The Annals of Statistics
2014, Vol. 42, No. 2, 757–788
DOI: 10.1214/14-AOS1204

PIVOTAL ESTIMATION VIA SQUARE-ROOT LASSO IN NONPARAMETRIC REGRESSION

BY ALEXANDRE BELLONI, VICTOR CHERNOZHUKOV AND LIE WANG

Duke University, Massachusetts Institute of Technology and Massachusetts Institute of Technology

We propose a self-tuning $\sqrt{\text{Lasso}}$ method that simultaneously resolves three important practical problems in high-dimensional regression analysis, namely it handles the unknown scale, heteroscedasticity and (drastic) non-Gaussianity of the noise. In addition, our analysis allows for badly behaved designs, for example, perfectly collinear regressors, and generates sharp bounds even in extreme cases, such as the infinite variance case and the noiseless case, in contrast to Lasso. We establish various nonasymptotic bounds for $\sqrt{\text{Lasso}}$ including prediction norm rate and sparsity. Our analysis is based on new impact factors that are tailored for bounding prediction norm. In order to cover heteroscedastic non-Gaussian noise, we rely on moderate deviation theory for self-normalized sums to achieve Gaussian-like results under weak conditions. Moreover, we derive bounds on the performance of ordinary least square (ols) applied to the model selected by $\sqrt{\text{Lasso}}$ accounting for possible misspecification of the selected model. Under mild conditions, the rate of convergence of ols post $\sqrt{\text{Lasso}}$ is as good as $\sqrt{\text{Lasso}}$'s rate. As an application, we consider the use of $\sqrt{\text{Lasso}}$ and ols post $\sqrt{\text{Lasso}}$ as estimators of nuisance parameters in a generic semiparametric problem (nonlinear moment condition or Z-problem), resulting in a construction of $\sqrt{n}$-consistent and asymptotically normal estimators of the main parameters.

1. Introduction.

We consider a nonparametric regression model:

$$y_i = f(z_i) + \sigma\varepsilon_i, \qquad i = 1, \ldots, n, \tag{1.1}$$

where y_i's are the outcomes, z_i's are vectors of fixed basic covariates, ε_i's are independent noise, f is the regression function and σ is an unknown scaling parameter. The goal is to recover the values $(f_i)_{i=1}^n = (f(z_i))_{i=1}^n$ of the regression function f at z_i's. To achieve this goal, we use linear combinations of technical regressors $x_i = P(z_i)$ to approximate f, where $P(z_i)$ is a dictionary of p-vector of transformations of z_i. We are interested in the high dimension low sample size case, where we potentially use $p > n$, to obtain a flexible approximation. In particular, we are interested in cases where the regression function can be well approximated by a sparse linear function of x_i.

Received September 2012; revised December 2013.
MSC2010 subject classifications. Primary 62G05, 62G08; secondary 62G35.
Key words and phrases. Pivotal, square-root Lasso, model selection, non-Gaussian heteroscedastic, generic semiparametric problem, nonlinear instrumental variable, Z-estimation problem, $\sqrt{n}$-consistency and asymptotic normality after model selection.

The model above can be written as $y_i = x_i'\beta_0 + r_i + \sigma\varepsilon_i$, where $f_i = f(z_i)$ and $r_i := f_i - x_i'\beta_0$ is the approximation error. The vector β_0 is defined as a solution of an optimization problem to compute the oracle risk, which balances bias and variance (see Section 2). The cardinality of the support of β_0 is denoted by $s := \|\beta_0\|_0$. It is well known that ordinary least squares (ols) is generally inconsistent when $p > n$. However, the sparsity assumption, namely that $s \ll n$, makes it possible to estimate these models effectively by searching for approximately the right set of the regressors. In particular, ℓ_1-penalization has played a central role [14, 15, 18, 35, 40, 47, 52, 54]. It was demonstrated that ℓ_1-penalized least squares estimators can achieve the rate $\sigma\sqrt{s/n}\sqrt{\log p}$, which is very close to the oracle rate $\sigma\sqrt{s/n}$ achievable when the true model is known. Importantly, in the context of linear regression, these ℓ_1-regularized problems can be cast as convex optimization problems which make them computationally efficient (computable in polynomial time). We refer to [14–17, 27, 38, 39, 42, 47] for a more detailed review of the existing literature which has focused on the homoscedastic case.

In this paper, we attack the problem of nonparametric regression under non-Gaussian, heteroscedastic errors ε_i, having an unknown scale σ. We propose to use a self-tuning $\sqrt{\text{Lasso}}$ which is pivotal with respect to the scaling parameter σ, and which handles non-Gaussianity and heteroscedasticity in the errors. The resulting rates and performance guarantees are very similar to the Gaussian case, due to the use of self-normalized moderate deviation theory. Such results and properties,[1] particularly the pivotality with respect to the scale, are in contrast to the previous results and methods on others ℓ_1-regularized methods, for example, Lasso and Dantzig selector that use penalty levels that depend linearly on the unknown scaling parameter σ.

There is now a growing literature on high-dimensional linear models[2] allowing for unknown scale σ. Städler et al. [43] propose a ℓ_1-penalized maximum likelihood estimator for parametric Gaussian regression models. Belloni et al. [12] consider $\sqrt{\text{Lasso}}$ for a parametric homoscedastic model with both Gaussian and non-Gaussian errors and establish that the choice of the penalty parameter in $\sqrt{\text{Lasso}}$ becomes pivotal with respect to σ. van de Geer [49] considers an equivalent formulation of the (homoscedastic) $\sqrt{\text{Lasso}}$ to establish finite sample results and derives results in the parametric homoscedastic Gaussian setting. Chen and Dalalyan [21] consider scaled fused Dantzig selector to allow for different sparsity patterns and provide results under homoscedastic Gaussian errors. Belloni and

[1] Earlier literature, for example, in bounded designs [15], provides bounds using refinements of Nemirovski's inequality; see [26]. These results provide rates as good as in the Gaussian case. However, when the design is unbounded (e.g., regressors generated as realizations of independent Gaussian random variables), the rates of convergence provided by these techniques are no longer sharp. The use of self-normalized moderate deviations in the present context allows to handle the latter cases, with sharp rates.

[2] There is also a literature on penalized median regression, which can be used in the case of symmetric errors, since these methods are independent of the unknown σ, cf. [5, 53].

Chernozhukov [6] study Lasso with a plug-in estimator of the noise level based on Lasso iterations in a parametric homoscedastic setting. Chrétien and Darses [24] study plug-in estimators and a trade-off penalty choice between fit and penalty in the parametric case with homoscedastic Gaussian errors under random support assumption (similar to [19]) using coherence condition. In a trace regression model for recovery of a matrix, [34] proposes and analyses a version of the $\sqrt{\text{Lasso}}$ under homoscedasticity. A comprehensive review is given in [30]. All these works rely essentially on the restricted eigenvalue condition [14] and homoscedasticity and do not differentiate penalty levels across components.

In order to address the nonparametric, heteroscedastic and non-Gaussian cases, we develop covariate-specific penalty loadings. To derive a practical and theoretically justified choice of penalty level and loadings, we need to account for the impact of the approximation error. We rely on moderate deviation theory for self-normalized sums of [33] to achieve Gaussian-like results in many non-Gaussian cases provided $\log p = o(n^{1/3})$, improving upon results derived in the parametric case that required $\log p \lesssim \log n$, see [12]. (In the context of standard Lasso, the self-normalized moderate deviation theory was first employed in [3].)

Our first contribution is the proposal of new design and noise impact factors, in order to allow for more general designs. Unlike previous conditions, these factors are tailored for establishing performance bounds with respect to the prediction norm, which is appealing in nonparametric problems. In particular, collinear designs motivate our new condition. In studying their properties, we further exploit the oracle based definition of the approximating function. The analysis based on these impact factors complements the analysis based on restricted eigenvalue proposed in [14] and compatibility condition in [48], which are more suitable for establishing rates for ℓ_k-norms.

The second contribution is a set of finite sample upper bounds and lower bounds for estimation errors under prediction norm, and upper bounds on the sparsity of the $\sqrt{\text{Lasso}}$ estimator. These results are "geometric," in that they hold conditional on the design and errors provided some key events occur. We further develop primitive sufficient conditions that allow for these results to be applied to heteroscedastic non-Gaussian errors. We also give results for other norms in the supplementary material [2].

The third contribution develops properties of the estimator that applies ordinary least squares (ols) to the model selected by $\sqrt{\text{Lasso}}$. Our focus is on the case that $\sqrt{\text{Lasso}}$ fails to achieve perfect model selection, including cases where the oracle model is not completely selected by $\sqrt{\text{Lasso}}$. This is usually the case in a nonparametric setting. This estimator intends to remove the potentially significant bias toward zero introduced by the ℓ_1-norm regularization employed in the $\sqrt{\text{Lasso}}$ estimator.

The fourth contribution is to study two extreme cases: (i) parametric noiseless case and (ii) nonparametric infinite variance case. $\sqrt{\text{Lasso}}$ has interesting theoretical properties for these two extreme cases. For case (i), $\sqrt{\text{Lasso}}$ can achieves

exact recovery in sharp contrast to Lasso. For case (ii), $\sqrt{\text{Lasso}}$ estimator can still be consistent with penalty choice that does not depend on the scale of the noise. We develop the necessary modifications of the penalty loadings and derive finite-sample bounds for the case of symmetric noise. When noise is Student's t-distribution with 2 degrees of freedom, we recover Gaussian-noise rates up to a multiplicative factor of $\log^{1/2} n$.

The final contribution is to provide an application of $\sqrt{\text{Lasso}}$ methods to a generic semiparametric problem, where some low-dimensional parameters are of interest and $\sqrt{\text{Lasso}}$ methods are used to estimate nonparametric nuisance parameters. These results extend the $\sqrt{n}$ consistency and asymptotic normality results of [3, 8] on a rather specific linear model to a generic nonlinear problem, which covers smooth frameworks in statistics and in econometrics, where the main parameters of interest are defined via nonlinear instrumental variable/moment conditions or Z-conditions containing unknown nuisance functions (as in [20]). This and all the above results illustrate the wide applicability of the proposed estimation procedure.

Notation. To make asymptotic statements, we assume that $n \to \infty$ and $p = p_n \to \infty$, and we allow for $s = s_n \to \infty$. In what follows, all parameters are indexed by the sample size n, but we omit the index whenever it does not cause confusion. We work with i.n.i.d., independent but not necessarily identically distributed data, $(w_i)_{i=1}^n$, with k-dimensional real vectors w_i containing $y_i \in \mathbb{R}$ and $z_i \in \mathbb{R}^{p_z}$, the latter taking values in a set $\mathcal{Z}$. We use the notation $(a)_+ = \max\{a, 0\}$, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. The ℓ_2-norm is denoted by $\|\cdot\|$, the ℓ_1-norm is denoted by $\|\cdot\|_1$, the ℓ_∞-norm is denoted by $\|\cdot\|_\infty$, and the ℓ_0-"norm" $\|\cdot\|_0$ denotes the number of nonzero components of a vector. The transpose of a matrix A is denoted by A'. Given a vector $\delta \in \mathbb{R}^p$, and a set of indices $T \subset \{1, \ldots, p\}$, we denote by δ_T the vector in which $\delta_{Tj} = \delta_j$ if $j \in T$, $\delta_{Tj} = 0$ if $j \notin T$, and by $|T|$ the cardinality of T. For a measurable function $f : \mathbb{R}^k \to \mathbb{R}$, the symbol $\mathrm{E}[f(w_i)]$ denotes the expected value of $f(w_i)$; $\mathbb{E}_n[f(w)]$ denotes the average $n^{-1}\sum_{i=1}^n f(w_i)$; $\bar{\mathbf{E}}[f(w)]$ denotes the average expectation $n^{-1}\sum_{i=1}^n \mathrm{E}[f(w_i)]$; and $\mathbb{G}_n(f(w))$ denotes $n^{-1/2}\sum_{i=1}^n (f(w_i) - \mathrm{E}[f(w_i)])$. We will work with regressor values $(x_i)_{i=1}^n$ generated via $x_i = P(z_i)$, where $P(\cdot) : \mathcal{Z} \mapsto \mathbb{R}^p$ is a measurable dictionary of transformations, where p is potentially larger than n. We define the prediction norm of a vector $\delta \in \mathbb{R}^p$ as $\|\delta\|_{2,n} = \{\mathbb{E}_n[(x'\delta)^2]\}^{1/2}$, and given values $y_1, \ldots, y_n$ we define $\widehat{Q}(\beta) = \mathbb{E}_n[(y - x'\beta)^2]$. We use the notation $a \lesssim b$ to denote $a \leq Cb$ for some constant $C > 0$ that does not depend on n (and, therefore, does not depend on quantities indexed by n like p or s); and $a \lesssim_P b$ to denote $a = O_P(b)$. Φ denotes the cumulative distribution of a standard Gaussian distribution and Φ^{-1} its inverse function.

2. Setting and estimators. Consider the nonparametric regression model:

$$
\begin{aligned}
&y_i = f(z_i) + \sigma\varepsilon_i, \qquad \varepsilon_i \sim F_i, \\
&\mathrm{E}[\varepsilon_i] = 0, \qquad i = 1, \ldots, n, \qquad \bar{\mathbf{E}}[\varepsilon^2] = 1,
\end{aligned} \tag{2.1}
$$

where z_i are vectors of fixed regressors, ε_i are independent errors, and σ is the scaling factor of the errors. In order to recover the regression function f, we consider linear combinations of the covariates $x_i = P(z_i)$ which are p-vectors of transformation of z_i normalized so that $\mathbb{E}_n[x_j^2] = 1$ $(j = 1, \ldots, p)$.

The goal is to estimate the value of the nonparametric regression function f at the design points, namely the values $(f_i)_{i=1}^n := (f(z_i))_{i=1}^n$. In the nonparametric settings, the regression functions f are generically nonsparse. However, often they can be well approximated by a sparse model $x'\beta_0$. One way to find such approximating model is to let β_0 be a solution of the following risk minimization problem:

$$\min_{\beta \in \mathbb{R}^p} \mathbb{E}_n\big[(f - x'\beta)^2\big] + \frac{\sigma^2 \|\beta\|_0}{n}. \tag{2.2}$$

The problem (2.2) yields the so called oracle risk—an upper bound on the risk of the best k-sparse least squares estimator in the case of homoscedastic Gaussian errors, that is, the best estimator among all least squares estimators that use k out of p components of x_i to estimate f_i. The solution β_0 achieves a balance between the mean square of the approximation error $r_i := f_i - x_i'\beta_0$ and the variance, where the latter is determined by the complexity $\|\beta_0\|_0$ of the model (number of nonzero components of β_0).

In what follows, we call β_0 the target parameter value, $T := \operatorname{supp}(\beta_0)$ the oracle model, $s := |T| = \|\beta_0\|_0$ the dimension of the oracle model, and $x_i'\beta_0$ the oracle or the best sparse approximation to f_i. We note that T is generally unknown. We summarize the preceding discussion as follows.

