

Massachusetts Institute of Technology
Department of Economics
Working Paper Series

**SPARSE MODELS AND METHODS FOR OPTIMAL
INSTRUMENTS WITH AN APPLICATION TO EMINENT
DOMAIN**

Alexandre Belloni
Daniel Chen
Victor Chernozhukov
Christian Hansen

Working Paper 11-19
July 12, 2011

Room E52-251
50 Memorial Drive
Cambridge, MA 02142

This paper can be downloaded without charge from the
Social Science Research Network Paper Collection at
<http://ssrn.com/abstract=1910169>

SPARSE MODELS AND METHODS FOR OPTIMAL INSTRUMENTS WITH AN APPLICATION TO EMINENT DOMAIN

A. BELLONI, D. CHEN, V. CHERNOZHUKOV, AND C. HANSEN

ABSTRACT. We develop results for the use of LASSO and Post-LASSO methods to form first-stage predictions and estimate optimal instruments in linear instrumental variables (IV) models with many instruments, p , that apply even when p is much larger than the sample size, n . We rigorously develop asymptotic distribution and inference theory for the resulting IV estimators and provide conditions under which these estimators are asymptotically oracle-efficient. In simulation experiments, the LASSO-based IV estimator with a data-driven penalty performs well compared to recently advocated many-instrument-robust procedures. In an empirical example dealing with the effect of judicial eminent domain decisions on economic outcomes, the LASSO-based IV estimator substantially reduces estimated standard errors allowing one to draw much more precise conclusions about the economic effects of these decisions.

Optimal instruments are conditional expectations; and in developing the IV results, we also establish a series of new results for LASSO and Post-LASSO estimators of non-parametric conditional expectation functions which are of independent theoretical and practical interest. Specifically, we develop the asymptotic theory for these estimators that allows for non-Gaussian, heteroscedastic disturbances, which is important for econometric applications. By innovatively using moderate deviation theory for self-normalized sums, we provide convergence rates for these estimators that are as sharp as in the homoscedastic Gaussian case under the weak condition that $\log p = o(n^{1/3})$. Moreover, as a practical innovation, we provide a fully data-driven method for choosing the user-specified penalty that must be provided in obtaining LASSO and Post-LASSO estimates and establish its asymptotic validity under non-Gaussian, heteroscedastic disturbances.

Key Words: Instrumental Variables, Optimal Instruments, LASSO, Post-LASSO, Sparsity, Eminent Domain, Data-Driven Penalty, Heteroscedasticity, non-Gaussian errors, moderate deviations for self-normalized sums

Date: First version: June 2009, this version October 28, 2010. Preliminary results of this paper were first presented at Chernozhukov's invited Cowles Foundation lecture at the Northern American meetings of the Econometric society in June of 2009. We thank seminar participants at Brown, Columbia, Harvard-MIT, the Dutch Econometric Study Group, Fuqua School of Business, and NYU for helpful comments. We also thank Denis Chetverikov, JB Doyle, and Joonhwan Lee for thorough reading of this paper and helpful feedback.

1. INTRODUCTION

Instrumental variables (IV) techniques are widely used in applied economic research. While these methods provide a useful tool for identifying structural effects of interest, their application often results in imprecise inference. One way to improve the precision of instrumental variables estimators is to use many instruments or to try to approximate the optimal instruments as in [1], [15], and [37]. Estimation of optimal instruments which will generally be done nonparametrically and thus implicitly makes use of many constructed instruments such as polynomials. The promised improvement in efficiency is appealing, but recent work in econometrics has also demonstrated that IV estimators that make use of many instruments may have very poor properties; see, for example, [3], [18], [27], and [28] which propose solutions for this problem based on “many-instrument” asymptotics.¹

In this paper, we contribute to the literature on IV estimation with many instruments by considering the use of LASSO-based methods, namely LASSO and Post-LASSO, for estimating the first-stage regression of endogenous variables on the instruments, and deriving the asymptotic estimation and inferential properties of the resulting second-stage IV estimators. LASSO is a widely used method that acts both as an estimator of regression functions and a model selection device. LASSO solves for regression coefficients by minimizing the usual least squares objective function subject to a penalty for model size through the sum of the absolute values of the coefficients. The resulting LASSO estimator selects only the most relevant instruments and estimates the first-stage regression coefficients via a shrinkage procedure. The Post-LASSO estimator discards the LASSO coefficient estimates and only uses the parsimonious set of data-dependent instruments selected by LASSO to refit the first stage regression via OLS thereby eliminating the LASSO’s shrinkage bias.² For theoretical and simulation evidence regarding LASSO’s performance see, for example, [8], [12], [13], [29], [11], [14], [32], [33], [34], [35], [41], [43], [45], [46], [47], and [4], among many others; for analogous results on Post-LASSO see [4].

¹It is important to note that the precise definition of “many-instrument” is $p \propto n$ with $p < n$ as in [3] where p is the number of instruments and n is the sample size. The current paper expressly allows for this case and also for “very many-instrument” asymptotics where $p \gg n$.

²[16] considers an alternate shrinkage estimator in the context of IV estimation with many instruments. [2] considers IV estimation with many instruments based on principal components analysis and variable selection via boosting, and [39] provides results for Ridge regression.

The use of LASSO-based methods to form first-stage predictions for use in IV estimation provides a practical approach to obtaining the efficiency gains available from using optimal instruments while dampening the problems associated with many instruments. We show that LASSO-based procedures produce first-stage predictions that approximate the optimal instruments and perform well when the optimal instrument may be well-approximated using a small, but unknown, set of the available instruments even when the number of potential instruments is allowed to be much larger than the sample size.³ We derive the asymptotic distribution of the resulting IV estimator and provide conditions under which it achieves the semi-parametric efficiency bound; i.e. it is oracle efficient. We provide a consistent asymptotic variance estimator that allow one to perform inference using the derived asymptotic distribution. Thus, our results considerably generalize and extend the classical IV procedure of [37] based on conventional series approximation of the optimal instruments.

Our paper also contributes to the growing literature on the theoretical properties of LASSO-based methods by providing results for LASSO-based estimators of nonparametric conditional expectations. We provide rates of convergence allowing for non-Gaussian, heteroscedastic disturbances. Our results generalize most LASSO and Post-LASSO results which assume both homoscedasticity and Gaussianity. These results are important for applied economic analysis where researchers are very concerned about heteroscedasticity and non-normality in their data. By innovatively using moderate deviation theory for self-normalized sums, we provide convergence rates for LASSO and Post-LASSO that are as sharp as in the homoscedastic Gaussian case under the weak condition that $\log p = o(n^{1/3})$. We provide a fully data-driven method for choosing the user-specified penalty that must be provided to obtain LASSO and Post-LASSO estimates, and we establish its asymptotic validity allowing for non-Gaussian, heteroscedastic disturbances. Ours is the first paper to provide such a data-driven penalty which was previously not available even in the Gaussian case.⁴ These results are of independent interest in a wide variety of theoretical and applied settings.

³This is in contrast to the variable selection method of [22] which relies on a *a priori* knowledge that allows one to order the instruments in terms of instrument strength.

⁴One exception is the work of [7], where the square-root-LASSO estimators are considered that allow one to uses pivotal penalty choices; those results however strongly rely on homoscedasticity.

We illustrate the performance of LASSO-based IV through a series of simulation experiments. In these experiments, we find that a feasible LASSO-based procedure that uses our data-driven penalty performs well across a wide range of simulation designs. In terms of estimation risk, it outperforms both LIML and its modification due to [25] (FULL)⁵ which have been advocated as procedures that are robust to using many instruments (e.g. [27]). In terms of inference based on 5% level tests, the LASSO-based IV estimator performs comparably to LIML and FULL in the majority of cases. Overall, the simulation results are very favorable to the proposed LASSO-based IV procedures.

Finally, we demonstrate the potential gains of the LASSO-based procedure in an application where there are many available instruments among which there is not a clear *a priori* way to decide which instruments to use. In particular, we look at the effect of judicial decisions at the federal circuit court level regarding the government’s exercise of eminent domain on house prices and state-level GDP as in [20]. We follow the identification strategy of [20] who use the random assignment of judges to three judge panels that are then assigned to eminent domain cases to justify using the demographic characteristics of the judges on the realized panels as instruments for their decision. This strategy produces a situation in which there are many potential instruments in that all possible sets of characteristics of the three judge panel are valid instruments. We find that the LASSO-based estimates using the data-dependent penalty produce much larger first-stage F-statistics and have substantially smaller estimated second stage standard errors than estimates obtained using the baseline instruments of [20]. This improvement of precision clearly allows one to draw more precise conclusions about the effects of the judicial decisions on economic outcomes relative to the benchmark case.

Notation. In what follows, we allow for the models to change with the sample size, i.e. we allow for array asymptotics, so all parameters are implicitly indexed by the sample size n , but we omit the index to simplify notation. We use array asymptotics to better capture some finite-sample phenomena. We also use the following empirical process notation,

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

⁵Note that these procedures are only applicable when the number of instruments p is less than the sample size n . As mentioned earlier, procedures developed in this paper allow for p to be much larger n .

and

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

The l_2 -norm is denoted by $\|\cdot\|$, and the l_0 -norm, $\|\cdot\|_0$, denotes the number of non-zero components of a vector. The empirical $L^2(\mathbb{P}_n)$ norm of a random variable W_i is defined as

$$\|W_i\|_{2,n} := \sqrt{\mathbb{E}_n[W_i^2]}.$$

Given a vector $\delta \in \mathbb{R}^p$, and a set of indices $T \subset \{1, \dots, p\}$, we denote by δ_T the vector in which $\delta_{Tj} = \delta_j$ if $j \in T$, $\delta_{Tj} = 0$ if $j \notin T$. We use the notation $(a)_+ = \max\{a, 0\}$, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. We also use the notation $a \lesssim b$ to denote $a \leq cb$ for some constant $c > 0$ that does not depend on n ; and $a \lesssim_P b$ to denote $a = O_P(b)$. For an event E , we say that E wp $\rightarrow 1$ when E occurs with probability approaching one as n grows. We say $X_n =_d Y_n + o_P(1)$ to mean that X_n has the same distribution as Y_n up to a term $o_P(1)$ that vanishes in probability. Such statements are needed to accommodate asymptotics for models that change with n . When Y_n is a fixed random vector, that does not change with n , i.e. $Y_n = Y$, this notation is equivalent to $X_n \rightarrow_d Y$.

2. SPARSE MODELS AND METHODS FOR OPTIMAL INSTRUMENTAL VARIABLES

In this section of the paper, we present the model and provide an overview of the main results. Sections 3 and 4 provide a technical presentation that includes a set of sufficient regularity conditions, discusses their plausibility, and establishes the main formal results of the paper.

2.1. The IV Model and Statement of The Problem. The model is $y_i = d_i' \alpha_0 + \epsilon_i$, where y_i is the response variable, and d_i is a finite k_d -vector of variables, whose first k_e elements contain endogenous variables. The disturbance ϵ_i obeys

$$\mathbb{E}[\epsilon_i | x_i] = 0,$$

where α_0 denotes the true value of a vector-valued parameter α and x_i are instrumental variables. As a motivation, suppose that the structural disturbance is conditionally homoscedastic, namely

$$\mathbb{E}[\epsilon_i^2 | x_i] = \sigma^2.$$

Given a k_d -vector of instruments $A(x_i)$, the standard IV estimator of α_0 is given by $\hat{\alpha} = (\mathbb{E}_n[A(x_i)d_i'])^{-1} \mathbb{E}_n[A(x_i)y_i]$, where $\{(x_i, d_i, y_i), i = 1, \dots, n\}$ is an i.i.d. sample from the IV model

above. For a given $A(x_i)$, $\sqrt{n}(\hat{\alpha} - \alpha_0) =_d N(0, Q_0^{-1}\Omega_0Q_0^{-1'}) + o_P(1)$, where $Q_0 = E[A(x_i)d_i']$ and $\Omega_0 = \sigma^2 E[A(x_i)A(x_i)']$ under the standard conditions. Setting

$$A(x_i) = D(x_i) = E[d_i|x_i]$$

minimizes the limit variance which becomes $\Lambda^* = \sigma^2\{E[D(x_i)D(x_i)']\}^{-1}$, the semi-parametric efficiency bound for estimating α_0 ; see [1], [15], and [37]. In practice, the optimal instrument $D(x_i)$ is an unknown non-parametric function and has to be estimated. In what follows, we investigate the use of sparse methods – namely LASSO and Post-LASSO – for use in estimating the optimal instruments. The resulting IV estimator is as efficient as the infeasible optimal IV estimator above.

Note that if d_i contains exogenous components w_i , then $d_i = (d_1, \dots, d_{k_e}, w_i)'$ where the first k_e variables are endogenous. Since the rest of the components w_i are exogenous, they appear in $x_i = (w_i', \tilde{x}_i)$. It follows that

$$D_i := D(x_i) := E[d_i|x_i] = (E[d_1|x_i], \dots, E[d_{k_e}|x_i], w_i)';$$

i.e. the estimator of w_i is simply w_i . Therefore, we discuss estimation of conditional expectation functions:

$$D_{il} := D_l(x_i) := E[d_l|x_i], \quad l = 1, \dots, k_e.$$

2.2. Sparse Models for Optimal Instruments and Other Conditional Expectations.

Suppose there is a very large list of technical instruments,

$$f_i := (f_{i1}, \dots, f_{ip})' := (f_1(x_i), \dots, f_p(x_i))', \quad (2.1)$$

to be used in estimation of conditional expectations $D_l(x_i)$, $l = 1, \dots, k_e$, where

the number of instruments p is possibly much larger than the sample size n .

For example, high-dimensional instruments f_i could arise as in the following two cases:

- **Many Instruments Case.** The list of available instruments is large, in which case we have $f_i = x_i$ as in e.g. [1] and [3].
- **Many Series Instruments Case.** The list f_i consists of a large number of series terms with respect to some elementary regressor vector x_i , e.g., f_i could be composed of B-splines, dummies, polynomials, and various interactions as in e.g. [37].

We mainly use the term “series instruments” and contrast our results with those in the seminal work of [37], though our results are not limited to canonical series regressors as in [37]. The most important feature of our approach is that by allowing p to be much larger than the sample size, we are able to consider many more series instruments than in [37] to approximate the optimal instruments.

The key assumption that allows effective use of this large set of instruments is sparsity. To fix ideas, consider the case where $D_l(x_i)$ is a function of only $s \ll n$ instruments:

$$D_l(x_i) = f'_i \beta_{l0}, \quad l = 1, \dots, k_e, \tag{2.2}$$

$$\max_{1 \leq l \leq k_e} \|\beta_{l0}\|_0 = \max_{1 \leq l \leq k_e} \sum_{j=1}^p 1\{\beta_{lj} \neq 0\} \leq s \ll n. \tag{2.3}$$

This simple sparsity model substantially generalizes the classical parametric model of optimal instruments of [1] by letting the identities of the relevant instruments $T_l = \text{support}(\beta_{l0}) = \{j \in \{1, \dots, p\} : |\beta_{lj}| > 0\}$ be unknown. This generalization is useful in practice since it is unrealistic to assume we know the identities of the relevant instruments in many examples.

The previous model is simple and allows us to convey the essence of the approach. However, it is unrealistic in that it presumes exact sparsity. We make no formal use of this model, but instead use a much more general, approximately sparse or nonparametric model:

Condition AS.(Approximately Sparse Optimal Instrument). *Each optimal instrument function $D_l(x_i)$ is well approximated by a function of unknown $s \ll n$ instruments:*

$$D_l(x_i) = f'_i \beta_{l0} + a_l(x_i), \quad l = 1, \dots, k_e, \tag{2.4}$$

$$\max_{1 \leq l \leq k_e} \|\beta_{l0}\|_0 \leq s = o(n), \quad \max_{1 \leq l \leq k_e} [\mathbb{E}_n a_l(x_i)^2]^{1/2} \leq c_s \lesssim_P \sqrt{s/n}. \tag{2.5}$$

This model generalizes the nonparametric model of the optimal instrument of [37] by letting the identities of the most important series terms

$$T_l = \text{support}(\beta_{l0})$$

be unknown and potentially different for $l = 1, \dots, k_e$. The number s is defined so that the approximation error is of the same order as the estimation error, $\sqrt{s/n}$, of the oracle estimator. This rate generalizes the rate for the optimal number s of series terms in [37] by not relying on knowledge of what s series terms to include. Knowing the identities of the most important

series terms is unrealistic in many examples in practice. Indeed, the most important series terms need not be the first s terms, and the optimal number of series terms to consider is also unknown. Moreover, an optimal series approximation to the instrument could come from the combination of completely different bases e.g by using both polynomials and B-splines. LASSO and Post-LASSO use the data to estimate the set of the most relevant series terms in a manner that allows the resulting IV estimator to achieve good performance if the following key growth condition holds:

$$\frac{s^2(\log p)^2}{n} \rightarrow 0 \quad (2.6)$$

along with other more technical conditions. This condition requires the optimal instruments to be sufficiently smooth so that a small (relative to n) number of series terms can be used to approximate them well, ensuring that the impact of instrument estimation on the IV estimator is asymptotically negligible.

