Massachusetts Institute of Technology
Department of Economics
Working Paper Series

HIGH DIMENSIONAL SPARSE ECONOMETRIC MODELS: AN INTRODUCTION

Alexandre Belloni
Victor Chernozhukov

Working Paper 11-17
June 26, 2011

Room E52-251
50 Memorial Drive
Cambridge, MA 02142

High Dimensional Sparse Econometric Models: An Introduction

Alexandre Belloni and Victor Chernozhukov

Abstract In this chapter we discuss conceptually high dimensional sparse econometric models as well as estimation of these models using ℓ_1-penalization and post-ℓ_1-penalization methods. Focusing on linear and nonparametric regression frameworks, we discuss various econometric examples, present basic theoretical results, and illustrate the concepts and methods with Monte Carlo simulations and an empirical application. In the application, we examine and confirm the empirical validity of the Solow-Swan model for international economic growth.

1 The High Dimensional Sparse Econometric Model

We consider linear, high dimensional sparse (HDS) regression models in econometrics. The HDS regression model has a large number of regressors p, possibly much larger than the sample size n, but only a relatively small number $s < n$ of these regressors are important for capturing accurately the main features of the regression function. The latter assumption makes it possible to estimate these models effectively by searching for approximately the right set of the regressors, using ℓ_1-based penalization methods. In this chapter we will review the basic theoretical properties of these procedures, established in the works of [8, 10, 18, 17, 7, 15, 13, 27, 26], among others (see [20, 7] for a detailed literature review). In this section, we review the modeling foundations as well as motivating examples for these procedures, with emphasis on applications in econometrics.

Let us first consider an exact or parametric HDS regression model, namely,

Alexandre Belloni
Duke University, Fuqua School of Business, 100 Fuqua Drive, Durham, NC, e-mail: abn5@duke.edu

Victor Chernozhukov
Massachusetts Institute of Technology, Department of Economics, 50 Memorial Drive, Cambridge, MA e-mail: vchern@mit.edu

$$y_i = x_i'\beta_0 + \varepsilon_i, \;\; \varepsilon_i \sim N(0,\sigma^2), \;\; \beta_0 \in \mathbb{R}^p, \; i = 1,\ldots,n, \tag{1}$$

where y_i's are observations of the response variable, x_i's are observations of p-dimensional fixed regressors, and ε_i's are i.i.d. normal disturbances, where possibly $p \geqslant n$. The key assumption of the exact model is that the true parameter value β_0 is sparse, having only $s < n$ non-zero components with support denoted by

$$T = \text{ support}(\beta_0) \subset \{1,\ldots,p\}. \tag{2}$$

Next let us consider an approximate or nonparametric HDS model. To this end, let us introduce the regression model

$$y_i = f(z_i) + \varepsilon_i, \;\; \varepsilon_i \sim N(0,\sigma^2), \;\; i = 1,\ldots,n, \tag{3}$$

where y_i is the outcome, z_i is a vector of elementary fixed regressors, $z \mapsto f(z)$ is the true, possibly non-linear, regression function, and ε_i's are i.i.d. normal disturbances. We can convert this model into an approximate HDS model by writing

$$y_i = x_i'\beta_0 + r_i + \varepsilon_i, \;\; i = 1,\ldots,n, \tag{4}$$

where $x_i = P(z_i)$ is a p-dimensional regressor formed from the elementary regressors by applying, for example, polynomial or spline transformations, β is a conformable parameter vector, whose "true" value β_0 has only $s < n$ non-zero components with support denoted as in (2), and $r_i := r(z_i) = f(z_i) - x_i'\beta_0$ is the approximation error. We shall define the true value β_0 more precisely in the next section. For now, it is important to note only that we assume there exists a value β_0 having only s non-zero components that sets the approximation error r_i to be small.

Before considering estimation, a natural question is whether exact or approximate HDS models make sense in econometric applications. In order to answer this question it is helpful to consider the following example, in which we abstract from estimation completely and only ask whether it is possible to accurately describe some structural econometric function $f(z)$ using a low-dimensional approximation of the form $P(z)'\beta_0$. In particular, we are interested in improving upon the conventional low-dimensional approximations.

Example 1: Sparse Models for Earning Regressions. In this example we consider a model for the conditional expectation of log-wage y_i given education z_i, measured in years of schooling. Since measured education takes on a finite number of years, we can expand the conditional expectation of wage y_i given education z_i:

$$E[y_i|z_i] = \sum_{j=1}^{p} \beta_{0j} P_j(z_i), \tag{5}$$

using some dictionary of approximating functions $P_1(z_i),\ldots,P_p(z_i)$, such as polynomial or spline transformations in z_i and/or indicator variables for levels of z_i. In fact, since we can consider an overcomplete dictionary, the representation of the function may not be unique, but this is not important for our purposes.

A conventional sparse approximation employed in econometrics is, for example,

$$f(z_i) := E[y_i|z_i] = \tilde{\beta}_1 P_1(z_i) + \cdots + \tilde{\beta}_s P_s(z_i) + \tilde{r}_i, \tag{6}$$

where the P_j's are low-order polynomials or splines, with typically $s = 4$ or 5 terms, but there is no guarantee that the approximation error $\tilde{r}_i$ in this case is small, or that these particular polynomials form the best possible s-dimensional approximation. Indeed, we might expect the function $E[y_i|z_i]$ to exhibit oscillatory behavior near the schooling levels associated with advanced degrees, such as MBA or MD. Low-degree polynomials may not be able to capture this behavior very well, resulting in large approximation errors $\tilde{r}_i$'s.

Therefore, the question is: With the same number of parameters, can we find a much better approximation? In other words, can we find some higher-order terms in the expansion (5) which will provide a higher-quality approximation? More specifically, can we construct an approximation

$$f(z_i) := E[y_i|z_i] = \beta_{k_1} P_{k_1}(z_i) + \cdots + \beta_{k_s} P_{k_s}(z_i) + r_i, \tag{7}$$

for some regressor indices $k_1, \ldots, k_s$ selected from $\{1, \ldots, p\}$, that is accurate and much better than (6), in the sense of having a much smaller approximation error r_i?

Obviously the answer to the latter question depends on how complex the behavior of the true regression function (5) is. If the behavior is not complex, then low-dimensional approximation should be accurate. Moreover, it is clear that the second approximation (7) is weakly better than the first (6), and can be much better if there are some important high-order terms in (5) that are completely missed by the first approximation. Indeed, in the context of the earning function example, such important high-order terms could capture abrupt positive changes in earning associated with advanced degrees such as MBA or MD. Thus, the answer to the question depends strongly on the empirical context.

Consider for example the earnings of prime age white males in the 2000 U.S. Census (see e.g., Angrist, Chernozhukov and Fernandez-Val [2]). Treating this data as the population data, we can then compute $f(z_i) = E[y_i|z_i]$ without error. Figure 1 plots this function. (Of course, such a strategy is not generally available in the empirical work, since the population data are generally not available.) We then construct two sparse approximations and also plot them in Figure 1: the first is the conventional one, of the form (6), with $P_1, \ldots, P_s$ representing an $(s-1)$-degree polynomial, and the second is an approximation of the form (7), with P_{k_1}, ..., P_{k_s} consisting of a constant, a linear term, and two linear splines terms with knots located at 16 and 19 years of schooling (in the case of $s = 5$ a third knot is located at 17). In fact, we find the latter approximation automatically using ℓ_1-penalization methods, although in this special case we could construct such an approximation just by eye-balling Figure 1 and noting that most of the function is described by a linear function, with a few abrupt changes that can be captured by linear spline terms that induce large changes in slope near 17 and 19 years of schooling. Note that an exhaustive search for a low-dimensional approximation requires looking at a very large set of models. We avoided this exhaustive search by using ℓ_1-penalized

least squares (LASSO), which penalizes the size of the model through the sum of absolute values of regression coefficients. Table 1 quantifies the performance of the different sparse approximations. (Of course, a simple strategy of eye-balling also works in this simple illustrative setting, but clearly does not apply to more general examples with several conditioning variables z_i, for example, when we want to condition on education, experience, and age.) □

Sparse Approximation	s	L_2 error	L_∞ error
Conventional	4	0.1212	0.2969
Conventional	5	0.1210	0.2896
LASSO	4	0.0865	0.1443
LASSO	5	0.0752	0.1154
Post-LASSO	4	0.0586	0.1334
Post-LASSO	5	0.0397	0.0788

Table 1 Errors of Conventional and the LASSO-based Sparse Approximations of the Earning Function. The LASSO estimator minimizes the least squares criterion plus the ℓ_1-norm of the coefficients scaled by a penalty parameter λ. As shown later, it turns out to have only a few non-zero components. The Post-LASSO estimator minimizes the least squares criterion over the non-zero components selected by the LASSO estimator.

The next two applications are natural examples with large sets of regressors among which we need to select some smaller sets to be used in further estimation and inference. These examples illustrate the potential wide applicability of HDS modeling in econometrics, since many classical and new data sets have naturally multi-dimensional regressors. For example, the American Housing Survey records prices and multi-dimensional features of houses sold, and scanner data-sets record prices and multi-dimensional information on products sold at a store or on the internet.

Example 2: Instrument Selection in Angrist and Krueger Data. The second example we consider is an instrumental variables model, as in Angrist and Krueger [3]

$$\begin{aligned} y_{i1} &= \theta_0 + \theta_1 y_{i2} + w_i'\gamma + v_i, & E[v_i|w_i,x_i] &= 0, \\ y_{i2} &= x_i'\beta + w_i'\delta + \varepsilon_i, & E[\varepsilon_i|w_i,x_i] &= 0, \end{aligned}$$

where, for person i, y_{i1} denotes wage, y_{i2} denotes education, w_i denotes a vector of control variables, and x_i denotes a vector of instrumental variables that affect education but do not directly affect the wage. The instruments x_i come from the quarter-of-birth dummies, and from a very large list, total of 180, formed by interacting quarter-of-birth dummies with control variables w_i. The interest focuses on measuring the coefficient θ_1, which summarizes the causal impact of education on earnings, via instrumental variable estimators.

There are two basic options used in the literature: one uses just the quarter-of-birth dummies, that is, the leading 3 instruments, and another uses all 183 instru-

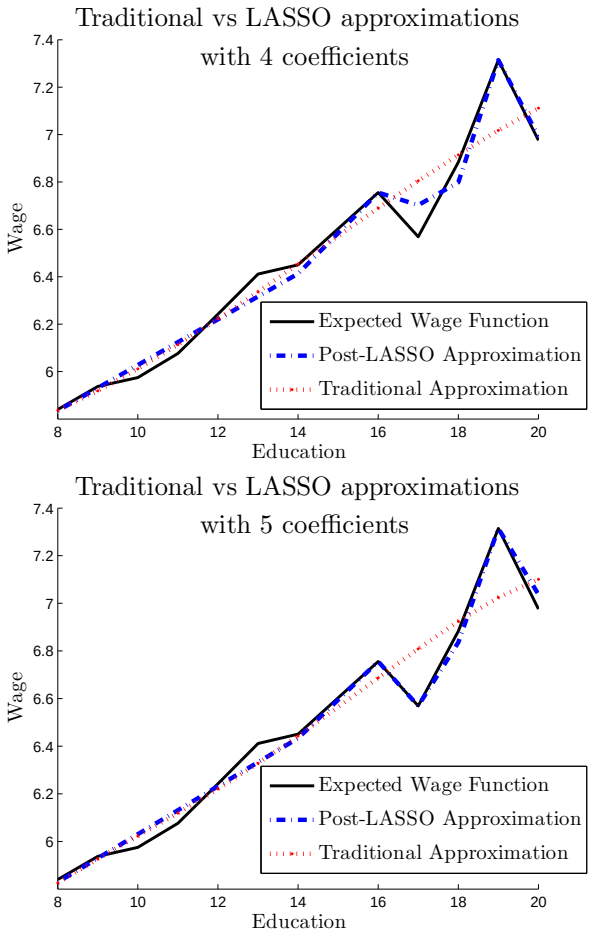

Fig. 1 The figures illustrates the Post-LASSO sparse approximation and the traditional (low degree polynomial) approximation of the wage function. The top figure uses $s = 4$ and the bottom figure uses $s = 5$.

ments. It is well known that using just 3 instruments results in estimates of the schooling coefficient θ_1 that have a large variance and small bias, while using 183 instruments results in estimates that have a much smaller variance but (potentially) large bias, see, e.g., [14]. It turns out that, under some conditions, by using ℓ_1-based estimation of the first stage, we can construct estimators that also have a nearly efficient variance and at the same time small bias. Indeed, as shown in Table 2, using the LASSO estimator induced by different penalty levels defined in Section 2, it is possible to find just 37 instruments that contain nearly all information in the first stage equation. Limiting the number of the instruments from 183 to just 37 reduces the bias of the final instrumental variable estimator. For a further analysis of IV estimates based on LASSO-selected instruments, we refer the reader to [6].