CONDITION ASM. We have data $\{(y_i, z_i) : i = 1, \ldots, n\}$ that for each n obey the regression model (2.1), where y_i are the outcomes, z_i are vectors of fixed basic covariates, the regressors $x_i := P(z_i)$ are transformations of z_i, and ε_i are i.n.i.d. errors. The vector β_0 is defined by (2.2) where the regressors x_i are normalized so that $\mathbb{E}_n[x_j^2] = 1$, $j = 1, \ldots, p$. We let

$$T := \operatorname{supp}(\beta_0), \qquad s := |T|, \qquad r_i := f_i - x_i'\beta_0 \quad \text{and} \quad c_s^2 := \mathbb{E}_n[r^2]. \tag{2.3}$$

REMARK 1 (Targeting $x_i'\beta_0$ is the same as targeting f_i's). We focus on estimating the oracle model $x_i'\beta_0$ using estimators of the form $x_i'\widehat{\beta}$, and we seek to bound estimation errors with respect to the prediction norm $\|\widehat{\beta} - \beta_0\|_{2,n} := \{\mathbb{E}_n[(x'\beta_0 - x'\widehat{\beta})^2]\}^{1/2}$. The bounds on estimation errors for the ultimate target f_i then follow from the triangle inequality, namely

$$\sqrt{\mathbb{E}_n\big[(f - x'\widehat{\beta})^2\big]} \le \|\widehat{\beta} - \beta_0\|_{2,n} + c_s. \tag{2.4}$$

REMARK 2 (Bounds on the approximation error). The approximation errors typically satisfy $c_s \le K\sigma\sqrt{(s \vee 1)/n}$ for some fixed constant K, since the optimization problem (2.2) balances the (squared) norm of the approximation error

(the norm of the bias) and the variance; see [4, 6, 45]. In particular, this condition holds for wide classes of functions; see Example S of Section 4 dealing with Sobolev classes and Section C.2 of supplementary material [2].

2.1. *Heteroscedastic* $\sqrt{\text{Lasso}}$. In this section, we formally define the estimators which are tailored to deal with heteroscedasticity.

We propose to define the $\sqrt{\text{Lasso}}$ estimator as

$$\widehat{\beta} \in \arg\min_{\beta\in\mathbb{R}^p} \sqrt{\widehat{Q}(\beta)} + \frac{\lambda}{n}\|\Gamma\beta\|_1, \tag{2.5}$$

where $\widehat{Q}(\beta) = \mathbb{E}_n[(y - x'\beta)^2]$, $\Gamma = \operatorname{diag}(\gamma_1, \ldots, \gamma_p)$ is a diagonal matrix of penalty loadings. The scaled ℓ_1-penalty allows component specific adjustments to more efficiently deal with heteroscedasticity.[3] Throughout, we assume $\gamma_j \geq 1$ for $j = 1, \ldots, p$.

In order to reduce the shrinkage bias of $\sqrt{\text{Lasso}}$, we consider the post model selection estimator that applies ordinary least squares (ols) to a model $\widehat{T}$ that contains the model selected by $\sqrt{\text{Lasso}}$. Formally, let $\widehat{T}$ be such that

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

and define the ols post $\sqrt{\text{Lasso}}$ estimator $\widetilde{\beta}$ associated with $\widehat{T}$ as

$$\widetilde{\beta} \in \arg\min_{\beta\in\mathbb{R}^p} \sqrt{\widehat{Q}(\beta)} : \beta_j = 0 \qquad \text{if } j \notin \widehat{T}. \tag{2.6}$$

A sensible choice for $\widehat{T}$ is simply to set $\widehat{T} = \operatorname{supp}(\widehat{\beta})$. Moreover, we allow for additional components (potentially selected through an arbitrary data-dependent procedure) to be added, which is relevant for practice.

2.2. *Typical conditions on the Gram matrix.* The Gram matrix $\mathbb{E}_n[xx']$ plays an important role in the analysis of estimators in this setup. When $p > n$, the smallest eigenvalue of the Gram matrix is 0, which creates identification problems. Thus, to restore identification, one needs to restrict the type of deviation vectors δ corresponding to the potential deviations of the estimator from the target value β_0. Because of the ℓ_1-norm regularization, the following restricted set is important:

$$\Delta_{\bar{c}} = \{\delta \in \mathbb{R}^p : \|\Gamma\delta_{T^c}\|_1 \leq \bar{c}\|\Gamma\delta_T\|_1, \delta \neq 0\} \qquad \text{for } \bar{c} \geq 1.$$

The restricted eigenvalue $\kappa_{\bar{c}}$ of the Gram matrix $\mathbb{E}_n[xx']$ is defined as

$$\kappa_{\bar{c}} := \min_{\delta\in\Delta_{\bar{c}}} \frac{\sqrt{s}\|\delta\|_{2,n}}{\|\Gamma\delta_T\|_1}. \tag{2.7}$$

[3]When errors are homoscedastic, we can set $\Gamma = I_p$. In the heteroscedastic case, using $\Gamma = I_p$ may require setting λ too conservatively, leading to over-penalization and worse performance bounds. In the paper, we develop data-dependent choice of Γ that allows us to avoid over-penalization thereby improving the performance.

The restricted eigenvalues can depend on n, T, and Γ, but we suppress the dependence in our notation. The restricted eigenvalues (2.7) are variants of the restricted eigenvalue introduced in [14] and of the compatibility condition in [48] that accommodate the penalty loadings Γ. They were proven to be useful for many designs of interest specially for establishing ℓ_k-norm rates. Below we suggest their generalizations that are useful for deriving rates in prediction norm.

The minimal and maximal m-sparse eigenvalues of a matrix M,

$$\begin{aligned}\phi_{\min}(m, M) &:= \min_{\|\delta_{T^c}\|_0 \le m, \delta \neq 0} \frac{\delta' M \delta}{\|\delta\|^2}, \\ \phi_{\max}(m, M) &:= \max_{\|\delta_{T^c}\|_0 \le m, \delta \neq 0} \frac{\delta' M \delta}{\|\delta\|^2}.\end{aligned} \tag{2.8}$$

Typically, we consider $M = \mathbb{E}_n[xx']$ or $M = \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1}$. When M is not specified, we mean $M = \mathbb{E}_n[xx']$, that is, $\phi_{\min}(m) = \phi_{\min}(m, \mathbb{E}_n[xx'])$ and $\phi_{\max}(m) = \phi_{\max}(m, \mathbb{E}_n[xx'])$. These quantities play an important role in the sparsity and post model selection analysis. Moreover, sparse eigenvalues provide a simple sufficient condition to bound restricted eigenvalues; see [14].

3. Finite-sample analysis of $\sqrt{\text{Lasso}}$. Next, we establish several finite-sample results regarding the $\sqrt{\text{Lasso}}$ estimator. Importantly, these results are based on new impact factors that can be very well behaved under repeated (i.e., collinear) regressors, and which strictly generalize the restricted eigenvalue (2.7) and compatibility constants.

The following event plays a central role in the analysis:

$$\lambda/n \ge c\|\Gamma^{-1}\widetilde{S}\|_\infty \qquad \text{where } \widetilde{S} := \mathbb{E}_n[x(\sigma\varepsilon + r)]/\sqrt{\mathbb{E}_n[(\sigma\varepsilon + r)^2]} \tag{3.1}$$

is the score of $\widehat{Q}^{1/2}$ at β_0 ($\widetilde{S} = 0$ if $\mathbb{E}_n[(\sigma\varepsilon + r)^2] = 0$). Throughout the section, we assume such event holds. Later we provide choices of λ and Γ based on primitive conditions such that the event in (3.1) holds with a high probability.

3.1. *New noise and design impact factors.* We define the following *noise* and *design* impact factors for a constant $c > 1$:

$$\varrho_c := \sup_{\|\delta\|_{2,n} > 0, \delta \in R_c} \frac{|\widetilde{S}'\delta|}{\|\delta\|_{2,n}}, \tag{3.2}$$

$$\bar{\kappa} := \inf_{\|\Gamma(\beta_0+\delta)\|_1 < \|\Gamma\beta_0\|_1} \frac{\sqrt{s}\|\delta\|_{2,n}}{\|\Gamma\beta_0\|_1 - \|\Gamma(\beta_0 + \delta)\|_1}, \tag{3.3}$$

where $R_c := \{\delta \in \mathbb{R}^p : \|\Gamma\delta\|_1 \ge c(\|\Gamma(\beta_0 + \delta)\|_1 - \|\Gamma\beta_0\|_1)\}$. For the case $\beta_0 = 0$, we define $\varrho_c = 0$ and $\bar{\kappa} = \infty$. These quantities depend on n, β_0 and Γ, albeit we suppress this when convenient.

An analysis based on the quantities ϱ_c and $\bar{\kappa}$ will be more general than the one relying only on restricted eigenvalues (2.7). This follows because (2.7) yields one possible way to bound both $\bar{\kappa}$ and ϱ_c, namely,

$$\bar{\kappa} \geq \underline{\kappa} := \inf_{\delta \in \text{int}(\Delta_1)} \frac{\sqrt{s}\|\delta\|_{2,n}}{\|\Gamma\delta_T\|_1 - \|\Gamma\delta_{T^c}\|_1} \geq \min_{\delta \in \Delta_1} \frac{\sqrt{s}\|\delta\|_{2,n}}{\|\Gamma\delta_T\|_1} \geq \min_{\delta \in \Delta_{\bar{c}}} \frac{\sqrt{s}\|\delta\|_{2,n}}{\|\Gamma\delta_T\|_1} = \kappa_{\bar{c}},$$

$$\varrho_c \leq \sup_{\delta \in \Delta_{\bar{c}}} \frac{\|\Gamma^{-1}\widetilde{S}\|_\infty \|\Gamma\delta\|_1}{\|\delta\|_{2,n}} \leq \sup_{\delta \in \Delta_{\bar{c}}} \frac{\|\Gamma^{-1}\widetilde{S}\|_\infty (1+\bar{c})\|\Gamma\delta_T\|_1}{\|\delta\|_{2,n}}$$

$$\leq \frac{(1+\bar{c})\sqrt{s}}{\kappa_{\bar{c}}} \|\Gamma^{-1}\widetilde{S}\|_\infty,$$

where $c > 1$ and $\bar{c} := (c+1)/(c-1) > 1$. The quantities $\bar{\kappa}$ and ϱ_c can be well behaved (i.e., $\bar{\kappa} > 0$ and $\varrho_c < \infty$) even in the presence of repeated (i.e., collinear) regressors (see Remark 4 for a simple example), while restricted eigenvalues and compatibility constants would be zero in that case.

The design impact factor $\bar{\kappa}$ in (3.3) strictly generalizes the original restricted eigenvalue (2.7) proposed in [14] and the compatibility constants proposed in [48] and in [46].[4] The design conditions based on these concepts are relatively weak, and hence (3.3) is a useful concept.

The noise impact factor ϱ_c also plays an important role. It depends on the noise, design and approximation errors, and can be controlled via empirical process methods. Note that under (3.1), the deviation $\widehat{\delta} = \widehat{\beta} - \beta_0$ of the $\sqrt{\text{Lasso}}$ estimator from β_0 obeys $\widehat{\delta} \in R_c$, explaining its appearance in the definition of ϱ_c. The lemmas below summarize the above discussion.

LEMMA 1 (Bounds on and invariance of design impact factor). *Under Condition ASM, we have $\bar{\kappa} \geq \underline{\kappa} \geq \kappa_1 \geq \kappa_{\bar{c}}$. Moreover, if copies of regressors are included with the same corresponding penalty loadings, the lower bound $\underline{\kappa}$ on $\bar{\kappa}$ does not change.*

LEMMA 2 (Bounds on and invariance of noise impact factor). *Under Condition ASM, we have $\varrho_c \leq (1+\bar{c})\sqrt{s}\|\Gamma^{-1}\widetilde{S}\|_\infty/\kappa_{\bar{c}}$. Moreover, if copies of regressors with indices $j \in T^c$ are included with the same corresponding penalty loadings, ϱ_c does not change (see also Remark 4).*

[4]The compatibility condition defined in [46] is defined as: $\exists \nu(T) > 0$ such that

$$\inf_{\delta \in \Delta_3} \frac{\sqrt{s}\|\delta\|_{2,n}}{(1+\nu(T))\|\Gamma\delta_T\|_1 - \|\Gamma\delta_{T^c}\|_1} > 0.$$

We have that $\bar{\kappa} \geq \underline{\kappa}$, where $\underline{\kappa}$ corresponds to setting $\nu(T) = 0$ and using Δ_1 in place of Δ_3, which strictly weakens [46]'s definition. Allowing for $\nu(T) = 0$ is necessary for allowing collinear regressors.

LEMMA 3 (Estimators belong to restricted sets). *Assume that for some* $c > 1$ *we have* $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$, *then* $\widehat{\delta} \in R_c$. *The latter condition implies that* $\widehat{\delta} \in \Delta_{\bar{c}}$ *for* $\bar{c} = (c+1)/(c-1)$.

3.2. *Finite-sample bounds on* $\sqrt{\text{Lasso}}$. In this section, we derive finite-sample bounds for the prediction norm of the $\sqrt{\text{Lasso}}$ estimator. These bounds are established under heteroscedasticity, without knowledge of the scaling parameter σ, and using the impact factors proposed in Section 3.1. For $c > 1$, let $\bar{c} = (c+1)/(c-1)$ and consider the conditions

$$\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty \quad \text{and} \quad \bar{\zeta} := \lambda\sqrt{s}/(n\bar{\kappa}) < 1. \tag{3.4}$$

THEOREM 1 (Finite sample bounds on estimation error). *Under Condition* ASM *and* (3.4), *we have*

$$\|\widehat{\beta} - \beta_0\|_{2,n} \leq 2\sqrt{\widehat{Q}(\beta_0)}B_n, \qquad B_n := \frac{\varrho_c + \bar{\zeta}}{1 - \bar{\zeta}^2}.$$

We recall that the choice of λ does not depend on the scaling parameter σ. The impact of σ in the bound of Theorem 1 comes through the factor $\widehat{Q}^{1/2}(\beta_0) \leq \sigma\sqrt{\mathbb{E}_n[\varepsilon^2]} + c_s$ where c_s is the size of the approximation error defined in Condition ASM. Moreover, under typical conditions that imply $\kappa_{\bar{c}}$ to be bounded away from zero, for example, under Condition P of Section 4 and standard choice of penalty, we have with a high probability

$$B_n \lesssim \sqrt{\frac{s\log(p \vee n)}{n}} \quad \Longrightarrow \quad \|\widehat{\beta} - \beta_0\|_{2,n} \lesssim \sigma\sqrt{\frac{s\log(p \vee n)}{n}}.$$

Thus, Theorem 1 generally leads to the same rate of convergence as in the case of the Lasso estimator that knows σ since $\mathbb{E}_n[\varepsilon^2]$ concentrates around 1 under (2.1) and provided a law of large numbers holds. We derive performance bounds for other norms of interest in the supplementary material [2].

