lambda\sqrt{s}}{n\kappa(\bar{c})}+2c_s\right)+1\{T\not\subseteq\widetilde{T}\}$$
$$\times\sqrt{\left[K_\varepsilon\sigma\sqrt{\frac{\widetilde{m}\log p+(\widetilde{m}+s)\log(\mathrm{e}\mu(\widetilde{m}))}{n}}+2c_s\right]\left(\gamma_t+\frac{1+c}{c}\frac{\lambda\sqrt{s}}{n\kappa(\bar{c})}+2c_s\right)},$$

where $\gamma_t \le t\sqrt{\phi(m_t)m_t}$. *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* $1-\alpha-\varepsilon-\tau$,

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

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) the model selection ability of the thresholding scheme. Generally, this bound may be worse than the bound for Lasso, because the OLS post-thresholded Lasso potentially uses too much thresholding, resulting in large goodness-of-fit losses, γ_t. We provide further theoretical comparisons below and computational examples in Section 4 of the supplemental article [2].

Remark 5.3 (Comparison of the performance of OLS post-thresholded Lasso, Lasso, and OLS post-Lasso). In this work, we also assume conditions in (2.5) and (3.2) presented in the foregoing formal comparisons. Under these conditions, OLS post-thresholded Lasso obeys the bound

$$\|\widetilde{f}-f\|_{\mathbb{P}_n,2}\lesssim_P \sigma\sqrt{\frac{\widetilde{m}\log p}{n}}+\sigma\sqrt{\frac{s}{n}}+c_s+1\{T\not\subseteq\widetilde{T}\}\left(\gamma_t\vee\sigma\sqrt{\frac{s\log p}{n}}\right). \tag{5.3}$$

In this case, we have $\widetilde{m}\vee m_t\le s+\widehat{m}\lesssim_P s$ by Theorem 3. In general, the foregoing rate cannot improve on Lasso's rate of convergence given in Lemma 1.

As expected, the choice of t, which controls γ_t via the bound $\gamma_t\le t\sqrt{\phi(m_t)m_t}$, can have a significant effect on the performance bounds. If

$$t\lesssim\sigma\sqrt{\frac{\log p}{n}} \quad \text{then } \|\widetilde{f}-f\|_{\mathbb{P}_n,2}\lesssim_P \sigma\sqrt{\frac{s\log p}{n}}+c_s. \tag{5.4}$$

The choice (5.4), suggested by [13] and Theorem 3, is theoretically sound, because it guarantees that OLS post-thresholded Lasso achieves the near-oracle rate of Lasso. Note that to implement the choice (5.4) in practice, we suggest setting $t = \lambda/n$, given that the separation of the coefficients from 0 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 in the rate of convergence of Lasso are possible. Specifically, if $\widetilde{m} = o_P(s)$ and $T \subseteq \widetilde{T}$ wp $\to 1$, then OLS post-thresholded Lasso strictly improves on the Lasso's rate. Furthermore, if $\widetilde{m} = o_P(\widehat{m})$ and $T \subseteq \widetilde{T}$ wp $\to 1$, then OLS post-thresholded Lasso also outperforms OLS post-Lasso:

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

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

Appendix: Proofs

A.1. Proofs for Section 3

Proof of Theorem 1. The bound in $\|\cdot\|_{2,n}$ norm follows by the same steps specified by [4], and thus we defer 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 $1 - \tau$, because $c' \geq c/\ell$. 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 that $\|\delta_{T^c}\|_1 \leq 2\bar{c}\|\delta_T\|_1$. In this case, by the 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 $\|\delta\|_{2,n}$ because $\bar{c} > 1$. On the other hand, consider the case where $\|\delta_{T^c}\|_1 > 2\bar{c}\|\delta_T\|_1$. Here 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 (2.3) in the supplemental article [2], 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,

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

because $\|\mathbb{E}_n[x_\bullet r_\bullet]\|_\infty \le \min\{\frac{\sigma}{\sqrt{n}}, c_s\}$ by step 2 below.

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

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

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

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

The result follows by Theorem 1 to bound $\|\delta\|_1$ and the arguments of [4] and [13] to show that the bound on the correlations imply that for any $C > 0$,