Table 2 Instrumental Variable Estimates of Return to Schooling in Angrist and Krueger Data

Instruments	Return to Schooling	Robust Std Error
3	0.1077	0.0201
180	0.0928	0.0144
LASSO-selected		
5	0.1062	0.0179
7	0.1034	0.0175
17	0.0946	0.0160
37	0.0963	0.0143

□

Example 3: Cross-country Growth Regression. One of the central issues in the empirical growth literature is estimating the effect of an initial (lagged) level of GDP (Gross Domestic Product) per capita on the growth rates of GDP per capita. In particular, a key prediction from the classical Solow-Swan-Ramsey growth model is the hypothesis of convergence, which states that poorer countries should typically grow faster and therefore should tend to catch up with the richer countries. Such a hypothesis implies that the effect of the initial level of GDP on the growth rate should be negative. As pointed out in Barro and Sala-i-Martin [5], this hypothesis is rejected using a simple bivariate regression of growth rates on the initial level of GDP. (In this data set, linear regression yields an insignificant positive coefficient of 0.0013.) In order to reconcile the data and the theory, the literature has focused on estimating the effect *conditional* on the pertinent characteristics of countries. Covariates that describe such characteristics can include variables measuring education and science policies, strength of market institutions, trade openness, savings rates and others [5]. The theory then predicts that for countries with similar other characteristics the effect of the initial level of GDP on the growth rate should be negative ([5]). Thus, we are interested in a specification of the form:

$$y_i = \alpha_0 + \alpha_1 \log G_i + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i, \tag{8}$$

where y_i is the growth rate of GDP over a specified decade in country i, G_i is the initial level of GDP at the beginning of the specified period, and the X_{ij}'s form a long list of country i's characteristics at the beginning of the specified period. We are interested in testing the hypothesis of convergence, namely that $\alpha_1 < 0$.

Given that in standard data-sets, such as Barro and Lee data [4], the number of covariates p we can condition on is large, at least relative to the sample size n, covariate selection becomes a crucial issue in this analysis ([16], [22]). In particular, previous findings came under severe criticism for relying on ad hoc procedures for covariate selection. In fact, in some cases, all of the previous findings have been questioned ([16]). Since the number of covariates is high, there is no simple way to resolve the model selection problem using only classical tools. Indeed the number of possible lower-dimensional models is very large, although [16] and [22] attempt to search over several millions of these models. We suggest ℓ_1-penalization and post-

ℓ_1-penalization methods to address this important issue. In Section 8, using these methods we estimate the growth model (8) and indeed find rather strong support for the hypothesis of convergence, thus confirming the basic implication of the Solow-Swan model. □

Notation. In what follows, all parameter values are indexed by the sample size n, but we omit the index whenever this does not cause confusion. In making asymptotic statements, we assume that $n \to \infty$ and $p = p_n \to \infty$, and we also allow for $s = s_n \to \infty$. 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 $\|\cdot\|$ 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 $j \notin T$. We also use standard notation in the empirical process literature,

$$\mathbb{E}_n[f] = \mathbb{E}_n[f(w_i)] = \sum_{i=1}^{n} f(w_i)/n,$$

and we use the notation $a \lesssim b$ to denote $a \leqslant cb$ for some constant $c > 0$ that does not depend on n; and $a \lesssim_P b$ to denote $a = O_P(b)$. Moreover, for two random variables X, Y we say that $X =_d Y$ if they have the same probability distribution. We also define the prediction norm associated with the empirical Gram matrix $\mathbb{E}_n[x_i x_i']$ as

$$\|\delta\|_{2,n} = \sqrt{\mathbb{E}_n[(x_i'\delta)^2]}.$$

2 The Setting and Estimators

2.1 The Model

Throughout the rest of the chapter we consider the nonparametric model introduced in the previous section:

$$y_i = f(z_i) + \varepsilon_i, \;\; \varepsilon_i \sim N(0, \sigma^2), \;\; i = 1, \ldots, n, \tag{9}$$

where y_i is the outcome, z_i is a vector of fixed regressors, and ε_i's are i.i.d. disturbances. Define $x_i = P(z_i)$, where $P(z_i)$ is a p-vector of transformations of z_i, including a constant, and $f_i = f(z_i)$. For a conformable sparse vector β_0 to be defined below, we can rewrite (9) in an approximately parametric form:

$$y_i = x_i'\beta_0 + u_i, \;\; u_i = r_i + \varepsilon_i, \;\; i = 1, \ldots, n, \tag{10}$$

where $r_i := f_i - x_i'\beta_0$, $i = 1, \ldots, n$, are approximation errors. We note that in the parametric case, we may naturally choose $x_i'\beta_0 = f_i$ so that $r_i = 0$ for all $i = 1, \ldots, n$. In the nonparametric case, we shall choose $x_i'\beta_0$ as a sparse parametric model that yields a good approximation to the true regression function f_i in equation (9).

Given (10), our target in estimation will become the parametric function $x_i'\beta_0$. Here we emphasize that the ultimate target in estimation is, of course, f_i, while $x_i'\beta_0$ is a convenient intermediate target, introduced so that we can approach the estimation problem as if it were parametric. Indeed, the two targets are equal up to approximation errors r_i's that will be set smaller than estimation errors. Thus, the problem of estimating the parametric target $x_i'\beta_0$ is equivalent to the problem of estimating the non-parametric target f_i modulo approximation errors.

With that in mind, we choose our target or "true" β_0, with the corresponding cardinality of its support

$$s = \|\beta_0\|_0,$$

as any solution to the following ideal risk minimization or oracle problem:

$$\min_{\beta \in \mathbb{R}^p} \mathbb{E}_n[(f_i - x_i'\beta)^2] + \sigma^2 \frac{\|\beta\|_0}{n}. \tag{11}$$

We call this problem the oracle problem for the reasons explained below, and we call

$$T = \text{support}(\beta_0)$$

the oracle or the "true" model. Note that we necessarily have that $s \leqslant n$.

The oracle problem (11) balances the approximation error $\mathbb{E}_n[(f_i - x_i'\beta)^2]$ over the design points with the variance term $\sigma^2\|\beta\|_0/n$, where the latter is determined by the number of non-zero coefficients in β. Letting

$$c_s^2 := \mathbb{E}_n[r_i^2] = \mathbb{E}_n[(f_i - x_i'\beta_0)^2]$$

denote the average square error from approximating values f_i by $x_i'\beta_0$, the quantity $c_s^2 + \sigma^2 s/n$ is the optimal value of (11). Typically, the optimality in (11) would balance the approximation error with the variance term so that for some absolute constant $K \geqslant 0$

$$c_s \leqslant K\sigma\sqrt{s/n}, \tag{12}$$

so that $\sqrt{c_s^2 + \sigma^2 s/n} \lesssim \sigma\sqrt{s/n}$. Thus, the quantity $\sigma\sqrt{s/n}$ becomes the ideal goal for the rate of convergence. If we knew the oracle model T, we would achieve this rate by using the oracle estimator, the least squares estimator based on this model, but we in general do not know T, since we do not observe the f_i's to attempt to solve the oracle problem (11). Since T is unknown, we will not be able to achieve the exact oracle rates of convergence, but we can hope to come close to this rate.

We consider the case of fixed design, namely we treat the covariate values $x_1, \ldots, x_n$ as fixed. This includes random sampling as a special case; indeed, in this case $x_1, \ldots, x_n$ represent a realization of this sample on which we condition throughout. Without loss of generality, we normalize the covariates so that

$$\widehat{\sigma}_j^2 = \mathbb{E}_n[x_{ij}^2] = 1 \text{ for } j = 1, \ldots, p. \tag{13}$$

We summarize the setup as the following condition.

Condition ASM. *We have data $\{(y_i, z_i), i = 1, \ldots, n\}$ that for each n obey the regression model (9), which admits the approximately sparse form (10) induced by (11) with the approximation error satisfying (12). The regressors $x_i = P(z_i)$ are normalized as in (13).*

Remark 1 (On the Oracle Problem). Let us now briefly explain what is behind problem (11). Under some mild assumptions, this problem directly arises as the (infeasible) oracle risk minimization problem. Indeed, consider an OLS 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 $\widehat{\beta}[\widetilde{T}] = \mathbb{E}_n[x_i[\widetilde{T}]x_i[\widetilde{T}]']^{-}\mathbb{E}_n[x_i[\widetilde{T}]y_i]$. The expected risk of this estimator $\mathbb{E}_n E[f_i - x_i[\widetilde{T}]'\widehat{\beta}[\widetilde{T}]]^2$ is equal to

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

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

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

by choosing the best model or the oracle model T. This problem is in fact equivalent to (11), provided that $\mathrm{rank}\,(\mathbb{E}_n[x_i[T]x_i[T]']) = \|\beta_0\|_0$, i.e. full rank. Thus, in this case the value β_0 solving (11) is the expected value of the oracle least squares estimator $\widehat{\beta}_T = \mathbb{E}_n[x_i[T]x_i[T]']^{-1}\mathbb{E}_n[x_i[T]y_i]$, i.e. $\beta_0 = \mathbb{E}_n[x_i[T]x_i[T]']^{-1}\mathbb{E}_n[x_i[T]\,f_i]$. 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.

2.2 LASSO and Post-LASSO Estimators

Having introduced the model (10) with the target parameter defined via (11), our task becomes to estimate β_0. We will focus on deriving rate of convergence results in the *prediction norm*, which measures the accuracy of predicting $x_i'\beta_0$ over the design points $x_1, \ldots, x_n$,

$$\|\delta\|_{2,n} = \sqrt{\mathbb{E}_n[x_i'\delta]^2}.$$

In what follows δ will denote deviations of the estimators from the true parameter value. Thus, e.g., for $\delta = \widehat{\beta} - \beta_0$, the quantity $\|\delta\|_{2,n}^2$ denotes the average of the

square errors $x_i'\widehat{\beta} - x_i'\beta_0$ resulting from using the estimate $x_i'\widehat{\beta}$ instead of $x_i'\beta_0$. Note that once we bound $\widehat{\beta} - \beta_0$ in the prediction norm, we can also bound the empirical risk of predicting values f_i by $x_i'\widehat{\beta}$ via the triangle inequality:

$$\sqrt{\mathbb{E}_n[(x_i'\widehat{\beta} - f_i)^2]} \leqslant \|\widehat{\beta} - \beta_0\|_{2,n} + c_s. \tag{14}$$

In order to discuss estimation consider first the classical ideal AIC/BIC type estimator ([1, 23]) that solves the empirical (feasible) analog of the oracle problem:

$$\min_{\beta \in \mathbb{R}^p} \widehat{Q}(\beta) + \frac{\lambda}{n}\|\beta\|_0,$$

where $\widehat{Q}(\beta) = \mathbb{E}_n[(y_i - x_i'\beta)^2]$ and $\|\beta\|_0 = \sum_{j=1}^p 1\{|\beta_j| > 0\}$ is the ℓ_0-norm and λ is the penalty level. This estimator has very attractive theoretical properties, but unfortunately it is computationally prohibitive, since the solution to the problem may require solving $\sum_{k \leqslant n} \binom{p}{k}$ least squares problems (generically, the complexity of this problem is NP-hard [19, 12]).

One way to overcome the computational difficulty is to consider a convex relaxation of the preceding problem, namely to employ the closest convex penalty – the ℓ_1 penalty – in place of the ℓ_0 penalty. This construction leads to the so called LASSO estimator:[1]

$$\widehat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \widehat{Q}(\beta) + \frac{\lambda}{n}\|\beta\|_1, \tag{15}$$

where as before $\widehat{Q}(\beta) = \mathbb{E}_n[(y_i - x_i'\beta)^2]$ and $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$. The LASSO estimator minimizes a convex function. Therefore, from a computational complexity perspective, (15) is a computationally efficient (i.e. solvable in polynomial time) alternative to AIC/BIC estimator.