The next result deals with $\widehat{Q}(\widehat{\beta})$ as an estimator for $\widehat{Q}(\beta_0)$ and σ^2.

THEOREM 2 (Estimation of σ). *Under Condition* ASM *and* (3.4)

$$-2\varrho_c\sqrt{\widehat{Q}(\beta_0)}B_n \leq \sqrt{\widehat{Q}(\widehat{\beta})} - \sqrt{\widehat{Q}(\beta_0)} \leq 2\bar{\zeta}\sqrt{\widehat{Q}(\beta_0)}B_n.$$

Under only Condition ASM, *we have*

$$|\sqrt{\widehat{Q}(\widehat{\beta})} - \sigma| \leq \|\widehat{\beta} - \beta_0\|_{2,n} + c_s + \sigma|\mathbb{E}_n[\varepsilon^2] - 1|.$$

We note that further bounds on $|\mathbb{E}_n[\varepsilon^2] - 1|$ are implied by von Bahr–Esseen's and Markov's inequalities, or by self-normalized moderate deviation (SNMD) theory as in Lemma 4. As a result, the theorem implies consistency $|\widehat{Q}^{1/2}(\widehat{\beta}) - \sigma| = o_P(1)$ under mild moment conditions; see Section 4. Theorem 2 is also useful for establishing the following sparsity properties.

THEOREM 3 (Sparsity bound for $\sqrt{\text{Lasso}}$). *Suppose Condition* ASM, (3.4), $\widehat{Q}(\beta_0) > 0$, *and* $2\varrho_c B_n \leq 1/(c\bar{c})$. *Then we have*

$$|\operatorname{supp}(\widehat{\beta})| \leq s \cdot 4\bar{c}^2(B_n/\bar{\zeta}\bar{\kappa})^2 \min_{m\in\mathcal{M}} \phi_{\max}\big(m, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1}\big),$$

where $\mathcal{M} = \{m \in \mathbb{N} : m > s\phi_{\max}(m, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1}) \cdot 8\bar{c}^2(B_n/(\bar{\zeta}\bar{\kappa}))^2\}$. *Moreover, if* $\kappa_{\bar{c}} > 0$ *and* $\bar{\zeta} < 1/\sqrt{2}$ *we have*

$$|\operatorname{supp}(\widehat{\beta})| \leq s \cdot \big(4\bar{c}^2/\kappa_{\bar{c}}\big)^2 \min_{m\in\mathcal{M}^*} \phi_{\max}\big(m, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1}\big),$$

where $\mathcal{M}^* = \{m \in \mathbb{N} : m > s\phi_{\max}(m, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1}) \cdot 2(4\bar{c}^2/\kappa_{\bar{c}})^2\}$.

REMARK 3 (On the sparsity bound). Section 4 will show that under minimal and maximal sparse eigenvalues of order $s \log n$ bounded away from zero and from above, Theorem 3 implies that with a high probability

$$|\operatorname{supp}(\widehat{\beta})| \lesssim s := |\operatorname{supp}(\beta_0)|.$$

That is, the selected model's size will be of the same order as the size of the oracle model. We note, however, that the former condition is merely a sufficient condition. The bound $|\operatorname{supp}(\widehat{\beta})| \lesssim s$ will apply for other designs of interest. This can be the case even if $\kappa_{\bar{c}} = 0$ (e.g., in the aforementioned design, if we change it by adding a single repeated regressor).

REMARK 4 (Maximum sparse eigenvalue and sparsity). Consider the case of $f(z) = z$ with p repeated regressors $x_i = (z_i, \ldots, z_i)'$ where $|z| \leq K$. In this case, one could set $\Gamma = I \cdot K$. In this setting, there is a sparse solution for $\sqrt{\text{Lasso}}$, but there is also a solution which has all p nonzero coefficients. Nonetheless, the bound for the prediction error rate will be well behaved since $\bar{\kappa}$ and $\bar{\zeta}$ are invariant to the addition of copies of z and

$$\bar{\kappa} \geq 1/K \quad \text{and} \quad \varrho_c = |\mathbb{E}_n[\varepsilon z]|/\{\mathbb{E}_n[\varepsilon^2]\mathbb{E}_n[z^2]\}^{1/2} \lesssim_P 1/\sqrt{n}$$

under mild moment conditions on the noise (e.g., $\bar{\mathbf{E}}[|\varepsilon|^3] \leq C$). In this case, $\phi_{\max}(m, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1}) = (m+1)\mathbb{E}_n[z^2]/K^2$ and the set $\mathcal{M}$ only contains integers larger than p, leading to the trivial bound $\widehat{m} \leq p$.

3.3. *Finite-sample bounds on ols post* $\sqrt{\text{Lasso}}$. Next, we consider the ols estimator applied to the model $\widehat{T}$ that was selected by $\sqrt{\text{Lasso}}$ or includes such model (plus other components that the data analyst may wish to include), namely $\operatorname{supp}(\widehat{\beta}) \subseteq \widehat{T}$. We are interested in the case when model selection does not work perfectly, as occurs in applications.

The following result establishes performance bounds for the ols post $\sqrt{\text{Lasso}}$ estimator. Following [6], the analysis accounts for the data-driven choice of components and for the possibly misspecified selected model (i.e., $T \not\subseteq \widehat{T}$).

THEOREM 4 (Performance of ols post $\sqrt{\text{Lasso}}$). *Under Condition* ASM *and* (3.4), *let* $\operatorname{supp}(\widehat{\beta}) \subseteq \widehat{T}$, *and* $\widehat{m} = |\widehat{T} \setminus T|$. *Then we have that the ols post* $\sqrt{\text{Lasso}}$ *estimator based on* $\widehat{T}$ *satisfies*

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \leq \frac{\sigma\sqrt{s+\widehat{m}}\|\mathbb{E}_n[x\varepsilon]\|_\infty}{\sqrt{\phi_{\min}(\widehat{m})}} + 2c_s + 2\sqrt{\widehat{Q}(\beta_0)}B_n.$$

The result is derived from the sparsity of the model $\widehat{T}$ and from its approximating ability. Note the presence of the new term $\|\mathbb{E}_n[x\varepsilon]\|_\infty$. Bounds on $\|\mathbb{E}_n[x\varepsilon]\|_\infty$ can be derived using the same tools used to justify the penalty level λ, via moderate deviation theory for self-normalized sums [33], Gaussian approximations to empirical processes [22, 23] or empirical process inequalities as in [5]. Under mild conditions, we have $\|\mathbb{E}_n[x\varepsilon]\|_\infty \leq C\sqrt{\log(pn)/n}$ with probability $1 - o(1)$.

3.4. *Two extreme cases.* *Case* (i): *Parametric noiseless case.* Consider the case that $\sigma = 0$ and $c_s = 0$. Therefore, the regression function is exactly sparse, $f(z_i) = x_i'\beta_0$. In this case, $\sqrt{\text{Lasso}}$ can exactly recover the f and even β_0 under weak conditions under a broad range of penalty levels.

THEOREM 5 (Exact recovery for the parametric noiseless case). *Under Condition* ASM, *let* $\sigma = 0$ *and* $c_s = 0$. *Suppose that* $\lambda > 0$ *obeys the growth restriction* $\bar{\zeta} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$. *Then we have* $\|\widehat{\beta} - \beta_0\|_{2,n} = 0$, *and if, moreover,* $\kappa_1 > 0$, *then* $\widehat{\beta} = \beta_0$.

REMARK 5 (Perfect recovery and Lasso). It is worth mentioning that for any $\lambda > 0$, unless $\beta_0 = 0$, Lasso cannot achieve exact recovery. Moreover, it is not obvious how to properly set the penalty level for Lasso even if we knew a priori that it is a parametric noiseless model. In contrast, $\sqrt{\text{Lasso}}$ can automatically adapt to the noiseless case.

Case (ii): *Nonparametric infinite variance*. We conclude this section with the infinite variance case. The finite sample theory does not rely on $\mathrm{E}[\varepsilon^2] < \infty$. Instead it relies on the choice of penalty level and penalty loadings to satisfy $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$. Under symmetric errors, we exploit the sub-Gaussian property of self-normalized sums [25] to develop a choice of penalty level λ and loadings $\Gamma = \operatorname{diag}(\gamma_j, j = 1, \ldots, p)$, where

$$\lambda = (1+u_n)c\sqrt{n}\{1 + \sqrt{2\log(2p/\alpha)}\} \quad \text{and} \quad \gamma_j = \max_{1\leq i\leq n}|x_{ij}|, \tag{3.5}$$

where u_n is defined below and typically we can select $u_n = o(1)$.

THEOREM 6 ($\sqrt{\text{Lasso}}$ prediction norm for symmetric errors). *Consider a nonparametric regression model with data* $(y_i, z_i)_{i=1}^n$, $y_i = f(z_i) + \sigma\varepsilon_i$, $x_i = P(z_i)$

such that $\mathbb{E}_n[x_j^2] = 1$ $(j = 1, \ldots, p)$, ε_i*'s are independent symmetric errors, and* β_0 *defined as any solution to* (2.2). *Let the penalty level and loadings as in* (3.5). *Assume that there exist sequences of constants* $\eta_1 \geq 0$ *and* $\eta_2 \geq 0$ *both converging to* 0 *and a sequence of constants* $0 \leq u_n \leq 1$ *such that* $P(\mathbb{E}_n[\sigma\varepsilon^2] > (1+u_n)\mathbb{E}_n[(\sigma\varepsilon + r)^2]) \leq \eta_1$ *and* $P(\mathbb{E}_n[\varepsilon^2] \leq \{1+u_n\}^{-1}) \leq \eta_2$ *for all* n. *If* $\bar{\zeta} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$, *then with probability at least* $1 - \alpha - \eta_1 - \eta_2$ *we have* $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$ *and*

$$\|\widehat{\beta} - \beta_0\|_{2,n} \leq 2B_n\Big(c_s + \sigma\sqrt{\mathbb{E}_n[\varepsilon^2]}\Big).$$

The rate of convergence will be affected by how fast $\mathbb{E}_n[\varepsilon^2]$ diverges. That is, the final rate will depend on the particular tail properties of the distribution of the noise. The rate also depends on u_n through λ. In many examples, u_n can be chosen as a constant or even a sequence going to zero sufficiently slowly, as in the next corollary where ε_i follows a t distribution with 2 degrees of freedom, that is, $\varepsilon_i \sim t(2)$.

COROLLARY 1 [$\sqrt{\text{Lasso}}$ prediction norm for $\varepsilon_i \sim t(2)$]. *Under the setting of Theorem* 6, *suppose that* $\varepsilon_i \sim t(2)$ *and are i.i.d. for all* i. *Then for any* $\tau \in (0, 1/2)$, *with probability at least* $1 - \alpha - \frac{3}{2}\tau - \frac{2\log(4n/\tau)}{nu_n/[1+u_n]} - \frac{72\log^2 n}{n^{1/2}(\log n - 6)^2}$, *we have* $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$ *and, if* $\bar{\zeta} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$, *we have*

$$\|\widehat{\beta} - \beta_0\|_{2,n} \leq 2\big(c_s + \sigma\sqrt{\log(4n/\tau) + 2\sqrt{2}/\tau}\big)B_n.$$

REMARK 6 [Asymptotic performance in $t(2)$ case]. Provided that regressors are uniformly bounded and satisfy the sparse eigenvalues condition (4.3), we have that the restricted eigenvalue $\kappa_{\bar{c}}$ is bounded away from zero for the specified choice of Γ. Because Corollary 1 ensures $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$ with the stated probability, by Lemmas 1 and 2 we have

$$\varrho_c + \bar{\zeta} \lesssim \frac{\lambda\sqrt{s}}{n\kappa_{\bar{c}}} \lesssim (1+u_n)\sqrt{\frac{s\log(p\vee n)}{n}} \quad \Longrightarrow \quad B_n \lesssim \sqrt{\frac{s\log(p\vee n)}{n}}.$$

Therefore, under these design conditions, assuming that $s\log(p/\alpha) = o(n)$ and that σ is fixed, and setting $1/\alpha = o(\log n)$, we can select $u_n = 1/2$ and $\tau = 1/\log n$ in Corollary 1, to conclude that the $\sqrt{\text{Lasso}}$ estimator satisfies

$$\|\widehat{\beta} - \beta_0\|_{2,n} \lesssim (c_s + \sigma\sqrt{\log n})\sqrt{\frac{s\log(p\vee n)}{n}}, \tag{3.6}$$

with probability $1 - \alpha(1 + o(1))$. Despite the infinite variance, the bound (3.6) differs from the Gaussian noise case only by a $\sqrt{\log n}$ factor.

4. Asymptotics analysis under primitive conditions. In this section, we formally state an algorithm to compute the estimators and we provide rates of convergence results under simple primitive conditions.

We propose setting the penalty level as

$$\lambda = c\sqrt{n}\Phi^{-1}(1-\alpha/2p), \tag{4.1}$$

where α controls the confidence level, and $c > 1$ is a slack constant similar to [14], and the penalty loadings according to the following iterative algorithm.

ALGORITHM 1 (Estimation of square-root Lasso loadings). *Choose* $\alpha \in (1/n, 1/2]$ *and a constant* $K \geq 1$ *as an upper bound on the number of iterations.* (0) *Set* $k = 0$, λ *as in* (4.1), *and* $\widehat{\gamma}_{j,0} = \max_{1\leq i\leq n} |x_{ij}|$ *for each* $j = 1, \ldots, p$. (1) *Compute the* $\sqrt{\text{Lasso}}$ *estimator* $\widehat{\beta}$ *based on the current penalty loadings* $\Gamma = \widehat{\Gamma}_k = \text{diag}\{\widehat{\gamma}_{j,k}, j = 1, \ldots, p\}$. (2) *Set*

$$\widehat{\gamma}_{j,k+1} := 1 \vee \sqrt{\mathbb{E}_n\big[x_j^2\big(y - x'\widehat{\beta}\big)^2\big]}\Big/\sqrt{\mathbb{E}_n\big[\big(y - x'\widehat{\beta}\big)^2\big]}.$$

(3) *If* $k > K$, *stop*; *otherwise set* $k \leftarrow k + 1$ *and go to step* 1.

REMARK 7 (Parameters of the algorithm). The parameter $1 - \alpha$ is a confidence level which guarantees near-oracle performance with probability at least $1 - \alpha$; we recommend $\alpha = 0.05/\log n$. The constant $c > 1$ is the slack parameter used as in [14]; we recommend $c = 1.01$. In order to invoke moderate deviation theorem for self-normalized sums, we need to be able to bound with a high probability:

$$\sqrt{\mathbb{E}_n\big[x_j^2\varepsilon^2\big]}\Big/\sqrt{\mathbb{E}_n\big[\varepsilon^2\big]} \leq \gamma_{j,0}. \tag{4.2}$$

The choice of $\widehat{\gamma}_{j,0} = \max_{1\leq i\leq n} |x_{ij}|$ automatically achieves (4.2). Nonetheless, we recommend iterating the procedure to avoid unnecessary over-penalization, since at each iteration more precise estimates of the penalty loadings are achieved. These recommendations are valid either in finite or large samples under the conditions stated below. They are also supported by the numerical experiments (see Section G of supplementary material [2]).