Remark 1.1(Plausibility, Generality, and Usefulness of Condition AS) It is clear from the statement of Condition AS that this expansion incorporates both substantial generalizations and improvements over the conventional series approximation of optimal instruments in [37] and [38]. In order to explain this consider the case of $k_e = 1$ and the set $\{f_j(x), j \geq 1\}$ of orthonormal basis functions on $[0, 1]^d$, e.g. B-splines or orthopolynomials, with respect to the Lebesgue measure. Suppose x_i have a uniform distribution on $[0, 1]^d$ for simplicity.⁶ Since $ED_l^2(x_i) < \infty$ by assumption, we can represent D_l via a Fourier expansion, $D_l(x) = \sum_{j=1}^{\infty} \delta_j f_j(x)$, where $\{\delta_j, j \geq 1\}$ are the Fourier coefficients such that $\sum_{j=1}^{\infty} \delta_j^2 < \infty$.

Suppose that Fourier coefficients feature a polynomial decay $\delta_j \propto j^{-a}$, where a is a measure of smoothness of D_l . Consider the conventional series expansion that uses the first K terms for approximation, $D_l(x) = \sum_{j=1}^K \beta_{l0j} f_j(x) + a_l^c(x)$, with $\beta_{l0j} = \delta_j$. Here $a_l^c(x_i)$ is the approximation error that obeys $\sqrt{\mathbb{E}_n[a_l^c(x_i)]} \lesssim_P \sqrt{\mathbb{E}[a_l^c(x_i)]} \lesssim K^{-\frac{2a+1}{2}}$. Balancing the order $K^{-\frac{2a+1}{2}}$ of approximation error with the order $\sqrt{K/n}$ of the estimation error gives the oracle-rate-optimal number of series terms $s = K \propto n^{1/2a}$, and the resulting oracle series estimator, which knows s , will estimate D_l at the oracle rate of $n^{\frac{1-2a}{4a}}$. This also gives us the identity of the most important series terms $T_l = \{1, \dots, s\}$, which are simply the first s terms. We conclude that Condition AS holds for the sparse approximation $D_l(x) = \sum_{j=1}^p \beta_{l0j} f_j(x) + a_l(x)$, with $\beta_{l0j} = \delta_j$ for $j \leq s$ and

⁶The discussion in this example continues to apply when x_i has a density that is bounded above and away from zero on $[0, 1]^d$.

$\beta_{l0j} = 0$ for $s + 1 \leq j \leq p$, and $a_l(x_i) = a_l^c(x_i)$, which coincides with the conventional series approximation above, so that $\sqrt{\mathbb{E}_n[a_l^2(x_i)]} \lesssim_P \sqrt{s/n}$ and $\|\beta_{l0}\|_0 \leq s$; moreover, the key growth condition (2.6) required for IV estimation holds if the smoothness index $a > 1$ is sufficiently high so that $n^{1/a}(\log p)^2/n \rightarrow 0$. Note that, despite not knowing the most relevant series terms or the optimal number of terms s , the LASSO-based estimators of the next section will match the oracle rate for estimating $D_l(x)$ up to logarithmic terms in p .

Next suppose that Fourier coefficients feature the following pattern $\delta_j = 0$ for $j \leq M$ and $\delta_j \propto (j - M)^{-a}$ for $j > M$. Clearly in this case the standard series approximation based on the first $K \leq M$ terms, $\sum_{j=1}^K \delta_j f_j(x)$, fails completely to provide any predictive power for $D_l(x)$, and the corresponding standard series estimator based on K terms therefore also fails completely.⁷ In sharp contrast, Condition AS allows for an approximation that performs at an oracle level. Indeed, if $\log M \lesssim \log n$ and $M \gg n^{\frac{1}{2a}}$, we can use first p series terms such that $M + n^{\frac{1}{2a}} = o(p)$ in the approximation $D_l(x) = \sum_{j=1}^p \beta_{l0j} f_j(x) + a_l(x)$, where for $s \propto n^{\frac{1}{2a}}$ we set $\beta_{l0j} = 0$ for $j \leq M$ and $j > M + s$, and $\beta_{l0j} = \delta_j$ for $M + 1 \leq j \leq M + s$. Hence $\|\beta_{l0}\|_0 \leq s$ and we have that $\sqrt{\mathbb{E}_n[a_l^2(x_i)]} \lesssim_P \sqrt{\mathbb{E}[a_l^2(x_i)]} \lesssim s^{-\frac{2a+1}{2}}(1 + o(1)) \lesssim \sqrt{s/n} \lesssim n^{\frac{1-2a}{4a}}$. Note again that the LASSO-based estimators of the next section will match the oracle rate for estimating $D_l(x)$ up to logarithmic terms in p despite not relying on knowledge of the most relevant series terms or the optimal number of terms s . \square

2.3. LASSO-Based Estimation Methods for Optimal Instruments and Other Conditional Expectation Functions. Let us write the first-stage regression equations as

$$d_{il} = D_l(x_i) + v_{il}, \quad \mathbb{E}[v_{il}|x_i] = 0, \quad l = 1, \dots, k_e. \tag{2.7}$$

Given the sample $\{(d_{il}, l = 1, \dots, k_e), x_i, i = 1, \dots, n\}$, we consider estimators of the optimal instrument $D_{il} = D_l(x_i)$ that take the form

$$\widehat{D}_{il} := \widehat{D}_l(x_i) = f_i' \widehat{\beta}_l, \quad l = 1, \dots, k_e, \tag{2.8}$$

where $\widehat{\beta}_l$ is the LASSO or Post-LASSO estimator obtained by using d_{il} as the dependent variables and f_i as regressors.

⁷This is not merely a finite sample phenomenon but is also accommodated in the asymptotics since we expressly allow for the array asymptotics; i.e. the underlying true model could change with n . Recall that we omit the indexing by n for ease of notation.

Consider the usual least squares criterion function:

$$\widehat{Q}_l(\beta) := \mathbb{E}_n[(d_{il} - f'_i \beta)^2]$$

The LASSO estimator [43] is defined as a solution of the following optimization program:

$$\widehat{\beta}_{lL} \in \arg \min_{\beta \in \mathbb{R}^p} \widehat{Q}_l(\beta) + \frac{\lambda}{n} \|\widehat{\Upsilon}_l \beta\|_1 \quad (2.9)$$

where λ is the penalty level, and $\widehat{\Upsilon}_l = \text{diag}(\widehat{\gamma}_{l1}, \dots, \widehat{\gamma}_{lp})$ is a diagonal matrix specifying penalty loadings.

We develop two options for setting the penalty level and the loadings:

$$\begin{aligned} \text{initial} \quad & \widehat{\gamma}_{lj} = \sqrt{\mathbb{E}_n[f_{ij}^2 (d_{il} - \bar{d}_l)^2]}, \quad \lambda = 2c \sqrt{2n \log(2pk_e)}, \\ \text{refined} \quad & \widehat{\gamma}_{lj} = \sqrt{\mathbb{E}_n[f_{ij}^2 \widehat{v}_{il}^2]}, \quad \lambda = 2c \sqrt{2n \log(2pk_e)}, \end{aligned} \quad (2.10)$$

where $c > 1$ is a constant and $\bar{d}_l := \mathbb{E}_n[d_{il}]$. We can use the initial option for penalty loadings to compute pilot LASSO and/or Post-LASSO estimates and then use the residuals \widehat{v}_{il} in the refined option. We can iterate on the latter step a bounded number of times. In practice, we recommend to set the constant $c = 1.1$, which we prove to be an asymptotically valid choice under the conditions stated.

The Post-LASSO estimator is defined as the ordinary least square regression applied to the model \widehat{T}_l selected by the LASSO. Formally, set

$$\widehat{T}_l = \text{support}(\widehat{\beta}_{lL}) = \{j \in \{1, \dots, p\} : |\widehat{\beta}_{lLj}| > 0\}, \quad l = 1, \dots, k_e,$$

and define the Post-LASSO estimator $\widehat{\beta}_{lPL}$ as

$$\widehat{\beta}_{lPL} \in \arg \min_{\beta \in \mathbb{R}^p: \beta_{\widehat{T}_l^c} = 0} \widehat{Q}_l(\beta), \quad l = 1, \dots, k_e. \quad (2.11)$$

In words, this estimator is ordinary least squares (OLS) applied to the data after removing the instruments/regressors that were not selected by LASSO.

LASSO and Post-LASSO are motivated by the desire to fit the target function well without overfitting. Clearly, the OLS estimator is not consistent for estimating β_{l0} in the setting with $p > n$. Some classical approaches based on BIC-penalization of model size are consistent but computationally infeasible. The LASSO estimator [43] resolves these difficulties by penalizing the model size through the sum of absolute parameter values. The LASSO estimator is computationally attractive because it minimizes a convex function. Moreover, under suitable

conditions, this estimator achieves near-optimal rates in estimating the model $D_l(x_i)$; see discussion and references below. The estimator achieves these rates by adapting to the unknown smoothness or sparsity of the regression function $D_l(x_i)$. Nonetheless, the estimator has an important drawback: The regularization by the l_1 -norm employed in (2.9) naturally lets the LASSO estimator avoid overfitting the data, but it also shrinks the fitted coefficients towards zero causing a potentially significant bias.

In order to remove some of this bias, we consider the Post-LASSO estimator. If the model selection by LASSO works perfectly – that is, when it selects exactly all “relevant” instruments – then the resulting Post-LASSO estimator is simply the standard oracle OLS estimator, and the resulting optimal IV estimator $\hat{\alpha}$ is simply the standard series estimator of the optimal instruments of [37] whose properties are well-known. In cases where perfect selection does not occur, Post-LASSO estimates of coefficients will still tend to be less biased than LASSO.

We contribute to the broad LASSO literature cited in the introduction by showing that under possibly heteroscedastic and non-Gaussian reduced form errors the LASSO and Post-LASSO estimators obey the following near-oracle performance bounds:

$$\max_{1 \leq l \leq k_e} \|\hat{D}_{il} - D_{il}\|_{2,n} \lesssim_P \sqrt{\frac{s \log p}{n}} \quad (2.12)$$

$$\max_{1 \leq l \leq k_e} \|\hat{\beta}_l - \beta_{l0}\|_1 \lesssim_P \sqrt{\frac{s^2 \log p}{n}}. \quad (2.13)$$

The performance bounds in (2.12)-(2.13) are called near-oracle because they coincide up to a $\sqrt{\log p}$ factor with the bounds achievable when the the ideal series terms T_l for each of the k_e regressions equations in (2.4) are known. Our results extend those of [8] for LASSO with Gaussian errors and those of [4] for Post-LASSO with Gaussian errors. Notably, these bounds are as sharp as the results for the Gaussian case under the weak condition $\log p = o(n^{1/3})$. They are also the first results in the literature that allow for data-driven choice of the penalty level. We prove the above results in part through an innovative use of the moderate deviation theory for self-normalized sums.

2.4. The Instrumental Variable Estimator based on LASSO and Post-LASSO constructed Optimal Instrument. Given smoothness assumption AS, we take advantage of the approximate sparsity by using LASSO and Post-LASSO methods to construct estimates of

$D_l(x_i)$ of the form

$$\widehat{D}_l(x_i) = f'_i \widehat{\beta}_l, l = 1, \dots, k_e, \quad (2.14)$$

and then set

$$\widehat{D}_i = (\widehat{D}_1(x_i), \dots, \widehat{D}_l(x_i), w'_i)'. \quad (2.15)$$

The resulting IV estimator takes the form

$$\widehat{\alpha}^* = \mathbb{E}_n[\widehat{D}_i d'_i]^{-1} \mathbb{E}_n[\widehat{D}_i y_i]. \quad (2.16)$$

The main result of this paper is to show that, despite the possibility of p being very large, LASSO and Post-LASSO can select a relatively small data-dependent set of effective instruments to produce estimates of the optimal instruments \widehat{D}_i such that the resulting IV estimator achieves the efficiency bound asymptotically:

$$\sqrt{n}(\widehat{\alpha}^* - \alpha_0) =_d N(0, \Lambda^*) + o_P(1). \quad (2.17)$$

That is, the LASSO-based and Post-LASSO based IV estimator asymptotically achieves oracle performance. Thus the estimator matches the performance of the classical/standard series-based IV estimator of [37] with the following advantages:

- **Adaptivity to Unknown Smoothness/Sparsity.** The LASSO-based procedures automatically adapt to the unknown smoothness/sparsity of the true optimal instrument D_i and automatically choose the (nearly) optimal number of series terms. This is in contrast to the standard series procedure that does not adapt to the unknown smoothness and can fail if the incorrect number of terms is chosen. In order for the standard procedure to perform well one needs to use cross-validation or other methods for choosing the optimal number of series terms. Note that both methods still rely on the sufficient smoothness of the optimal instrument.
- **Enhanced Approximation of the Optimal Instrument by Considering Very Many Series Terms.** The LASSO-based procedures can estimate optimal instruments more precisely than the standard series procedures that use just the first $K = o(n)$ series terms both in finite samples and under array asymptotics. Indeed the LASSO-based procedures use very many series instruments, with the total number of series instruments p being possibly much larger than the sample size n , and select amongst

them find the best (relatively) small set for approximating the optimal instruments. To illustrate this potential advantage with an extreme example, suppose that the optimal instrument obeys $D_{il} = \sum_{j=1}^p \delta_{lj} f_j$, with $\delta_{lj} = 0$ for $j \leq K$, and $\delta_{lj} \propto j^{-a}$, then the standard series procedure based on the first K series terms will fail to approximate the optimal instrument completely, but the LASSO-based procedure will estimate the optimal instrument at the near oracle rate, ensuring asymptotic oracle-optimality for the final IV estimator, if the smoothness index a is not too low.⁸

We also show that the IV estimator with LASSO-based optimal instruments continues to be root- n consistent and asymptotically normal in the presence of heteroscedasticity:

$$\sqrt{n}(\hat{\alpha} - \alpha_0) =_d N(0, Q^{-1}\Omega Q^{-1}) + o_P(1), \tag{2.18}$$

where $\Omega := E[\epsilon_i^2 D(x_i) D(x_i)']$ and $Q := E[D(x_i) D(x_i)']$. A consistent estimator for the asymptotic variance is

$$\hat{Q}^{-1} \hat{\Omega} \hat{Q}^{-1}, \quad \hat{\Omega} := \mathbb{E}_n[\hat{\epsilon}_i^2 \hat{D}(x_i) \hat{D}(x_i)'], \quad \hat{Q} := \mathbb{E}_n[\hat{D}(x_i) \hat{D}(x_i)']. \tag{2.19}$$

Using (2.19) permits us to perform robust inference.

Finally, we remark that our results for the IV estimator do not rely on the LASSO and LASSO-based procedure specifically; we provide the properties of the IV estimator for any generic sparsity-based procedure that achieves the near-oracle performance bounds (2.12)-(2.13).

2.5. Implementation Algorithms. It is useful to organize the precise implementation details into the following two algorithms. We establish the asymptotic validity of these algorithms in the subsequent sections. Let $K \geq 1$ denote a bounded number of iterations.

Algorithm 2.1 (IV Estimation and Inference Using LASSO Estimates of Optimal Instrument).

- (1) For each $l = 1, \dots, k_e$, specify penalty loadings according to the initial option in (2.10) and compute the LASSO estimator $\hat{\beta}_{lL}$ via (2.9) and the residuals $\hat{v}_{il} = d_{il} - f_i' \hat{\beta}_{lL}$, $i = 1, \dots, n$.

⁸ These finite-sample differences translate into asymptotic differences, as we do allow for the true models to change with the sample size n to better approximate such finite-sample phenomena. (See also Remark 1.1.)

- (2) For each $l = 1, \dots, k_e$, update the penalty loadings according to the refined option in (2.10) and update the LASSO estimator $\widehat{\beta}_{lL}$ via (2.9) and the residuals $\widehat{v}_{il} = d_{li} - f'_i \widehat{\beta}_{lL}$, $i = 1, \dots, n$.
- (3) Repeat the previous step K times, where K is bounded. Compute the estimates of the optimal instrument $\widehat{D}_{il} = f'_i \widehat{\beta}_{lL}$ for $i = 1, \dots, n$ and each $l = 1, \dots, k_e$. Then compute the IV estimator $\widehat{\alpha}$ defined in (2.16).
- (4) Compute the robust estimates (2.19) of the asymptotic variance matrix, and proceed to perform conventional inference using the normality result (2.18).

The first algorithm involves the use of LASSO in the first two steps.

Algorithm 2.2 (IV Estimation and Inference Using Post-LASSO Estimates of the Optimal Instrument).