In order to describe the choice of λ, we highlight that the following key quantity determining this choice:

$$S = 2\mathbb{E}_n[x_i\varepsilon_i],$$

which summarizes the noise in the problem. We would like to choose the smaller penalty level so that

$$\lambda \geqslant cn\|S\|_\infty \text{ with probability at least } 1-\alpha, \tag{16}$$

where $1-\alpha$ needs to be close to one, and c is a constant such that $c > 1$. Following [7] and [8], respectively, we consider two choices of λ that achieve the above:

$$X\text{-independent penalty:} \quad \lambda := 2c\sigma\sqrt{n}\Phi^{-1}(1-\alpha/2p), \tag{17}$$

$$X\text{-dependent penalty:} \quad \lambda := 2c\sigma\Lambda(1-\alpha|X), \tag{18}$$

where $\alpha \in (0,1)$ and $c > 1$ is constant, and

[1] The abbreviation LASSO stands for Least Absolute Shrinkage and Selection Operator, c.f. [24].

$$\Lambda(1-\alpha|X) := (1-\alpha)-\text{quantile of } n\|S/(2\sigma)\|_\infty,$$

conditional on $X = (x_1,\ldots,x_n)'$. Note that

$$\|S/(2\sigma)\|_\infty =_d \max_{1\leqslant j\leqslant p} |\mathbb{E}_n[x_{ij}g_i]|, \text{ where } g_i\text{'s are i.i.d. } N(0,1),$$

conditional on X, so we can compute $\Lambda(1-\alpha|X)$ simply by simulating the latter quantity, given the fixed design matrix X. Regarding the choice of α and c, asymptotically we require $\alpha \to 0$ as $n \to \infty$ and $c > 1$. Non-asymptotically, in our finite-sample experiments, $\alpha = .1$ and $c = 1.1$ work quite well. The noise level σ is unknown in practice, but we can estimate it consistently using the approach of Section 6. We recommend the X-dependent rule over the X-independent rule, since the former by construction adapts to the design matrix X and is less conservative than the latter in view of the following relationship that follows from Lemma 8:

$$\Lambda(1-\alpha|X) \leqslant \sqrt{n}\Phi^{-1}(1-\alpha/2p) \leqslant \sqrt{2n\log(2p/\alpha)}. \tag{19}$$

Regularization by the ℓ_1-norm employed in (15) naturally helps the LASSO estimator to avoid overfitting the data, but it also shrinks the fitted coefficients towards zero, causing a potentially significant bias. In order to remove some of this bias, let us consider the Post-LASSO estimator that applies ordinary least squares regression to the model $\widehat{T}$ selected by LASSO. Formally, set

$$\widehat{T} = \text{support}(\widehat{\beta}) = \{j \in \{1,\ldots,p\} \ : \ |\widehat{\beta}_j| > 0\},$$

and define the Post-LASSO estimator $\widetilde{\beta}$ as

$$\widetilde{\beta} \in \arg\min_{\beta\in\mathbb{R}^p} \widehat{Q}(\beta) : \beta_j = 0 \text{ for each } j \in \widehat{T}^c, \tag{20}$$

where $\widehat{T}^c = \{1,...,p\} \setminus \widehat{T}$. In words, the estimator is ordinary least squares applied to the data after removing the regressors that were not selected by LASSO. If the model selection works perfectly – that is, $\widehat{T} = T$ – then the Post-LASSO estimator is simply the oracle estimator whose properties are well known. However, perfect model selection might be unlikely for many designs of interest, so we are especially interested in the properties of Post-LASSO in such cases, namely when $\widehat{T} \neq T$, especially when $T \nsubseteq \widehat{T}$.

2.3 Intuition and Geometry of LASSO and Post-LASSO

In this section we discuss the intuition behind LASSO and Post-LASSO estimators defined above. We shall rely on a dual interpretation of the LASSO optimization problem to provide some geometrical intuition for the performance of LASSO. Indeed, it can be seen that the LASSO estimator also solves the following optimization

program:

$$\min_{\beta\in\mathbb{R}^p} \|\beta\|_1 : \widehat{Q}(\beta) \leqslant \gamma \tag{21}$$

for some value of $\gamma \geqslant 0$ (that depends on the penalty level λ). Thus, the estimator minimizes the ℓ_1-norm of coefficients subject to maintaining a certain goodness-of-fit; or, geometrically, the LASSO estimator searches for a minimal ℓ_1-ball – the diamond– subject to the diamond having a non-empty intersection with a fixed lower contour set of the least squares criterion function – the ellipse.

In Figure 2 we show an illustration for the two-dimensional case with the true parameter value (β_{01},β_{02}) equal $(1,0)$, so that $T = \text{support}(\beta_0) = \{1\}$ and $s = 1$. In the figure we plot the diamonds and ellipses. In the top figure, the ellipse represents a lower contour set of the population criterion function $Q(\beta) = E[(y_i - x_i'\beta)^2]$ in the zero noise case or the infinite sample case. In the bottom figures the ellipse represents a contour set of the sample criterion function $\widehat{Q}(\beta) = \mathbb{E}_n[(y_i - x_i'\beta)^2]$ in the non-zero noise or the finite sample case. The set of optimal solutions $\widehat{\beta}$ for LASSO is then given by the intersection of the minimal diamonds with the ellipses. Finally, recall that Post-LASSO is computed as the ordinary least square solution using covariates selected by LASSO. Thus, Post-LASSO estimate $\widetilde{\beta}$ is given by the center of the ellipse intersected with the linear subspace selected by LASSO.

In the zero-noise case or in population (top figure), LASSO easily recovers the correct sparsity pattern of β_0. Note that due to the regularization, in spite of the absence of noise, the LASSO estimator has a large bias towards zero. However, in this case Post-LASSO $\widetilde{\beta}$ removes the bias and recovers β_0 perfectly.

In the non-zero noise case (middle and bottom figures), the contours of the criterion function and its center move away from the population counterpart. The empirical error in the middle figure moves the center of the ellipse to a non-sparse point. However, LASSO correctly sets $\widehat{\beta}_2 = 0$ and $\widehat{\beta}_1 \neq 0$ recovering the sparsity pattern of β_0. Using the selected support, Post-LASSO $\widetilde{\beta}$ becomes the oracle estimator which drastically improves upon LASSO. In the case of the bottom figure, we have large empirical errors that push the center of the lower contour set further away from the population counterpart. These large empirical errors make the LASSO estimator non-sparse, incorrectly setting $\widehat{\beta}_2 \neq 0$. Therefore, Post-LASSO uses $\widehat{T} = \{1,2\}$ and does not use the exact support $T = \{1\}$. Thus, Post-LASSO is not the oracle estimator in this case.

All three figures also illustrate the shrinkage bias towards zero in the LASSO estimator that is introduced by the ℓ_1-norm penalty. The Post-LASSO estimator is motivated as a solution to remove (or at least alleviate) this shrinkage bias. In cases where LASSO achieves a good sparsity pattern, Post-LASSO can drastically improve upon LASSO.

2.4 Primitive conditions

In both the parametric and non-parametric models described above, whenever $p > n$, the empirical Gram matrix $\mathbb{E}_n[x_i x_i']$ does not have full rank and hence it is not well-behaved. However, we only need good behavior of certain moduli of continuity of the Gram matrix called restricted sparse eigenvalues. We define the minimal restricted sparse eigenvalue

$$\kappa(m)^2 := \min_{\|\delta_{T^c}\|_0 \leq m, \delta \neq 0} \frac{\|\delta\|_{2,n}^2}{\|\delta\|^2}, \tag{22}$$

and the maximal restricted sparse eigenvalue as

$$\phi(m) := \max_{\|\delta_{T^c}\|_0 \leq m, \delta \neq 0} \frac{\|\delta\|_{2,n}^2}{\|\delta\|^2}, \tag{23}$$

where m is the upper bound on the number of non-zero components outside the support T. To assume that $\kappa(m) > 0$ requires that all empirical Gram submatrices formed by any m components of x_i in addition to the components in T are positive definite. It will be convenient to define the following sparse *condition* number associated with the empirical Gram matrix:

$$\mu(m) = \frac{\sqrt{\phi(m)}}{\kappa(m)}. \tag{24}$$

In order to state simplified asymptotic statements, we shall also invoke the following condition.

Condition RSE. *Sparse eigenvalues of the empirical Gram matrix are well behaved, in the sense that for* $m = m_n = s \log n$

$$\mu(m) \lesssim 1, \quad \phi(m) \lesssim 1, \quad 1/\kappa(m) \lesssim 1. \tag{25}$$

This condition holds with high probability for many designs of interest under mild conditions on s. For example, as shown in Lemma 1, when the covariates are Gaussians, the conditions in (25) are true with probability converging to one under the mild assumption that $s \log p = o(n)$. Condition RSE is likely to hold for other regressors with jointly light-tailed distributions, for instance log-concave distribution. As shown in Lemma 2, the conditions in (25) also hold for general bounded regressors under the assumption that $s(\log^4 n) \log(p \vee n) = o(n)$. Arbitrary bounded regressors often arise in non-parametric models, where regressors x_i are formed as spline, trigonometric, or polynomial transformations $P(z_i)$ of some elementary bounded regressors z_i.

Lemma 1 (Gaussian design). *Suppose* $\tilde{x}_i$, $i = 1, \ldots, n$, *are i.i.d. zero-mean Gaussian random vectors, such that the population design matrix* $E[\tilde{x}_i \tilde{x}_i']$ *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 $x_{ij} = \tilde{x}_{ij}/\sqrt{\mathbb{E}_n[\tilde{x}_{ij}^2]}$. Then for any $m \leqslant (s\log(n/e)) \wedge (n/[16\log p])$, with probability at least $1 - 2\exp(-n/16)$,

$$\phi(m) \leqslant 8\varphi, \;\; \kappa(m) \geqslant \kappa/6\sqrt{2}, \;\; \text{and} \;\; \mu(m) \leqslant 24\sqrt{\varphi}/\kappa.$$

Lemma 2 (Bounded design). *Suppose $\tilde{x}_i$, $i = 1,\ldots,n$, are i.i.d. vectors, such that the population design matrix $E[\tilde{x}_i\tilde{x}_i']$ 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 $x_{ij} = \tilde{x}_{ij}/(\mathbb{E}_n[\tilde{x}_{ij}^2])^{1/2}$. 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 $m \geq 0$ such that $m + s \leq s\log n$, we have that as $n \to \infty$*

$$\phi(m) \leqslant 4\varphi, \;\; \kappa(m) \geqslant \kappa/2, \;\; \text{and} \;\; \mu(m) \leqslant 4\sqrt{\varphi}/\kappa,$$

with probability approaching 1.

For proofs, see [7]; the first lemma builds upon results in [26] and the second builds upon results in [21].

3 Analysis of LASSO

In this section we discuss the rate of convergence of LASSO in the prediction norm; our exposition follows mainly [8].

The key quantity in the analysis is the following quantity called "score":

$$S = S(\beta_0) = 2\mathbb{E}_n[x_i\varepsilon_i].$$

The score is the effective "noise" in the problem. Indeed, defining $\delta := \widehat{\beta} - \beta_0$, note that by the Hölder's inequality

$$\begin{aligned} \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) - \|\delta\|_{2,n}^2 &= -2\mathbb{E}_n[\varepsilon_i x_i'\delta] - 2\mathbb{E}_n[r_i x_i'\delta] \\ &\geqslant -\|S\|_\infty\|\delta\|_1 - 2c_s\|\delta\|_{2,n}. \end{aligned} \tag{26}$$

Intuition suggests that we need to majorize the "noise term" $\|S\|_\infty$ by the penalty level λ/n, so that the bound on $\|\delta\|_{2,n}^2$ will follow from a relation between the prediction norm $\|\cdot\|_{2,n}$ and the penalization norm $\|\cdot\|_1$ on a suitable set. Specifically, for any $c > 1$, it will follow that if

$$\lambda \geqslant cn\|S\|_\infty$$

and $\|\delta\|_{2,n} \geqslant 2c_s$, the vector δ will also satisfy

$$\|\delta_{T^c}\|_1 \leqslant \bar{c}\|\delta_T\|_1, \tag{27}$$

where $\bar{c} = (c+1)/(c-1)$. That is, in this case the error in the regularization norm outside the true support does not exceed $\bar{c}$ times the error in the true support. (In the case $\|\delta\|_{2,n} \leqslant 2c_s$ the inequality (27) may not hold, but the bound $\|\delta\|_{2,n} \leqslant 2c_s$ is already good enough.)

Consequently, the analysis of the rate of convergence of LASSO relies on the so-called restricted eigenvalue $\kappa_{\bar{c}}$, introduced in [8], which controls the modulus of continuity between the prediction norm $\|\cdot\|_{2,n}$ and the penalization norm $\|\cdot\|_1$ over the set of vectors $\delta \in \mathbb{R}^p$ that satisfy (27):

$$\kappa_{\bar{c}} := \min_{\|\delta_{T^c}\|_1 \leqslant \bar{c}\|\delta_T\|_1, \delta_T \neq 0} \frac{\sqrt{s}\|\delta\|_{2,n}}{\|\delta_T\|_1}, \qquad (\text{RE}(c))$$

where $\kappa_{\bar{c}}$ can depend on n. The constant $\kappa_{\bar{c}}$ is a crucial technical quantity in our analysis and we need to bound it away from zero. In the leading cases that condition RSE holds this will in fact be the case as the sample size grows, namely

$$1/\kappa_{\bar{c}} \lesssim 1. \tag{28}$$

Indeed, we can bound $\kappa_{\bar{c}}$ from below by

$$\kappa_{\bar{c}} \geqslant \max_{m \geqslant 0} \kappa(m)\left(1 - \mu(m)\bar{c}\sqrt{s/m}\right) \geqslant \kappa(s\log n)\left(1 - \mu(s\log n)\bar{c}\sqrt{1/\log n}\right)$$

by Lemma 10 stated and proved in the appendix. Thus, under the condition RSE, as n grows, $\kappa_{\bar{c}}$ is bounded away from zero since $\kappa(s\log n)$ is bounded away from zero and $\phi(s\log n)$ is bounded from above as in (25). Several other primitive assumptions can be used to bound $\kappa_{\bar{c}}$. We refer the reader to [8] for a further detailed discussion of lower bounds on $\kappa_{\bar{c}}$.

We next state a non-asymptotic performance bound for the LASSO estimator.