REMARK 8 (Alternative estimation of loadings). Algorithm 1 relies on the $\sqrt{\text{Lasso}}$ estimator $\widehat{\beta}$. Another possibility is to use the ols post $\sqrt{\text{Lasso}}$ estimator $\widetilde{\beta}$. This leads to similar theoretical and practical results. Moreover, we can define the initial penalty loading as $\widehat{\gamma}_{j,0} = W\{\mathbb{E}_n[x_j^4]\}^{1/4}$ where the kurtosis parameter $W > \{\bar{\mathbf{E}}[\varepsilon^4]\}^{1/4}/\{\bar{\mathbf{E}}[\varepsilon^2]\}^{1/2}$ is pivotal with respect to the scaling parameter σ, but we need to assume an upper bound for this quantity. The purpose of this parameter is to bound the kurtosis of the marginal distribution of errors, namely that of $\bar{F}_\varepsilon(v) = n^{-1}\sum_{i=1}^n P(\varepsilon_i \leq v)$. We recommend $W = 2$, which permits a wide class

of marginal distributions of errors, in particular it allows $\bar{F}_\varepsilon$ to have tails as heavy as those of $t(a)$ with $a > 5$. This method also achieves (4.2); see Section C.1 of supplementary material [2].

The following is a set of simple sufficient conditions which yields practical corollaries. Let $\ell_n \nearrow \infty$ be a sequence of positive constants.

CONDITION P. The noise and regressors obey $\sup_{n\geq 1} \bar{\mathbf{E}}[|\varepsilon|^q] < \infty$, $q > 4$, $\inf_{n\geq 1} \min_{1\leq j\leq p} \mathbb{E}_n[x_j^2 \mathrm{E}[\varepsilon^2]] > 0$, $\sup_{n\geq 1} \max_{1\leq j\leq p} \mathbb{E}_n[|x_j|^3 \mathrm{E}[|\varepsilon|^3]] < \infty$ and

$$\sup_{n\geq 1} \phi_{\max}\big(s\ell_n, \mathbb{E}_n[xx']\big)/\phi_{\min}\big(s\ell_n, \mathbb{E}_n[xx']\big) < \infty. \tag{4.3}$$

Moreover, we have that $\max_{i\leq n, j\leq p} |x_{ij}|^2/\ell_n = o(1)$, $\log p \leq C(n/\log^2 n)^{1/3}$, $\ell_n^2 s \log(p \vee n) \leq Cn/\log n$, $s \geq 1$, and $c_s^2 \leq C\sigma^2(s\log(p\vee n)/n)$.

Condition P imposes conditions on moments that allow us to use results of the moderate deviation theory for self-normalized sums, weak requirements on (s, p, n), well behaved sparse eigenvalues as a sufficient condition on the design to bound the impact factors and a mild condition on the approximation errors (see Remark 2 for a discussion and references).

The proofs in this section rely on the following result due to [33].

LEMMA 4 (Moderate deviations for self-normalized sums). *Let $X_1, \ldots, X_n$ be independent, zero-mean random variables and $\delta \in (0, 1]$. Let $S_{n,n} = n\mathbb{E}_n[X]$, $V_{n,n}^2 = n\mathbb{E}_n[X^2]$, $M_n = \{\bar{\mathbf{E}}[|X|^{2+\delta}]\}^{1/\{2+\delta\}}/\{\bar{\mathbf{E}}[X^2]\}^{1/2} < \infty$ and $\jmath_n \leq n^{\delta/(2(2+\delta))} M_n^{-1}$. For some absolute constant A, uniformly on $0 \leq |x| \leq n^{\delta/(2(2+\delta))} M_n^{-1}/\jmath_n - 1$, we have*

$$\left| \frac{P(S_{n,n}/V_{n,n} \geq x)}{(1-\Phi(x))} - 1 \right| \leq \frac{A}{\jmath_n^{2+\delta}}.$$

The following theorem summarizes the asymptotic performance of $\sqrt{\text{Lasso}}$, based upon Algorithm 1, for commonly used designs.

THEOREM 7 (Performance of $\sqrt{\text{Lasso}}$ and ols post $\sqrt{\text{Lasso}}$ under Condition P). *Suppose Conditions ASM and P hold. Let $\alpha \in (1/n, 1/\log n)$, $c \geq 1.01$, the penalty level λ be set as in (4.1) and the penalty loadings as in Algorithm 1. Then for all $n \geq n_0$, with probability at least $1 - \alpha\{1 + \bar{C}/\log n\} - \bar{C}\{n^{-1/2}\log n + n^{1-q/4}\}$ we have*

$$\|\widehat{\beta} - \beta_0\|_{2,n} \leq \sigma\bar{C}\sqrt{\frac{s\log(n \vee (p/\alpha))}{n}},$$

$$\sqrt{\mathbb{E}_n\big[\big(f - x'\widehat{\beta}\big)^2\big]} \le \sigma \bar{C}\sqrt{\frac{s\log(n \vee (p/\alpha))}{n}},$$

$$\|\widehat{\beta} - \beta_0\|_1 \le \sigma \bar{C}\sqrt{\frac{s^2\log(n \vee (p/\alpha))}{n}} \quad \textit{and} \quad \big|\operatorname{supp}(\widehat{\beta})\big| \le \bar{C}s,$$

where n_0 and $\bar{C}$ depend only on the constants in Condition P. *Moreover, the ols post* $\sqrt{\text{Lasso}}$ *estimator satisfies with the same probability for all $n \ge n_0$,*

$$\|\widetilde{\beta} - \beta_0\|_{2,n} \le \sigma \bar{C}\sqrt{\frac{s\log(n \vee (p/\alpha))}{n}},$$

$$\sqrt{\mathbb{E}_n\big[\big(f - x'\widetilde{\beta}\big)^2\big]} \le \sigma \bar{C}\sqrt{\frac{s\log(n \vee (p/\alpha))}{n}} \quad \textit{and}$$

$$\|\widehat{\beta} - \beta_0\|_1 \le \sigma \bar{C}\sqrt{\frac{s^2\log(n \vee (p/\alpha))}{n}}.$$

REMARK 9 (Gaussian-like performance and normalization assumptions). Theorem 7 yields bounds on the estimation errors that are "Gaussian-like," namely the factor $\sqrt{\log(p/\alpha)}$ and other constants in the performance bound are the same as if errors were Gaussian, but the probabilistic guarantee is not $1 - \alpha$ but rather $1 - \alpha + o(1)$, which together with mildly more restrictive growth conditions is the cost of non-Gaussianity. We also note that the normalization $\mathbb{E}_n[x_j^2] = 1$, $j = 1, \ldots, p$ is not used in the construction of the estimator, and the results of the theorem hold under the condition: $C_1 \le \mathbb{E}_n[x_j^2] \le C_2$, $j = 1, \ldots, p$ uniformly for all $n \ge n_0$, for some positive, finite constants C_1 and C_2.

The results above establish that $\sqrt{\text{Lasso}}$ achieves the same near oracle rate of convergence as Lasso despite not knowing the scaling parameter σ. They allow for heteroscedastic errors with mild restrictions on its moments. Moreover, it allows for an arbitrary number of iterations. The results also establish that the upper bounds on the rates of convergence of $\sqrt{\text{Lasso}}$ and ols post $\sqrt{\text{Lasso}}$ coincide under these conditions. This is confirmed also by Monte–Carlo experiments reported in the supplementary material [2], with ols post $\sqrt{\text{Lasso}}$ performing no worse and often outperforming $\sqrt{\text{Lasso}}$ due to having a much smaller bias. Notably, this theoretical and practical performance occurs despite the fact that $\sqrt{\text{Lasso}}$ may in general fail to correctly select the oracle model T as a subset and potentially select variables not in T.

EXAMPLE S (Performance for Sobolev balls and p-rearranged Sobolev balls). In this example, we show how our results apply to an important class of Sobolev functions, and illustrates how modern selection drastically reduces the dependency on knowing the order of importance of the basis functions.

Suppose that z_i's are generated as i.i.d. from Uniform$(0,1)$, x_i's are formed as $(x_{ij})_{j=1}^p$ with $x_{ij} = P_j(z_i)$, $\sigma = 1$, and $\varepsilon_i \sim N(0,1)$. Following [45], consider an orthonormal bounded basis $\{P_j(\cdot)\}_{j=1}^\infty$ in $L^2[0,1]$, consider functions $f(z) = \sum_{j=1}^\infty \theta_j P_j(z)$ in a Sobolev space $\mathcal{S}(\alpha, L)$ for some $\alpha \geq 1$ and $L > 0$. This space consists of functions whose Fourier coefficients θ satisfy $\sum_{j=1}^\infty |\theta_j| < \infty$ and

$$\theta \in \Theta(\alpha, L) = \left\{ \theta \in \ell^2(\mathbb{N}) : \sum_{j=1}^\infty j^{2\alpha} \theta_j^2 \leq L^2 \right\}.$$

We also consider functions in a p-rearranged Sobolev space $\mathcal{RS}(\alpha, p, L)$. These functions take the form $f(z) = \sum_{j=1}^\infty \theta_j P_j(z)$ such that $\sum_{j=1}^\infty |\theta_j| < \infty$ and $\theta \in \Theta^R(\alpha, p, L)$, where

$$\Theta^R(\alpha, p, L) = \left\{ \theta \in \ell^2(\mathbb{N}) : \begin{array}{c} \exists \text{ permutation} \Upsilon : \{1, \ldots, p\} \to \{1, \ldots, p\} \\ \sum_{j=1}^p j^{2\alpha} \theta_{\Upsilon(j)}^2 + \sum_{j=p+1}^\infty j^{2\alpha} \theta_j^2 \leq L^2 \end{array} \right\}.$$

Note that $\mathcal{S}(\alpha, L) \subset \mathcal{RS}(\alpha, p, L)$.

In the supplementary material [2], we show that the rate-optimal choice for the size of the support of the oracle model β_0 is $s \lesssim n^{1/[2\alpha+1]}$. This rate can be achieved with the support consisting of indices j that correspond to the s largest coefficients $|\theta_j|$. The oracle projection estimator $\widehat{\beta}^{\text{or}}$ that uses these "ideal" s components achieves optimal prediction error rate uniformly over the regression functions $f \in \mathcal{S}(\alpha, L)$ or $f \in \mathcal{RS}(\alpha, p, L)$: $(\mathbb{E}_n[\{f - \sum_{j=1}^\infty \widehat{\beta}_j^{\text{or}} P_j(z)\}^2])^{1/2} \lesssim_P n^{-\alpha/[2\alpha+1]}$. Under mild regularity conditions, as in Theorem 7, $\sqrt{\text{Lasso}}$ estimator $\widehat{\beta}$ that uses $x_i = (P_1(z_i), \ldots, P_p(z))'$ achieves a near-optimal rate uniformly over the regression functions $f \in \mathcal{S}(\alpha, L)$ or $f \in \mathcal{RS}(\alpha, p, L)$:

$$\sqrt{\mathbb{E}_n\big[\big(f - x'\widehat{\beta}\big)^2\big]} \lesssim_P n^{-\alpha/[2\alpha+1]} \sqrt{\log(n \vee p)},$$

without knowing the "ideal" s components among x_i. The same statement also holds for the ols post $\sqrt{\text{Lasso}}$ estimator $\widetilde{\beta}$.

Therefore, the $\sqrt{\text{Lasso}}$ and ols post $\sqrt{\text{Lasso}}$ estimators achieve near oracle rates uniformly over rearranged Sobolev balls under mild conditions. As a contrast, consider the "naive oracle" series projection estimator that uses the first s components of the basis, assuming that the parameter space is $\mathcal{S}(\alpha, L)$. This estimator achieves the optimal rate for the Sobolev space $\mathcal{S}(\alpha, L)$, but fails to be uniformly consistent over p-rearranged Sobolev space $\mathcal{RS}(\alpha, p, L)$, since we can select a model $f \in \mathcal{RS}(\alpha, p, L)$ such that its first s Fourier coefficients are zero, and the remaining coefficients are nonzero, therefore, the "naive oracle" fit will be 0 plus a centered noise, and the estimator will be inconsistent for this f.

We proceed to state a result on estimation of σ^2 under the asymptotic framework.

COROLLARY 2 (Estimation of σ^2 under asymptotics). *Suppose Conditions* ASM *and* P *hold. Let* $\alpha \in (1/n, 1/\log n)$, $c \geq 1.01$, *the penalty level* λ *be set as in* (4.1) *and the penalty loadings as in Algorithm* 1. *Then for all* $n \geq n_0$, *with probability at least* $1 - \alpha\{1 + \bar{C}/\log n\} - \bar{C}\{n^{-1/2}\log n + n^{1-q/4}\} - 2\delta$,

$$|\widehat{Q}(\widehat{\beta}) - \sigma^2| \leq \frac{\sigma^2 \bar{C} s \log(n \vee (p/\alpha))}{n} + \frac{\sigma^2 \bar{C}\sqrt{s \log(p \vee n)}}{\sqrt{\delta} n^{1-1/q}} + \frac{\sigma^2 \bar{C}}{\sqrt{\delta n}}.$$

Moreover, provided further that $s^2 \log^2(p \vee n) \leq Cn/\log n$, *we have that*

$$\{\sigma^2 \xi_n\}^{-1} n^{1/2}(\widehat{Q}(\widehat{\beta}) - \sigma^2) \Rightarrow N(0, 1),$$

where $\xi_n^2 = \bar{\mathbf{E}}[\{\varepsilon^2 - \mathrm{E}[\varepsilon^2]\}^2]$.

This result extends [6, 44] to the heteroscedastic, non-Gaussian cases.

5. An application to a generic semi-parametric inference problem. In this section, we present a generic application of the methods of this paper to semiparametric problems, where some lower-dimensional structural parameter is of interest and the √Lasso or ols post √Lasso are used to estimate the high-dimensional nuisance function. We denote the true value of the target parameter by $\theta_0 \in \Theta \subset \mathbb{R}^d$, and assume that it satisfies the following moment condition:

$$\mathrm{E}[\psi(w_i, \theta_0, h_0(z_i))] = 0, \qquad i = 1, \ldots, n, \tag{5.1}$$

where w_i is a random vector taking values in $\mathcal{W}$, containing vector z_i taking values in $\mathcal{Z}$ as a subcomponent; the function $(w, \theta, t) \mapsto \psi(w, \theta, t) = (\psi_j(w, \theta, t))_{j=1}^d$ is a measurable map from an open neighborhood of $\mathcal{W} \times \Theta \times T$, a subset of $\mathbb{R}^{d_w + d + d_t}$, to $\mathbb{R}^d$, and $z \mapsto h_0(z) = (h_{0m}(z))_{m=1}^M$ is the nuisance function mapping $\mathcal{Z}$ to $T \subset \mathbb{R}^M$. We note that M and d are fixed and do not depend on n in what follows.