- (1) For each $l = 1, \dots, k_e$, specify penalty loadings according to the initial option in (2.10) and compute the LASSO estimator via (2.9), then compute the Post-LASSO estimator $\widehat{\beta}_{lPL}$ via (2.11) and the residuals as $\widehat{v}_{il} = d_{li} - f'_i \widehat{\beta}_{lPL}$, $i = 1, \dots, n$.
- (2) For each $l = 1, \dots, k_e$, update the penalty loadings according to the refined option in (2.10) and update the LASSO estimator via (2.9), then update the Post-LASSO estimator $\widehat{\beta}_{lPL}$ via (2.11) and the residuals as $\widehat{v}_{il} = d_{li} - f'_i \widehat{\beta}_{lPL}$, $i = 1, \dots, n$.
- (3) Repeat the previous step K times, where K is bounded. Compute the estimates of the optimal instrument $\widehat{D}_{il} = f'_i \widehat{\beta}_{lPL}$ for $i = 1, \dots, n$ and each $l = 1, \dots, k_e$. Then compute the IV estimator $\widehat{\alpha}$ defined in (2.16).
- (4) Compute the robust estimates (2.19) of the asymptotic variance matrix, and proceed to perform conventional inference using the normality result (2.18).

The second algorithm involves the use of Post-LASSO in the first two steps and is our preferred algorithm. Our results allow for hybrids between this and the previous algorithm.

3. MAIN RESULTS ON LASSO AND POST-LASSO ESTIMATORS OF THE CONDITIONAL EXPECTATION FUNCTIONS UNDER HETEROSCEDASTIC, NON-GAUSSIAN ERRORS

In this section, we present our main results on LASSO and Post-LASSO estimators of conditional expectation functions under non-classical assumptions and data-driven penalty choices.

The problem we are analyzing in this section is of general interest, having many applications well-outside the IV framework of the present paper.

3.1. Regularity Conditions for Estimating Conditional Expectations. The key condition concerns the behavior of the empirical Gram matrix $\mathbb{E}_n[f_i f_i']$. This matrix is necessarily singular when $p > n$, so in principle it is not well-behaved. However, we only need good behavior of certain moduli of continuity of the Gram matrix. The first modulus of continuity is called the restricted eigenvalues and is needed for LASSO. The second modulus is called the sparse eigenvalue and is needed for Post-LASSO.

In order to define the restricted eigenvalue, first define the restricted set:

$$\Delta_{C,T} = \{\delta \in \mathbb{R}^p : \|\delta_{T^c}\|_1 \leq C\|\delta_T\|_1, \delta \neq 0\},$$

then the restricted eigenvalues of a Gram matrix $M = \mathbb{E}_n[f_i f_i']$ takes the form:

$$\kappa_C^2(M) := \min_{\delta \in \Delta_{C,T}, |T| \leq s} \frac{\delta' M \delta}{\|\delta_T\|_1^2} \text{ and } \tilde{\kappa}_C^2(M) := \min_{\delta \in \Delta_{C,T}, |T| \leq s} \frac{\delta' M \delta}{\|\delta\|_2^2}. \quad (3.20)$$

These restricted eigenvalues can depend on n , but we suppress the dependence in our notation.

In making simplified asymptotic statements involving the LASSO estimator, we will invoke the following condition:

Condition RE. *For any $C > 0$, there exist finite constants $n_0 > 0$ and $\kappa > 0$, which can depend on C , such that the restricted eigenvalues obey with probability approaching one $\kappa_C(\mathbb{E}_n[f_i f_i']) \geq \kappa$ and $\tilde{\kappa}_C(\mathbb{E}_n[f_i f_i']) \geq \kappa$ as $n \rightarrow \infty$.*

The restricted eigenvalue (3.20) is a variant of the restricted eigenvalues introduced in Bickel, Ritov and Tsybakov [8] to analyze the properties of LASSO in the classical Gaussian regression model. Even though the minimal eigenvalue of the empirical Gram matrix $\mathbb{E}_n[f_i f_i']$ is zero whenever $p \geq n$, [8] show that its restricted eigenvalues can in fact be bounded away from zero. Lemmas 1 and 2 below contain sufficient conditions for this. Many more sufficient conditions are available from the literature; see [8]. Consequently, we take the restricted eigenvalues as primitive quantities and Condition RE as a primitive condition. Note also that the restricted eigenvalues are tightly tailored to the ℓ_1 -penalized estimation problem.

In order to define the sparse eigenvalues, let us define the m -sparse subset of a unit sphere as

$$\Delta(m) = \{\delta \in \mathbb{R}^p : \|\delta\|_0 \leq m, \|\delta\|_2 = 1\},$$

and also define the minimal and maximal m -sparse eigenvalue of the Gram matrix $M = \mathbb{E}_n [f_i f_i']$ as

$$\phi_{\min}(m)(M) = \min_{\delta \in \Delta(m)} \delta' M \delta \quad \text{and} \quad \phi_{\max}(m)(M) = \max_{\delta \in \Delta(m)} \delta' M \delta. \quad (3.21)$$

To simplify asymptotic statements for Post-LASSO, we use the following condition:

Condition SE. *For any $C > 0$, there exists constants $0 < \kappa' < \kappa'' < \infty$ that do not depend on n but can depend on C , such that with probability approaching one, as $n \rightarrow \infty$, $\kappa' \leq \phi_{\min}(Cs)(\mathbb{E}_n[f_i f_i']) \leq \phi_{\max}(Cs)(\mathbb{E}_n[f_i f_i']) \leq \kappa''$.*

Recall that the empirical Gram matrix $\mathbb{E}_n[f_i f_i']$ is necessarily singular when $p > n$, so in principle it is not well-behaved. However, Condition SE requires only that certain “small” $m \times m$ submatrices of the large $p \times p$ empirical Gram matrix are well-behaved, which is a reasonable assumption and which will be sufficient for the results that follow. Moreover, Condition SE implies Condition RE by the argument given in [8].

The following lemmas show that Conditions RE and SE are plausible for both many-instrument and many series-instrument settings.

Lemma 1 (Plausibility of RE and SE under Many Gaussian Instruments). *Suppose f_i , $i = 1, \dots, n$, are i.i.d. zero-mean Gaussian random vectors. Further suppose that the population Gram matrix $\mathbb{E}[f_i f_i']$ has diagonal entries bounded above and away from zero and that its $s \log n$ -sparse eigenvalues are bounded from above and away from zero. Then if $s \log n = o(n/\log p)$, Conditions RE and SE hold.*

Lemma 2 (Plausibility of RE and SE under Many Series Instruments). *Suppose f_i $i = 1, \dots, n$, are i.i.d. bounded zero-mean random vectors with $\|f_i\|_\infty \leq K_B$ a.s. Further suppose that the population Gram matrix $\mathbb{E}[f_i f_i']$ has diagonal entries bounded above and away from zero and that its $s \log n$ -sparse eigenvalues are bounded from above and away from zero. Then if $\sqrt{n}/K_B \rightarrow \infty$ and $s \log n = o((1/K_B)\sqrt{n/\log p})$, Conditions RE and SE hold.*

Recall that a standard assumption in econometric research is to assume that the the population Gram matrix $\mathbb{E}[f_i f_i']$ has eigenvalues bounded from above and below, see e.g. [38]. The

lemmas above allow for this and even much more general behavior, requiring only that the sparse eigenvalues of the population Gram matrix $E[f_i f_i']$ are bounded from below and from above. The latter is important for allowing functions f_i to be formed as a combination of elements from different bases, e.g. a combination of B-splines with polynomials. The lemmas above further show that under some restrictions on the growth of s in relation to the sample size n , the good behavior of the population sparse eigenvalues translates into a good behavior of empirical sparse eigenvalues, which ensures that Conditions RE and SE are satisfied in large samples.

We also impose the following moment conditions on the reduced form errors v_{il} and regressors f_i , where we let $\tilde{d}_{il} := d_{il} - E[d_{il}]$.

Condition RF. (i) *The following growth conditions hold*

$$\log p = o(n^{1/3}) \quad \text{and} \quad s \log p/n \rightarrow 0.$$

(ii) *The moments $E[\tilde{d}_{il}^8]$ and $E[v_{il}^8]$ are bounded uniformly in $1 \leq l \leq k_e$ and in n . (iii) The regressors f_i obey: $\max_{1 \leq j \leq p} \mathbb{E}_n[f_{ij}^8] \lesssim_P 1$ and $\max_{1 \leq i \leq n, 1 \leq j \leq p} |f_{ij}^2| \frac{s \log p}{n} \rightarrow_P 0$. (iv) The moments $E[f_{ij}^2 v_{il}^2]$ are bounded away from zero and from above uniformly in $1 \leq j \leq p$, $1 \leq l \leq k_e$, uniformly in n , and the moments $E[f_{ij}^6 \tilde{d}_{il}^6]$, $E[f_{ij}^6 v_{il}^6]$, $E[|f_{ij}|^3 |v_{il}|^3]$ are bounded, uniformly in $1 \leq j \leq p$, $1 \leq l \leq k_e$, uniformly in n .*

We emphasize that the condition given above is only one possible set of sufficient conditions, which are presented in a manner that reduces the complexity of the exposition. The proofs contain a more refined set of conditions.

The following lemma shows that the population and empirical moment conditions appearing in Condition RF (iii)-(iv) are plausible for both many-instrument and many series-instrument settings. Note that we say that a random variable g_i has uniformly bounded conditional moments of order K if for some positive constants $0 < B_1 < B_2 < \infty$:

$$B_1 \leq E\left[|g_i|^k \middle| x_i\right] \leq B_2 \quad \text{with probability 1, for } k = 1, \dots, K.$$

Lemma 3 (Plausibility of RF(iii)-(iv)). (1) *If the regressors f_i are Gaussian as in Lemma 1, then Condition RF(iii) holds under Condition RF (i) and under $s(\log p)^2/n \rightarrow 0$. (2) If the regressors f_i are arbitrary i.i.d. vectors with bounded entries as in Lemma 2, then Condition RF(iii) holds under Condition RF(i). Suppose that \tilde{d}_{il} and v_{il} have uniformly bounded conditional moments*

of order 6 uniformly in $l = 1, \dots, k_e$, then Condition $RF(iv)$ holds (3) if the regressors f_i are Gaussian or (4) if the regressors f_i are arbitrary i.i.d. vectors with bounded entries.

3.2. Main Results on LASSO and Post-LASSO under Non-Gaussian, Heteroscedastic Errors. We consider LASSO and Post-LASSO estimators defined in equations (2.9) and (2.11) in the system of k_e non-parametric regression equations (2.7) with non-Gaussian and heteroscedastic errors. These results extend the previous results of [8] for LASSO and of [4] for Post-LASSO with classical i.i.d. errors. In addition, we account for the fact that we are simultaneously estimating k_e regressions and account for the dependence of our results on k_e .

Our analysis will first employ the following “ideal” penalty loadings:

$$\hat{\Upsilon}_l^0 = \text{diag}(\hat{\gamma}_{l1}^0, \dots, \hat{\gamma}_{lp}^0), \quad \hat{\gamma}_{lj}^0 = \sqrt{\mathbb{E}_n[f_{ij}^2 v_{il}^2]}, \quad j = 1, \dots, p.$$

We use these penalty loadings to develop basic results and then verify that the results continue to hold for feasible, data-driven penalty loadings.

In the analysis of LASSO, the following quantity, that we refer to as the score,

$$S_l = 2\mathbb{E}_n[(\hat{\Upsilon}_l^0)^{-1} f_i v_{il}],$$

plays a key role. The score represents the noise in the problem. Accordingly, we select the penalty level λ/n to dominate the noise for all k_e regression problems simultaneously, specifically so that

$$\mathbb{P} \left(\lambda \geq cn \max_{1 \leq l \leq k_e} \|S_l\|_\infty \right) \rightarrow 1, \quad (3.22)$$

for some constant $c > 1$. Indeed, using moderate deviation theory for self-normalized sums, we show that any choice of the form

$$\lambda = 2c\sqrt{2n \log(2pk_e)}, \quad (3.23)$$

implements (3.22) as $p \rightarrow \infty$.

The following theorem derives the properties of LASSO. Let us call asymptotically valid any penalty loadings $\hat{\Upsilon}_l$ that obey a.s.

$$\ell \hat{\Upsilon}_l^0 \leq \hat{\Upsilon}_l \leq u \hat{\Upsilon}_l^0, \quad (3.24)$$

with $0 < \ell \leq 1 \leq u$ such that $\ell \rightarrow_P 1$ and $u \rightarrow_P u'$ with $u' \geq 1$.

Theorem 1 (Rates for LASSO under Non-Gaussian and Heteroscedastic Errors). *Suppose that in the regression model (2.7) Conditions RE and RF hold. Suppose the penalty level is specified as in (3.23), and consider any asymptotically valid penalty loadings $\widehat{\Upsilon}$. Then, the LASSO estimator $\widehat{\beta}_l = \widehat{\beta}_{lL}$ and the LASSO fit $\widehat{D}_{il} = f_i' \widehat{\beta}_{lL}$, $l = 1, \dots, k_e$, satisfy*

$$\begin{aligned} \max_{1 \leq l \leq k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n} &\lesssim_P \sqrt{\frac{s \log p}{n}}, \\ \max_{1 \leq l \leq k_e} \|\widehat{\beta}_l - \beta_{l0}\|_2 &\lesssim_P \sqrt{\frac{s \log p}{n}}, \\ \max_{1 \leq l \leq k_e} \|\widehat{\beta}_l - \beta_{l0}\|_1 &\lesssim_P \sqrt{\frac{s^2 \log p}{n}}. \end{aligned}$$

The following theorem derives the properties of Post-LASSO.

Theorem 2 (Rates for Post-LASSO under Non-Gaussian and Heteroscedastic Errors). *Suppose that in the regression model (2.7) Conditions SE and RF hold. Suppose the penalty level for the LASSO estimator is specified as in (3.23), and that LASSO's penalty loadings $\widehat{\Upsilon}$ are asymptotically valid. Then, the Post-LASSO estimator $\widehat{\beta}_l = \widehat{\beta}_{lPL}$ and the Post-LASSO fit $\widehat{D}_{il} = f_i' \widehat{\beta}_{lPL}$, $l = 1, \dots, k_e$, satisfy*

$$\begin{aligned} \max_{1 \leq l \leq k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n} &\lesssim_P \sqrt{\frac{s \log p}{n}}, \\ \max_{1 \leq l \leq k_e} \|\widehat{\beta}_l - \beta_{l0}\|_2 &\lesssim_P \sqrt{\frac{s \log p}{n}}, \\ \max_{1 \leq l \leq k_e} \|\widehat{\beta}_l - \beta_{l0}\|_1 &\lesssim_P \sqrt{\frac{s^2 \log p}{n}}. \end{aligned}$$

Finally, we show that the data-driven penalty loadings that we have proposed in (2.10) are asymptotic valid. We believe that this result is of a major practical interest and has many applications well outside the IV framework of this paper.

Before stating the result, we recall that to obtain the penalty loadings under the refined option, we can use the residuals \widehat{v}_{il} from either LASSO or Post-LASSO computed using the penalty loadings under the basic option. To obtain the penalty loadings under the K -th iteration of the refined option, we can use the residuals \widehat{v}_{il} from either LASSO or Post-LASSO computed using the penalty loadings under the $(K - 1)$ -th iteration of the refined option. The number of iterations K is assumed to be bounded.

Theorem 3 (Asymptotic Validity of the Data-Driven Penalty Loadings). *Under either conditions of Theorem 1 or 2, the penalty loadings $\hat{\Upsilon}$ specified in (2.10), obtained under the basic, the refined, and the K -step refined option based on residuals obtained from LASSO or Post-LASSO are asymptotically valid. (In particular, for the refined options $u' = 1$).*

4. MAIN RESULTS ON THE IV ESTIMATION WITH THE OPTIMAL IV ESTIMATED BY LASSO, POST-LASSO, AND A GENERIC SPARSITY-BASED ESTIMATOR

In this section we present our main inferential results on the instrumental variable estimators.

4.1. Regularity Conditions on the Structural Equation. We shall impose the following moment conditions on the instruments and the structural errors and regressors.

Condition SM. (i) *The disturbance ϵ_i has conditional variance $E[\epsilon_i^2|x_i]$ that is bounded uniformly from above and away from zero, uniformly in n . Given this assumption, without loss of generality, we normalize the instruments so that $E[f_{ij}^2\epsilon_i^2] = 1$ for each $1 \leq j \leq p$ and for all n .* (ii) *$E[\|D_i\|^q]$ and $E[\|d_i\|^q]$ and $E[|\epsilon_i|^{q_\epsilon}]$ are bounded uniformly in n , where $q_\epsilon > 4$ and $q > 4$.* (iii) *The moments $E[\epsilon_i^4\|D_i\|^2]$ and $E[|f_{ij}|^3|\epsilon_i|^3]$ are bounded uniformly in $1 \leq j \leq p$, uniformly in n .* (iv) *The following growth conditions hold:*

$$(a) \frac{s \log p}{n} n^{2/q_\epsilon} \rightarrow 0 \quad \text{and} \quad (b) \frac{s^2(\log p)^2}{n} \rightarrow 0$$

Condition SM(i) requires that structural errors are boundedly heteroscedastic. Given this we make a normalization assumption on the instruments. This entails no loss of generality, since this is equivalent to suitably rescaling the parameter space for coefficients $\beta_{l0}, l = 1, \dots, k_e$, via an isomorphic transformation. Moreover, we only need this normalization to simplify notation in the proofs, and we do not use it in the construction of the estimators. Condition SM(ii) imposes some mild moment assumptions. Condition SM(iv) strengthens the growth requirement $s \log p/n \rightarrow 0$ needed for estimating conditional expectations. However, the restrictiveness of Condition SM(iv)(a) rapidly decreases as the number of bounded moments of the structural error increases. Condition SM(iv)(b) indirectly requires the optimal instruments in Condition AS to be sufficiently smooth, so that the number of unknown series terms s needed to approximate

them well is not too large; this condition ensures that the impact of the instrument estimation on the IV estimator is asymptotically negligible.