Theorem 1 (Non-Asymptotic Bound for LASSO). *Under condition ASM, the event $\lambda \geqslant cn\|S\|_\infty$ implies*

$$\|\widehat{\beta} - \beta_0\|_{2,n} \leqslant \left(1 + \frac{1}{c}\right)\frac{\lambda\sqrt{s}}{n\kappa_{\bar{c}}} + 2c_s, \tag{29}$$

where $c_s = 0$ in the parametric case, and $\bar{c} = (c+1)/(c-1)$. Thus, if $\lambda \geqslant cn\|S\|_\infty$ with probability at least $1-\alpha$, as guaranteed by either X-independent or X-dependent penalty levels (17) and (17), then the bound (29) occurs with probability at least $1-\alpha$.

The proof of Theorem 1 is given in the appendix. The theorem also leads to the following useful asymptotic bounds.

Corollary 1 (Asymptotic Bound for LASSO). *Suppose that conditions ASM and RSE hold. If λ is chosen according to either the X-independent or X-dependent rule*

specified in (17) and (18) with $\alpha = o(1)$, $\log(1/\alpha) \lesssim \log p$, *or more generally so that*

$$\lambda \lesssim_P \sigma\sqrt{n \log p} \text{ and } \lambda \geqslant c' n \|S\|_\infty \text{ wp } \to 1, \tag{30}$$

for some $c' > 1$, *then the following asymptotic bound holds:*

$$\|\widehat{\beta} - \beta_0\|_{2,n} \lesssim_P \sigma\sqrt{\frac{s \log p}{n}} + c_s.$$

The non-asymptotic and asymptotic bounds for the empirical risk immediately follow from the triangle inequality:

$$\sqrt{\mathbb{E}_n[(f_i - x_i'\widehat{\beta})^2]} \leqslant \|\widehat{\beta} - \beta_0\|_{2,n} + c_s. \tag{31}$$

Thus, the rate of convergence of $x_i'\widehat{\beta}$ to f_i coincides with the rate of convergence of the oracle estimator $\sqrt{c_s^2 + \sigma^2 s/n}$ up to a logarithmic factor of p. Nonetheless, the performance of LASSO can be considered optimal in the sense that under general conditions the oracle rate is achievable only up to logarithmic factor of p (see Donoho and Johnstone [11] and Rigollet and Tsybakov [20]), apart from very exceptional, stringent cases, in which it is possible to perform perfect or near-perfect model selection.

4 Model Selection Properties and Sparsity of LASSO

The purpose of this section is, first, to provide bounds (sparsity bounds) on the dimension of the model selected by LASSO, and, second, to describe some special cases where the model selected by LASSO perfectly matches the "true" (oracle) model.

4.1 Sparsity Bounds

Although perfect model selection can be seen as unlikely in many designs, sparsity of the LASSO estimator has been documented in a variety of designs. Here we describe the sparsity results obtained in [7]. Let us define

$$\widehat{m} := |\widehat{T} \setminus T| = \|\widehat{\beta}_{T^c}\|_0,$$

which is the number of unnecessary components or regressors selected by LASSO.

Theorem 2 (Non-Asymptotic Sparsity Bound for LASSO). *Suppose condition ASM holds. The event* $\lambda \geqslant cn\|S\|_\infty$ *implies that*

$$\widehat{m} \leqslant s \cdot \left[\min_{m \in \mathscr{M}} \phi(m \wedge n) \right] \cdot L,$$

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

Under Conditions ASM and RSE, for n sufficiently large we have $1/\kappa_{\bar{c}} \lesssim 1$, $c_s \lesssim \sigma\sqrt{s/n}$, and $\phi(s\log n) \lesssim 1$; and under the conditions of Corollary 1, $\lambda \geq c\sigma\sqrt{n}$ with probability approaching one. Therefore, we have that $L \lesssim_P 1$ and

$$s\log n > s\phi(s\log n) \cdot 2L, \text{ that is, } s\log n \in \mathscr{M}$$

with probability approaching one as n grows. Therefore, under these conditions we have

$$\min_{m \in \mathscr{M}} \phi(m \wedge n) \lesssim_P 1.$$

Corollary 2 (Asymptotic Sparsity Bound for LASSO). *Under the conditions of Corollary 1, we have that*

$$\widehat{m} \lesssim_P s. \tag{32}$$

Thus, using a penalty level that satisfies (30) LASSO's sparsity is asymptotically of the same order as the oracle sparsity, namely

$$\widehat{s} := |\widehat{T}| \leqslant s + \widehat{m} \lesssim_P s. \tag{33}$$

We note here that Theorem 2 is particularly helpful in designs in which $\min_{m \in \mathscr{M}} \phi(m) \ll \phi(n)$. This allows Theorem 2 to sharpen the sparsity bound of the form $\widehat{s} \lesssim_P s\phi(n)$ considered in [8] and [18]. The bound above is comparable to the bounds in [26] in terms of order of magnitude, but Theorem 2 requires a smaller penalty level λ which also does not depend on the unknown sparse eigenvalues as in [26].

4.2 Perfect Model Selection Results

The purpose of this section is to describe very special cases where perfect model selection is possible. Most results in the literature for model selection have been developed for the parametric case only ([18],[17]). Below we provide some results for the nonparametric models, which cover the parametric models as a special case.

Lemma 3 (Cases with Perfect Model Selection by Thresholded LASSO). *Suppose condition ASM holds. (1) If the non-zero coefficients of the oracle model are well separated from zero, that is*

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

then the oracle model is a subset of the selected model,

$$T := \text{support}(\beta_0) \subseteq \widehat{T} := \text{support}(\widehat{\beta}).$$

Moreover the oracle model T can be perfectly selected by applying hard-thresholding of level t to the LASSO estimator $\widehat{\beta}$:

$$T = \left\{ j \in \{1,\dots,p\} \ : \ |\widehat{\beta}_j| > t \right\}.$$

(2) In particular, if $\lambda \geqslant cn\|S\|_\infty$, then for $\widehat{m} = |\widehat{T} \setminus T| = \|\widehat{\beta}_{T^c}\|_0$ we have

$$\zeta \leqslant \left(1 + \frac{1}{c}\right) \frac{\lambda\sqrt{s}}{n\kappa_{\bar{c}}\kappa(\widehat{m})} + \frac{2c_s}{\kappa(\widehat{m})}.$$

(3) In particular, if $\lambda \geqslant cn\|S\|_\infty$, and there is a constant $U > 5\bar{c}$ such that the empirical Gram matrix satisfies $|\mathbb{E}_n[x_{ij}x_{ik}]| \leqslant 1/[Us]$ for all $1 \leqslant j < k \leqslant p$, then

$$\zeta \leqslant \frac{\lambda}{n} \cdot \frac{U + \bar{c}}{U - 5\bar{c}} + \min\left\{ \frac{\sigma}{\sqrt{n}}, c_s \right\} + \frac{6\bar{c}}{U - 5\bar{c}} \frac{c_s}{\sqrt{s}} + \frac{4\bar{c}}{U} \frac{n}{\lambda} \frac{c_s^2}{s}.$$

Thus, we see from parts (1) and (2) that perfect model selection is possible under strong assumptions on the coefficients' separation away from zero. We also see from part (3) that the strong separation of coefficients can be considerably weakened in exchange for a strong assumption on the maximal pairwise correlation of regressors. These results generalize to the nonparametric case the results of [17] and [18] for the parametric case in which $c_s = 0$.

Finally, 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, $\text{sign}(v) = v/|v|$ 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 $T \subset \{1,...,p\}$, let us denote $x_i[T] := \{x_{ij}, j \in T\}$.

Lemma 4 (Cases with Perfect Model Selection by LASSO). *Suppose condition ASM holds. We have perfect model selection for LASSO, $\widehat{T} = T$, if and only if*

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

The result follows immediately from the first order optimality conditions, see [25]. [27] and [9] provides further primitive sufficient conditions for perfect model selection for the parametric case in which $u_i = \varepsilon_i$. The conditions above might typically require a slightly larger choice of λ than (17) and larger separation from zero of the minimal non-zero coefficient $\min_{j\in T}|\beta_{0j}|$.

5 Analysis of Post-LASSO

Next we study the rate of convergence of the Post-LASSO estimator. Recall that for $\widehat{T} = \text{ support } (\widehat{\beta})$, the Post-LASSO estimator solves

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

It is clear that if the model selection works perfectly (as it will under some rather stringent conditions discussed in Section 4.2), that is, $T = \widehat{T}$, then this estimator is simply the oracle least squares estimator whose properties are well known. However, if the model selection does not work perfectly, that is, $T \neq \widehat{T}$, the resulting performance of the estimator faces two different perils: First, in the case where LASSO selects a model $\widehat{T}$ that does not fully include the true model T, we have a specification error in the second step. Second, if LASSO includes additional regressors outside T, these regressors were not chosen at random and are likely to be spuriously correlated with the disturbances, so we have a data-snooping bias in the second step.

It turns out that despite of the possible poor selection of the model, and the aforementioned perils this causes, the Post-LASSO estimator still performs well theoretically, as shown in [7]. Here we provide a proof similar to [6] which is easier generalize to non-Gaussian cases.

Theorem 3 (Non-Asymptotic Bound for Post-LASSO). *Suppose condition ASM holds. If $\lambda \geqslant cn\|S\|_\infty$ holds with probability at least $1-\alpha$, then for any $\gamma > 0$ there is a constant K_γ independent of n such that with probability at least $1-\alpha-\gamma$*

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \leqslant \frac{K_\gamma \sigma}{\kappa(\widehat{m})} \sqrt{\frac{s + \widehat{m}\log p}{n}} + 2c_s + 1\{T \not\subseteq \widehat{T}\} \sqrt{\frac{\lambda\sqrt{s}}{n\kappa_{\bar{c}}^-} \cdot \left(\frac{(1+c)\lambda\sqrt{s}}{cn\kappa_{\bar{c}}^-} + 2c_s \right)}.$$

This theorem provides a performance bound for Post-LASSO as a function of LASSO's sparsity characterized by $\widehat{m}$, LASSO's rate of convergence, and LASSO's model selection ability. For common designs this bound implies that Post-LASSO performs at least as well as LASSO, but it can be strictly better in some cases, and has a smaller shrinkage bias by construction.

Corollary 3 (Asymptotic Bound for Post-LASSO). *Suppose conditions of Corollary 1 hold. Then*

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \lesssim_P \sigma \sqrt{\frac{s \log p}{n}} + c_s. \tag{34}$$

If further $\widehat{m} = o(s)$ and $T \subseteq \widehat{T}$ with probability approaching one, then

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \lesssim_P \sigma \left[\sqrt{\frac{o(s)\log p}{n}} + \sqrt{\frac{s}{n}} \right] + c_s. \tag{35}$$

If $\widehat{T} = T$ with probability approaching one, then Post-LASSO achieves the oracle performance

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \lesssim_P \sigma\sqrt{s/n} + c_s. \tag{36}$$

It is also worth repeating here that finite-sample and asymptotic bounds in other norms of interest immediately follow by the triangle inequality and by definition of $\kappa(\widehat{m})$:

$$\sqrt{\mathbb{E}_n[(x_i'\widetilde{\beta} - f_i)^2]} \leqslant \|\widetilde{\beta} - \beta_0\|_{2,n} + c_s \text{ and } \|\widetilde{\beta} - \beta_0\| \leqslant \|\widetilde{\beta} - \beta_0\|_{2,n}/\kappa(\widehat{m}). \tag{37}$$

The corollary above shows that Post-LASSO achieves the same near-oracle rate as LASSO. Notably, this occurs despite the fact that LASSO may in general fail to correctly select the oracle model T as a subset, that is $T \not\subseteq \widehat{T}$. The intuition for this result is that any components of T that LASSO misses cannot be very important. This corollary also shows that in some special cases Post-LASSO strictly improves upon LASSO's rate. Finally, note that Corollary 3 follows by observing that under the stated conditions,

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \lesssim_P \sigma\left[\sqrt{\frac{\widehat{m}\log p}{n}} + \sqrt{\frac{s}{n}} + 1\{T \not\subseteq \widehat{T}\}\sqrt{\frac{s\log p}{n}}\right] + c_s. \tag{38}$$

6 Estimation of Noise Level

Our specification of penalty levels (18) and (17) require the practitioner to know the noise level σ of the disturbances or at least estimate it. The purpose of this section is to propose the following method for estimating σ. First, we use a conservative estimate $\widehat{\sigma}^0 = \sqrt{\mathrm{Var}_n[y_i]} := \sqrt{\mathbb{E}_n[(y_i - \bar{y})^2]}$, where $\bar{y} = \mathbb{E}_n[y_i]$, in place of σ^2 to obtain the initial LASSO and Post-LASSO estimates, $\widehat{\beta}$ and $\widetilde{\beta}$. The estimate $\widehat{\sigma}^0$ is conservative since $\widehat{\sigma}^0 = \sigma^0 + o_P(1)$ where $\sigma^0 = \sqrt{\mathrm{Var}[y_i]} \geqslant \sigma$, since x_i contains a constant by assumption. Second, we define the refined estimate $\widehat{\sigma}$ as

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

in the case of LASSO and

$$\widehat{\sigma} = \sqrt{\frac{n}{n-\widehat{s}} \cdot \widehat{Q}(\widetilde{\beta})}$$

in the case of Post-LASSO. In the latter case we employ the standard degree-of-freedom correction with $\widehat{s} = \|\widetilde{\beta}\|_0 = |\widehat{T}|$, and in the former case we need no additional corrections, since the LASSO estimate is already sufficiently regularized. Third, we use the refined estimate $\widehat{\sigma}^2$ to obtain the refined LASSO and Post-LASSO estimates $\widehat{\beta}$ and $\widetilde{\beta}$. We can stop here or further iterate on the last two steps.