Perhaps the simplest, that is linear, example of this kind arises in the instrumental variable (IV) regression problem in [3, 8], where $\psi(w_i, \theta_0, h_0(z_i)) = (u_i - \theta_0 d_i)h_0(z_i)$, where u_i is the response variable, d_i is the endogenous variable, z_i is the instrumental variable, $h_0(z_i) = \mathrm{E}[d_i|z_i]$ is the optimal instrument, and $\mathrm{E}[(u_i - \theta_0 d_i)|z_i] = 0$. Other examples include partially linear models, heterogeneous treatment effect models, nonlinear instrumental variable, Z-problems as well as many others (see, e.g., [1, 3, 7, 9–11, 13, 20, 28, 29, 31, 32, 49, 55, 56]), which all give rise to nonlinear moment conditions with respect to the nuisance functions.

We assume that the nuisance functions h_0 arise as conditional expectations of some variables that can be modeled and estimated in the approximately sparse framework, as formally described below. For instance, in the example mentioned above, the function h_0 is indeed a conditional expectation of the endogenous variable given the instrumental variable. We let $\widehat{h} = (\widehat{h}_m)_{m=1}^M$ denote the estimator of h_0, which obeys conditions stated below. The estimator $\widehat{\theta}$ of θ_0 is constructed as

any approximate ε_n-solution in Θ to a sample analog of the moment condition above:

$$\|\mathbb{E}_n[\psi(w,\widehat{\theta},\widehat{h}(z))]\| \le \varepsilon_n \qquad \text{where } \varepsilon_n = o(n^{-1/2}). \tag{5.2}$$

The key condition needed for regular estimation of θ_0 is the orthogonality condition:

$$\mathrm{E}[\partial_t \psi(w_i,\theta_0,h_0(z_i))|z_i] = 0, \qquad i = 1,\dots,n, \tag{5.3}$$

where here and below we use the symbol ∂_t to abbreviate $\frac{\partial}{\partial t'}$. For instance, in the IV example this condition holds, since $\partial_t \psi(w_i,\theta_0,h_0(z_i)) = (u_i - \theta_0 d_i)$ and $\mathrm{E}[(u_i - \theta_0 d_i)|z_i] = 0$ by assumption. In other examples, it is important to construct the scores that have this orthogonality property. Generally, if we have a score, which identifies the target parameter but does not have the orthogonality property, we can construct the score that has the required property by projecting the original score onto the orthocomplement of the tangent space for the nuisance parameter; see, for example, [36, 50, 51] for detailed discussions. This often results in a semiparametrically efficient score function.

The orthogonality condition reduces sensitivity to "crude" estimation of the nuisance function h_0. Indeed, under appropriate sparsity assumptions stated below, the estimation errors for h_0, arising as sampling, approximation, and model selection errors, will be of order $o_P(n^{-1/4})$. The orthogonality condition together with other conditions will guarantee that these estimation errors do not impact the first-order asymptotic behavior of the estimating equations, so that

$$\sqrt{n}\mathbb{E}_n[\psi(w,\widehat{\theta},\widehat{h}(z))] = \sqrt{n}\mathbb{E}_n[\psi(w,\widehat{\theta},h_0(z))] + o_P(1). \tag{5.4}$$

This leads us to a regular estimation problem, despite $\widehat{h}$ being highly nonregular.

In what follows, we shall denote by c and C some positive constants, and by L_n a sequence of positive constants that may grow to infinity as $n \to \infty$.

CONDITION SP. For each n, we observe the independent data vectors $(w_i)_{i=1}^n$ with law determined by the probability measure $\mathrm{P} = \mathrm{P}_n$. Uniformly, for all n the following conditions hold. (i) The true parameter values θ_0 obeys (5.1) and is interior relative to Θ, namely there is a ball of fixed positive radius centered at θ_0 contained in Θ, where Θ is a fixed compact subset of $\mathbb{R}^d$. (ii) The map $\nu \mapsto \psi(w,\nu)$ is twice continuously differentiable with respect to $\nu = (\nu_k)_{k=1}^K = (\theta,t)$ for all $\nu \in \Theta \times T$, where T is convex, with $\sup_{\nu\in\Theta\times T} |\partial_{\nu_k}\partial_{\nu_r}\psi_j(w_i,\nu)| \le L_n$ a.s., for all $k,r \le K$, $j \le d$, and $i \le n$. The conditional second moments of the first derivatives are bounded as follows: P-a.s. $\mathrm{E}(\sup_{\nu\in\Theta\times T} |\partial_{\nu_k}\psi_j(w_i,\nu)|^2|z_i) \le C$ for each k, j and i. (iii) The orthogonality condition (5.3) holds. (iv) The following identifiability condition holds: for all $\theta \in \Theta$, $\|\bar{\mathbf{E}}[\psi(w,\theta,h_0(z))]\| \ge 2^{-1}(\|J_n(\theta-\theta_0)\| \wedge c)$, where $J_n := \bar{\mathbf{E}}[\partial_\theta \psi(w,\theta_0,h_0(z))]$ has singular values bounded away from zero and above. (v) $\bar{\mathbf{E}}[\|\psi(w,\theta_0,h_0(z))\|^3]$ is bounded from above.

In addition to the previous conditions, Condition SP imposes standard identifiability and certain smoothness on the problem, requiring second derivatives to be

bounded by L_n, which is allowed to grow with n subject to restrictions specified below. It is possible to allow for nondifferentiable ψ at the cost of a more complicated argument; see [11]. In what follows, let $\delta_n \searrow 0$ be a sequence of constants approaching zero from above.

CONDITION AS. The following conditions hold for each n. (i) The function $h_0 = (h_{0m})_{m=1}^M : \mathcal{Z} \mapsto T$ is approximately sparse, namely, for each m, $h_{0m}(\cdot) = \sum_{l=1}^p P_{ml}(\cdot)\beta_{0ml} + r_m(\cdot)$, where $P_{ml} : \mathcal{Z} \mapsto \mathbb{R}$ are approximating functions, $\beta_{0m} = (\beta_{0ml})_{l=1}^p$ obeys $|\operatorname{supp}(\beta_{0m})| \leq s$, $s \geq 1$, and the approximation errors $(r_m)_{m=1}^M : \mathcal{Z} \to \mathbb{R}$ obey $\bar{\mathbf{E}}[r_m^2(z)] \leq Cs\log(p \vee n)/n$. There is an estimator $\widehat{h}_m(\cdot) = \sum_{l=1}^p P_{ml}(\cdot)\widehat{\beta}_{ml}$ of h_{0m} such that, with probability at least $1 - \delta_n$, $\widehat{h} = (\widehat{h}_m)_{m=1}^M$ maps $\mathcal{Z}$ into T, $\widehat{\beta}_m = (\widehat{\beta}_{ml})_{l=1}^p$ satisfies $\|\widehat{\beta}_m - \beta_{0m}\|_1 \leq C\sqrt{s^2\log(p \vee n)/n}$ and $\mathbb{E}_n[(\widehat{h}_m(z) - h_{0m}(z))^2] \leq Cs\log(p \vee n)/n$ for all m. (ii) The scalar random variables $\dot{\psi}_{mjl}(w_i) := \partial_{t_m}\psi_j(w_i, \theta_0, h_0(z))P_{ml}(z_i)$ obey $\max_{m,j,l}\mathbb{E}_n[|\dot{\psi}_{mjl}(w)|^2] \leq L_n^2$ with probability at least $1 - \delta_n$ and $\max_{m,j,l}(\bar{\mathbf{E}}[|\dot{\psi}_{mjl}(w)|^3])^{1/3}/(\bar{\mathbf{E}}[|\dot{\psi}_{mjl}(w)|^2])^{1/2} \leq M_n$. (iii) Finally, the following growth restrictions hold as $n \to \infty$:

$$L_n^2 s^2 \log^2(p \vee n)/n \to 0 \quad \text{and} \quad \log(p \vee n)n^{-1/3}M_n^2 \to 0. \tag{5.5}$$

The assumption records a formal sense in which approximate sparsity is used, as well as requires reasonable behavior of the estimator $\widehat{h}$. In the previous sections, we established primitive conditions under which this behavior occurs in problems where h_0 arise as conditional expectation functions. By virtue of (5.5) the assumption implies that $\{\mathbb{E}_n(\widehat{h}_m(z) - h_{0m}(z))^2\}^{1/2} = o_P(n^{-1/4})$. It is standard that the square of this term multiplied by $\sqrt{n}$ shows up as a linearization error for $\sqrt{n}(\widehat{\theta} - \theta_0)$ and, therefore, this term does not affect its first-order behavior. Moreover, the assumption implies by virtue of (5.5) that $\|\widehat{\beta}_m - \beta_{0m}\|_1 = o_P(L_n^{-1}(\log(p \vee n))^{-1})$, which is used to control another key term in the linearization as follows:

$$\sqrt{n}\max_{j,m,l}|\mathbb{E}_n[\dot{\psi}_{mjl}(w)]|\|\widehat{\beta}_m - \beta_{0m}\|_1 \lesssim_P L_n\sqrt{\log(p \vee n)}\|\widehat{\beta}_m - \beta_{0m}\|_1 = o_P(1),$$

where the bound follows from an application of the moderate deviation inequalities for self-normalized sums (Lemma 4). The idea for this type of control is borrowed from [3], who used it in the IV model above.

THEOREM 8. *Under Conditions* SP *and* AS, *the estimator* $\widehat{\theta}$ *that obeys equation* (5.2) *and* $\widehat{\theta} \in \Theta$ *with probability approaching* 1, *satisfies* $\sqrt{n}(\widehat{\theta} - \theta_0) = -J_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi(w_i, \theta_0, h_0(z_i)) + o_P(1)$. *Furthermore, provided* $\Omega_n = \bar{\mathbf{E}}[\psi(w, \theta_0, h_0(z))\psi(w, \theta_0, h_0(z))']$ *has eigenvalues bounded away from zero,*

$$\Omega_n^{-1/2}J_n\sqrt{n}(\widehat{\theta} - \theta_0) \Rightarrow N(0, I).$$

This theorem extends the analogous result in [3, 8] for a specific linear problem to a generic nonlinear setting, and could be of independent interest in many problems cited above. The theorem allows for the probability measure $\mathrm{P} = \mathrm{P}_n$ to change with n, which implies that the confidence bands based on the result have certain uniform validity ("honesty") with respect to P, as formalized in [10], thereby constructively addressing [37]'s critique. See [11] for a generalization of the result above to the case $\dim(\theta_0) \gg n$.

EXAMPLE PL. It is instructive to conclude this section by inspecting the example of approximately sparse partially linear regression [9, 10], which also nests the sparse linear regression model [55]. The partially linear model of [41] is

$$y_i = d_i\theta_0 + g(z_i) + \varepsilon_i, \qquad \mathrm{E}[\varepsilon_i|z_i, d_i] = 0,$$
$$d_i = m(z_i) + v_i, \qquad \mathrm{E}[v_i|z_i] = 0.$$

The target is the real parameter θ_0, and an orthogonal score function ψ for this parameter is $\psi(w_i, \theta, t) = (y_i - \theta(d_i - t_2) - t_1)(d_i - t_2)$, where $t = (t_1, t_2)'$ and $w_i = (y_i, d_i, z_i')'$. Let $\ell(z_i) := \theta_0 m(z_i) + g(z_i)$, and $h_0(z_i) := (\ell(z_i), m(z_i))' = (\mathrm{E}[y_i|z_i], \mathrm{E}[d_i|z_i])'$. Note that

$$\mathrm{E}\big[\psi\big(w_i, \theta_0, h_0(z_i)\big)|z_i\big] = 0 \quad \text{and} \quad \mathrm{E}\big[\partial_t\psi\big(w_i, \theta, h_0(z_i)\big)|z_i\big] = 0,$$

so the orthogonality condition holds. If the regression functions $\ell(z_i)$ and $m(z_i)$ are approximately sparse with respect to $x_i = P(z_i)$, we can estimate them by $\sqrt{\text{Lasso}}$ or ols post $\sqrt{\text{Lasso}}$ regression of y_i on x_i and d_i on x_i, respectively. The resulting estimator $\widehat{\theta}$ of θ_0, defined as a solution to (5.2), is a $\sqrt{\text{Lasso}}$ analog of Robinson's [41] estimator. If assumptions of Theorem 8 hold, $\widehat{\theta}$ obeys

$$\Omega_n^{-1/2} J_n \sqrt{n}(\widehat{\theta} - \theta_0) \Rightarrow N(0, 1)$$

for $J_n = \bar{\mathrm{E}}[v^2]$ and $\Omega_n = \bar{\mathrm{E}}[\varepsilon^2 v^2]$. In a homoscedastic model, $\widehat{\theta}$ is semiparametrically efficient, since its asymptotic variance Ω_n/J_n^2 reduces to the efficiency bound $\mathrm{E}[\varepsilon^2]/\mathrm{E}[v^2]$ of Robinson [41]; as pointed out in [9, 10]. In the linear regression model, this estimator is first-order equivalent to, but different in finite samples from, a one-step correction from the scaled Lasso proposed in [55]; in the partially linear model, it is equivalent to the post-double selection method of [9, 10].

APPENDIX A: PROOFS OF SECTION 3

PROOF OF LEMMA 1. The first result holds by the inequalities given in the main text.