The following lemma shows that moment assumptions in Condition SM (iii) are plausible for both many-instrument and many series-instrument settings.

Lemma 4 (Plausibility of SM(iii)). *Suppose that the structural disturbance ϵ_i has uniformly bounded conditional moments of order 4 uniformly in n , then Condition SM(iii) holds, for example, if (1) the regressors f_i are Gaussian as in Lemma 1 or (2) the regressors f_i are arbitrary i.i.d. vectors with bounded entries as in Lemma 2.*

4.2. Main Results on IV Estimators. The first result describes the properties of the IV estimator with the optimal IV constructed using LASSO or Post-LASSO in the setting of the standard model with homoscedastic structural errors. In these settings the estimator achieves the efficiency bound asymptotically. The result also provides a consistent estimator for the asymptotic variance of this estimator.

Theorem 4 (Inference with Optimal IV Estimated by LASSO or Post-LASSO). *Suppose that data (y_i, x_i, d_i) are i.i.d. and obey the linear IV model described in Section 2, and that the structural error ϵ_i is homoscedastic conditional on x_i , that is, $E[\epsilon_i^2|x_i] = \sigma^2$ a.s. Suppose also that Conditions AS, RF, and SM hold. Suppose also that Condition RE holds in the case of using LASSO to construct the estimate of the optimal instrument, and Condition SE holds in the case of using Post-LASSO to construct the estimate of the optimal instrument. Then, the resulting IV estimator $\hat{\alpha}$, based on either LASSO or Post-LASSO estimates of the optimal instrument, is root- n consistent, asymptotically normal, and achieves the efficiency bound:*

$$(\Lambda^*)^{-1/2}\sqrt{n}(\hat{\alpha} - \alpha_0) \rightarrow_d N(0, I),$$

where $\Lambda^* := \sigma^2 Q^{-1}$ for $Q = E[D(x_i)D(x_i)']$, provided that the variance σ^2 is bounded away from zero and from above, uniformly in n , and the eigenvalues of Q are bounded away from zero and from above. Moreover, the result above continues to hold with Λ^* replaced by $\hat{\Lambda}^* := \hat{\sigma}^2 \hat{Q}^{-1}$, where $\hat{Q} = E_n[\hat{D}(x_i)\hat{D}(x_i)']$ and $\hat{\sigma}^2 = E_n[(y_i - d_i'\hat{\alpha})^2]$.

The second result below describes the properties of the IV estimator with the optimal instrument estimated by LASSO or Post-LASSO in the setting of the standard model with heteroscedastic structural errors. In this case, the estimator does not achieve the efficiency bound, but we can expect it to be close to achieving the bound if heteroscedasticity is mild. The result also provides a consistent estimator for the asymptotic variance of this estimator under heteroscedasticity, which allows us to perform inference that is robust to the presence of heteroscedasticity.

Theorem 5 (Robust Inference with IV Constructed by LASSO or Post-LASSO). *Suppose conditions of Theorem 1 hold, except that now the structural errors ϵ_i can be heteroscedastic conditional on x_i . Then the IV estimator $\hat{\alpha}$, based on either LASSO or Post-LASSO estimates of the optimal instrument, is root- n consistent and asymptotically normal:*

$$(Q^{-1}\Omega Q^{-1})^{-1/2}\sqrt{n}(\hat{\alpha} - \alpha_0) \rightarrow_d N(0, I),$$

for $\Omega := \mathbb{E}[\epsilon_i^2 D(x_i)D(x_i)']$ and $Q := \mathbb{E}[D(x_i)D(x_i)']$, provided that the eigenvalues of the latter matrices are bounded away from zero and from above, uniformly in n . Moreover, the result above continues to hold with Ω replaced by $\hat{\Omega} := \mathbb{E}_n[\hat{\epsilon}_i^2 \hat{D}(x_i)\hat{D}(x_i)']$ for $\hat{\epsilon}_i = y_i - d_i'\hat{\alpha}$, and Q replaced by $\hat{Q} := \mathbb{E}_n[\hat{D}(x_i)\hat{D}(x_i)']$.

The final result of this section extends the previous two results to any IV-estimator with a generic sparse estimator of the optimal instruments.

Theorem 6 (Inference with IV Constructed by a Generic Sparsity-Based Procedure). *Suppose that conditions AS, RF, SM hold and suppose now that the fitted values of the optimal instrument, $\hat{D}_{il} = f_i'\hat{\beta}_l$, are constructed using any estimator $\hat{\beta}_l$ such that*

$$\max_{1 \leq l \leq k_e} \|\hat{D}_{il} - D_{il}\|_{2,n} \lesssim_P \sqrt{\frac{s \log p}{n}} \quad (4.25)$$

$$\max_{1 \leq l \leq k_e} \|\hat{\beta}_l - \beta_{l0}\|_1 \lesssim_P \sqrt{\frac{s^2 \log p}{n}}. \quad (4.26)$$

then the conclusions reached in Theorem 5 or Theorem 6 continue to apply in this case.

This result shows that the previous two theorems continue to apply if the first-stage estimator attains the near-oracle performance given in (4.25)-(4.26). Examples of other sparse estimators covered by this theorem are

- Dantzig and Gauss-Dantzig, [14]
- $\sqrt{\text{LASSO}}$ and post- $\sqrt{\text{LASSO}}$, [7] and [6],
- thresholded LASSO and post-thresholded LASSO, [4]
- grouped LASSO and post-grouped LASSO, [29], [34]
- adaptive versions of the above, [29].

Verification of the near-oracle performance (4.25)-(4.26) can be done on a case by case basis using the best current and future conditions in the literature.⁹ Moreover, our results extend to LASSO-type estimators under alternative forms of regularity conditions that fall outside the framework of Conditions RE and Conditions SM, for example, permitting potentially highly correlated regressors. As stated above, all that is required is the near-oracle performance of the kind (4.25)-(4.26).

5. SIMULATION EXPERIMENT

The previous sections' results suggest that using LASSO for fitting first-stage regressions should result in IV estimators with good estimation and inference properties. In this section, we provide simulation regarding these properties in a situation where there are many possible instruments though only a small number are very informative about the endogenous regressor. We also compare the performance of the developed LASSO-based estimators to many-instrument robust estimators that are available in the literature.

Our simulations are based on a simple instrumental variables model:

$$\begin{aligned} y_i &= \beta x_i + e_i \\ x_i &= z_i' \Pi + v_i \end{aligned} \quad (e_i, v_i) \sim N \left(0, \begin{pmatrix} \sigma_e^2 & \sigma_{ev} \\ \sigma_{ev} & \sigma_v^2 \end{pmatrix} \right) \text{ i.i.d.}$$

where $\beta = 1$ is the parameter of interest, and $z_i = (z_{i1}, z_{i2}, \dots, z_{i100})' \sim N(0, \Sigma_Z)$ is a 100 x 1 vector with $E[z_{ih}^2] = \sigma_z^2$ and $\text{Corr}(z_{ih}, z_{ij}) = .5^{|j-h|}$. In all simulations, we set $\sigma_e^2 = 2$ and $\sigma_z^2 = 0.3$.

For the other parameters, we consider various settings. We provide results for sample sizes, n , of 100, 250, and 500; and we consider three different values for $\text{Corr}(e, v)$: 0, .3, and .6.

⁹Note also that the post- ℓ_1 -penalized procedures have only been analyzed for the case of LASSO and $\sqrt{\text{LASSO}}$, [4] and [6], but we expect that similar results carry over to other procedures listed above, namely Dantzig and grouped LASSO.

We also consider three values of σ_v^2 which are chosen to benchmark three different strengths of instruments. The three values of σ_v^2 are found as $\sigma_v^2 = \frac{n\Pi'\Sigma_Z\Pi}{F^*\Pi'\Pi}$ for three different values of F^* : 10, 40, and 160.¹⁰ Finally, we use two different settings for the first stage coefficients, Π . The first sets the first five elements of Π equal to one and the remaining elements equal to zero. We refer to this design as the “cut-off” design. The second model sets the coefficient on $z_{ih} = .7^{h-1}$ for $h = 1, \dots, 100$. We refer to this design as the “exponential” design. In the cut-off case, the first-stage has an exact sparse representation, while in the exponential design, the model is not literally sparse although the majority of explanatory power is contained in the first few instruments.

For each setting of the simulation parameter values, we report results from five different estimation procedures. A simple possibility when presented with many instrumental variables is to just estimate the model using 2SLS and all of the available instruments. It is well-known that this will result in poor-finite sample properties unless there are many more observations than instruments; see, for example, [3]. The limited information maximum likelihood estimator (LIML) and its modification by [25] (FULL)¹¹ are both robust to many instruments as long as the presence of many instruments is accounted for when constructing standard errors for the estimators; see [3] and [27] for example. We report results for these estimators in rows labeled 2SLS(100), LIML(100), and FULL(100) respectively.¹² In addition, we report estimates based on the LASSO IV estimator (LASSO) and the Post-LASSO IV estimator (Post-LASSO) using the refined data-dependent penalty loadings given in (2.10).¹³ For each estimator, we report root-mean-squared-error (RMSE), median bias (Med. Bias), mean absolute deviation (MAD),

¹⁰These values were chosen by roughly benchmarking to the first-stage F-statistics calculated using clustered standard errors and the instruments used by [20] from the empirical example in Section 6. Using these instruments yields F’s between 29 and 119. We took 40 as a rough intermediate value and then multiplied and divided by four to obtain weaker and stronger identification.

¹¹[25] requires a user-specified parameter. We set this parameter equal to one which produces a higher-order unbiased estimator. See [26] for additional discussion.

¹²With $n = 100$, we randomly select 99 instruments for use in FULL(100) and LIML(100).

¹³Specifically, we used the initial option to construct loadings and then found a corresponding value of λ for which LASSO selected only one instrument. We used this instrument to construct an initial set of residuals for use in defining the refined penalty loadings and reestimated LASSO using these loadings and the value of λ in (2.10). We then computed another set of residuals and used these residuals to recompute the loadings. We report results based on this final LASSO-step.

and rejection frequencies for 5% level tests (rp(.05)). For computing rejection frequencies, we estimate conventional 2SLS standard errors for 2SLS(100), LASSO, and Post-LASSO, and the many instrument robust standard errors of [27] for LIML(100) and FULL(100).

Simulation results are presented in Tables 1-6. Tables 1-3 give results for the cut-off design with $n = 100$, $n = 250$, and $n = 500$ respectively; and Tables 4-6 give results for the exponential design with $n = 100$, $n = 250$, and $n = 500$ respectively. As expected, 2SLS(100) does extremely poorly along all dimensions except in the case with no correlation between e and v . In this case, there is no endogeneity bias and OLS would be the optimal estimator. 2SLS(100) performs well here since the many instrument bias moves 2SLS toward the OLS estimator and is thus favorable to the performance of the estimator; see, e.g. [3]. Of course, this bias works against 2SLS(100) once one moves away from the exogenous case, and we see that the performance of 2SLS(100) rapidly declines in the strength of the correlation of the errors. It is also not surprising that the performance of LASSO and post-LASSO are very similar. Given this, we refer only to the LASSO estimator for simplicity but the discussion is essentially unchanged replacing LASSO with post-LASSO.

There is interesting variation in the performance of the other estimators across the simulation parameters. LASSO performs very well in terms of RMSE and MAD, having weakly smaller RMSE and MAD than any of the other considered estimators in each design with $\text{Corr}(e, v) \neq 0$, and beating all estimators but 2SLS(100) when the two error terms are independent. LIML(100) and FULL(100) are only competitive based on these metrics when the instruments are strong ($F^* = 160$) and the sample is large ($n = 500$). This result can intuitively be associated to the “amount of information” available in the instruments relative to the sample size. With weak instruments, there is relatively little useful information available for forming the first stage prediction, and LASSO selects a relatively strong instrument and, by keeping the number of instruments small, keeps bias small as well.¹⁴ LIML(100) and FULL(100) use all of the

¹⁴It is interesting to note that LASSO with the advocated penalty choices may be overly conservative with weak instruments as it selects no instruments in many cases. For reporting the simulation results, we calculate RMSE, Med. Bias, and MAD conditional on the number of simulation replications in which LASSO selects a non-empty set of instruments. For rejection frequencies, we consider the cases where LASSO does not pick instruments as failures to reject the null hypothesis. The exact numbers of cases for each design where LASSO selected no instruments may be found in Tables 1-6.

instruments, the majority of which contribute little if anything other than noise, and so do not perform as well as LASSO. On the other hand, there is a substantial “amount of information” contained in several of the instruments when the first-stage relationship is stronger. LIML(100) and FULL(100) use this information as well as the noise in the remaining less informative instruments. When the sample is small, the noise from the additional instruments results in poor performance of LIML(100) and FULL(100) though it has relatively small impact when the sample is larger. LASSO, on the other hand, adapts to the “amount of information” available for forming the first-stage prediction and outperforms the less adaptive procedures across the designs considered.

Considering next median bias, we see that LIML(100) and FULL(100) perform relatively well with $n = 250$ or $n = 500$ as theory predicts; see, for example, [26]. However, both estimators have substantial bias with $n = 100$ in which case there are as many potential instruments as observations and the many instrument theory provides a poor guide to finite sample performance. In most cases, the bias of the LASSO IV estimator is quite small, though the bias does increase more rapidly than that of LIML(100) or FULL(100) as the strength of the correlation between e and v increases.

Finally, we see that LASSO does quite well in terms of rejection frequencies of 5% level tests. As with median bias, there is no procedure that is uniformly dominant based on this metric, though 2SLS(100) performs very poorly except when there is no endogeneity. Using this metric, LASSO is very competitive with inference based on FULL(100) or LIML(100) with many-instrument-robust standard errors. Sizes of LASSO based tests are uniformly better in the cut-off design and are better in the exponential design except for some cases with $\text{corr}(e, v) = .6$. Whether one would be willing to trade this modest deterioration in testing performance using LASSO versus FULL(100) or LIML(100) in these scenarios for the sizeable improvements in RMSE and MAD of the estimator will of course depend on the preferences of the researcher.¹⁵

Overall, the evidence from the simulations is quite favorable to LASSO-based IV methods. The LASSO IV estimator dominates the other estimators considered based on RMSE or MAD

¹⁵Another possibility that we do not consider is to use LASSO coupled with a many-instrument robust procedure such as LIML or FULL or a weak-instrument robust procedure as in [31], for example. We are investigating these possibilities as well as other approaches to instrument selection, e.g. based on [24], in the on-going research.

and dominates on all metrics considered when $p = n$. It also produces an estimator with relatively small finite sample bias, though the higher-order unbiased FULL(100) estimator and the approximately median-unbiased LIML(100) do slightly better based on this metric in the cases we consider with $p < n$. Finally, the LASSO-based procedures produce tests with size close to the nominal level that are not dominated by FULL(100) and LIML(100) with many-instrument robust standard errors. While the results do not show that the LASSO-based procedures uniformly dominate the many-instrument robust procedures in situations with $p < n$, they do show that the simple LASSO-based procedures are highly competitive, having uniformly lower risk (as measured by RMSE and MAD) and not being dominated in terms of testing performance.

6. THE IMPACT OF EMINENT DOMAIN ON ECONOMIC OUTCOMES

As an example of the potential application of LASSO to select instruments, we consider IV estimation of the effects of federal appellate court decisions regarding eminent domain on a variety of economic outcomes. The study of the economic consequences of eminent domain, when the government “takes” the property rights of one or more individuals, either physically or through regulation, is important for a variety of reasons. Takings are often justified based on “public use” arguments such as removing economic blight and/or promoting economic development through private commercial development. People worried about eminent domain, on the other hand, express two major concerns. First, wealth from groups with little political power could be redistributed to those with more political power. Second, eminent domain may induce distortions in the efficient investment of capital. Scholars are of two minds regarding the likely distortions to investment. Uncertainty in whether the government will appropriate an investment could lead to underinvestment; see [40]. On the other hand, overinvestment could stem from individuals trying to deter the government from exercising eminent domain or from individuals anticipating “just compensation” at full value since full compensation provides full-coverage insurance to the private investor against the risk of taking with no premium paid; see [9], [36], [30], and [44].