Thus, the algorithm for estimating σ using LASSO is as follows:

Algorithm 1 (Estimation of σ using LASSO iterations) *Set $\widehat{\sigma}^0 = \sqrt{Var_n[y_i]}$ and $k = 0$, and specify a small constant $\nu > 0$, the tolerance level, and a constant $I > 1$, the upper bound on the number of iterations. (1) Compute the LASSO estimator $\widehat{\beta}$ based on $\lambda = 2c\widehat{\sigma}^k \Lambda(1-\alpha|X)$. (2) Set*

$$\widehat{\sigma}^{k+1} = \sqrt{\widehat{Q}(\widehat{\beta})}.$$

(3) If $|\widehat{\sigma}^{k+1} - \widehat{\sigma}^k| \leqslant \nu$ or $k+1 \geqslant I$, then stop and report $\widehat{\sigma} = \widehat{\sigma}^{k+1}$; otherwise set $k \leftarrow k+1$ and go to (1).

And the algorithm for estimating σ using Post-LASSO is as follows:

Algorithm 2 (Estimation of σ using Post-LASSO iterations) *Set $\widehat{\sigma}^0 = \sqrt{Var_n[y_i]}$ and $k = 0$, and specify a small constant $\nu \geqslant 0$, the tolerance level, and a constant $I > 1$, the upper bound on the number of iterations. (1) Compute the Post-LASSO estimator $\widetilde{\beta}$ based on $\lambda = 2c\widehat{\sigma}^k \Lambda(1-\alpha|X)$. (2) Set*

$$\widehat{\sigma}^{k+1} = \sqrt{\frac{n}{n-\widehat{s}} \cdot \widehat{Q}(\widetilde{\beta})},$$

where $\widehat{s} = \|\widetilde{\beta}\|_0 = |\widehat{T}|$. (3) If $|\widehat{\sigma}^{k+1} - \widehat{\sigma}^k| \leqslant \nu$ or $k+1 \geqslant I$, then stop and report $\widehat{\sigma} = \widehat{\sigma}^{k+1}$; otherwise, set $k \leftarrow k+1$ and go to (1).

We can also use $\lambda = 2c\widehat{\sigma}^k \sqrt{n}\Phi^{-1}(1-\alpha/2p)$ in place of X-dependent penalty. We note that using LASSO to estimate σ it follows that the sequence $\widehat{\sigma}^k$, $k \geqslant 2$, is monotone, while using Post-LASSO the estimates $\widehat{\sigma}^k$, $k \geqslant 1$, can only assume a finite number of different values.

The following theorem shows that these algorithms produce consistent estimates of the noise level, and that the LASSO and Post-LASSO estimators based on the resulting data-driven penalty continue to obey the asymptotic bounds we have derived previously.

Theorem 4 (Validity of Results with Estimated σ). *Suppose conditions ASM and RES hold. Suppose that $\sigma \leqslant \widehat{\sigma}^0 \lesssim \sigma$ with probability approaching 1 and $s \log p/n \to 0$. Then $\widehat{\sigma}$ produced by either Algorithm 1 or 2 is consistent*

$$\widehat{\sigma}/\sigma \to_P 1$$

so that the penalty levels $\lambda = 2c\widehat{\sigma}^k \Lambda(1-\alpha|X)$ and $\lambda = 2c\widehat{\sigma}^k \sqrt{n}\Phi^{-1}(1-\alpha/2p)$ with $\alpha = o(1)$, and $\log(1/\alpha) \lesssim \log p$, satisfy the condition (30) of Corollary 1, namely

$$\lambda \lesssim_P \sigma\sqrt{n \log p} \ \text{ and } \ \lambda \geqslant c' n \|S\|_\infty \ wp \ \to 1, \tag{39}$$

for some $1 < c' < c$. Consequently, the LASSO and Post-LASSO estimators based on this penalty level obey the conclusions of Corollaries 1, 2, and 3.

7 Monte Carlo Experiments

In this section we compare the performance of LASSO, Post-LASSO, and the ideal oracle linear regression estimators. The oracle estimator applies ordinary least square to the true model. (Such an estimator is not available outside Monte Carlo experiments.)

We begin by considering the following regression model:

$$y = x'\beta_0 + \varepsilon, \quad \beta_0 = (1, 1, 1/2, 1/3, 1/4, 1/5, 0, \ldots, 0)',$$

where $x = (1, z')'$ consists of an intercept and covariates $z \sim N(0, \Sigma)$, and the errors ε are independently and identically distributed $\varepsilon \sim N(0, \sigma^2)$. The dimension p of the covariates x is 500, the dimension s of the true model is 6, and the sample size n is 100. We set λ according to the X-dependent rule with $1 - \alpha = 90\%$. The regressors are correlated with $\Sigma_{ij} = \rho^{|i-j|}$ and $\rho = 0.5$. We consider two levels of noise: Design 1 with $\sigma^2 = 1$ (higher level) and Design 2 with $\sigma^2 = 0.1$ (lower level). For each repetition we draw new vectors x_i's and errors ε_i's.

We summarize the model selection performance of LASSO in Figures 3 and 4. In the left panels of the figures, we plot the frequencies of the dimensions of the selected model; in the right panels we plot the frequencies of selecting the correct regressors. From the left panels we see that the frequency of selecting a much larger model than the true model is very small in both designs. In the design with a larger noise, as the right panel of Figure 3 shows, LASSO frequently fails to select the entire true model, missing the regressors with small coefficients. However, it almost always includes the most important three regressors with the largest coefficients. Notably, despite this partial failure of the model selection Post-LASSO still performs well, as we report below. On the other hand, we see from the right panel of Figure 4 that in the design with a lower noise level LASSO rarely misses any component of the true support. These results confirm the theoretical results that when the non-zero coefficients are well-separated from zero, the penalized estimator should select a model that includes the true model as a subset. Moreover, these results also confirm the theoretical result of Theorem 2, namely, that the dimension of the selected model should be of the same stochastic order as the dimension of the true model. In summary, the model selection performance of the penalized estimator agrees very well with the theoretical results.

We summarize the results on the performance of estimators in Table 3, which records for each estimator $\breve{\beta}$ the mean ℓ_0-norm $E[\|\breve{\beta}\|_0]$, the norm of the bias $\|E\breve{\beta} - \beta_0\|$ and also the prediction error $E[\mathbb{E}_n[|x_i'(\breve{\beta} - \beta_0)|^2]^{1/2}]$ for recovering the regression function. As expected, LASSO has a substantial bias. We see that Post-LASSO drastically improves upon the LASSO, particularly in terms of reducing the bias, which also results in a much lower overall prediction error. Notably, despite that under the higher noise level LASSO frequently fails to recover the true model, the Post-LASSO estimator still performs well. This is because the penalized estimator always manages to select the most important regressors. We also see that the prediction error of the Post-LASSO is within a factor $\sqrt{\log p}$ of the prediction error

of the oracle estimator, as we would expect from our theoretical results. Under the lower noise level, Post-LASSO performs almost identically to the ideal oracle estimator. We would expect this since in this case LASSO selects the model especially well making Post-LASSO nearly the oracle.

Monte Carlo Results

Design 1 ($\sigma^2 = 1$)

	Mean ℓ_0-norm	Bias	Prediction Error
LASSO	5.41	0.4136	0.6572
Post-LASSO	5.41	0.0998	0.3298
Oracle	6.00	0.0122	0.2326

Design 2 ($\sigma^2 = 0.1$)

	Mean ℓ_0-norm	Bias	Prediction Error
LASSO	6.3640	0.1395	0.2183
Post-LASSO	6.3640	0.0068	0.0893
Oracle	6.00	0.0039	0.0736

Table 3 The table displays the average ℓ_0-norm of the estimators as well as mean bias and prediction error. We obtained the results using 1000 Monte Carlo repetitions for each design.

The results above used the true value of σ in the choice of λ. Next we illustrate how σ can be estimated in practice. We follow the iterative procedure described in the previous section. In our experiments the tolerance was 10^{-8} times the current estimate for σ, which is typically achieved in less than 15 iterations.

We assess the performance of the iterative procedure under the design with the larger noise, $\sigma^2 = 1$ (similar results hold for $\sigma^2 = 0.1$). The histograms in Figure 5 show that the model selection properties are very similar to the model selection when σ is known. Figure 6 displays the distribution of the estimator $\widehat{\sigma}$ of σ based on (iterative) Post-LASSO, (iterative) LASSO, and the initial estimator $\widehat{\sigma}^0 = \sqrt{\mathrm{Var}_n[y_i]}$. As we expected, estimator $\widehat{\sigma}$ based on LASSO produces estimates that are somewhat higher than the true value. In contrast, the estimator $\widehat{\sigma}$ based on Post-LASSO seems to perform very well in our experiments, giving estimates $\widehat{\sigma}$ that bunch closely near the true value σ.

8 Application to Cross-Country Growth Regression

In this section we apply LASSO and Post-LASSO to an international economic growth example. We use the Barro and Lee [4] data consisting of a panel of 138 countries for the period of 1960 to 1985. We consider the national growth rates in GDP per capita as a dependent variable y for the periods 1965-75 and 1975-85.[2] In

[2] The growth rate in GDP over a period from t_1 to t_2 is commonly defined as $\log(GDP_{t_2}/GDP_{t_1})$.

our analysis, we will consider a model with $p = 62$ covariates, which allows for a total of $n = 90$ complete observations. Our goal here is to select a subset of these covariates and briefly compare the resulting models to the standard models used in the empirical growth literature (Barro and Sala-i-Martin [5]).

Let us now turn to our empirical results. We performed covariate selection using LASSO, where we used our data-driven choice of penalty level λ in two ways. First we used an upper bound on σ being $\widehat{\sigma}^0$ and decreased the penalty to estimate different models with λ, $\lambda/2$, $\lambda/3$, $\lambda/4$, and $\lambda/5$. Second, we applied the iterative procedure described in the previous section to define λ^{it} (which is computed based on $\widehat{\sigma}^{it}$ obtained using the iterative Post-LASSO procedure).

The initial choice of the first approach led us to select no covariates, which is consistent with over-regularization since an upper bound for σ was used. We then proceeded to slowly decrease the penalty level in order to allow for some covariates to be selected. We present the model selection results in Table 5. With the first relaxation of the choice of λ, we select the black market exchange rate premium (characterizing trade openness) and a measure of political instability. With a second relaxation of the choice of λ we select an additional set of variables reported in the table. The iterative approach led to a model with only the black market exchange premium. We refer the reader to [4] and [5] for a complete definition and discussion of each of these variables.

We then proceeded to apply ordinary linear regression to the selected models and we also report the standard confidence intervals for these estimates. Table 8 shows these results. We find that in all models with additional selected covariates, the linear regression coefficients on the initial level of GDP is always negative and the standard confidence intervals do not include zero. We believe that these empirical findings firmly support the hypothesis of (conditional) convergence derived from the classical Solow-Swan-Ramsey growth model.[3] Finally, our findings also agree with and thus support the previous findings reported in Barro and Sala-i-Martin [5], which relied on ad-hoc reasoning for covariate selection.

Acknowledgements We would like to thank Denis Chetverikov and Brigham Fradsen for thorough proof-reading of several versions of this paper and their detailed comments that helped us considerably improve the paper. We also would like to thank Eric Gautier, Alexandre Tsybakov, and two anonymous referees for their comments that also helped us considerably improve the chapter. We would also like to thank the participants of seminars in Cowles Foundation Lecture at the Econometric Society Summer Meeting, Duke University, Harvard-MIT, and the Stats in the Chateau.

[3] The inferential method used here is actually valid under certain conditions, despite the fact that the model has been selected; this is demonstrated in a work in progress.

Confidence Intervals after Model Selection for the International Growth Regressions

Penalization Parameter $\lambda = 2.7870$	Real GDP per capita (log) Coefficient	90% Confidence Interval
$\lambda^{it} = 2.3662$	-0.0112	$[-0.0219, -0.0007]$
$\lambda/2$	-0.0120	$[-0.0225, -0.0015]$
$\lambda/3$	-0.0153	$[-0.0261, -0.0045]$
$\lambda/4$	-0.0221	$[-0.0346, -0.0097]$
$\lambda/5$	-0.0370	$[-0.0556, -0.0184]$

Table 4 The table above displays the coefficient and a 90% confidence interval associated with each model selected by the corresponding penalty level. The selected models are displayed in Table 5.