To show the next statement, note that T does not change by including repeated regressors (indeed, since T is selected by the oracle (2.2), T will not contain repeated regressors). Let R denote the set of repeated regressors and $\tilde{x}_i = (x_i', z_i')'$ where $x_i \in \mathbb{R}^p$ is the vector of original regressors and $z_i \in \mathbb{R}^{|R|}$ the vector of repeated regressors. We denote by $\widetilde{\Gamma}$ and $\|\cdot\|_{2,\tilde{n}}$ the penalty loadings and the prediction norm associated with $(\tilde{x}_i)_{i=1}^n$. Let $\tilde{\delta} = (\delta^{1'}, \delta^{2'})'$, where $\delta^1 \in \mathbb{R}^p$ and $\delta^2 \in \mathbb{R}^{|R|}$, define $\bar{\delta}^2 \in \mathbb{R}^p$ so that $\bar{\delta}^2_j = \delta^2_j$ if $j \in R$, and $\bar{\delta}^2_j = 0$ if $j \notin R$, and denote $\delta = \delta^1 + \bar{\delta}^2$. It follows that

$$\underline{\kappa} \geq \frac{\|\tilde{\delta}\|_{2,\tilde{n}}}{\|\widetilde{\Gamma}\tilde{\delta}_T\|_1 - \|\widetilde{\Gamma}\tilde{\delta}_{T^c}\|_1} = \frac{\|\delta\|_{2,n}}{\|\Gamma\delta^1_T\|_1 - \|\Gamma\delta^1_{T^c}\|_1 - \|\Gamma\bar{\delta}^2_T\|_1 - \|\Gamma\bar{\delta}^2_{T^c}\|_1},$$

which is minimized in the case that $\delta^1 = \delta$ and $\bar{\delta}^2 = 0$. Thus, the worst case for $\bar{\kappa}$ correspond to $\bar{\delta}^2 = 0$ which corresponds to ignoring the repeated regressors. □

PROOF OF LEMMA 2. The first part is shown in the main text. The second part is proven in supplementary material [2]. □

PROOF OF LEMMA 3. By definition of $\widehat{\beta}$, $\sqrt{\widehat{Q}(\widehat{\beta})} - \sqrt{\widehat{Q}(\beta_0)} \leq \frac{\lambda}{n}\|\Gamma\beta_0\|_1 - \frac{\lambda}{n}\|\Gamma\widehat{\beta}\|_1$. By convexity of $\sqrt{\widehat{Q}}$, by $-\widetilde{S} \in \partial\sqrt{\widehat{Q}}(\beta_0)$, and by $\lambda/n \geq cn\|\Gamma^{-1}\widetilde{S}\|_\infty$, we have $\sqrt{\widehat{Q}(\widehat{\beta})} - \sqrt{\widehat{Q}(\beta_0)} \geq -\widetilde{S}'\widehat{\delta} \geq -\|\Gamma^{-1}\widetilde{S}\|_\infty\|\Gamma\widehat{\delta}\|_1 \geq -\frac{\lambda}{cn}\|\Gamma\widehat{\delta}\|_1$ where $\widehat{\delta} = \widehat{\beta} - \beta_0$. Combining the lower and upper bounds yields $\|\Gamma\widehat{\delta}\|_1 \geq c(\|\Gamma(\beta_0 + \widehat{\delta})\|_1 - \|\Gamma\beta_0\|_1)$. Thus, $\widehat{\delta} \in R_c$; that $\widehat{\delta} \in \Delta_{\bar{c}}$ follows by a standard argument based on elementary inequalities. □

PROOF OF THEOREM 1. First, note that by Lemma 3 we have $\widehat{\delta} := \widehat{\beta} - \beta_0 \in R_c$. By optimality of $\widehat{\beta}$ and definition of $\bar{\kappa}$, $\bar{\zeta} = \lambda\sqrt{s}/[n\bar{\kappa}]$, we have

$$\sqrt{\widehat{Q}(\widehat{\beta})} - \sqrt{\widehat{Q}(\beta_0)} \leq \frac{\lambda}{n}\|\Gamma\beta_0\|_1 - \frac{\lambda}{n}\|\Gamma\widehat{\beta}\|_1 \leq \bar{\zeta}\|\widehat{\delta}\|_{2,n}. \tag{A.1}$$

Multiplying both sides by $\sqrt{\widehat{Q}(\widehat{\beta})} + \sqrt{\widehat{Q}(\beta_0)}$ and since $(a+b)(a-b) = a^2 - b^2$

$$\|\widehat{\delta}\|^2_{2,n} \leq 2\mathbb{E}_n[(\sigma\varepsilon + r)x'\widehat{\delta}] + \big(\sqrt{\widehat{Q}(\widehat{\beta})} + \sqrt{\widehat{Q}(\beta_0)}\big)\bar{\zeta}\|\widehat{\delta}\|_{2,n}. \tag{A.2}$$

From (A.1), we have $\sqrt{\widehat{Q}(\widehat{\beta})} \leq \sqrt{\widehat{Q}(\beta_0)} + \bar{\zeta}\|\widehat{\delta}\|_{2,n}$ so that

$$\|\widehat{\delta}\|^2_{2,n} \leq 2\mathbb{E}_n[(\sigma\varepsilon + r)x'\widehat{\delta}] + 2\sqrt{\widehat{Q}(\beta_0)}\bar{\zeta}\|\widehat{\delta}\|_{2,n} + \bar{\zeta}^2\|\widehat{\delta}\|^2_{2,n}.$$

Since $|\mathbb{E}_n[(\sigma\varepsilon + r)x'\widehat{\delta}]| = \sqrt{\widehat{Q}(\beta_0)}|\widetilde{S}'\widehat{\delta}| \leq \sqrt{\widehat{Q}(\beta_0)}\varrho_c\|\widehat{\delta}\|_{2,n}$, we obtain

$$\|\widehat{\delta}\|^2_{2,n} \leq 2\sqrt{\widehat{Q}(\beta_0)}\varrho_c\|\widehat{\delta}\|_{2,n} + 2\sqrt{\widehat{Q}(\beta_0)}\bar{\zeta}\|\widehat{\delta}\|_{2,n} + \bar{\zeta}^2\|\widehat{\delta}\|^2_{2,n},$$

and the result follows provided $\bar{\zeta} < 1$. $\square$

PROOF OF THEOREM 2. We have $\widehat{\delta} := \widehat{\beta} - \beta_0 \in R_c$ under the condition that $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$ by Lemma 3. We also have $\bar{\zeta} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$ by assumption.

First, we establish the upper bound. By the previous proof, we have inequality (A.1). The bound follows from Theorem 1 to bound $\|\widehat{\delta}\|_{2,n}$. To establish the lower bound, by convexity of $\sqrt{\widehat{Q}}$ and the definition of ϱ_c we have $\sqrt{\widehat{Q}(\widehat{\beta})} - \sqrt{\widehat{Q}(\beta_0)} \geq -\widetilde{S}'\widehat{\delta} \geq -\varrho_c\|\widehat{\delta}\|_{2,n}$. Thus, by Theorem 1 we obtain $\sqrt{\widehat{Q}(\widehat{\beta})} - \sqrt{\widehat{Q}(\beta_0)} \geq -2\sqrt{\widehat{Q}(\beta_0)}\varrho_c B_n$.

Moreover, by the triangle inequality

$$|\sqrt{\widehat{Q}(\widehat{\beta})} - \sigma| \leq |\sqrt{\widehat{Q}(\widehat{\beta})} - \sigma\{\mathbb{E}_n[\varepsilon^2]\}^{1/2}| + \sigma|\{\mathbb{E}_n[\varepsilon^2]\}^{1/2} - 1|$$

and the right-hand side is bounded by $\|\widehat{\beta} - \beta_0\|_{2,n} + c_s + \sigma|\mathbb{E}_n[\varepsilon^2] - 1|$. $\square$

PROOF OF THEOREM 3. For notational convenience, we denote $\phi_n(m) = \phi_{\max}(m, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1})$. We shall rely on the following lemma, whose proof is given after the proof of this theorem.

LEMMA 5 (Relating sparsity and prediction norm). *Under Condition* ASM, *let* $G \subseteq \operatorname{supp}(\widehat{\beta})$. *For any* $\lambda > 0$, *we have*

$$\frac{\lambda}{n}\sqrt{\widehat{Q}(\widehat{\beta})}\sqrt{|G|} \leq \sqrt{|G|}\|\Gamma^{-1}\widetilde{S}\|_\infty\sqrt{\widehat{Q}(\beta_0)} + \sqrt{\phi_{\max}(|G \setminus T|, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1})}\|\widehat{\beta} - \beta_0\|_{2,n}.$$

Define $\widehat{m} := |\operatorname{supp}(\widehat{\beta}) \setminus T|$. In the event $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$, by Lemma 5

$$\left(\sqrt{\frac{\widehat{Q}(\widehat{\beta})}{\widehat{Q}(\beta_0)}} - \frac{1}{c}\right)\frac{\lambda}{n}\sqrt{\widehat{Q}(\beta_0)}\sqrt{|\operatorname{supp}(\widehat{\beta})|} \leq \sqrt{\phi_n(\widehat{m})}\|\widehat{\beta} - \beta_0\|_{2,n}. \tag{A.3}$$

Under the condition $\bar{\zeta} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$, we have by Theorems 1 and 2 that

$$\left(1 - 2\varrho_c B_n - \frac{1}{c}\right)\frac{\lambda}{n}\sqrt{\widehat{Q}(\beta_0)}\sqrt{|\operatorname{supp}(\widehat{\beta})|} \leq \sqrt{\phi_n(\widehat{m})}2\sqrt{\widehat{Q}(\beta_0)}B_n,$$

where $B_n = \frac{\varrho_c + \bar{\zeta}}{1 - \bar{\zeta}^2}$. Since we assume $2\varrho_c B_n \leq 1/(c\bar{c})$, we have

$$\sqrt{|\operatorname{supp}(\widehat{\beta})|} \leq 2\bar{c}\sqrt{\phi_n(\widehat{m})}\frac{n}{\lambda}B_n = \sqrt{s}\sqrt{\phi_n(\widehat{m})}2\bar{c}B_n/(\bar{\zeta}\bar{\kappa}),$$

where the last equality follows from $\bar{\zeta} = \lambda\sqrt{s}/[n\bar{\kappa}]$.

Let $L := 2\bar{c}B_n/(\bar{\zeta}\bar{\kappa})$. Consider any $m \in \mathcal{M}$, and suppose $\widehat{m} > m$. Therefore, by the sublinearity of maximum sparse eigenvalues (see Lemma 3 in [6]), $\phi_n(\ell m) \leq$

$\lceil \ell \rceil \phi_n(m)$ for $\ell \geq 1$, and $\widehat{m} \leq |\operatorname{supp}(\widehat{\beta})|$ we have $\widehat{m} \leq s \cdot \lceil \frac{\widehat{m}}{m} \rceil \phi_n(m) L^2$. Thus, since $\lceil k \rceil < 2k$ for any $k \geq 1$ we have $m < s \cdot 2\phi_n(m)L^2$ which violates the condition of $m \in \mathcal{M}$ and s. Therefore, we must have $\widehat{m} \leq m$. Repeating the argument once more with $\widehat{m} \leq m$ we obtain $\widehat{m} \leq s \cdot \phi_n(m)L^2$. The result follows by minimizing the bound over $m \in \mathcal{M}$.

To show the second part, by Lemma 2 and $\lambda/n \geq c\|\Gamma^{-1}\widehat{S}\|_\infty$, we have $\varrho_c \leq \frac{\lambda\sqrt{s}}{n\kappa_{\bar{c}}} \frac{1+\bar{c}}{c}$. Lemma 1 yields $\bar{\kappa} \geq \kappa_{\bar{c}}$ and recall $\bar{\zeta} = \lambda\sqrt{s}/(n\bar{\kappa})$. Therefore,

$$
\begin{aligned}
B_n/(\bar{\zeta}\bar{\kappa}) &\leq \frac{1 + \{(\lambda\sqrt{s}/(n\kappa_{\bar{c}}))((1+\bar{c})/c)\}\{n\bar{\kappa}/(\lambda\sqrt{s})\}}{\bar{\kappa}(1-\bar{\zeta}^2)} \\
&\leq \frac{1 + (1+\bar{c})/c}{\kappa_{\bar{c}}(1-\bar{\zeta}^2)} = \frac{\bar{c}}{\kappa_{\bar{c}}(1-\bar{\zeta}^2)} \leq \frac{2\bar{c}}{\kappa_{\bar{c}}},
\end{aligned}
$$

where the last inequality follows from the condition $\bar{\zeta} \leq 1/\sqrt{2}$. Thus, it follows that $4\bar{c}^2(B_n/(\bar{\zeta}\bar{\kappa}))^2 \leq (4\bar{c}^2/\kappa_{\bar{c}})^2$ which implies $\mathcal{M}^* \subseteq \mathcal{M}$. □

PROOF OF LEMMA 5. Recall that $\Gamma = \operatorname{diag}(\gamma_1, \ldots, \gamma_p)$. $\widehat{\beta}$ is the solution of a conic optimization problem (see Section H.1 of supplementary material [2]). Let $\widehat{a}$ denote the solution to its dual problem: $\max_{a \in \mathbb{R}^n} \mathbb{E}_n[ya] : \|\Gamma^{-1}\mathbb{E}_n[x_j a]\|_\infty \leq \lambda/n$, $\|a\| \leq \sqrt{n}$. By strong duality $\mathbb{E}_n[y\widehat{a}] = \frac{\|Y - X\widehat{\beta}\|}{\sqrt{n}} + \frac{\lambda}{n}\sum_{j=1}^p \gamma_j |\widehat{\beta}_j|$. Moreover, by the first-order optimality conditions, $\mathbb{E}_n[x_j\widehat{a}]\widehat{\beta}_j = \lambda\gamma_j|\widehat{\beta}_j|/n$ holds for every $j = 1, \ldots, p$. Thus, we have

$$
\mathbb{E}_n[y\widehat{a}] = \frac{\|Y - X\widehat{\beta}\|}{\sqrt{n}} + \sum_{j=1}^p \mathbb{E}_n[x_j\widehat{a}]\widehat{\beta}_j = \frac{\|Y - X\widehat{\beta}\|}{\sqrt{n}} + \mathbb{E}_n\left[\widehat{a}\sum_{j=1}^p x_j\widehat{\beta}_j\right].
$$

Rearranging the terms, we have $\mathbb{E}_n[(y - x'\widehat{\beta})\widehat{a}] = \|Y - X\widehat{\beta}\|/\sqrt{n}$.

If $\|Y - X\widehat{\beta}\| = 0$, we have $\sqrt{\widehat{Q}(\widehat{\beta})} = 0$ and the statement of the lemma trivially holds. If $\|Y - X\widehat{\beta}\| > 0$, since $\|\widehat{a}\| \leq \sqrt{n}$ the equality can only hold for $\widehat{a} = \sqrt{n}(Y - X\widehat{\beta})/\|Y - X\widehat{\beta}\| = (Y - X\widehat{\beta})/\sqrt{\widehat{Q}(\widehat{\beta})}$.