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

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

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

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

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

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

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

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

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

using the definition of $\phi(\widehat{m})$ and the Holder inequality,

$$\begin{aligned}\left\|\left(X'(R + X\beta_0 - X\widehat{\beta})\right)_{\widehat{T}}\right\| &\le \sup_{\|\alpha_{T^c}\|_0 \le \widehat{m}, \|\alpha\| \le 1} |\alpha' X'(R + X\beta_0 - X\widehat{\beta})| \\ &\le \sup_{\|\alpha_{T^c}\|_0 \le \widehat{m}, \|\alpha\| \le 1} \|\alpha' X'\| \|R + X\beta_0 - X\widehat{\beta}\| \\ &= \sup_{\|\alpha_{T^c}\|_0 \le \widehat{m}, \|\alpha\| \le 1} \sqrt{|\alpha' X' X \alpha|} \|R + X\beta_0 - X\widehat{\beta}\| \\ &= n\sqrt{\phi(\widehat{m})}\left(\mathbb{E}_n[(x_\bullet'\widehat{\beta} - f_\bullet)^2]\right)^{1/2}.\end{aligned}$$

Because $\lambda/c \ge n\|S\|_\infty$, we have

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

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

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

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

Proof of Theorem 3. By Lemma 2, $\sqrt{\widehat{m}} \le \sqrt{\phi(\widehat{m})} \cdot 2\bar{c}\sqrt{s}/\kappa(\bar{c}) + 3(\bar{c} + 1)\sqrt{\phi(\widehat{m})} \cdot nc_s/\lambda$, which, by letting $L_n = (\frac{2\bar{c}}{\kappa(\bar{c})} + 3(\bar{c} + 1)\frac{nc_s}{\lambda\sqrt{s}})^2$, can be rewritten as

$$\widehat{m} \le s \cdot \phi(\widehat{m})L_n. \tag{A.2}$$

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

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

Thus, because $\lceil k \rceil < 2k$ for any $k \ge 1$, we have $M < s \cdot 2\phi(M)L_n$, which violates the condition of $M \in \mathcal{M}$. Therefore, we must have $\widehat{m} \le M$. In turn, applying (A.2) once more with $\widehat{m} \le (M \wedge n)$, we obtain $\widehat{m} \le s \cdot \phi(M \wedge n)L_n$. 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 the definition of the second-step estimator, it follows that $\widehat{Q}(\widetilde{\beta}) \le \widehat{Q}(\widehat{\beta})$ and $\widehat{Q}(\widetilde{\beta}) \le \widehat{Q}(\beta_{0\widehat{T}})$. Thus,

$$\widehat{Q}(\widetilde{\beta}) - \widehat{Q}(\beta_0) \le \big(\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0)\big) \wedge \big(\widehat{Q}(\beta_{0\widehat{T}}) - \widehat{Q}(\beta_0)\big) \le B_n \wedge C_n.$$

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

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

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

Proof of Lemma 4. The proof of part (1) follows from the relation

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

and 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 $K_\varepsilon = 6\sqrt{2}\log^{1/2}\max\{e, D, 1/(\mathrm{e}^s\varepsilon[1 - 1/\mathrm{e}])\}$. The proof part (2) also follows from Lemma 5, but applying it with $s = 0$, $p = s$ (because 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 $\|\delta\| = 1$ because the function is homogenous of degree 0.

Step 1. For each nonnegative integer $m \le n$ and each set $\widetilde{T} \subset \{1, \dots, p\}$, with $|\widetilde{T} \setminus T| \le m$, define the class of functions

$$\mathcal{G}_{\widetilde{T}} = \{\epsilon_i x_i'\delta/\|\delta\|_{2,n} : \mathrm{support}(\delta) \subseteq \widetilde{T}, \|\delta\| = 1\}. \tag{A.3}$$

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

$$P\Big(\sup_{f\in\mathcal{F}_m} |\mathbb{G}_n(f)| \ge e_n(m,\eta)\Big) \le \binom{p}{m} \max_{|\widetilde{T}\setminus T|\le m} P\Big(\sup_{f\in\mathcal{G}_{\widetilde{T}}} |\mathbb{G}_n(f)| \ge e_n(m,\eta)\Big). \tag{A.4}$$

We apply the Samorodnitsky–Talagrand inequality (Proposition A.2.7 of van der Vaart and Wellner [23]) to bound the right-hand side of (A.4). Let

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

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

$$N(\varepsilon, \mathcal{G}_{\widetilde{T}}, \rho) \leq \left(6\sigma\mu(m)/\varepsilon\right)^{m+s} \qquad \text{for each } 0 < \varepsilon \leq \sigma, \tag{A.5}$$

and $\sigma^2(\mathcal{G}_{\widetilde{T}}) := \max_{f\in\mathcal{G}_{\widetilde{T}}} \mathrm{E}[\mathbb{G}_n(f)]^2 = \sigma^2$. Then, by the Samorodnitsky–Talagrand inequality,

$$P\Big(\sup_{f\in\mathcal{G}_{\widetilde{T}}} |\mathbb{G}_n(f)| \geq e_n(m,\eta)\Big) \leq \left(\frac{D\sigma\mu(m)e_n(m,\eta)}{\sqrt{m+s}\sigma^2}\right)^{m+s} \bar{\Phi}\big(e_n(m,\eta)/\sigma\big) \tag{A.6}$$

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 $P(\sup_{f\in\mathcal{G}_{\widetilde{T}}} |\mathbb{G}_n(f)| \geq e_n(m,\eta)) \leq \eta e^{-m-s}/\binom{p}{m}$ by simple substitution into (A.6).Then,

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

which proves the claim.