Model Selection Results for the International Growth Regressions

Penalization Parameter $\lambda = 2.7870$	Real GDP per capita (log) is included in all models Additional Selected Variables
λ	-
λ^{it}	Black Market Premium (log)
$\lambda/2$	Black Market Premium (log) Political Instability
$\lambda/3$	Black Market Premium (log) Political Instability Ratio of nominal government expenditure on defense to nominal GDP Ratio of import to GDP
$\lambda/4$	Black Market Premium (log) Political Instability Ratio of nominal government expenditure on defense to nominal GDP
$\lambda/5$	Black Market Premium (log) Political Instability Ratio of nominal government expenditure on defense to nominal GDP Ratio of import to GDP Exchange rate % of “secondary school complete” in male population Terms of trade shock Measure of tariff restriction Infant mortality rate Ratio of real government “consumption” net of defense and education Female gross enrollment ratio for higher education

Table 5 The models selected at various levels of penalty.

Appendix

9 Proofs

Proof (Theorem 1). Proceeding similarly to [8], by optimality of $\widehat{\beta}$ we have that

$$\widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) \leqslant \frac{\lambda}{n}\|\beta_0\|_1 - \frac{\lambda}{n}\|\widehat{\beta}\|_1. \tag{40}$$

To prove the result we make the use of the following relations: for $\delta = \widehat{\beta} - \beta_0$, if $\lambda \geqslant cn\|S\|_\infty$

$$\begin{aligned} \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) - \|\delta\|_{2,n}^2 &= -2\mathbb{E}_n[\varepsilon_i x_i'\delta] - 2\mathbb{E}_n[r_i x_i'\delta] && (41)\\ &\geqslant -\|S\|_\infty\|\delta\|_1 - 2c_s\|\delta\|_{2,n} \\ &\geqslant -\frac{\lambda}{cn}(\|\delta_T\|_1 + \|\delta_{T^c}\|_1) - 2c_s\|\delta\|_{2,n}, && (42) \end{aligned}$$

$$\|\beta_0\|_1 - \|\widehat{\beta}\|_1 = \|\beta_{0T}\|_1 - \|\widehat{\beta}_T\|_1 - \|\widehat{\beta}_{T^c}\|_1 \leqslant \|\delta_T\|_1 - \|\delta_{T^c}\|_1. \tag{43}$$

Thus, combining (40) with (41)–(43) implies that

$$-\frac{\lambda}{cn}(\|\delta_T\|_1 + \|\delta_{T^c}\|_1) + \|\delta\|_{2,n}^2 - 2c_s\|\delta\|_{2,n} \leqslant \frac{\lambda}{n}(\|\delta_T\|_1 - \|\delta_{T^c}\|_1). \tag{44}$$

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 $\|\delta\|_{2,n}^2 - 2c_s\|\delta\|_{2,n} \geqslant 0$ we get

$$\|\delta_{T^c}\|_1 \leqslant \frac{c+1}{c-1}\cdot\|\delta_T\|_1 = \bar{c}\|\delta_T\|_1, \tag{45}$$

and therefore δ satisfies the condition to invoke RE(c). From (44) and using RE(c), $\|\delta_T\|_1 \leqslant \sqrt{s}\|\delta\|_{2,n}/\kappa_{\bar{c}}$, we get

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

which gives the result on the prediction norm.

Lemma 5 (Empirical pre-sparsity for LASSO). *In either the parametric model or the nonparametric model, let $\widehat{m} = |\widehat{T} \setminus T|$ and $\lambda \geqslant c\cdot n\|S\|_\infty$. We have*

$$\sqrt{\widehat{m}} \leqslant \sqrt{s}\sqrt{\phi(\widehat{m})}\,2\bar{c}/\kappa_{\bar{c}} + 3(\bar{c}+1)\sqrt{\phi(\widehat{m})}\,nc_s/\lambda,$$

where $c_s = 0$ in the parametric model.

Proof. We have from the optimality conditions that

$$2\mathbb{E}_n[x_{ij}(y_i - x_i'\widehat{\beta})] = \text{sign}(\widehat{\beta}_j)\lambda/n \text{ for each } j \in \widehat{T} \setminus T.$$

Therefore we have for $R = (r_1, \ldots, r_n)'$, $X = [x_1, ..., x_n]'$, and $Y = (y_1, ..., y_n)'$

$$\begin{aligned}\sqrt{\widehat{m}}\lambda &= 2\|(X'(Y - X\widehat{\beta}))_{\widehat{T}\setminus T}\| \\ &\leqslant 2\|(X'(Y - R - X\beta_0))_{\widehat{T}\setminus T}\| + 2\|(X'R)_{\widehat{T}\setminus T}\| + 2\|(X'X(\beta_0 - \widehat{\beta}))_{\widehat{T}\setminus T}\| \\ &\leqslant \sqrt{\widehat{m}} \cdot n\|S\|_\infty + 2n\sqrt{\phi(\widehat{m})}c_s + 2n\sqrt{\phi(\widehat{m})}\|\widehat{\beta} - \beta_0\|_{2,n},\end{aligned}$$

where we used that

$$\begin{aligned}\|(X'X(\beta_0 - \widehat{\beta}))_{\widehat{T}\setminus T}\| &\leqslant \sup_{\|v_{T^c}\|_0 \leqslant \widehat{m}, \|v\| \leqslant 1} |v'X'X(\beta_0 - \widehat{\beta})| \\ &\leqslant \sup_{\|v_{T^c}\|_0 \leqslant \widehat{m}, \|v\| \leqslant 1} \|v'X'\| \|X(\beta_0 - \widehat{\beta})\| \\ &= \sup_{\|v_{T^c}\|_0 \leqslant \widehat{m}, \|v\| \leqslant 1} \sqrt{|v'X'Xv|} \|X(\beta_0 - \widehat{\beta})\| \\ &= n\sqrt{\phi(\widehat{m})}\|\beta_0 - \widehat{\beta}\|_{2,n},\end{aligned}$$

and similarly $\|(X'R)_{\widehat{T}\setminus T}\| \leqslant n\sqrt{\phi(\widehat{m})}c_s$.

Since $\lambda/c \geqslant n\|S\|_\infty$, and by Theorem 1, $\|\beta_0 - \widehat{\beta}\|_{2,n} \leqslant \left(1 + \frac{1}{c}\right)\frac{\lambda\sqrt{s}}{n\kappa_{\bar{c}}} + 2c_s$, we have

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

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

Proof (Proof of Theorem 2). Since $\lambda \geqslant c \cdot n\|S\|_\infty$ by Lemma 5 we have

$$\sqrt{\widehat{m}} \leqslant \sqrt{\phi(\widehat{m})} \cdot 2\bar{c}\sqrt{s}/\kappa_{\bar{c}} + 3(\bar{c} + 1)\sqrt{\phi(\widehat{m})} \cdot nc_s/\lambda,$$

which, by letting $L = \left(\frac{2\bar{c}}{\kappa_{\bar{c}}} + 3(\bar{c} + 1)\frac{nc_s}{\lambda\sqrt{s}}\right)^2$, can be rewritten as

$$\widehat{m} \leqslant s \cdot \phi(\widehat{m})L. \tag{46}$$

Note that $\widehat{m} \leqslant n$ by optimality conditions. Consider any $M \in \mathscr{M}$, and suppose $\widehat{m} > M$. Therefore by Lemma 9 on sublinearity of sparse eigenvalues

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

Thus, since $\lceil k \rceil < 2k$ for any $k \geqslant 1$ we have

$$M < s \cdot 2\phi(M)L$$

which violates the condition of $M \in \mathscr{M}$ and s. Therefore, we must have $\widehat{m} \leqslant M$.

In turn, applying (46) once more with $\widehat{m} \leqslant (M \wedge n)$ we obtain

$$\widehat{m} \leqslant s \cdot \phi(M \wedge n)L.$$

The result follows by minimizing the bound over $M \in \mathscr{M}$.

Proof (Lemma 3, part (1)). The result follows immediately from the assumptions.

Proof (Lemma 3, part (2)). Let $\widehat{m} = |\widehat{T} \setminus T| = \|\widehat{\beta}_{T^c}\|_0$. Then, note that $\|\delta\|_\infty \leqslant \|\delta\| \leqslant \|\delta\|_{2,n}/\kappa(\widehat{m})$. The result follows from Theorem 1.

Proof (Lemma 3, part (3)). Let $\delta := \widehat{\beta} - \beta_0$. Note that by the first order optimality conditions of $\widehat{\beta}$ and the assumption on λ

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

since $\|\mathbb{E}_n[x_i r_i]\|_\infty \leqslant \min\left\{\frac{\sigma}{\sqrt{n}}, c_s\right\}$ by Lemma 6 below.

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

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

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

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

The result follows by Lemma 7 to bound $\|\delta\|_1$ and the arguments in [8] and [17] to show that the bound on the correlations imply that for any $C > 0$

$$\kappa_C \geqslant \sqrt{1 - s(1+2C)\|\mathbb{E}_n[x_i x_i' - I]\|_\infty}$$

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

Lemma 6. *Under condition ASM, we have that*

$$\|\mathbb{E}_n[x_i r_i]\|_\infty \leqslant \min\left\{\frac{\sigma}{\sqrt{n}}, c_s\right\}.$$

Proof. First note that for every $j = 1, \ldots, p$, we have $|\mathbb{E}_n[x_{ij} r_i]| \leqslant \sqrt{\mathbb{E}_n[x_{ij}^2]\mathbb{E}_n[r_i^2]} = c_s$.

Next, by definition of β_0 in (11), for $j \in T$ we have

$$\mathbb{E}_n[x_{ij}(f_i - x_i'\beta_0)] = \mathbb{E}_n[x_{ij} r_i] = 0$$

since β_0 is a minimizer over the support of β_0. For $j \in T^c$ we have that for any $t \in \mathbb{R}$

$$\mathbb{E}_n[(f_i - x_i'\beta_0)^2] + \sigma^2\frac{s}{n} \leqslant \mathbb{E}_n[(f_i - x_i'\beta_0 - t x_{ij})^2] + \sigma^2\frac{s+1}{n}.$$

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

$$-\sigma^2/n \leqslant \mathbb{E}_n[(f_i - x_i'\beta_0 - t x_{ij})^2] - \mathbb{E}_n[(f_i - x_i'\beta_0)^2] = -2t\mathbb{E}_n[x_{ij}(f_i - x_i'\beta_0)] + t^2\mathbb{E}_n[x_{ij}^2].$$

Taking the minimum over t in the right hand side at $t^* = \mathbb{E}_n[x_{ij}(f_i - x_i'\beta_0)]$ we obtain

$$-\sigma^2/n \leqslant -(\mathbb{E}_n[x_{ij}(f_i - x_i'\beta_0)])^2$$

or equivalently, $|\mathbb{E}_n[x_{ij}(f_i - x_i'\beta_0)]| \leqslant \sigma/\sqrt{n}$.

Lemma 7. *If $\lambda \geqslant cn\|S\|_\infty$, then for $\bar{c} = (c+1)/(c-1)$ we have*

$$\|\widehat{\beta} - \beta_0\|_1 \leqslant \frac{(1+2\bar{c})\sqrt{s}}{\kappa_{2\bar{c}}}\left[\left(1+\frac{1}{c}\right)\frac{\lambda\sqrt{s}}{n\kappa_{\bar{c}}} + 2c_s\right] + \left(1+\frac{1}{2\bar{c}}\right)\frac{2c}{c-1}\frac{n}{\lambda}c_s^2,$$

where $c_s = 0$ in the parametric case.

Proof. First, assume $\|\delta_{T^c}\|_1 \leqslant 2\bar{c}\|\delta_T\|_1$. In this case, by definition of the restricted eigenvalue, we have

$$\|\delta\|_1 \leqslant (1+2\bar{c})\|\delta_T\|_1 \leqslant (1+2\bar{c})\sqrt{s}\|\delta\|_{2,n}/\kappa_{2\bar{c}}$$

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 $\|\delta_{T^c}\|_1 > 2\bar{c}\|\delta_T\|_1$ which would already imply $\|\delta\|_{2,n} \leqslant 2c_s$. Moreover, the relation (44) implies that

$$\begin{aligned}\|\delta_{T^c}\|_1 &\leqslant \bar{c}\|\delta_T\|_1 + \tfrac{c}{c-1}\tfrac{n}{\lambda}\|\delta\|_{2,n}(2c_s - \|\delta\|_{2,n}) \\ &\leqslant \bar{c}\|\delta_T\|_1 + \tfrac{c}{c-1}\tfrac{n}{\lambda}c_s^2 \\ &\leqslant \tfrac{1}{2}\|\delta_{T^c}\|_1 + \tfrac{c}{c-1}\tfrac{n}{\lambda}c_s^2.\end{aligned}$$

Thus,

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

The result follows by adding the bounds on each case and invoking Theorem 1 to bound $\|\delta\|_{2,n}$.