Next, note that for any $j \in \operatorname{supp}(\widehat{\beta})$ we have $\mathbb{E}_n[x_j\widehat{a}] = \operatorname{sign}(\widehat{\beta}_j)\lambda\gamma_j/n$. Therefore, for any subset $G \subseteq \operatorname{supp}(\widehat{\beta})$ we have

$$
\begin{aligned}
&\sqrt{\widehat{Q}(\widehat{\beta})}\sqrt{|G|}\lambda \\
&\quad = \|\Gamma^{-1}(X'(Y - X\widehat{\beta}))_G\| \\
&\quad \leq \|\Gamma^{-1}(X'(Y - X\beta_0))_G\| + \|\Gamma^{-1}(X'X(\beta_0 - \widehat{\beta}))_G\| \\
&\quad \leq \sqrt{|G|}n\|\Gamma^{-1}\mathbb{E}_n[x(\sigma\varepsilon + r)]\|_\infty \\
&\qquad + n\sqrt{\phi_{\max}(|G \setminus T|, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1})}\|\widehat{\beta} - \beta_0\|_{2,n}
\end{aligned}
$$

$$= \sqrt{|G|}n\sqrt{\widehat{Q}(\beta_0)}\|\Gamma^{-1}\widetilde{S}\|_\infty + n\sqrt{\phi_{\max}(|G\setminus T|, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1})}\|\widehat{\beta}-\beta_0\|_{2,n},$$

where we used

$$\|\Gamma^{-1}(X'X(\widehat{\beta}-\beta_0))_G\| \le \sup_{\|\alpha_{T^c}\|_0\le|G\setminus T|,\|\alpha\|\le1}|\alpha'\Gamma^{-1}X'X(\widehat{\beta}-\beta_0)| \le \sup_{\|\alpha_{T^c}\|_0\le|G\setminus T|,\|\alpha\|\le1}\|\alpha'\Gamma^{-1}X'\|\|X(\widehat{\beta}-\beta_0)\| \le n\sqrt{\phi_{\max}(|G\setminus T|, \Gamma^{-1}\mathbb{E}_n[xx']\Gamma^{-1})}\|\widehat{\beta}-\beta_0\|_{2,n}. \quad \square$$

PROOF OF THEOREM 4. In this proof, let $f=(f_1,\dots,f_n)'$, $R=(r_1,\dots,r_n)'$, $\varepsilon=(\varepsilon_1,\dots,\varepsilon_n)'$ (n-vectors) and $X=[x_1;\dots;x_n]'$ (an $n\times p$ matrix). For a set of indices $S\subset\{1,\dots,p\}$, define $\mathcal{P}_S = X[S](X[S]'X[S])^-X[S]'$, where we interpret $\mathcal{P}_S$ as a null operator if S is empty. We have that $f - X\widetilde{\beta} = (I-\mathcal{P}_{\widehat{T}})f - \sigma\mathcal{P}_{\widehat{T}}\varepsilon$, where I is the identity operator. Therefore,

$$\begin{aligned}\sqrt{n}\|\beta_0-\widetilde{\beta}\|_{2,n} &= \|X\beta_0 - X\widetilde{\beta}\| \\ &= |f - X\widetilde{\beta} - R\| \\ &= \|(I-\mathcal{P}_{\widehat{T}})f - \sigma\mathcal{P}_{\widehat{T}}\varepsilon - R\| \\ &\le \|(I-\mathcal{P}_{\widehat{T}})f\| + \sigma\|\mathcal{P}_{\widehat{T}}\varepsilon\| + \|R\|\end{aligned} \tag{A.4}$$

where $\|R\|\le\sqrt{n}c_s$. Since for $\widehat{m}=|\widehat{T}\setminus T|$, we have

$$\|X[\widehat{T}](X[\widehat{T}]'X[\widehat{T}])^-\|_{\mathrm{op}} \le \sqrt{1/\phi_{\min}(\widehat{m},\mathbb{E}_n[xx'])} = \sqrt{1/\phi_{\min}(\widehat{m})},$$

[where the bound is interpreted as $+\infty$ if $\phi_{\min}(\widehat{m})=0$], the term $\|\mathcal{P}_{\widehat{T}}\varepsilon\|$ in (A.4) satisfies

$$\|\mathcal{P}_{\widehat{T}}\varepsilon\| \le \sqrt{1/\phi_{\min}(\widehat{m})}\|X[\widehat{T}]'\varepsilon/\sqrt{n}\| \le \sqrt{|\widehat{T}|/\phi_{\min}(\widehat{m})}\|X'\varepsilon/\sqrt{n}\|_\infty.$$

Therefore, we have $\|\widetilde{\beta}-\beta_0\|_{2,n} \le \frac{\sigma\sqrt{s+\widehat{m}}\|\mathbb{E}_n[x\varepsilon]\|_\infty}{\sqrt{\phi_{\min}(\widehat{m})}} + c_s + c_{\widehat{T}}$, where $c_{\widehat{T}} = \min_{\beta\in\mathbb{R}^p}\sqrt{\mathbb{E}_n[(f-x'\beta_{\widehat{T}})^2]}$. Since $\operatorname{supp}(\widehat{\beta})\subseteq\widehat{T}$ and (3.4) holds,

$$c_{\widehat{T}} = \min_{\beta\in\mathbb{R}^p}\{\mathbb{E}_n[(f-x'\beta_{\widehat{T}})^2]\}^{1/2} \le \{\mathbb{E}_n[(f-x'\widehat{\beta})^2]\}^{1/2} \le c_s + \|\beta_0-\widehat{\beta}\|_{2,n} \le c_s + 2\sqrt{\widehat{Q}(\beta_0)}B_n,$$

where we have used Theorem 1. $\square$

PROOF OF THEOREM 5. Note that because $\sigma=0$ and $c_s=0$, we have $\sqrt{\widehat{Q}(\beta_0)}=0$ and $\sqrt{\widehat{Q}(\widehat{\beta})}=\|\widehat{\beta}-\beta_0\|_{2,n}$. Thus, by optimality of $\widehat{\beta}$ we have $\|\widehat{\beta}-\beta_0\|_{2,n}+\frac{\lambda}{n}\|\Gamma\widehat{\beta}\|_1 \le \frac{\lambda}{n}\|\Gamma\beta_0\|_1$ which implies $\|\Gamma\widehat{\beta}\|_1\le\|\Gamma\beta_0\|_1$. Moreover,

$\delta = \widehat{\beta} - \beta_0$ satisfies $\|\delta\|_{2,n} \leq \frac{\lambda}{n}(\|\Gamma\beta_0\|_1 - \|\Gamma\widehat{\beta}\|_1) \leq \bar{\zeta}\|\delta\|_{2,n}$, where $\bar{\zeta} = \frac{\lambda\sqrt{s}}{n\bar{\kappa}} < 1$. Hence, $\|\delta\|_{2,n} = 0$.

Since $\|\Gamma\widehat{\beta}\|_1 \leq \|\Gamma\beta_0\|_1$ implies $\delta \in \Delta_1$, it follows that $0 = \sqrt{s}\|\delta\|_{2,n} \geq \|\Gamma\delta_T\|_1/\kappa_1 \geq \frac{1}{2}\|\Gamma\delta\|_1/\kappa_1$, which implies that $\delta = 0$ if $\kappa_1 > 0$. □

PROOF OF THEOREM 6. If $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$ and $\bar{\zeta} = \lambda\sqrt{s}/[n\bar{\kappa}] < 1$, by Theorem 1 we have $\|\widehat{\beta} - \beta_0\|_{2,n} \leq 2\sqrt{\widehat{Q}(\beta_0)}B_n$, and the bound on the prediction norm follows by $\sqrt{\widehat{Q}(\beta_0)} \leq c_s + \sigma\sqrt{\mathbb{E}_n[\varepsilon^2]}$.

Thus, we need to show that the choice of λ and Γ ensures the event $\lambda/n \geq c\|\Gamma^{-1}\widetilde{S}\|_\infty$ with probability no less than $1 - \alpha - \eta_1 - \eta_2$. Since $\gamma_j = \max_{1\leq i\leq n}|x_{ij}| \geq \mathbb{E}_n[x_j^2] = 1$, by the choice of u_n we have

$$P\left(c\|\Gamma^{-1}\widetilde{S}\|_\infty > \frac{\lambda}{n}\right) \leq P\left(\max_{1\leq j\leq p}\frac{c|\mathbb{E}_n[(\sigma\varepsilon + r)x_j]|}{\gamma_j\sqrt{\mathbb{E}_n[(\sigma\varepsilon)^2]}} > \frac{\lambda}{n(1+u_n)^{1/2}}\right) + \eta_1$$
$$\leq I + II + \eta_1,$$
$$I := P\left(\max_{1\leq j\leq p}\frac{|\mathbb{E}_n[\varepsilon x_j]|}{\gamma_j\sqrt{\mathbb{E}_n[\varepsilon^2]}} > \frac{\sqrt{2\log(2p/\alpha)}}{\sqrt{n}}\right),$$
$$II := P\left(\frac{\|\mathbb{E}_n[rx]\|_\infty}{\sqrt{\mathbb{E}_n[(\sigma\varepsilon)^2]}} > \frac{(1+u_n)^{1/2}}{\sqrt{n}}\right).$$

We invoke the following lemma, which is proven in [6]—see step 2 of the proof of [6]'s Theorem 2; for completeness, supplementary material [2] also provides the proof.

LEMMA 6. *Under Condition ASM, we have* $\|\mathbb{E}_n[xr]\|_\infty \leq \min\{\frac{\sigma}{\sqrt{n}}, c_s\}$.

By Lemma 6, $\|\mathbb{E}_n[rx]\|_\infty \leq \sigma/\sqrt{n}$ and $P(\mathbb{E}_n[\varepsilon^2] \leq \{1+u_n\}^{-1}) \leq \eta_2$, we have $II \leq P(\sqrt{\mathbb{E}_n[(\varepsilon)^2]} \leq \{1+u_n\}^{-1/2}) \leq \eta_2$. Also,

$$I \leq p\max_{1\leq j\leq p}P\left(\frac{\sqrt{n}|\mathbb{E}_n[\varepsilon x_j]|}{\sqrt{\mathbb{E}_n[x_j^2\varepsilon^2]}} > \sqrt{2\log(2p/\alpha)}\right) \leq \alpha$$

where we used that $\gamma_j\sqrt{\mathbb{E}_n[\varepsilon^2]} \geq \sqrt{\mathbb{E}_n[x_j^2\varepsilon^2]}$, the union bound, and the subGaussian inequality for self-normalized sums stated in Theorem 2.15 of [25], since ε_i's are independent and symmetric by assumption. □

PROOF OF COROLLARY 1. See supplementary material [2]. □

APPENDIX B: PROOFS OF SECTION 4

PROOF OF THEOREM 7. The proof is given in supplementary material [2] and follows from Theorems 1–3 with the help of Lemma 7 in supplementary material [2]. □

PROOF OF COROLLARY 2. See supplementary material [2]. □

APPENDIX C: PROOFS FOR SECTION 5

PROOF OF THEOREM 8. Throughout the proof, we use the notation

$$B(w) := \max_{j,k} \sup_{\nu \in \Theta \times T} \left|\partial_{\nu_k} \psi_j(w, \nu)\right|, \qquad \tau_n := \sqrt{s \log(p \vee n)/n}.$$

Step 1. (A preliminary rate result.) In this step, we claim that $\|\widehat{\theta} - \theta_0\| \lesssim_P \tau_n$. By definition, $\|\mathbb{E}_n[\psi(w, \widehat{\theta}, \widehat{h}(z))]\| \leq \varepsilon_n$ and $\widehat{\theta} \in \Theta$ with probability $1 - o(1)$, which implies via triangle inequality that with the same probability:

$$\left\|\bar{\mathbf{E}}\big[\psi\big(w, \theta, h_0(z)\big)\big]|_{\theta=\widehat{\theta}}\right\| \leq \varepsilon_n + I_1 + I_2 \lesssim_P \tau_n,$$

where I_1 and I_2 are defined in step 2 below, and the last bound also follows from step 2 below and from the numerical tolerance obeying $\varepsilon_n = o(n^{-1/2})$ by assumption. Since by Condition SP(iv), $2^{-1}(\|J_n(\widehat{\theta} - \theta_0)\| \wedge c)$ is weakly smaller than the left-hand side of the display, we conclude that $\|\widehat{\theta} - \theta_0\| \lesssim_P \tau_n$, using that singular values of J_n are bounded away from zero uniformly in n by Condition SP(v).

Step 2. (Define and bound I_1 and I_2.) We claim that:

$$I_1 := \sup_{\theta \in \Theta} \left\|\mathbb{E}_n \psi\big(w, \theta, \widehat{h}(z)\big) - \mathbb{E}_n \psi\big(w, \theta, h_0(z)\big)\right\| \lesssim_P \tau_n,$$

$$I_2 := \sup_{\theta \in \Theta} \left\|\mathbb{E}_n \psi\big(w, \theta, h_0(z)\big) - \bar{\mathbf{E}} \psi\big(w, \theta, h_0(z)\big)\right\| \lesssim_P n^{-1/2}.$$

Using Taylor's expansion, for $\tilde{h}(z; \theta, j)$ denoting a point on a line connecting vectors $h_0(z)$ and $h(z)$, which can depend on θ and j,

$$\begin{aligned} I_1 &\leq \sum_{j=1}^{d} \sum_{m=1}^{M} \sup_{\theta \in \Theta} \left|\mathbb{E}_n\big[\partial_{t_m} \psi_j\big(w, \theta, \tilde{h}(z; \theta, j)\big)\big(\widehat{h}_m(z) - h_{0m}(z)\big)\big]\right| \\ &\leq dM\big\{\mathbb{E}_n B^2(w)\big\}^{1/2} \max_m \big\{\mathbb{E}_n\big(\widehat{h}_m(z) - h_{0m}(z)\big)^2\big\}^{1/2}, \end{aligned}$$

where the last inequality holds by definition of $B(w)$ given earlier and Hölder's inequality. Since $\bar{\mathbf{E}} B^2(w) \leq C$ by Condition SP(ii), $\mathbb{E}_n B^2(w) \lesssim_P 1$ by Markov's inequality. By this, by Condition AS(i), by d and M fixed, conclude that $I_1 \lesssim_P \tau_n$.

Using Jain–Marcus' theorem, as stated in Example 2.11.13 in [51], we conclude that $\sqrt{n}I_2 \lesssim_P 1$. Indeed the hypotheses of that example follow from the assumption that Θ is a fixed compact subset of $\mathbb{R}^d$, and from the Lipschitz property, $\|\psi(w,\theta,h_0(z)) - \psi(w,\tilde{\theta},h_0(z))\| \le \sqrt{d}B(w)\|\tilde{\theta}-\theta\|$ holding uniformly for all θ and $\tilde{\theta}$ in Θ, with $\bar{\mathbf{E}}B^2(w) \le C$.

Step 3. (Main step.) We have that $\sqrt{n}\|\mathbb{E}_n\psi(w,\widehat{\theta},\widehat{h}(z))\| \le \varepsilon_n\sqrt{n}$. Application of Taylor's theorem and the triangle inequality gives

$$\|\sqrt{n}\mathbb{E}_n\psi(w,\theta_0,h_0(z)) + J_n\sqrt{n}(\widehat{\theta}-\theta_0)\| \le \varepsilon\sqrt{n} + \|II_1\| + \|II_2\| + \|II_3\| = o_P(1),$$

where $J_n = \bar{\mathbf{E}}\partial_\theta\psi(w,\theta_0,h_0(z))$, the terms II_1, II_2 and II_3 are defined and bounded below in step 4; the $o_P(1)$ bound follows from step 4 and from $\varepsilon_n\sqrt{n} = o(1)$ holding by assumption. Conclude using Condition SP(iv) that

$$\|J_n^{-1}\sqrt{n}\mathbb{E}_n\psi(w,\theta_0,h_0(z)) + \sqrt{n}(\widehat{\theta}-\theta_0)\| = o_P(1),$$

which verifies the first claim of the theorem. Application of Liapunov's central limit theorem in conjunction with Condition SP(v) and the conditions on Ω_n imposed by the theorem imply the second claim.