To be concrete about how judicial decisions might impact economic outcomes, consider the following results of actual eminent domain decisions. 1) The government must pay monetary compensation for an unlawful taking, even a temporary one. Simply removing the regulation that justified the taking is not sufficient redress for having property unlawfully taken. 2) In imposing conditions that must be met by a private developer in order for a development project

(such as a commercial shopping center) to be approved, the government must show that there is a public burden imposed by the development that is closely linked to the conditions imposed. 3) The magnitude of a condition imposed on a development project (such as a fee or a mandated setting aside of part of the development for public use) must be “roughly proportional” to the size of the public impact that the conditions are intended to mitigate. Each of these judicial decisions expands the ability of private property owners to seek compensation from the government for regulations and for conditions imposed on them as a requirement for approval for development projects. Thus, these decisions could be expected to make it more difficult for the government to exercise eminent domain. To the extent that the threat of the government’s exercise of eminent domain influences economic decisions, these judicial decisions should have impacts on economic quantities such as housing prices or other measures of investment.

Despite the amount of theoretical and doctrinal writings that speculate about the effects of takings law on economic outcomes, little empirical evidence exists. One notable recent exception is [20] which provides a careful empirical analysis of the effect of appellate court decisions regarding takings law on economic outcomes. Understanding five aspects of the US legal system are important for the development and understanding of [20] as well as our results below. First, the US has a common law system where American judges not only apply the law but also make the law. This effective making of law occurs since a judge’s decisions in current cases become precedent for use in decisions in future cases. Second, there are three layers of courts in the US judicial system. District courts are the general trial courts. When cases are appealed, they go to appellate courts, which typically decide issues of new law or look to see if the district court was in error. A small portion of these cases are appealed again to the Supreme Court. Therefore appellate courts are quite active in shaping law. They handle the vast majority of cases deciding issues of new law and provide new interpretations or distinctions of pre-existing precedents or statutes. Third, there are 12 appellate (Circuit) courts, each in charge of a geographic region of the US. Appellate decisions are only binding precedent in the Circuit, meaning that the district courts with a Circuit and the Circuit court itself must follow the precedent set by prior appellate decisions with the Circuit. Note that appellate decisions in one circuit do not establish precedent that other circuits must follow. Fourth, judges are randomly assigned to cases in appellate courts. Some judges take a reduced caseload but all are randomly assigned by a computer and are typically not revealed to the litigating parties until after they file their briefs and sometimes

only a few days before the hearing, if there is a hearing. Fifth, appellate courts assign three judges to a case and a Circuit can have twenty to forty judges in the pool of judges available to be assigned. Therefore the number of possible combinations of judges or combinations of judicial demographic characteristics on a panel is very large. Studies have shown that demographic characteristics of a judge relate to the decisions of the judge and decisions of the panel. These two facts allow us to construct a natural experiment using random variation in the establishment of precedent across different regions of the US. The results we provide in this section complement the analysis of [20] by taking their estimation strategy and augmenting it with the use of LASSO to choose instruments.

To try to uncover the relationship between takings law and economic outcomes, we estimate structural models of the form

$$y_{ict} = \alpha_c + \alpha_t + \gamma_{ct} + \beta \text{ Takings Law}_{ct} + W'_{ct} \delta + \epsilon_{ict} \quad (6.27)$$

where y_{ict} is an economic outcome for area i in circuit c at time t , Takings Law_{ct} represents the number of pro-plaintiff appellate takings decisions in circuit c and year t ; W_{ct} are judicial pool characteristics,¹⁶ a dummy for whether there were no cases in that circuit-year, and the number of takings appellate decisions; and α_c , α_t , and γ_{ct} are respectively circuit-specific effects, time-specific effects, and circuit-specific time trends. An appellate court decision is coded as pro-plaintiff if the court ruled that a taking was unlawful, thus overturning the government's seizure of the property in favor of the private owner. We construe pro-plaintiff decisions to indicate a regime that is more protective of individual property rights. The parameter of interest, β , thus represents the effect of an additional decision upholding individual property rights on an economic outcome.

We provide results using four different economic outcomes as dependent variables: the log of three home-price-indices and log(GDP). The three different home-price-indices we consider are the quarterly, weighted, repeat-sales FHFA/OFHEO house price index that tracks single-family house prices at the state level for metro (FHFA) and non-metro (Non-Metro) areas and the Case-Shiller home price index (Case-Shiller) by month for 20 metropolitan areas based on

¹⁶The judicial pool characteristics are the probability of a panel being assigned with the characteristics used to construct the instruments. There are 30, 33, 32, and 30 for FHFA house prices, non-metro house prices, Case-Shiller house prices, and GDP respectively.

repeat-sales residential housing prices. The total sample sizes are 5304, 1920, and 4320 for FHFA, Non-Metro, and Case-Shiller respectively. We also use state level GDP from the Bureau of Economic Analysis to form $\log(\text{GDP})$. The GDP regressions are based on 1326 observations.

The analysis of the effects of takings law is complicated by the possible endogeneity between governmental takings and takings law decisions and economic variables. For example, low property values may make it cheaper for the government to exercise eminent domain and seize property, while high property values may reveal the viability of a redevelopment or commercial project. Either of these channels may encourage judges to see the public use of a project and decide in favor of the government's taking a property, leading to over- or under-estimates of the unconfounded effect of takings decisions. Endogeneity may also be due to unobserved factors such as decisions in other areas of law that affect economic outcomes and also influence judicial decisions related to takings. [20] provide additional discussion of potential sources of endogeneity that motivate the use of an instrumental variables strategy.

To address the potential endogeneity of takings law, we employ the instrumental variables strategy based on the identification argument of [19] and [20] that relies on the random assignment of judges to appellate panels that decide federal appellate cases. Since judges are randomly assigned to three judge panels to decide appellate cases, the exact identity of the judges and, more importantly, their demographics are randomly assigned conditional on the distribution of characteristics of federal circuit court judges in a given circuit-year. Thus, once the distribution of characteristics is controlled for, the realized characteristics of the randomly assigned three judge panel should be unrelated to other factors besides judicial decisions that may be related to economic outcomes. There is also substantial evidence that the demographic characteristics of judges are related to their judicial decision making.¹⁷ Thus, the characteristics of judges on panels deciding eminent domain cases should provide valid instruments for learning about the effect of appellate court decisions regarding takings law on economic outcomes.

As noted above, there are many potential characteristics of three judge panels that may be used as instruments. While the basic identification argument suggests any set of characteristics of the three judge panel will be uncorrelated with the structural unobservable conditional on the set of controls in the model, there will clearly be some instruments which are more worthwhile

¹⁷See, e.g., [10], [17], [23], [19], and [20].

than others in obtaining precise second-stage estimates. For simplicity, we consider only the following demographics: gender, race, religion, political affiliation, whether the judge’s bachelor was obtained in-state, whether the bachelor is from a public university, whether the JD was obtained from a public university, and whether the judge was elevated from a district court along with various interactions. In total, we have 138, 143, 147, and 138 non-redundant potential instruments for FHFA prices, non-metro prices, Case-Shiller, and GDP respectively that we select among using LASSO. The exact description of the instrument set is available upon request.

Table 7 contains estimation results for β . We report OLS estimates and results based on two different sets of instruments. The first set of instruments, used in the rows labeled 2SLS, are the instruments adopted in [20]. We consider this the baseline. [20] used two variables, whether a panel was assigned an appointee who did not report a public religious affiliation and whether a panel was assigned an appointee who earned their first law degree from a public university, as instruments. The choice of these instruments is motivated on intuitive grounds. [20] note that judges who are strongly affiliated with a religious group are more likely to vote anti-government (pro-plaintiff) as many religious groups are populated by people who believe in smaller government and more private agency. They also argue that judges who attended public institutions to obtain their law degrees are more likely to hold populist positions that would also lead them to vote pro-plaintiff. Regardless of the intuitive justification, [20] find that these two variables do predict the number of pro-plaintiff decisions in first-stage regressions. The second set of instruments are those selected through LASSO using the refined data-driven penalty.¹⁸

¹⁸For GDP and FHFA, LASSO selects the number of panels with at least one appointee whose law degree is from a public university squared and the interaction of the number of panels with at least one appointee whose law degree is from a public university and the number of panels with at least one member who reports belonging to a mainline protestant religion. In the Case-Shiller data, LASSO selects these same two variables plus the number of panels with three members with a bachelor’s from a public university and the interaction of the number of panels with at least one democrat with the number of panels with at least one appointee whose law degree is from a public university. For non-metro prices, the selected instruments are the interaction of the number of panels with at least one democrat with the number of panels with at least one Jewish member, the number of panels with at least one appointee whose law degree is from a public university with the number of panels with at least one member who reports belonging to a mainline protestant religion, the interaction of the number of panels with at least one member whose bachelor is from within state with the number of panels with at least one non-white panel member, and the number of panels with at least one evangelical member with the number of panels with at least one member who reports belonging to a mainline protestant religion.

The number of instruments selected by LASSO is reported in the row “S”. In all cases, estimated standard errors are clustered at the circuit-year level. We use the Post-LASSO 2SLS estimator and report these results in the rows labeled “Post-LASSO”.

There are a number of key patterns that emerge when looking at the results in Table 7. Consistent with the view that any of the potential instruments are valid, we see that the point estimates produced using either set of instruments are broadly consistent with each other.¹⁹ For the FHFA-price-index and GDP, the point estimates are negative and not significantly different from zero. The estimated effects for non-metro prices and the Case-Shiller index are small, positive, and significantly different from zero suggesting that a pro-plaintiff decision in a circuit-year is associated with a small increase in these indices. Note that the average number of pro-plaintiff decisions in a circuit-year is around .15 in the sample. As a rough benchmark, this number and the Post-LASSO estimates for the Case-Shiller index, for example, suggest that pro-plaintiff decisions lead to an increase of somewhere between 0.2-1.0% in the Case-Shiller index in an “average” circuit-year.

The most interesting results are found by comparing first-stage F-statistics and estimated standard errors across the instrument sets. Interestingly, LASSO selects the same number of instruments, two, as [20] for two of the dependent variables considered and selects four instruments for the other. However, the LASSO instruments are much better first-stage predictors as measured by the first-stage F-statistic. The F-statistics increase for FHFA prices, non-metro prices, and GDP, and the associated p-values for testing the the coefficients on all of the selected instruments decrease for each dependent variable relative to the benchmark of the [20] instruments. This improved first-stage prediction is, unsurprisingly, associated with the resulting 2SLS estimator having smaller estimated standard errors than the benchmark case for each dependent variable. In no case would the decrease in standard errors lead one to draw different conclusions about testing the null that $\beta = 0$ at the 5% level, but the smaller standard errors obviously allow one to draw tighter conclusions about the range of plausible values for β for each dependent variable.

¹⁹The estimated standard error of the difference between $\hat{\beta}_{2SLS}$ and $\hat{\beta}_{Post-LASSO}$ is .0282, .0198, .0096, and .0170 for FHFA, Non-Metro, Case-Shiller, and GDP respectively.

In summary, we find evidence that the effect of takings law decisions on contemporaneous two of the property price indices is small but positive while there is little evidence of any appreciable effect on GDP. The results are consistent with the developed asymptotic theory in that the 2SLS point-estimates based on each set of instruments are similar while the Post-LASSO estimates are more precise. In this example, we see that the potential gains to first-stage prediction by using LASSO variable selection can be large. In each specification considered, the first-stage predictions are more precise, as measured by the first-stage F-statistic, with the variables selected with LASSO than the baseline instruments of [20]. The LASSO-based IV estimator also produces standard errors that are noticeably smaller than the benchmark in all specifications. Overall, the findings suggest that there is the potential for LASSO to be fruitfully employed to choose instruments in economic applications.

7. CONCLUSION

In this paper, we have considered the use of LASSO and Post-LASSO methods for forming first-stage predictions in a linear instrumental variables model with potentially many instruments. We note that two leading cases where this might arise are when a researcher has a small set of many-valued, possibly continuous, instruments and wishes to nonparametrically estimate the optimal instrument or when the set of potential basic instruments itself is large. We rigorously develop the theory for the resulting IV estimator and provide conditions under which the LASSO predictions approximate the optimal instruments. We also contribute to the LASSO literature by providing results for LASSO model selection allowing for non-Gaussian, heteroscedastic disturbances. This generalization is very important for applied economic analysis where researchers routinely have prior beliefs that heteroscedasticity is present and important and desire to use procedures that are robust to departures from the simple homoscedastic-Gaussian case.

We also consider the practical properties of the proposed procedures through simulation examples and an empirical application. In the simulations, we see that feasible LASSO procedures that use a data-dependent penalty perform very well across the range of simulation designs we consider. The LASSO-based IV performs as well as or better than recently advocated many-instrument robust procedures in the majority of designs and clearly dominates in a scenario with $p = n$. This performance suggests that it may be useful to use LASSO-based instrument

selection in conjunction with the many instrument robust procedures, and exploring this may be an interesting avenue for future research.

In the empirical example, we look at the effect of judicial decisions at the federal circuit court level regarding the government's exercise of eminent domain on house prices and state-level GDP as in [20]. We use the random assignment of judges to three judge panels who decide the outcomes of the case to justify using the demographic characteristics of the judges on the realized panels as instruments for their decision. This strategy produces a situation in which there are many potential instruments in that all possible sets of characteristics of the three judge panel are valid instruments. The results of the analysis suggest that judicial decisions positively affect contemporaneous house prices but have small, if any, impact on contemporaneous GDP. Relative to a baseline obtained by using the instruments of [20], we see that the LASSO-based results using the data-dependent penalty substantially reduce estimated standard errors and consequently allow one to draw more precise conclusions about the effects of the judicial decisions. Overall, the simulation and empirical example clearly demonstrate the potential benefits from using LASSO in conjunction with instrumental variables models, and we conjecture that this potential gain will also be realized for other sensible dimension reduction techniques.

APPENDIX A. TOOLS: MODERATE DEVIATIONS FOR SELF-NORMALIZED SUMS

We shall be using the following result – Theorem 7.4 in [21].

Let X_1, \dots, X_n be independent, mean-zero variables, and

$$S_n = \sum_{i=1}^n X_i, \quad V_n^2 = \sum_{i=1}^n X_i^2.$$

For $0 < \delta \leq 1$ set

$$B_n^2 = \sum_{i=1}^n \mathbb{E}X_i^2, \quad L_{n,\delta} = \sum_{i=1}^n \mathbb{E}|X_i|^{2+\delta}, \quad d_{n,\delta} = B_n/L_{n,\delta}^{1/(2+\delta)}.$$

Then for uniformly in $0 \leq x \leq d_{n,\delta}$,

$$\frac{\mathbb{P}(S_n/V_n \geq x)}{\bar{\Phi}(x)} = 1 + O(1) \left(\frac{1+x}{d_{n,\delta}} \right)^{2+\delta},$$

$$\frac{\mathbb{P}(S_n/V_n \leq -x)}{\Phi(-x)} = 1 + O(1) \left(\frac{1+x}{d_{n,\delta}} \right)^{2+\delta},$$

where the terms $O(1)$ are bounded in absolute value by a universal constant A , and $\bar{\Phi} := 1 - \Phi$.

Application of this result gives the following lemma:

Lemma 5 (Moderate Deviations for Self-Normalized Sums). *Let $X_{1,n}, \dots, X_{n,n}$ be the triangular array of i.i.d, zero-mean random variables. Suppose that*

$$M_n = \frac{(\mathbb{E}X_{1,n}^2)^{1/2}}{(\mathbb{E}|X_{1,n}|^3)^{1/3}} > 0$$

and that for some $\ell_n \rightarrow \infty$

$$n^{1/6}M_n/\ell_n \geq 1.$$

Then uniformly on $0 \leq x \leq n^{1/6}M_n/\ell_n - 1$, the quantities

$$S_{n,n} = \sum_{i=1}^n X_{i,n}, \quad V_{n,n}^2 = \sum_{i=1}^n X_{i,n}^2.$$

obey

$$\left| \frac{\mathbb{P}(|S_{n,n}/V_{n,n}| \geq x)}{2\Phi(x)} - 1 \right| \leq \frac{A}{\ell_n^3} \rightarrow 0.$$

Proof. This follows by the application of the quoted theorem to the i.i.d. case with $\delta = 1$ and $d_{n,1} = n^{1/6}M_n$. The calculated error bound follows from the triangular inequalities and conditions on ℓ_n and M_n . \square

APPENDIX B. PROOF OF THEOREM 1

The proof of Theorem 1 has four steps. The most important steps are the Steps 1-3. One half of Step 1 for bounding $\|\cdot\|_{2,n}$ -rate follows the strategy of [8], but accommodates data-driven penalty loadings. Another half of Step 1 for bounding the $\|\cdot\|_1$ -rate is completely new for the non-parametric case and does not follow any prior reference. Steps 2 and 3 innovatively use the moderate deviation theory for self-normalized sums which allows us to obtain sharp results for non-Gaussian and heteroscedastic errors as well as handle data-driven penalty loadings. These steps also do not follow any prior reference. Step 4 puts the results together to make conclusions.