Proof (Theorem 3). Let $\widetilde{\delta} := \widetilde{\beta} - \beta_0$. By definition of the Post-LASSO estimator, it follows that $\widehat{Q}(\widetilde{\beta}) \leqslant \widehat{Q}(\widehat{\beta})$ and $\widehat{Q}(\widetilde{\beta}) \leqslant \widehat{Q}(\beta_{0\widehat{T}})$. Thus,

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

The least squares criterion function satisfies

$$
\begin{aligned}
|\widehat{Q}(\widetilde{\beta}) - \widehat{Q}(\beta_0) - \|\widetilde{\delta}\|_{2,n}^2| &\leqslant |S'\widetilde{\delta}| + 2c_s\|\widetilde{\delta}\|_{2,n} \\
&\leqslant |S_T'\widetilde{\delta}| + |S_{T^c}'\widetilde{\delta}| + 2c_s\|\widetilde{\delta}\|_{2,n} \\
&\leqslant \|S_T\|\|\widetilde{\delta}\| + \|S_{T^c}\|_\infty\|\widetilde{\delta}_{T^c}\|_1 + 2c_s\|\widetilde{\delta}\|_{2,n} \\
&\leqslant \|S_T\|\|\widetilde{\delta}\| + \|S_{T^c}\|_\infty\sqrt{\widehat{m}}\|\widetilde{\delta}\| + 2c_s\|\widetilde{\delta}\|_{2,n} \\
&\leqslant \|S_T\|\frac{\|\widetilde{\delta}\|_{2,n}}{\kappa(\widehat{m})} + \|S_{T^c}\|_\infty\sqrt{\widehat{m}}\frac{\|\widetilde{\delta}\|_{2,n}}{\kappa(\widehat{m})} + 2c_s\|\widetilde{\delta}\|_{2,n}.
\end{aligned}
$$

Next, note that for any $j \in \{1,\ldots,p\}$ we have $E[S_j^2] = 4\sigma^2/n$, so that $E[\|S_T\|^2] \leqslant 4\sigma^2 s/n$. Thus, by Chebyshev inequality, for any $\tilde{\gamma} > 0$, there is a constant $A_{\tilde{\gamma}}$ such that $\|S_T\| \leqslant A_{\tilde{\gamma}}\sigma\sqrt{s/n}$ with probability at least $1-\tilde{\gamma}$. Moreover, using Lemma 8, $\|S_{T^c}\|_\infty \leqslant A'_{\tilde{\gamma}}2\sigma\sqrt{2\log p\,/n}$ with probability at least $1-\tilde{\gamma}$ for some constant $A'_{\tilde{\gamma}}$. Define $A_{\gamma,n} := K_\gamma\sigma\sqrt{(s+\widehat{m}\log p)/n}$ so that $A_{\gamma,n} \geqslant \|S_T\| + \sqrt{\widehat{m}}\|S_{T^c}\|_\infty$ with probability at least $1-\gamma$ for some constant $K_\gamma < \infty$ independent of n and p.

Combining these relations, with probability at least $1-\gamma$ we have

$$
\|\widetilde{\delta}\|_{2,n}^2 - A_{\gamma,n}\|\widetilde{\delta}\|_{2,n}/\kappa(\widehat{m}) - 2c_s\|\widetilde{\delta}\|_{2,n} \leqslant B_n \wedge C_n,
$$

solving which we obtain:

$$
\|\widetilde{\delta}\|_{2,n} \leqslant A_{\gamma,n}/\kappa(\widehat{m}) + 2c_s + \sqrt{(B_n)_+ \wedge (C_n)_+}. \tag{47}
$$

Note that by the optimality of $\widehat{\beta}$ in the LASSO problem, and letting $\widehat{\delta} = \widehat{\beta} - \beta_0$,

$$
B_n = \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) \leqslant \tfrac{\lambda}{n}(\|\beta_0\|_1 - \|\widehat{\beta}\|_1) \leqslant \tfrac{\lambda}{n}(\|\widehat{\delta}_T\|_1 - \|\widehat{\delta}_{T^c}\|_1). \tag{48}
$$

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

$$
B_n := \widehat{Q}(\widehat{\beta}) - \widehat{Q}(\beta_0) \leqslant \frac{\lambda}{n}\|\widehat{\delta}_T\|_1 \leqslant \frac{\lambda}{n}\frac{\sqrt{s}\|\widehat{\delta}\|_{2,n}}{\kappa_{\bar{c}}}. \tag{49}
$$

The choice of λ yields $\lambda \geqslant cn\|S\|_\infty$ with probability $1-\alpha$. Thus, by applying Theorem 1, which requires $\lambda \geqslant cn\|S\|_\infty$, we can bound $\|\widehat{\delta}\|_{2,n}$.

Finally, with probability $1-\alpha-\gamma$ we have that (47) and (49) with $\|\widehat{\delta}\|_{2,n} \leqslant (1+1/c)\lambda\sqrt{s}/n\kappa_{\bar{c}} + 2c_s$ hold, and the result follows since if $T \subseteq \widehat{T}$ we have $C_n = 0$ so that $B_n \wedge C_n \leqslant 1\{T \not\subseteq \widehat{T}\}B_n$.

Proof (Theorem 4). Consider the case of Post-LASSO; the proof for LASSO is similar. Consider the case with $k=1$, i.e. when $\widehat{\sigma} = \widehat{\sigma}^k$ for $k=1$. Then we have

$$\left|\frac{\widehat{Q}(\widetilde{\beta})}{\sigma^2}-\frac{\mathbb{E}_n[\varepsilon_i^2]}{\sigma^2}\right| \leqslant \frac{\|\widetilde{\beta}-\beta_0\|_{2,n}^2}{\sigma^2}+\frac{\|S\|_\infty\|\widetilde{\beta}-\beta_0\|_1}{\sigma^2}+ \\ +\frac{2c_s\|\widetilde{\beta}-\beta_0\|_{2,n}}{\sigma^2}+\frac{2c_s\sqrt{\mathbb{E}_n[\varepsilon_i^2]}}{\sigma^2}+\frac{c_s^2}{\sigma^2}=o_P(1).$$

since $\|\widetilde{\beta}-\beta_0\|_{2,n}\lesssim_P \sigma\sqrt{(s/n)\log p}$ by Corollary 3 and by assumption on $\widehat{\sigma}^0$, $\|S\|_\infty\lesssim_P \sigma\sqrt{(1/n)\log p}$ by Lemma 8, $\|\widetilde{\beta}-\beta_0\|_1\leqslant\sqrt{\widehat{s}}\|\widetilde{\beta}-\beta\|_2\lesssim_P\sqrt{\widehat{s}}\|\widetilde{\beta}-\beta\|_{2,n}$ by condition RSE, $\widehat{s}\lesssim_P s$ by Corollary 2 and $c_s\lesssim\sigma\sqrt{s/n}$ by condition ASM, and $s\log p/n\to 0$ by assumption, and $\frac{\mathbb{E}_n[\varepsilon_i^2]}{\sigma^2}-1\to_P 0$ by the Chebyshev inequality. Finally, $n/(n-\widehat{s})=1+o_P(1)$ since $\widehat{s}\lesssim_P s$ by Corollary 2 and $s\log p/n\to 0$. The result for $2\leqslant k\leqslant I-1$ follows by induction.

10 Auxiliary Lemmas

Recall that $\|S/(2\sigma)\|_\infty=\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[x_{ij}g_i]|$, where g_i are i.i.d. $N(0,1)$, for $i=1,...,n$, conditional on $X=[x_1',...,x_n']'$, and $\mathbb{E}_n[x_{ij}^2]=1$ for each $j=1,...,p$, and note that $P(n\|S/(2\sigma)\|_\infty\geqslant\Lambda(1-\alpha|X)|X)=\alpha$ by definition.

Lemma 8. *We have that for $t\geqslant 0$:*

$$P(n\|S/(2\sigma)\|_\infty\geqslant t\sqrt{n}|X)\leqslant 2p(1-\Phi(t))\leqslant 2p\frac{1}{t}\phi(t),$$
$$\Lambda(1-\alpha|X)\leqslant\sqrt{n}\Phi^{-1}(1-\alpha/2p)\leqslant\sqrt{2n\log(2p/\alpha)},$$
$$P(n\|S/(2\sigma)\|_\infty\geqslant\sqrt{2n\log(2p/\alpha)}|X)\leq\alpha.$$

Proof. To establish the first claim, note that $\sqrt{n}\|S/2\sigma\|_\infty=\max_{1\leqslant j\leqslant p}|Z_j|$, where $Z_j=\sqrt{n}\mathbb{E}_n[x_{ij}g_i]$ are $N(0,1)$ by g_i i.i.d. $N(0,1)$ conditional on X and by $\mathbb{E}_n[x_{ij}^2]=1$ for each $j=1,...,p$. Then the first claim follows by observing that for $z\geqslant 0$ by the union bound $P(\max_{1\leqslant j\leqslant p}|Z_j|>z)\leqslant pP(|Z_j|>z)=2p(1-\Phi(z))$ and by $(1-\Phi(z))=\int_z^\infty\phi(u)du\leqslant\int_z^\infty(u/z)\phi(u)dz\leqslant(1/z)\phi(z)$. The second and third claim follow by noting that $2p(1-\Phi(t'))=\alpha$ at $t'=\Phi^{-1}(1-\alpha/2p)$, and $2p\frac{1}{t''}\phi(t'')=\alpha$ at $t''\leqslant\sqrt{2\log(2p/\alpha)}$, so that, in view of the first claim, $\Lambda(1-\alpha|X)\leqslant\sqrt{n}t'\leqslant\sqrt{n}t''$.

Lemma 9 (Sub-linearity of restricted sparse eigenvalues). *For any integer $k\geqslant 0$ and constant $\ell\geqslant 1$ we have $\phi(\lceil\ell k\rceil)\leqslant\lceil\ell\rceil\phi(k)$.*

Proof. Let $W:=\mathbb{E}_n[x_ix_i']$ and $\bar{\alpha}$ be such that $\phi(\lceil\ell k\rceil)=\bar{\alpha}'W\bar{\alpha}$, $\|\bar{\alpha}\|=1$. We can decompose the vector $\bar{\alpha}$ so that

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

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

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

where we used that

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

Lemma 10. *Let $\bar{c} = (c+1)/(c-1)$ we have for any integer $m > 0$*

$$\kappa_{\bar{c}} \geqslant \kappa(m) \left(1 - \mu(m) \bar{c} \sqrt{\frac{s}{m}} \right).$$

Proof. We follow the proof in [8]. Pick an arbitrary vector δ such that $\|\delta_{T^c}\|_1 \leqslant \bar{c} \|\delta_T\|_1$. Let T^1 denote the m largest components of δ_{T^c}. Moreover, let $T^c = \cup_{k=1}^K T^k$ where $K = \lceil (p-s)/m \rceil$, $|T^k| \leqslant m$ and T^k corresponds to the m largest components of δ outside $T \cup (\cup_{d=1}^{k-1} T^d)$.

We have

$$\begin{aligned}
\|\delta\|_{2,n} \geqslant \|\delta_{T \cup T^1}\|_{2,n} - \|\delta_{(T \cup T^1)^c}\|_{2,n} &\geqslant \kappa(m) \|\delta_{T \cup T^1}\| - \sum_{k=2}^K \|\delta_{T^k}\|_{2,n} \\
&\geqslant \kappa(m) \|\delta_{T \cup T^1}\| - \sqrt{\phi(m)} \sum_{k=2}^K \|\delta_{T^k}\|.
\end{aligned}$$

Next note that

$$\|\delta_{T^{k+1}}\| \leqslant \|\delta_{T^k}\|_1 / \sqrt{m}.$$

Indeed, consider the problem $\max\{\|v\| / \|u\|_1 : v, u \in \mathbb{R}^m, \max_i |v_i| \leqslant \min_i |u_i|\}$. Given a v and u we can always increase the objective function by using $\tilde{v} = \max_i |v_i| (1, \ldots, 1)'$ and $\tilde{u}' = \min_i |u_i| (1, \ldots, 1)'$ instead. Thus, the maximum is achieved at $v^* = u^* = (1, \ldots, 1)'$, yielding $1/\sqrt{m}$.