Step 4. (Define and bound II_1, II_2 and II_3.) Let $II_1 := (II_{1j})_{j=1}^d$ and $II_2 = (II_{2j})_{j=1}^d$, where

$$II_{1j} := \sum_{m=1}^M \sqrt{n}\mathbb{E}_n\big[\partial_{t_m}\psi_j(w,\theta_0,h_0(z))\big(\widehat{h}_m(z) - h_{0m}(z)\big)\big],$$

$$II_{2j} := \sum_{r,k=1}^K \sqrt{n}\mathbb{E}_n\big[\partial_{v_k}\partial_{v_r}\psi_j(w,\tilde{\nu}(w;j))\big\{\widehat{\nu}_r(w) - \nu_{0r}(w)\big\}\big\{\widehat{\nu}_k(w) - \nu_{0k}(w)\big\}\big],$$

$$II_3 := \sqrt{n}\big(\mathbb{E}_n\partial_\theta\psi(w,\theta_0,h_0(z)) - J_n\big)(\widehat{\theta}-\theta_0),$$

where $\nu_0(w) := (\nu_{0k}(w))_{k=1}^K := (\theta_0', h_0(z)')'$; $K = d+M$; $\widehat{\nu}(w) := (\widehat{\nu}_k(w))_{k=1}^K := (\widehat{\theta}', \widehat{h}(z)')'$, and $\tilde{\nu}(w;j)$ is a vector on the line connecting $\nu_0(w)$ and $\widehat{\nu}(w)$ that may depend on j. We show in this step that $\|II_1\| + \|II_2\| + \|II_3\| \lesssim_P o(1)$.

The key portion of the proof is bounding II_1, which is very similar to the argument given in [3] (pages 2421–2423). We repeat it here for completeness. We split $II_1 = III_1 + III_2 = (III_{1j})_{j=1}^d + (III_{2j})_{j=1}^d$, where

$$III_{1j} := \sum_{m=1}^M \sqrt{n}\mathbb{E}_n\left[\partial_{t_m}\psi_j(w,\theta_0,h_0(z))\sum_{l=1}^p P_{ml}(z)(\widehat{\beta}_{ml} - \beta_{0ml})\right],$$

$$III_{2j} := \sum_{m=1}^M \sqrt{n}\mathbb{E}_n\big[\partial_{t_m}\psi_j(w,\theta_0,h_0(z))r_m(z)\big].$$

Using Hölder inequality, $\max_j |III_{1j}| \le M\max_{j,m,l}|\sqrt{n}\mathbb{E}_n\dot{\psi}_{mjl}(w)|\|\widehat{\beta}_m - \beta_{0m}\|_1$. By Condition AS(i) $\max_m \|\widehat{\beta}_m - \beta_{0m}\|_1 \le C\sqrt{s}\tau_n$ with probability at least $1-\delta_n$.

Moreover, using $\mathrm{E}\dot{\psi}_{mjl}(w_i) = 0$ for all i, which holds by the orthogonality property (5.3), and that $\max_{j,m,l} \mathbb{E}_n|\dot{\psi}_{mjl}(w)|^2 \leq L_n^2$ with probability at least $1 - \delta_n$ by Condition AS(ii), we can apply Lemma 4 on the moderate deviations for self-normalized sum, following the idea in [3], to conclude that $\max_{j,m,l} |\sqrt{n}\mathbb{E}_n\dot{\psi}_{mjl}(w)| \leq \sqrt{2\log(pn)}L_n$ with probability $1 - o(1)$. Note that this application requires the side condition $\sqrt{2\log(pn)}M_n n^{-1/6} = o(1)$ be satisfied for M_n defined in Condition AS(ii), which indeed holds by Condition AS(iii). We now recall the details of this calculation:

$$
\begin{aligned}
&P\Big(\max_{j,m,l}|\sqrt{n}\mathbb{E}_n\dot{\psi}_{mjl}(w)| > \sqrt{2\log(pn)}L_n\Big) \\
&\quad\leq P\Big(\max_{j,m,l}|\sqrt{n}\mathbb{E}_n\dot{\psi}_{mjl}(w)|/\sqrt{\mathbb{E}_n|\dot{\psi}_{mjl}(w)|^2} > \sqrt{2\log(pn)}\Big) + \delta_n \\
&\quad\leq dMp\max_{j,m,l} P\big(|\sqrt{n}\mathbb{E}_n\dot{\psi}_{mjl}(w)|/\sqrt{\mathbb{E}_n|\dot{\psi}_{mjl}(w)|^2} > \sqrt{2\log(pn)}\big) + \delta_n \\
&\quad\leq dMp2\big(1 - \Phi\big(\sqrt{2\log(pn)}\big)\big)\big(1 + o(1)\big) + \delta_n \leq dMp\frac{2}{pn}\big(1 + o(1)\big) + \delta_n \\
&\quad= o(1),
\end{aligned}
$$

where the penultimate inequality occurs due to the application of Lemma 4 on moderate deviations for self-normalized sums. Putting bounds together we conclude that $\|III_1\| \leq \sqrt{d}\max_j |III_{1j}| \lesssim_P L_n\sqrt{\log(p \vee n)}\sqrt{s}\tau_n = o(1)$, where $o(1)$ holds by the growth restrictions imposed in Condition AS(iii).

The bound on III_2 also follows similarly to [3]. III_{2j} is a sum of M terms, each having mean zero and variance of order $s\log(p \vee n)/n = o(1)$. Indeed, the mean zero occurs because

$$
n^{-1/2}\sum_{i=1}^{n}\mathrm{E}\big[\partial_{t_m}\psi_j\big(w_i, \theta_0, h_0(z_i)\big)r_m(z_i)\big] = n^{-1/2}\sum_{i=1}^{n}\mathrm{E}\big[0 \cdot r_m(z_i)\big] = 0
$$

for each mth term, which holds by $\mathrm{E}[\partial_{t_m}\psi_j(w_i, \theta_0, h_0(z_i))|z_i] = 0$, that is, the orthogonality property (5.3), and the law of iterated expectations. To derive the variance bound, note that for each mth term the variance is

$$
n^{-1}\sum_{i=1}^{n}\mathrm{E}\big[\big\{\partial_{t_m}\psi_j\big(w_i, \theta_0, h_0(z_i)\big)\big\}^2 r_m^2(z_i)\big] \leq C\bar{\mathbf{E}}\big[r_m^2(z)\big] \leq C^2 s\log(p \vee n)/n,
$$

which holds by $\mathrm{E}[\{\partial_{t_m}\psi_j(w_i, \theta_0, h_0(z_i))\}^2|z_i] \leq \mathrm{E}[B^2(w)|z_i] \leq C$ a.s. by virtue of Condition SP(iii), and the law iterated expectations; the last bound in the display holds by AS(i). Hence, $\mathrm{var}(III_{2j}) \leq M^2C^2 s\log(p \vee n)/n \lesssim s\log(p \vee n)/n = o(1)$. Therefore, $\|III_2\| \leq \sum_{j=1}^{d}|III_{2j}| \lesssim_P \sqrt{s\log(p \vee n)/n} = o(1)$ by Chebyshev's inequality.

To deduce that $\|II_2\| = o_P(1)$, we use Condition AS(i)–(iii), the claim of step 1, and Hölder inequalities, concluding that

$$\max_j |II_{2j}| \leq \sqrt{n} K^2 L_n \max_k \mathbb{E}_n\{\widehat{v}_k(w) - v_{0k}(w)\}^2 \lesssim_P \sqrt{n} L_n \tau_n^2 = o(1).$$

Finally, since $\|II_3\| \leq \sqrt{n}\|(\mathbb{E}_n \partial_\theta \psi(w, \theta_0, h_0(z)) - J_n)\|_{\mathrm{op}} \|\widehat{\theta} - \theta_0\|$ and since $\|\mathbb{E}_n \partial_\theta \psi(w, \theta_0, h_0(z)) - J_n\|_{\mathrm{op}} \lesssim_P n^{-1/2}$ by Chebyshev's inequality, using that $\bar{\mathbf{E}} B^2(w) \leq C$ by Condition AS(ii), and $\|\widehat{\theta} - \theta_0\| \lesssim_P \tau_n$ by step 1, conclude that $\|II_3\| \lesssim_P \tau_n = o(1)$. □

Acknowledgements. We are grateful to the Editors and two referees for thoughtful comments and suggestions, which helped improve the paper substantially. We also thank seminar participants at the 2011 Joint Statistical Meetings, 2011 INFORMS, Duke and MIT for many useful suggestions. We gratefully acknowledge research support from the NSF.

SUPPLEMENTARY MATERIAL

Supplementary material (DOI: 10.1214/14-AOS1204SUPP; .pdf). The material contains deferred proofs, additional theoretical results on convergence rates in ℓ_2, ℓ_1 and ℓ_∞, lower bound on the prediction rate, and Monte-Carlo simulations.

REFERENCES

[1] AMEMIYA, T. (1977). The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model. *Econometrica* **45** 955–968. MR0455253

[2] BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2014). Supplement to "Pivotal estimation via square-root Lasso in nonparametric regression." DOI:10.1214/14-AOS1204SUPP.

[3] BELLONI, A., CHEN, D., CHERNOZHUKOV, V. and HANSEN, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80** 2369–2429. MR3001131

[4] BELLONI, A. and CHERNOZHUKOV, V. (2011). High dimensional sparse econometric models: An introduction. In *Inverse Problems and High-Dimensional Estimation*. *Lect*. *Notes Stat*. *Proc*. **203** 121–156. Springer, Heidelberg. MR2868201

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

[6] BELLONI, A. and CHERNOZHUKOV, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547. MR3037163

[7] BELLONI, A., CHERNOZHUKOV, V., FERNANDEZ-VAL, I. and HANSEN, C. (2013). Program evaluation with high-dimensional data. Available at arXiv:1311.2645.

[8] BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2010). Lasso methods for Gaussian instrumental variables models. Available at arXiv:1012.1297.

[9] BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2011). Inference for high-dimensional sparse econometric models. In *Advances in Economics and Econometrics. 10th World Congress of Econometric Society. August 2010* **III** 245–295. Cambridge Univ. Press, New York.

[10] BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2013). Inference on treatment effects after selection amongst high-dimensional controls. *Rev. Econom. Stud.* DOI:10.1093/restud/rdt044.

[11] BELLONI, A., CHERNOZHUKOV, V. and KATO, K. (2012). Uniform post selection inference for LAD regression and other Z-estimation problems. Available at arXiv:1304.0282.

[12] BELLONI, A., CHERNOZHUKOV, V. and WANG, L. (2011). Square-root Lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806. MR2860324

[13] BELLONI, A., CHERNOZHUKOV, V. and WEI, Y. (2013). Honest confidence regions for a regression parameter in logistic regression with a large number of controls. Available at arXiv:1304.3969.

[14] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

[15] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

[16] BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. MR2312149

[17] BUNEA, F., TSYBAKOV, A. B. and WEGKAMP, M. H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35** 1674–1697. MR2351101

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

[19] CANDÈS, E. J. and PLAN, Y. (2009). Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.* **37** 2145–2177. MR2543688

[20] CHAMBERLAIN, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica* **60** 567–596. MR1162999

[21] CHEN, Y. and DALALYAN, A. S. (2012). Fused sparsity and robust estimation for linear models with unknown variance. *Adv. Neural Inf. Process. Syst.* **25** 1268–1276.

[22] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2012). Gaussian approximations of suprema of empirical processes. Available at arXiv:1212.6885.

[23] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448

[24] CHRÉTIEN, S. and DARSES, S. (2012). Sparse recovery with unknown variance: A Lasso-type approach. Available at arXiv:1101.0434.

[25] DE LA PEÑA, V. H., LAI, T. L. and SHAO, Q.-M. (2009). *Self-Normalized Processes*. Springer, Berlin.

[26] DÜMBGEN, L., VAN DE GEER, S. A., VERAAR, M. C. and WELLNER, J. A. (2010). Nemirovski's inequalities revisited. *Amer. Math. Monthly* **117** 138–160. MR2590193

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

[28] FARRELL, M. (2013). Robust inference on average treatment effects with possibly more covariates than observations. Available at arXiv:1309.4686.

[29] GAUTIER, E. and TSYBAKOV, A. (2011). High-dimensional instrumental variables rergession and confidence sets. Available at arXiv:1105.2454.

[30] GIRAUD, C., HUET, S. and VERZELEN, N. (2012). High-dimensional regression with unknown variance. *Statist. Sci.* **27** 500–518. MR3025131

[31] HANSEN, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054. MR0666123

[32] HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley, CA*, 1965/66), *Vol. I: Statistics* 221–233. Univ. California Press, Berkeley, CA. MR0216620

[33] JING, B.-Y., SHAO, Q.-M. and WANG, Q. (2003). Self-normalized Cramér-type large deviations for independent random variables. *Ann. Probab.* **31** 2167–2215. MR2016616

[34] KLOPP, O. (2011). High dimensional matrix estimation with unknown variance of the noise. Available at arXiv:1112.3055.

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

[36] KOSOROK, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York. MR2724368

[37] LEEB, H. and PÖTSCHER, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory* **24** 338–376. MR2422862

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

[39] LOUNICI, K., PONTIL, M., TSYBAKOV, A. B. and VAN DE GEER, S. (2010). Taking advantage of sparsity in multi-task learning. Available at arXiv:0903.1468.

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

[41] ROBINSON, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica* **56** 931–954. MR0951762

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

[43] STÄDLER, N., BÜHLMANN, P. and VAN DE GEER, S. (2010). ℓ_1-penalization for mixture regression models. *TEST* **19** 209–256. MR2677722

[44] SUN, T. and ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166

[45] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. MR2724359

[46] VAN DE GEER, S. A. (2007). The deterministic Lasso. In *JSM proceedings*.

[47] VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the Lasso. *Ann. Statist.* **36** 614–645. MR2396809

[48] VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316

[49] VAN DE GEER, S. A., BÜHLMANN, P. and RITOV, Y. (2013). On asymptotically optimal confidence regions and tests for high-dimensional models. Available at arXiv:1303.0518.

[50] VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

[51] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. Springer, New York. MR1385671

[52] WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873

[53] WANG, L. (2013). The L_1 penalized LAD estimator for high dimensional linear regression. *J. Multivariate Anal.* **120** 135–151. MR3072722

[54] ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448

[55] ZHANG, C.-H. and ZHANG, S. S. (2011). Confidence intervals for low-dimensional parameters with high-dimensional data. Available at arXiv:1110.2563.

[56] ZHAO, R., SUN, T., ZHANG, C.-H. and ZHOU, H. H. (2013). Asymptotic normality and optimalities in estimation of large Gaussian graphical model. Available at arXiv:1309.6024.

A. BELLONI
FUQUA SCHOOL OF BUSINESS
DUKE UNIVERSITY
100 FUQUA DRIVE
DURHAM, NORTH CAROLINA 27708
USA
E-MAIL: abn5@duke.edu

V. CHERNOZHUKOV
DEPARTMENT OF ECONOMICS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
52 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02142
USA
E-MAIL: vchern@mit.edu

L. WANG
DEPARTMENT OF MATHEMATICS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
77 MASSACHUSETTS AVENUE
CAMBRIDGE, MASSACHUSETTS 02139
USA
E-MAIL: lwang@mit.edu