Step 1. For $C > 0$ and each $l = 1, \dots, k_e$, consider the following weighted restricted eigenvalue

$$\kappa_C^l = \min_{\delta \in \mathbb{R}^p: \|\widehat{\Upsilon}_l^0 \delta_{T_l^c}\|_1 \leq C \|\widehat{\Upsilon}_l^0 \delta_{T_l}\|_1, \|\delta\| \neq 0} \frac{\sqrt{s} \|f'_l \delta\|_{2,n}}{\|\widehat{\Upsilon}_l^0 \delta_{T_l}\|_1}.$$

This quantity controls the modulus of continuity between the prediction norm and the l_1 -norm within a restricted region that depends on $l = 1, \dots, k_e$. Note that if

$$a \leq \min_{1 \leq l \leq k_e} \min_{1 \leq j \leq p} \hat{\Upsilon}_{lj}^0 \leq \max_{1 \leq l \leq k_e} \|\hat{\Upsilon}_l^0\|_\infty \leq b,$$

for every $C > 0$ we have

$$\min_{1 \leq l \leq k_e} \kappa_C^l \geq (1/b) \kappa_{(bC/a)}(\mathbb{E}_n[f_i f_i'])$$

where the latter is the restricted eigenvalue defined in (3.20). By Condition RF and by Step 3 of Appendix B below, we have a bounded away from zero and b bounded from above with probability approaching one as n increases. Therefore, $bC/a \lesssim_P C$, $b \lesssim_P 1$, and, by Condition RE, we have that $(1/b) \kappa_{(bC/a)}(\mathbb{E}_n[f_i f_i'])$ is bounded away from zero with probability approaching 1. Therefore, κ_C^l is also bounded away from zero with probability approaching 1.

The main result of this step is the following lemma:

Lemma 6. *If $\lambda/n \geq c \|S_l\|_\infty$, and $\hat{\Upsilon}_l$ satisfy (3.24) with $u \geq 1 \geq \ell > 1/c$ then*

$$\begin{aligned} \|f_i'(\hat{\beta}_l - \beta_{l0})\|_{2,n} &\leq \left(u + \frac{1}{c}\right) \frac{\lambda \sqrt{s}}{n \kappa_{c_0}^l} + 2c_s, \\ \|\hat{\Upsilon}_l^0(\hat{\beta}_l - \beta_{l0})\|_1 &\leq 3c_0 \frac{\sqrt{s}}{\kappa_{2c_0}^l} \left(\left(u + [1/c]\right) \frac{\lambda \sqrt{s}}{n \kappa_{c_0}^l} + 2c_s \right) + \frac{3c_0 n}{\lambda} c_s^2, \end{aligned}$$

where $c_0 = (uc + 1)/(\ell c - 1)$.

Proof of Lemma 6. Let $\delta_l := \hat{\beta}_l - \beta_{l0}$. By optimality of $\hat{\beta}_l$ we have

$$\hat{Q}_l(\hat{\beta}_l) - \hat{Q}_l(\beta_{l0}) \leq \frac{\lambda}{n} \left(\|\hat{\Upsilon}_l \beta_{l0}\|_1 - \|\hat{\Upsilon}_l \hat{\beta}_l\|_1 \right),$$

and we also have

$$\left| \hat{Q}_l(\hat{\beta}_l) - \hat{Q}_l(\beta_{l0}) - \|f_i' \delta_l\|_{2,n}^2 \right| \leq \|S_l\|_\infty \|\hat{\Upsilon}_l^0 \delta_l\|_1 + 2c_s \|f_i' \delta_l\|_{2,n} \quad (\text{B.28})$$

so that from $\lambda \geq cn \|S_l\|_\infty$ and the conditions imposed on $\hat{\Upsilon}_l$ in the statement of the theorem,

$$\begin{aligned} \|f_i' \delta_l\|_{2,n}^2 &\leq \frac{\lambda}{n} \left(\|\hat{\Upsilon}_l \delta_{lT_l}\|_1 - \|\hat{\Upsilon}_l \delta_{lT_l^c}\|_1 \right) + \|S_l\|_\infty \|\hat{\Upsilon}_l^0 \delta_l\|_1 + 2c_s \|f_i' \delta_l\|_{2,n} \\ &\leq \left(u + \frac{1}{c}\right) \frac{\lambda}{n} \|\hat{\Upsilon}_l^0 \delta_{lT_l}\|_1 - \left(\ell - \frac{1}{c}\right) \frac{\lambda}{n} \|\hat{\Upsilon}_l^0 \delta_{lT_l^c}\|_1 + 2c_s \|f_i' \delta_l\|_{2,n}. \end{aligned} \quad (\text{B.29})$$

To show the first statement of the Lemma we can assume $\|f'_i \delta_l\|_{2,n} \geq 2c_s$, otherwise we are done. This condition together with relation (B.29) implies that for $c_0 = (uc + 1)/(\ell c - 1)$ we have

$$\|\widehat{\Upsilon}_l^0 \delta_{lT_l^c}\|_1 \leq c_0 \|\widehat{\Upsilon}_l^0 \delta_{lT_l}\|_1.$$

Therefore, by definition of $\kappa_{c_0}^l$, we have

$$\|\widehat{\Upsilon}_l^0 \delta_{lT_l}\|_1 \leq \sqrt{s} \|f'_i \delta_l\|_{2,n} / \kappa_{c_0}^l.$$

Thus, relation (B.29) implies

$$\|f'_i \delta_l\|_{2,n}^2 \leq \left(u + \frac{1}{c}\right) \frac{\lambda \sqrt{s}}{n \kappa_{c_0}^l} \|f'_i \delta_l\|_{2,n} + 2c_s \|f'_i \delta_l\|_{2,n}$$

and the result follows.

To establish the second statement of the Lemma, we consider two cases. First, assume

$$\|\widehat{\Upsilon}_l^0 \delta_{lT_l^c}\|_1 \leq 2c_0 \|\widehat{\Upsilon}_l^0 \delta_{lT_l}\|_1.$$

In this case, by definition of $\kappa_{2c_0}^l$, we have

$$\|\widehat{\Upsilon}_l^0 \delta_l\|_1 \leq (1 + 2c_0) \|\widehat{\Upsilon}_l^0 \delta_{lT}\|_1 \leq (1 + 2c_0) \sqrt{s} \|f'_i \delta_l\|_{2,n} / \kappa_{2c_0}^l$$

and the result follows by applying the first bound to $\|f'_i \delta_l\|_{2,n}$.

On the other hand, consider the case that

$$\|\widehat{\Upsilon}_l^0 \delta_{lT_l^c}\|_1 > 2c_0 \|\widehat{\Upsilon}_l^0 \delta_{lT_l}\|_1 \tag{B.30}$$

which would already imply $\|f'_i \delta_l\|_{2,n} \leq 2c_s$ by (B.29). Moreover,

$$\begin{aligned} \|\widehat{\Upsilon}_l^0 \delta_{lT_l^c}\|_1 &\stackrel{(1)}{\leq} c_0 \|\widehat{\Upsilon}_l^0 \delta_{lT_l}\|_1 + \frac{c}{\ell c - 1} \frac{n}{\lambda} \|f'_i \delta_l\|_{2,n} (2c_s - \|f'_i \delta_l\|_{2,n}) \\ &\stackrel{(2)}{\leq} c_0 \|\widehat{\Upsilon}_l^0 \delta_{lT_l}\|_1 + \frac{c}{\ell c - 1} \frac{n}{\lambda} c_s^2 \\ &\stackrel{(3)}{\leq} \frac{1}{2} \|\widehat{\Upsilon}_l^0 \delta_{lT_l^c}\|_1 + \frac{c}{\ell c - 1} \frac{n}{\lambda} c_s^2, \end{aligned}$$

where (1) holds by (B.29), (2) holds since $\|f'_i \delta_l\|_{2,n} (2c_s - \|f'_i \delta_l\|_{2,n}) \leq \max_{x \geq 0} x(2c_s - x) \leq c_s^2$, and (3) follows from (B.30).

Thus,

$$\|\widehat{\Upsilon}_l^0 \delta_l\|_1 \leq \left(1 + \frac{1}{2c_0}\right) \|\widehat{\Upsilon}_l^0 \delta_{lT_l^c}\|_1 \leq \left(1 + \frac{1}{2c_0}\right) \frac{2c}{\ell c - 1} \frac{n}{\lambda} c_s^2,$$

and the result follows from noting that $c/(\ell c - 1) \leq c_0/u \leq c_0$ and $1 + 1/2c_0 \leq 3/2$. \square

Step 2. In this step we prove a lemma about the quantiles of the maximum of the scores

$$S_l = 2\mathbb{E}_n[(\hat{\Upsilon}_l^0)^{-1}f_i v_{il}],$$

and use it to pin down the level of the penalty.

Lemma 7. *For $\lambda = 2c\sqrt{2n\log(2pk_e)}$, we have that as $n \rightarrow \infty$ and $p \rightarrow \infty$*

$$\mathbb{P}\left(c \max_{1 \leq l \leq k_e} n \|S_l\|_\infty > \lambda\right) = o(1),$$

provided that for some $b_n \rightarrow \infty$ slowly

$$2\log(2pk_e) \leq \frac{n^{1/3}}{b_n} \min_{1 \leq j \leq p, 1 \leq l \leq k_e} M_{jl}^2, \quad M_{jl} := \frac{\mathbb{E}[f_{ij}^2 v_{il}^2]^{1/2}}{\mathbb{E}[|f_{ij}|^3 |v_{il}|^3]^{1/3}}.$$

Note that the last condition is satisfied under our conditions for large n for some $b_n \rightarrow \infty$, since k_e is fixed, $\log p = o(n^{1/3})$, and $\min_{1 \leq j \leq p, 1 \leq l \leq k_e} M_{jl}^2$ is bounded away from zero.

Proof of Lemma 7. The lemma follows from the following bound: as $n \rightarrow \infty$

$$\mathbb{P}\left(\max_{1 \leq l \leq k_e} \sqrt{n} \|S_l\|_\infty \geq 2\sqrt{2\log(2pk_e/a)}\right) \leq \frac{a(1+o(1))}{\sqrt{2\log(2pk_e/a)}}, \quad (\text{B.31})$$

uniformly for all $0 < a \leq 1$ and p and k_e such that

$$2\log(2pk_e/a) \leq \frac{n^{1/3}}{b_n} \min_{1 \leq j \leq p, 1 \leq l \leq k_e} M_{jl}^2.$$

To prove the bound, note that

$$\begin{aligned} & \mathbb{P}\left(\max_{1 \leq l \leq k_e} \sqrt{n} \|S_l\|_\infty \geq 2\sqrt{2\log(2pk_e/a)}\right) \\ & \leq_{(1)} pk_e \max_{1 \leq j \leq p, 1 \leq l \leq k_e} \mathbb{P}\left(\frac{|\mathbb{G}_n(f_{ij} v_{il})|}{\sqrt{\mathbb{E}_n[f_{ij}^2 v_{il}^2]}} > \sqrt{2\log(2pk_e/a)}\right) \\ & \leq_{(2)} pk_e 2\bar{\Phi}(\sqrt{2\log(2pk_e/a)})(1+o(1)) \\ & \leq_{(3)} \frac{a(1+o(1))}{\sqrt{2\log(2pk_e/a)}} = o(1), \end{aligned}$$

uniformly over the region specified above as $p \rightarrow \infty$. The bound (1) follows by the union bound; (2) follows by the moderate deviation theory for self-normalized sums, specifically Lemma 5; and (3) by $\bar{\Phi}(t) \leq \phi(t)/t$. Finally, boundedness of M_{jl} from below is immediate from Condition RF. \square

Step 3. The main result of this step is the following: Define the expected “ideal” penalty loadings

$$\Upsilon_l^0 := \text{diag} \left(\sqrt{\mathbb{E}[f_{i1}^2 v_{il}^2]}, \dots, \sqrt{\mathbb{E}[f_{ip}^2 v_{il}^2]} \right),$$

where the entries of Υ_l^0 are bounded away from zero and from above uniformly in n by Condition RF. Then the empirical “ideal” loadings converge to the expected “ideal” loadings:

$$\max_{1 \leq l \leq k_e} \|\widehat{\Upsilon}_l^0 - \Upsilon_l^0\|_\infty \rightarrow_P 0.$$

This follows from

$$\Delta := \max_{1 \leq l \leq k_e, 1 \leq j \leq p} |\mathbb{E}_n[f_{ij}^2 v_{il}^2] - \mathbb{E}[f_{ij}^2 v_{il}^2]| \lesssim_P \max_{1 \leq l \leq k_e, 1 \leq j \leq p} \sqrt{\mathbb{E}_n[f_{ij}^4 v_{il}^4]} \sqrt{\frac{\log(pk_e)}{n}} \lesssim_P \sqrt{\frac{\log p}{n}} \rightarrow_P 0.$$

Indeed, application of Lemma 5, gives us that

$$\begin{aligned} & \mathbb{P} \left(\Delta \geq \max_{1 \leq l \leq k_e, 1 \leq j \leq p} \sqrt{\mathbb{E}_n[f_{ij}^4 v_{il}^4]} \sqrt{2 \log(2pk_e/a)} \right) \\ & \leq pk_e \max_{1 \leq l \leq k_e, 1 \leq j \leq p} \mathbb{P} \left(\frac{|\mathbb{G}_n[f_{ij}^2 v_{il}^2]|}{\sqrt{\mathbb{E}_n[f_{ij}^4 v_{il}^4]}} \geq \sqrt{2 \log(2pk_e/a)} \right) \\ & \leq pk_e 2\bar{\Phi}(\sqrt{2 \log(2pk_e/a)})(1 + o(1)) \\ & \leq \frac{a(1 + o(1))}{\sqrt{2 \log(2pk_e/a)}}, \end{aligned}$$

uniformly in $0 < a \leq 1$ and p and k_e on the region where for some $b_n \rightarrow \infty$ slowly

$$2 \log(2pk_e/a) \leq \frac{n^{1/3}}{b_n} \min_{1 \leq j \leq p, 1 \leq l \leq k_e} W_{jl}^2, \quad W_{jl} := \frac{\mathbb{E}[f_{ij}^4 v_{il}^4]^{1/2}}{\mathbb{E}[f_{ij}^6 v_{il}^6]^{1/3}}.$$

Note that under our assumption on moments in Condition RF and Lyapunov moment inequality, the term $\min_{1 \leq j \leq p, 1 \leq l \leq k_e} W_{jl}^2$ is bounded away from zero, so the growth conditions holds for some $a \rightarrow 0$ and $b_n \rightarrow \infty$ under our condition $\log p = o(n^{1/3})$. Moreover,

$$\max_{1 \leq l \leq k_e, 1 \leq j \leq p} \mathbb{E}_n[f_{ij}^4 v_{il}^4] \leq \max_{1 \leq j \leq p} \sqrt{\mathbb{E}_n[f_{ij}^8]} \max_{1 \leq l \leq k_e} \sqrt{\mathbb{E}_n[v_{il}^8]} \lesssim_P 1,$$

where $\max_{1 \leq j \leq p} \sqrt{\mathbb{E}_n[f_{ij}^8]} \lesssim_P 1$ by assumption and $\max_{1 \leq l \leq k_e} \sqrt{\mathbb{E}_n[v_{il}^8]} \lesssim_P 1$ by the bounded k_e , Markov inequality, and the assumption that $\mathbb{E}[v_{il}^q]$ are uniformly bounded in n and l for $q \geq 8$.

Step 4. Combining the results of all the steps above, given that $\lambda = 2c\sqrt{2n \log(2pk_e)}$ and asymptotic valid penalty loadings $\widehat{\Upsilon}_l$, and using the bound $c_s \lesssim_P \sqrt{s/n}$ from Condition AS, we

obtain the conclusion that

$$\|f'_i(\widehat{\beta}_l - \beta_{l0})\|_{2,n} \lesssim_P \sqrt{\frac{s \log p}{n}} + \sqrt{\frac{s}{n}},$$

which gives, by the triangular inequality and by $\|D_i - f'_i \beta_{l0}\|_{2,n} \leq c_s \lesssim_P \sqrt{s/n}$ holding by Condition AS, the first result

$$\|\widehat{D}_i - D_i\|_{2,n} \lesssim_P \sqrt{\frac{s \log p}{n}}.$$

Then the second result follows from the definition of $\tilde{\kappa}_{c_0}(\mathbb{E}_n[f_i f'_i])$ which by Condition RE is bounded away from zero with probability approaching 1, thus

$$\|\widehat{\beta}_l - \beta_{l0}\|_2 \lesssim_P \sqrt{\frac{s \log p}{n}}.$$

Finally, we obtain the third result

$$\|\widehat{\beta}_l - \beta_{l0}\|_1 \lesssim_P \|(\widehat{\Upsilon}_l^0)^{-1}\|_\infty \left(\sqrt{s} \left(\sqrt{\frac{s \log p}{n}} + \sqrt{\frac{s}{n}} \right) + \frac{1}{\sqrt{\log p}} \frac{s}{\sqrt{n}} \right) \lesssim_P \sqrt{\frac{s^2 \log p}{n}},$$

which gives the result. \square

APPENDIX C. PROOF OF THEOREM 2

The proof proceeds in three steps. The general strategy of Step 1 follows [5, 4], but a major difference is the use of moderate deviation theory for self-normalized sums which allows us to obtain the results for non-Gaussian and heteroscedastic errors as well as handle data-driven penalty loadings. The sparsity proofs are motivated by [4] but adjusted for the data-driven penalty loadings that contain self-normalizing factors. The proof is divided in several steps.