Thus, by $\|\delta_{T^c}\|_1 \leqslant \bar{c} \|\delta_T\|_1$ and $|T| = s$

$$\sum_{k=2}^{K} \|\delta_{T^k}\| \leqslant \sum_{k=1}^{K-1} \frac{\|\delta_{T^k}\|_1}{\sqrt{m}} \leqslant \frac{\|\delta_{T^c}\|_1}{\sqrt{m}} \leqslant \bar{c}\|\delta_T\|\sqrt{\frac{s}{m}} \leqslant \bar{c}\|\delta_{T\cup T^1}\|\sqrt{\frac{s}{m}}.$$

Therefore, combining these relations with $\|\delta_{T\cup T^1}\| \geqslant \|\delta_T\| \geqslant \|\delta_T\|_1/\sqrt{s}$ we have

$$\|\delta\|_{2,n} \geqslant \frac{\|\delta_T\|_1}{\sqrt{s}} \kappa(m) \left(1 - \mu(m)\bar{c}\sqrt{s/m}\right)$$

which leads to

$$\frac{\sqrt{s}\|\delta\|_{2,n}}{\|\delta_T\|_1} \geqslant \kappa(m) \left(1 - \mu(m)\bar{c}\sqrt{s/m}\right).$$

References

1. Akaike, H.: A new look at the statistical model identification. MIEEE Transactions on Automatic Control **AC-19**, 716–723 (1974)
2. Angrist, J., Chernozhukov, V., Fernndez-Val, I.: Quantile regression under misspecification, with an application to the U.S. wage structure. Econometrica **74**(2), 539–563 (2006)
3. Angrist, J.D., Krueger, A.B.: Does compulsory school attendance affect schooling and earnings? The Quarterly Journal of Economics **106**(4), 979–1014 (1991)
4. Barro, R.J., Lee, J.W.: Data set for a panel of 139 countries. NBER, http://www.nber.org/pub/barro.lee.html (1994)
5. Barro, R.J., Sala–i–Martin, X.: Economic Growth. McGraw-Hill, New York (1995)
6. Belloni, A., Chen, D., Chernozhukov, V., Hansen, C.: Sparse models and methods for optimal instruments with an application to eminent domain. arXiv:[math.ST] (2010)
7. Belloni, A., Chernozhukov, V.: Post-ℓ_1-penalized estimators in high-dimensional linear regression models. arXiv:[math.ST] (2009)
8. Bickel, P.J., Ritov, Y., Tsybakov, A.B.: Simultaneous analysis of LASSO and Dantzig selector. Annals of Statistics **37**(4), 1705–1732 (2009)
9. Candès, E.J., Plan, Y.: Near-ideal model selection by l_1 minimization. Ann. Statist. **37**(5A), 2145–2177 (2009)
10. Candès, E.J., Tao, T.: The Dantzig selector: statistical estimation when p is much larger than n. Ann. Statist. **35**(6), 2313–2351 (2007)
11. Donoho, D.L., Johnstone, J.M.: Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**(3), 425–455 (1994)
12. Ge, D., Jiang, X., Ye, Y.: A note on complexity of l_p minimization. Stanford Working Paper (2010)
13. van de Geer, S.A.: High-dimensional generalized linear models and the lasso. Annals of Statistics **36**(2), 614–645 (2008)
14. Hansen, C., Hausman, J., Newey, W.K.: Estimation with many instrumental variables. Journal of Business and Economic Statistics **26**, 398–422 (2008)
15. Koltchinskii, V.: Sparsity in penalized empirical risk minimization. Ann. Inst. H. Poincar Probab. Statist. **45**(1), 7–57 (2009)
16. Levine, R., Renelt, D.: A sensitivity analysis of cross-country growth regressions. The American Economic Review **82**(4), 942–963 (1992)
17. Lounici, K.: Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. Electron. J. Statist. **2**, 90–102 (2008)
18. Meinshausen, N., Yu, B.: Lasso-type recovery of sparse representations for high-dimensional data. Annals of Statistics **37**(1), 2246–2270 (2009)
19. Natarajan, B.K.: Sparse approximate solutions to linear systems. SIAM Journal on Computing **24**, 227–234 (1995)

20. Rigollet, P., Tsybakov, A.B.: Exponential screening and optimal rates of sparse estimation. ArXiv:1003.2654 (2010)
21. Rudelson, M., Vershynin, R.: On sparse reconstruction from fourier and gaussian measurements. Communications on Pure and Applied Mathematics **61**, 10251045 (2008)
22. Sala–i–Martin, X.: I just ran two million regressions. The American Economic Review **87**(2), 178–183 (1997)
23. Schwarz, G.: Estimating the dimension of a model. Annals of Statistics pp. 461–464 (1978)
24. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B **58**, 267–288 (1996)
25. Wainwright, M.: 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 (2009)
26. Zhang, C.H., Huang, J.: The sparsity and bias of the lasso selection in high-dimensional linear regression. Ann. Statist. **36**(4), 1567–1594 (2008)
27. Zhao, P., Yu, B.: On model selection consistency of lasso. J. Machine Learning Research **7**, 2541–2567 (2006)

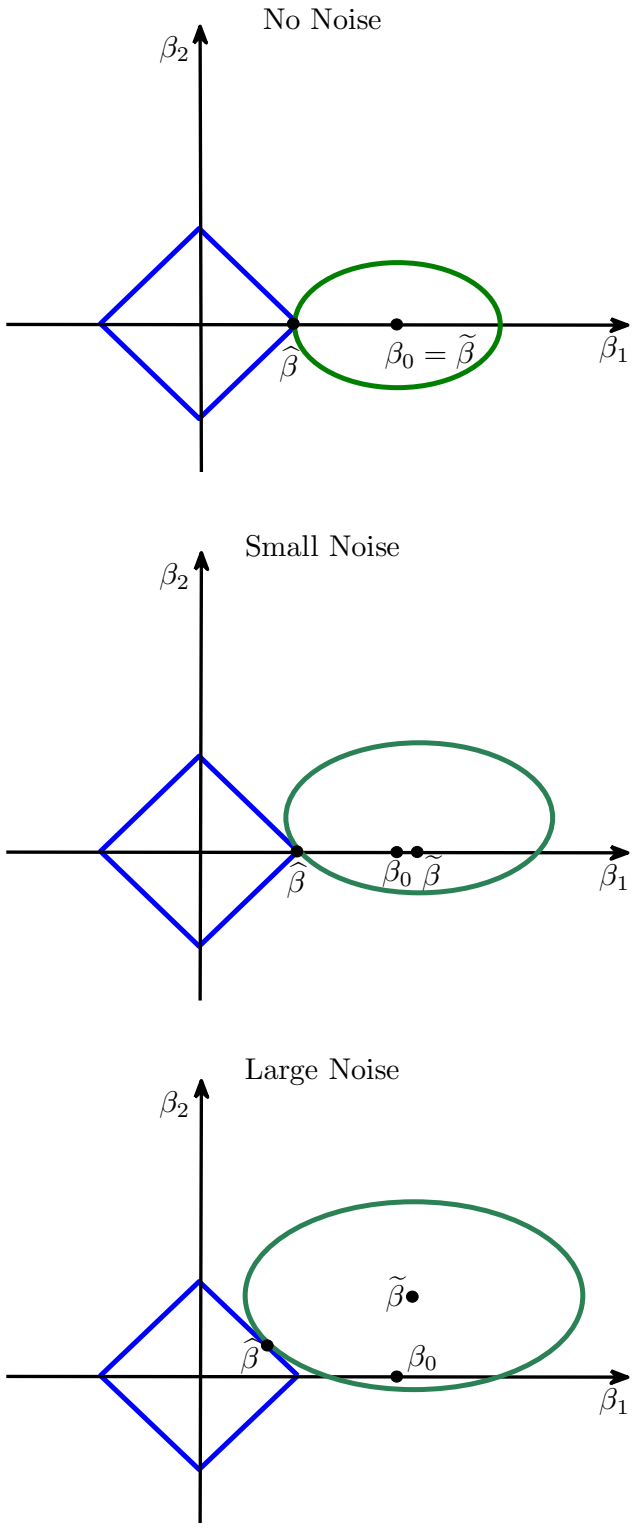

Fig. 2 The figures illustrate the geometry of LASSO and Post-LASSO estimator.

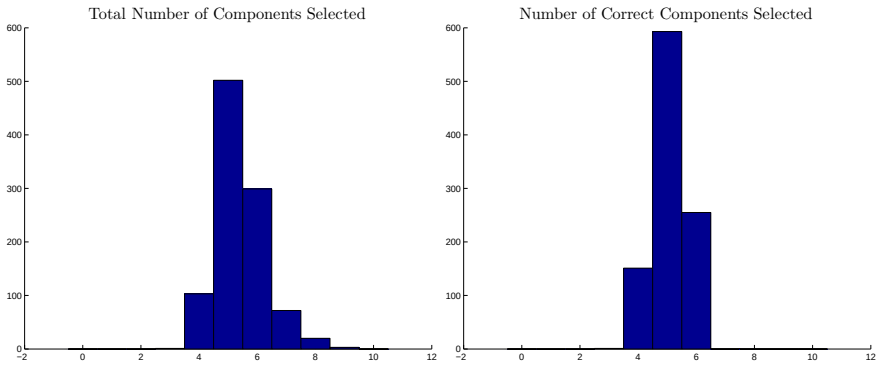

Fig. 3 The figure summarizes the covariate selection results for the design with $\sigma = 1$, based on 1000 Monte Carlo repetitions. The left panel plots the histogram for the number of covariates selected by LASSO out of the possible 500 covariates, $|\widehat{T}|$. The right panel plots the histogram for the number of significant covariates selected by LASSO, $|\widehat{T} \cap T|$; there are in total 6 significant covariates amongst 500 covariates. The sample size for each repetition was $n = 100$.

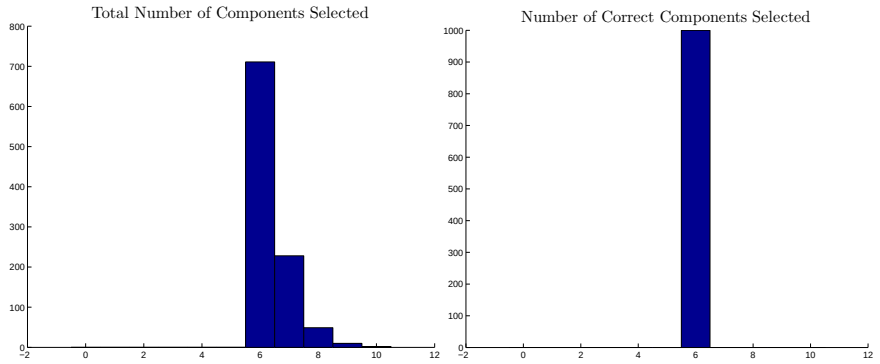

Fig. 4 The figure summarizes the covariate selection results for the design with $\sigma^2 = 0.1$, based on 1000 Monte Carlo repetitions. The left panel plots the histogram for the number of covariates selected out of the possible 500 covariates, $|\widehat{T}|$. The right panel plots the histogram for the number of significant covariates selected, $|\widehat{T} \cap T|$; there are in total 6 significant covariates amongst 500 covariates. The sample size for each repetition was $n = 100$.

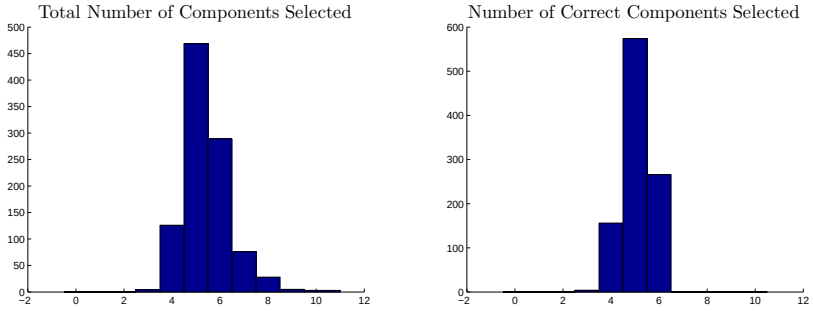

Fig. 5 The figure summarizes the covariate selection results for the design with $\sigma = 1$, when σ is estimated, based on 1000 Monte Carlo repetitions. The left panel plots the histogram for the number of covariates selected out of the possible 500 covariates. The right panel plots the histogram for the number of significant covariates selected; there are in total 6 significant covariates amongst 500 covariates. The sample size for each repetition was $n = 100$.

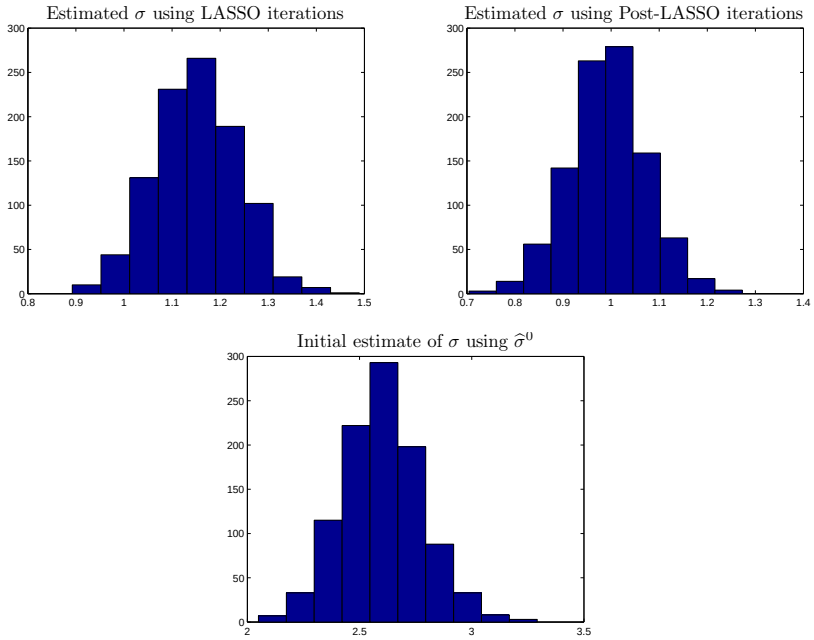

Fig. 6 The figure displays the distribution of the estimator $\widehat{\sigma}$ of σ based on (iterative) LASSO, (iterative) Post-LASSO, and the conservative initial estimator $\widehat{\sigma}^0 = \sqrt{\text{Var}_n[y_i]}$. The plots summarize the estimation performance for the design with $\sigma = 1$, based on 1000 Monte Carlo repetitions.