Step 2. This step establishes (A.5). 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

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

By definition of $\mathcal{G}_{\widetilde{T}}$ in (A.3), 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.3). Thus, by the definition of $RSE(m)$,

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

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

so that

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

Then the bound (A.5) follows from the bound of [23], 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 then 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 $\widehat{\delta} = \widehat{\beta} - \beta_0$,

$$B_n := \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) \leq \frac{\lambda}{n}(\|\beta_0\|_1 - \|\widehat{\beta}\|_1) \leq \frac{\lambda}{n}(\|\widehat{\delta}_T\|_1 - \|\widehat{\delta}_{T^c}\|_1). \tag{A.7}$$

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

$$B_n := \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) \leq \frac{\lambda}{n}\|\widehat{\delta}_T\|_1 \leq \frac{\lambda}{n}\frac{\sqrt{s}\|\widehat{\delta}\|_{2,n}}{\kappa(1)}. \tag{A.8}$$

The result follows by applying Theorem 1 to bound $\|\widehat{\delta}\|_{2,n}$, under the condition that $RE(1)$ holds, along with 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 also thank Denis Chetverikov, Brigham Fradsen, Joonhwan Lee, two refereers, and the associate editor for numerous suggestions that helped improve the articler. We thank Kengo Kato for pointing out the usefulness of the approach of [18] for bounding sparse eigenvalues of the empirical Gram matrix. We gratefully acknowledge the financial support from the National Science Foundation.

Supplementary Material

Supplementary material for Least squares after model selection in high-dimensional sparse models (DOI: 10.3150/11-BEJ410SUPP; .pdf). The online supplemental article [2] contains a

finite sample results for the estimation of σ, details regarding the oracle problem, omitted proofs, uniform control of sparse eigenvalues, and Monte Carlo experiments to access the performance of the estimators proposed in the paper.

References

[1] Belloni, A. and Chernozhukov, V. (2011). Supplement to "ℓ_1-penalized quantile regression in high-dimensional sparse models." DOI:10.1214/10-AOS827SUPP.

[2] Belloni, A. and Chernozhukov, V. (2012). Supplement to "Least squares after model selection in high-dimensional sparse models." DOI:10.3150/11-BEJ410SUPP.

[3] Belloni, A. and Chernozhukov, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.* **39** 82–130. MR2797841

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

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

[6] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2006). Aggregation and sparsity via l_1 penalized least squares. In *Learning Theory. Lecture Notes in Computer Science* **4005** 379–391. Berlin: Springer. MR2280619

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

[8] Bunea, F., Tsybakov, A.B. and Wegkamp, M.H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** 1674–1697. MR2351101

[9] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

[10] Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory, and Applications. Springer Series in Statistics*. New York: Springer. MR1705298

[11] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70** 849–911. MR2530322

[12] Koltchinskii, V. (2009). Sparsity in penalized empirical risk minimization. *Ann. Inst. Henri Poincaré Probab. Stat.* **45** 7–57. MR2500227

[13] Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.* **2** 90–102. MR2386087

[14] Lounici, K., Pontil, M., Tsybakov, A.B. and van de Geer, S. (2009). Taking advantage of sparsity in multi-task learning. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT 2009)* 73–82. Omnipress.

[15] Lounici, K., Pontil, M., Tsybakov, A.B. and van de Geer, S. (2012). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* To appear.

[16] Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351

[17] Rosenbaum, M. and Tsybakov, A.B. (2010). Sparse recovery under matrix uncertainty. *Ann. Statist.* **38** 2620–2651. MR2722451

[18] Rudelson, M. and Vershynin, R. (2008). On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.* **61** 1025–1045. MR2417886

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

[20] Tsybakov, A.B. (2008). *Introduction to Nonparametric Estimation*. Berlin: Springer.

[21] van de Geer, S.A. (2000). *Empirical Processes in M-Estimation*. Cambridge: Cambridge Univ. Press.
[22] van de Geer, S.A. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36** 614–645. MR2396809
[23] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes. Springer Series in Statistics*. New York: Springer. MR1385671
[24] Wainwright, M.J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873
[25] Wasserman, L. (2006). *All of Nonparametric Statistics. Springer Texts in Statistics*. New York: Springer. MR2172729
[26] Zhang, C.H. and Huang, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448
[27] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

Received April 2010 and revised June 2011