Step 1. Here we derive a general performance bound for Post-LASSO, that actually contains more information than the statement of the theorem.

Lemma 8 (Performance of the Post-LASSO Estimator). *Let \widehat{T}_l denote the support selected by $\widehat{\beta}_l = \widehat{\beta}_{lL}$, $\widehat{m}_l = |\widehat{T}_l \setminus T_l|$, $\widehat{\beta}_{lPL}$ be the Post-LASSO estimator, $\lambda/n \geq c \|S_l\|_\infty$, and $\widehat{\Upsilon}_l$ satisfies*

(3.24) with $u \geq 1 \geq \ell > 1/c$ in the first stage for LASSO for every $l = 1, \dots, k_e$. Then we have

$$\begin{aligned} \max_{1 \leq l \leq k_e} \|f'_i(\widehat{\beta}_{lPL} - \beta_{l0})\|_{2,n} &\lesssim_P \frac{\|\widehat{\Upsilon}_l^0\|_\infty}{\sqrt{\phi_{\min}(\widehat{m}_l + s)}} \left(\sqrt{k_e \wedge \log(sk_e)} \sqrt{\frac{s}{n}} + \sqrt{\frac{\widehat{m}_l \log(pk_e)}{n}} \right) + 2c_s + \\ &\quad + \max_{1 \leq l \leq k_e} \mathbf{1}\{T_l \not\subseteq \widehat{T}_l\} \left(\frac{4u^2 s \lambda^2}{n^2 (\kappa_{(u/\ell)}^l)^2} + \frac{2uc_s \sqrt{s} \lambda}{n \kappa_{(u/\ell)}^l} \right)^{1/2} \end{aligned}$$

$$\max_{1 \leq l \leq k_e} \|\widehat{\Upsilon}_l(\widehat{\beta}_{lPL} - \beta_{l0})\|_1 \leq \max_{1 \leq l \leq k_e} \left(\|\widehat{\Upsilon}_l^0\|_\infty + \|\widehat{\Upsilon}_l - \widehat{\Upsilon}_l^0\|_\infty \right) \frac{\sqrt{\widehat{m}_l + s}}{\sqrt{\phi_{\min}(\widehat{m}_l + s)}} \|f'_i(\widehat{\beta}_{lPL} - \beta_{l0})\|_{2,n}$$

Proof. Let $\delta_l := \widehat{\beta}_{lPL} - \beta_{l0}$. By definition of the Post-LASSO estimator, it follows that $\widehat{Q}_l(\widehat{\beta}_{lPL}) \leq \widehat{Q}_l(\widehat{\beta}_{lL})$ and $\widehat{Q}_l(\widehat{\beta}_{lPL}) \leq \widehat{Q}_l(\beta_{l0\widehat{T}_l})$. Thus,

$$\widehat{Q}_l(\widehat{\beta}_{lPL}) - \widehat{Q}_l(\beta_{l0}) \leq \left(\widehat{Q}_l(\widehat{\beta}_{lL}) - \widehat{Q}_l(\beta_{l0}) \right) \wedge \left(\widehat{Q}_l(\beta_{l0\widehat{T}_l}) - \widehat{Q}_l(\beta_{l0}) \right) =: B_{l,n} \wedge C_{l,n}.$$

Next note that the least squares criterion function satisfies

$$\begin{aligned} |\widehat{Q}_l(\widehat{\beta}_{lPL}) - \widehat{Q}_l(\beta_{l0}) - \|f'_i \delta_l\|_{2,n}^2| &\leq |S'_l \widehat{\Upsilon}_l^0 \delta_l| + 2c_s \|f'_i \delta_l\|_{2,n} \\ &\leq |S'_{lT_l} \widehat{\Upsilon}_l^0 \delta_l| + |S'_{lT_l^c} \widehat{\Upsilon}_l^0 \delta_l| + 2c_s \|f'_i \delta_l\|_{2,n} \\ &\leq \|S'_{lT_l} \widehat{\Upsilon}_l^0\|_2 \|\delta_l\|_2 + \|S_{lT_l^c}\|_\infty \|\widehat{\Upsilon}_l^0 \delta_{lT_l^c}\|_1 + 2c_s \|f'_i \delta_l\|_{2,n} \\ &\leq \|S'_{lT_l} \widehat{\Upsilon}_l^0\|_2 \|\delta_l\|_2 + \|S_{lT_l^c}\|_\infty \|\widehat{\Upsilon}_l^0\|_\infty \sqrt{\widehat{m}_l} \|\delta_{lT_l^c}\|_2 + 2c_s \|f'_i \delta_l\|_{2,n} \\ &\leq \frac{\|\widehat{\Upsilon}_l^0\|_\infty \|f'_i \delta_l\|_{2,n}}{\sqrt{\phi_{\min}(\widehat{m}_l + s)}} \left(\frac{\|S'_{lT_l} \widehat{\Upsilon}_l^0\|_2}{\|\widehat{\Upsilon}_l^0\|_\infty} + \sqrt{\widehat{m}_l} \|S_{lT_l^c}\|_\infty \right) + 2c_s \|f'_i \delta_l\|_{2,n}. \end{aligned}$$

We next bound the quantities $\max_{1 \leq l \leq k_e} \|S_{lT_l^c}\|_\infty$ and $\max_{1 \leq l \leq k_e} \|S_{lT_l}\|$.

By Lemma 7, we have

$$\|S_{lT_l^c}\|_\infty \leq \max_{1 \leq l \leq k_e} \|S_l\|_\infty \lesssim_P \sqrt{\log(pk_e)} / \sqrt{n}$$

provided $\log p = o(n^{1/3})$.

Next, note that for any $j \in T_l$ we have $\mathbb{E}[S_{lj}^2 \widehat{\Upsilon}_{lj}^{02}] = \mathbb{E}[(\mathbb{E}_n[f_{ij} v_{il}])^2] \lesssim \mathbb{E}[f_{ij}^2 v_{il}^2]/n$, recall that $\Upsilon_{lj}^0 = \sqrt{\mathbb{E}[f_{ij}^2 v_{il}^2]}$ was defined in Step 3 of Appendix B. So that

$$\mathbb{E} \left[\max_{1 \leq l \leq k_e} \|S'_{lT_l} \widehat{\Upsilon}_l^0\|_2 \right] \leq \sqrt{\sum_{l=1}^{k_e} \mathbb{E}[\|S'_{lT_l} \widehat{\Upsilon}_l^0\|_2^2]} \leq \sqrt{k_e s/n} \|\Upsilon_l^0\|_\infty.$$

Thus, by Chebyshev inequality, we have $\max_{1 \leq l \leq k_e} \|S'_{T_l} \hat{\Upsilon}_l^0\|_2 \lesssim_P \sqrt{k_e s/n} \|\Upsilon_l^0\|_\infty$ and by Step 3 of Appendix B we have $\|\Upsilon_l^0\|_\infty / \|\hat{\Upsilon}_l^0\|_\infty \lesssim_P 1$. On the other hand, $\max_{1 \leq l \leq k_e} \|S_{T_l}\|_2 \leq \max_{1 \leq l \leq k_e} \sqrt{s} \|S_{lT_l}\|_\infty \lesssim_P \sqrt{s \log(sk_e)} / \sqrt{n}$ by Lemma 7.

Combining these relations and letting $A_n = \sqrt{k_e \wedge \log(sk_e)}$, we have

$$\|f'_i \delta_l\|_{2,n}^2 \lesssim_P \frac{\|\hat{\Upsilon}_l^0\|_\infty \|f'_i \delta_l\|_{2,n}}{\sqrt{\phi_{\min}(\hat{m}_l + s)}} \left(A_n \sqrt{\frac{s}{n}} + \sqrt{\frac{\hat{m}_l \log(pk_e)}{n}} \right) + 2c_s \|f'_i \delta_l\|_{2,n} + B_{l,n} \wedge C_{l,n},$$

solving which we obtain the stated result:

$$\|f'_i \delta_l\|_{2,n} \lesssim_P \frac{\|\hat{\Upsilon}_l^0\|_\infty}{\sqrt{\phi_{\min}(\hat{m}_l + s)}} \left(A_n \sqrt{\frac{s}{n}} + \sqrt{\frac{\hat{m}_l \log(pk_e)}{n}} \right) + 2c_s + \sqrt{(B_{l,n})_+ \wedge (C_{l,n})_+}.$$

Next we bound the goodness of fit terms $B_{l,n}$ and $C_{l,n}$. If $T_l \subseteq \hat{T}_l$ we directly have $C_{l,n} \leq 0$. Next, letting $\delta_{lL} := \hat{\beta}_{lL} - \beta_{l0}$, by definition of LASSO and that $\ell \hat{\Upsilon}_l^0 \leq \hat{\Upsilon}_l \leq u \hat{\Upsilon}_l^0$, we have

$$\begin{aligned} B_{l,n} = \hat{Q}(\hat{\beta}_{lL}) - \hat{Q}(\beta_{l0}) &\leq \frac{\lambda}{n} \|\hat{\Upsilon}_l \beta_{l0}\|_1 - \frac{\lambda}{n} \|\hat{\Upsilon}_l \hat{\beta}_{lL}\|_1 \leq \frac{\lambda}{n} \left(\|\hat{\Upsilon}_l \delta_{lLT_l}\|_1 - \|\hat{\Upsilon}_l \delta_{lLT_l^c}\|_1 \right) \\ &\leq \frac{\lambda}{n} \left(u \|\hat{\Upsilon}_l^0 \delta_{lLT_l}\|_1 - \ell \|\hat{\Upsilon}_l^0 \delta_{lLT_l^c}\|_1 \right). \end{aligned}$$

If $\|\hat{\Upsilon}_l^0 \delta_{lLT_l^c}\|_1 > (u/\ell) \|\hat{\Upsilon}_l^0 \delta_{lLT_l}\|_1$, we have $\hat{Q}_l(\hat{\beta}_{lL}) - \hat{Q}_l(\beta_{l0}) \leq 0$. Otherwise, $\|\hat{\Upsilon}_l^0 \delta_{lLT_l^c}\|_1 \leq (u/\ell) \|\hat{\Upsilon}_l^0 \delta_{lLT_l}\|_1$ and we have $\|\hat{\Upsilon}_l^0 \delta_{lLT_l}\|_1 \leq \sqrt{s} \|f'_i \delta_{lL}\|_{2,n} / \kappa_{(u/\ell)}^l$ by definition of $\kappa_{(u/\ell)}^l$. Then, by $\lambda/n \geq c \|S_l\|_\infty$, and the conditions on the penalty loadings, we have by (B.29) that

$$\|f'_i \delta_{lL}\|_{2,n} \leq (u + [1/c]) \frac{\lambda \sqrt{s}}{n \kappa_{(u/\ell)}^l} + 2c_s$$

and the bound on $B_{l,n}$ follows.

The second statement of the theorem follows from

$$\|\hat{\Upsilon}_l \delta_l\|_1 \leq \|\hat{\Upsilon}_l\|_\infty \|\delta_l\|_1 \leq \|\hat{\Upsilon}_l\|_\infty \sqrt{\|\delta_l\|_0} \|\delta_l\|_2 \leq \|\hat{\Upsilon}_l\|_\infty \sqrt{\|\delta_l\|_0} \|f'_i \delta_l\|_{2,n} / \sqrt{\phi_{\min}(\|\delta_l\|_0)},$$

and noting that $\|\delta_l\|_0 \leq \hat{m}_l + s$ and $\|\hat{\Upsilon}_l\|_\infty \leq \|\hat{\Upsilon}_l^0\|_\infty + \|\hat{\Upsilon}_l - \hat{\Upsilon}_l^0\|_\infty$. \square

Step 2. In this step we provide a sparsity bound for LASSO, which is important for converting the previous result to a rate result. It relies on the following lemmas.

Lemma 9 (Empirical pre-sparsity for LASSO). *Let \hat{T}_l denote the support selected by the LASSO estimator, $\hat{m}_l = |\hat{T}_l \setminus T_l|$, and assume that $\lambda/n \geq c \cdot \|S_l\|_\infty$ and $u \geq 1 \geq \ell > 1/c$ as in Lemma*

6. Then, for $c_0 = (uc + 1)/(\ell c - 1)$ we have

$$\sqrt{\widehat{m}_l} \leq \sqrt{\phi_{\max}(\widehat{m}_l)} \|(\widehat{\Upsilon}_l^0)^{-1}\|_{\infty} c_0 \left[\frac{2\sqrt{s}}{\kappa_{c_0}^l} + \frac{6nc_s}{\lambda} \right].$$

Proof of Lemma 9. We have from the optimality conditions that the LASSO estimator $\widehat{\beta}_l = \widehat{\beta}_{lL}$ satisfies

$$2\mathbb{E}_n[\widehat{\Upsilon}_{lj}^{-1} f_{ij}(y_i - f'_i \widehat{\beta}_l)] = \text{sign}(\widehat{\beta}_{lj}) \lambda/n \quad \text{for each } j \in \widehat{T}_l \setminus T_l.$$

Therefore, noting that $\|\widehat{\Upsilon}_l^{-1} \widehat{\Upsilon}_l^0\|_{\infty} \leq 1/\ell$, we have for $R = (a_{l1}, \dots, a_{ln})'$ and F denoting the $n \times p$ matrix with rows f'_i , $i = 1, \dots, n$

$$\begin{aligned} \sqrt{\widehat{m}_l} \lambda &= 2 \|(\widehat{\Upsilon}_l^{-1} F'(Y - F\widehat{\beta}_l))_{\widehat{T}_l \setminus T_l}\|_2 \\ &\leq 2 \|(\widehat{\Upsilon}_l^{-1} F'(Y - R - F\beta_{l0}))_{\widehat{T}_l \setminus T_l}\|_2 + 2 \|(\widehat{\Upsilon}_l^{-1} F'R)_{\widehat{T}_l \setminus T_l}\|_2 + 2 \|(\widehat{\Upsilon}_l^{-1} F'F(\beta_{l0} - \widehat{\beta}_l))_{\widehat{T}_l \setminus T_l}\|_2 \\ &\leq \sqrt{\widehat{m}_l} \cdot n \|\widehat{\Upsilon}_l^{-1} \widehat{\Upsilon}_l^0\|_{\infty} \|S_l\|_{\infty} + 2n \sqrt{\phi_{\max}(\widehat{m}_l)} \|\widehat{\Upsilon}_l^{-1}\|_{\infty} c_s + 2n \sqrt{\phi_{\max}(\widehat{m}_l)} \|\widehat{\Upsilon}_l^{-1}\|_{\infty} \|\widehat{\beta}_l - \beta_{l0}\|_{2,n}, \\ &\leq \sqrt{\widehat{m}_l} \cdot (1/\ell) \cdot n \|S_l\|_{\infty} + 2n \sqrt{\phi_{\max}(\widehat{m}_l)} \frac{\|(\widehat{\Upsilon}_l^0)^{-1}\|_{\infty}}{\ell} c_s + 2n \sqrt{\phi_{\max}(\widehat{m}_l)} \frac{\|(\widehat{\Upsilon}_l^0)^{-1}\|_{\infty}}{\ell} \|\widehat{\beta}_l - \beta_{l0}\|_{2,n}, \end{aligned}$$

where we used that

$$\begin{aligned} \|(F'F(\beta_{l0} - \widehat{\beta}_l))_{\widehat{T}_l \setminus T_l}\|_2 &= \sup_{\|\delta\|_0 \leq \widehat{m}_l, \|\delta\|_2 \leq 1} |\delta' F' F(\beta_{l0} - \widehat{\beta}_l)| \\ &\leq \sup_{\|\delta\|_0 \leq \widehat{m}_l, \|\delta\|_2 \leq 1} \|\delta' F'\|_2 \|F(\beta_{l0} - \widehat{\beta}_l)\|_2 \\ &= \sup_{\|\delta\|_0 \leq \widehat{m}_l, \|\delta\|_2 \leq 1} \sqrt{|\delta' F' F \delta|} \|F(\beta_{l0} - \widehat{\beta}_l)\|_2 \\ &\leq n \sqrt{\phi_{\max}(\widehat{m}_l)} \|\beta_{l0} - \widehat{\beta}_l\|_{2,n}, \end{aligned}$$

and similarly

$$\begin{aligned} \|(F'R)_{\widehat{T}_l \setminus T_l}\|_2 &= \sup_{\|\delta\|_0 \leq \widehat{m}_l, \|\delta\|_2 \leq 1} |\delta' F' R| \\ &\leq \sup_{\|\delta\|_0 \leq \widehat{m}_l, \|\delta\|_2 \leq 1} \|\delta' F'\|_2 \|R\|_2 \\ &= \sup_{\|\delta\|_0 \leq \widehat{m}_l, \|\delta\|_2 \leq 1} \sqrt{|\delta' F' F \delta|} \|R\|_2 \\ &\leq n \sqrt{\phi_{\max}(\widehat{m}_l)} c_s, \end{aligned}$$

Since $\lambda/c \geq n \|S_l\|_{\infty}$, and by Lemma 6, $\|\widehat{\beta}_l - \beta_{l0}\|_{2,n} \leq (u + \frac{1}{c}) \frac{\lambda\sqrt{s}}{n\kappa_{c_0}^l} + 2c_s$ we have

$$\sqrt{\widehat{m}_l} \leq \frac{2\sqrt{\phi_{\max}(\widehat{m}_l)} \frac{\|(\widehat{\Upsilon}_l^0)^{-1}\|_{\infty}}{\ell} \left[(u + \frac{1}{c}) \frac{\sqrt{s}}{\kappa_{c_0}^l} + \frac{3nc_s}{\lambda} \right]}{(1 - \frac{1}{\ell c})}.$$

The result follows by noting that $(u + [1/c])/(1 - 1/[\ell c]) = c_0 \ell$ by definition of c_0 . \square

Lemma 10 (Sub-linearity of maximal sparse eigenvalues). *For any integer $k \geq 0$ and constant $\ell \geq 1$ we have $\phi_{\max}(\lceil \ell k \rceil) \leq \lceil \ell \rceil \phi_{\max}(k)$.*

The proof of Lemma 10 can be found in [5].

Lemma 11 (Sparsity bound for LASSO under data-driven penalty). *Consider the LASSO estimator $\widehat{\beta}_l = \widehat{\beta}_{lL}$ with $\lambda/n \geq c\|S_l\|_\infty$, and let $\widehat{m}_l := |\widehat{T}_l \setminus T_l|$. Consider the set*

$$\mathcal{M} = \left\{ m \in \mathbb{N} : m > s \cdot 2\phi_{\max}(m) \frac{\|(\widehat{\Upsilon}_l^0)^{-1}\|_\infty^2}{\ell^2} \left[\frac{2c_0}{\kappa_{c_0}^l} + \frac{6c_0nc_s}{\lambda\sqrt{s}} \right]^2 \right\}.$$

Then,

$$\widehat{m}_l \leq s \cdot \left(\min_{m \in \mathcal{M}} \phi_{\max}(m \wedge n) \right) \cdot \|(\widehat{\Upsilon}_l^0)^{-1}\|_\infty^2 \left(\frac{2c_0}{\kappa_{c_0}^l} + \frac{6c_0nc_s}{\lambda\sqrt{s}} \right)^2.$$

Proof of Lemma 11. Rewriting the conclusion in Lemma 9 we have

$$\widehat{m}_l \leq \phi_{\max}(\widehat{m}_l) \|(\widehat{\Upsilon}_l^0)^{-1}\|_\infty^2 \left[\frac{2c_0}{\kappa_{c_0}^l} + \frac{6c_0nc_s}{\lambda\sqrt{s}} \right]^2. \quad (\text{C.32})$$

Note that $\widehat{m} \leq n$ by optimality conditions. Consider any $M \in \mathcal{M}$, and suppose $\widehat{m} > M$. Therefore by Lemma 10 on sublinearity of sparse eigenvalues

$$\widehat{m}_l \leq s \cdot \left\lceil \frac{\widehat{m}_l}{M} \right\rceil \phi_{\max}(M) \|(\widehat{\Upsilon}_l^0)^{-1}\|_\infty^2 \left[\frac{2c_0}{\kappa_{c_0}^l} + \frac{6c_0nc_s}{\lambda\sqrt{s}} \right]^2.$$

Thus, since $\lceil k \rceil \leq 2k$ for any $k \geq 1$ we have

$$M \leq s \cdot 2\phi_{\max}(M) \|(\widehat{\Upsilon}_l^0)^{-1}\|_\infty^2 \left[\frac{2c_0}{\kappa_{c_0}^l} + \frac{6c_0nc_s}{\lambda\sqrt{s}} \right]^2$$

which violates the condition that $M \in \mathcal{M}$. Therefore, we have $\widehat{m} \leq M$.

In turn, applying (C.32) once more with $\widehat{m} \leq (M \wedge n)$ we obtain

$$\widehat{m} \leq s \cdot \phi_{\max}(M \wedge n) \|(\widehat{\Upsilon}_l^0)^{-1}\|_\infty^2 \left[\frac{2c_0}{\kappa_{c_0}^l} + \frac{6c_0nc_s}{\lambda\sqrt{s}} \right]^2.$$

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

Step 3. Next we combine the previous steps to establish Theorem 5. As in Step 3 of Appendix B, recall that $\max_{1 \leq l \leq k_e} \|\widehat{\Upsilon}_l^0 - \Upsilon_l^0\|_\infty \rightarrow_P 0$, and that $1/\kappa_{c_0}^l \lesssim_P 1$ by Step 1 of Appendix B. Moreover, under conditions RE and SE, as long as $\lambda/n \geq c \max_{1 \leq l \leq k_e} \|S_l\|_\infty$, $\ell \rightarrow_P 1$ and $c > 1$, by Lemma 11 we have for every $l = 1, \dots, k_e$ that

$$\widehat{m}_l \lesssim_P s \quad (\text{C.33})$$

since $c_s \lesssim_P \sqrt{s/n}$ leads to $nc_s/[\lambda\sqrt{s}] \lesssim_P 1$. Therefore, by Lemma 8 we have

$$\max_{1 \leq l \leq k_e} \|f'_i(\widehat{\beta}_{lPL} - \beta_{l0})\|_{2,n} \lesssim_P \sqrt{\frac{s \log p}{n}} + c_s + \max_{1 \leq l \leq k_e} \frac{\lambda\sqrt{s}}{n\kappa_{(u/l)}^l}.$$

By the choice of $\lambda = 2c\sqrt{2n \log(2pk_e)}$, obtained in Lemma 7, and that $1/\kappa_{(u/l)}^l \lesssim_P 1$ for $l = 1, \dots, k_e$ by Step 1 of Appendix B, we have

$$\|f'_i(\widehat{\beta}_{lPL} - \beta_{l0})\|_{2,n} \lesssim_P \sqrt{\frac{s \log p}{n}}$$

since the event $\lambda/n \geq c \max_{1 \leq l \leq k_e} \|S_l\|_\infty$ holds with probability approaching 1. That establishes the first inequality of Theorem 5. The second follows since the minimum $(\widehat{m}_l + s)$ -sparse eigenvalues of $\mathbb{E}_n[f_i f'_i]$ are bounded away from zero, and the third inequality follows from the sparsity bound (C.33). \square

APPENDIX D. PROOF OF THEOREM 3

The proof is original and exploits the use of moderate derivation theory for self-normalized sums. We divide the proof in several steps.

Step 1. Let us define $\widetilde{d}_{il} = d_{il} - \mathbb{E}[d_{il}]$. Here we consider the basic option, in which

$$\widehat{\gamma}_{jl}^2 = \mathbb{E}_n[f_{ij}^2(d_{il} - \mathbb{E}_n d_{il})^2].$$

Let $\widetilde{\gamma}_{jl}^2 = \mathbb{E}_n[f_{ij}^2 \widetilde{d}_{il}^2]$ and $\gamma_{jl}^2 = \mathbb{E}[f_{ij}^2 \widetilde{d}_{il}^2]$. We want to show that

$$\Delta_1 = \max_{1 \leq l \leq k_e, 1 \leq j \leq p} |\widehat{\gamma}_{jl}^2 - \widetilde{\gamma}_{jl}^2| \rightarrow_P 0, \quad \Delta_2 = \max_{1 \leq l \leq k_e, 1 \leq j \leq p} |\widetilde{\gamma}_{jl}^2 - \gamma_{jl}^2| \rightarrow_P 0,$$

which would imply that $\max_{1 \leq j \leq p, 1 \leq l \leq k_e} |\widehat{\gamma}_{jl}^2 - \gamma_{jl}^2| \rightarrow_P 0$ and then since γ_{jl}^2 's are uniformly bounded from above by Condition RF and bounded below by $\gamma_{jl}^{02} = \mathbb{E}[f_{ij}^2 v_{il}^2]$, which are bounded away from zero, the asymptotic validity of the basic option then follows.

We have that

$$\Delta_1 \leq \max_{1 \leq l \leq k_e, 1 \leq j \leq p} 2|\mathbb{E}_n[f_{ij}^2 \widetilde{d}_{il}] \mathbb{E}_n[\widetilde{d}_{il}]| + \max_{1 \leq l \leq k_e, 1 \leq j \leq p} |\mathbb{E}_n[f_{ij}^2] (\mathbb{E}_n \widetilde{d}_{il})^2| \rightarrow_P 0.$$

Indeed, we have for the first term that, $\max_{1 \leq l \leq k_e, 1 \leq j \leq p} |\mathbb{E}_n[f_{ij}^2 \widetilde{d}_{il}]| \leq \max_{1 \leq j \leq p} \sqrt{\mathbb{E}_n[f_{ij}^4]} \max_{1 \leq l \leq k_e} \sqrt{\mathbb{E}_n[\widetilde{d}_{il}^2]} \lesssim_P 1$ by the assumption on the empirical moments of f_{ij} and the Markov inequality and by $\text{Var}(d_{il})$ being uniformly bounded in n and l by assumption; also recall that k_e is fixed.

Moreover, $\max_{1 \leq l \leq k_e} |\mathbb{E}_n \tilde{d}_{il}| \lesssim_P \sqrt{k_e}/\sqrt{n}$ by the Chebyshev inequality and by $\text{Var}(d_{il})$ being uniformly bounded by Condition RF. Likewise, the second term vanishes by a similar argument.

Furthermore,

$$\Delta_2 \lesssim_P \max_{1 \leq l \leq k_e, 1 \leq j \leq p} \sqrt{\mathbb{E}_n [f_{ij}^4 \tilde{d}_{il}^4]} \sqrt{\frac{\log p}{n}} \lesssim_P \sqrt{\frac{\log p}{n}} \rightarrow 0.$$

Indeed, application of Lemma 5 gives us that

$$\begin{aligned} & \mathbb{P} \left(\Delta_2 \geq \max_{1 \leq l \leq k_e, 1 \leq j \leq p} \sqrt{\mathbb{E}_n [f_{ij}^4 \tilde{d}_{il}^4]} \sqrt{2 \log(2pk_e/a)} \right) \\ & \leq pk_e \max_{1 \leq l \leq k_e, 1 \leq j \leq p} \mathbb{P} \left(\frac{|\mathbb{G}_n [f_{ij}^2 \tilde{d}_{il}^2]|}{\sqrt{\mathbb{E}_n [f_{ij}^4 \tilde{d}_{il}^4]}} \geq \sqrt{2 \log(2pk_e/a)} \right) \\ & \leq pk_e 2\bar{\Phi}(\sqrt{2 \log(2pk_e/a)})(1 + o(1)) \\ & \leq \frac{a(1 + o(1))}{\sqrt{2 \log(2pk_e/a)}} = o(1), \end{aligned}$$

uniformly in $0 < a \leq 1$ and $p \rightarrow \infty$ and k_e on the region,

$$2 \log(2pk_e/a) \leq \frac{n^{1/3}}{b_n} \min_{1 \leq l \leq k_e, 1 \leq j \leq p} W_{jl}^2, \quad W_{jl} := \frac{\mathbb{E}[f_{ij}^4 \tilde{d}_{il}^4]^{1/2}}{\mathbb{E}[f_{ij}^6 \tilde{d}_{il}^6]^{1/3}},$$

for some $b_n \rightarrow \infty$ slowly. Note that under Condition RF, by Lyapunov moment inequality, and since $\mathbb{E}[\tilde{d}_{il}^2 | x_i] \geq \mathbb{E}[v_{il}^2 | x_i]$, we have that

$$\min_{1 \leq l \leq k_e, 1 \leq j \leq p} W_{jl} \geq \min_{1 \leq l \leq k_e, 1 \leq j \leq p} \frac{\mathbb{E}[f_{ij}^2 \tilde{d}_{il}^2]}{\mathbb{E}[f_{ij}^6 \tilde{d}_{il}^6]^{1/3}} \geq \min_{1 \leq l \leq k_e, 1 \leq j \leq p} \frac{\mathbb{E}[f_{ij}^2 \tilde{v}_{il}^2]}{\mathbb{E}[f_{ij}^6 \tilde{d}_{il}^6]^{1/3}},$$

where the last term is bounded away from zero by Condition RF, so the restriction above is satisfied for some $a \rightarrow 0$ and $b_n \rightarrow \infty$ under our condition $\log p = o(n^{1/3})$. Moreover,

$$\max_{1 \leq l \leq k_e, 1 \leq j \leq p} \mathbb{E}_n [f_{ij}^4 \tilde{d}_{il}^4] \leq \max_{1 \leq j \leq p} \sqrt{\mathbb{E}_n [f_{ij}^8]} \max_{1 \leq l \leq k_e} \sqrt{\mathbb{E}_n [\tilde{d}_{il}^8]} \lesssim_P 1,$$

where $\max_{1 \leq j \leq p} \sqrt{\mathbb{E}_n [f_{ij}^8]} \lesssim_P 1$ by assumption and $\max_{1 \leq l \leq k_e} \sqrt{\mathbb{E}_n [\tilde{d}_{il}^8]} \lesssim_P 1$ by the bounded k_e , Markov inequality, and the assumption that $\mathbb{E}[\tilde{d}_{il}^q]$ uniformly bounded in n for $q \geq 8$.

Step 2. Here we consider the refined option, in which

$$\hat{\gamma}_{jl}^2 = \mathbb{E}_n [f_{ij}^2 \hat{v}_{il}^2].$$

The residual here $\hat{v}_{il} = d_{il} - \hat{D}_{il}$ can be based on any estimator that obeys

$$\|\hat{D}_i - D_i\|_{2,n} \lesssim_P \sqrt{\frac{s \log p}{n}}. \quad (\text{D.34})$$

Such estimators include the LASSO and Post-LASSO estimators based on the basic option. Below we establish that the penalty levels, based on the refined option using any estimator obeying (D.34), are asymptotically valid. Thus by Theorem 4 and 5, the LASSO and Post-LASSO estimators based on the refined option also obey (D.34). This, establishes that we can iterate on the refined option a bounded number of times, without affecting the validity of the approach.

Recall that $\widehat{\gamma}_{jl}^{02} = \mathbb{E}_n[f_{ij}^2 v_{il}^2]$ and define $\gamma_{jl}^{02} := \mathbb{E}[f_{ij}^2 v_{il}^2]$, which is bounded away from zero and from above by assumption. Hence it suffices to show that $\max_{1 \leq j \leq p, 1 \leq l \leq k_e} |\widehat{\gamma}_{jl}^2 - \gamma_{jl}^{02}| \rightarrow_P 0$. This in turn follows from

$$\Delta_1 = \max_{1 \leq l \leq k_e, 1 \leq j \leq p} |\widehat{\gamma}_{jl}^2 - \widehat{\gamma}_{jl}^{02}| \rightarrow_P 0, \quad \Delta_2 = \max_{1 \leq l \leq k_e, 1 \leq j \leq p} |\widehat{\gamma}_{jl}^{02} - \gamma_{jl}^{02}|^2 \rightarrow_P 0,$$

which we establish below.

Now note that we have proven $\Delta_2 \rightarrow_P 0$ in the Step 3 of the proof of Theorem 1. As for Δ_1 we note that

$$\Delta_1 \leq 2 \max_{1 \leq l \leq k_e, 1 \leq j \leq p} |\mathbb{E}_n[f_{ij}^2 v_{jl}(\widehat{D}_{il} - D_{il})]| + \max_{1 \leq l \leq k_e, 1 \leq j \leq p} \mathbb{E}_n[f_{ij}^2 (\widehat{D}_{il} - D_{il})^2].$$

The first term is bounded by

$$\max_{1 \leq j \leq p} (\mathbb{E}_n[f_{ij}^8])^{1/4} \max_{1 \leq l \leq k_e} (\mathbb{E}_n[v_{il}^4])^{1/4} \max_{1 \leq l \leq k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n} \lesssim_P \sqrt{\frac{s \log p}{n}} \rightarrow 0.$$

since $\max_{1 \leq j \leq p} \sqrt{\mathbb{E}_n[f_{ij}^8]} \lesssim_P 1$ by Condition RF and $\max_{1 \leq l \leq k_e} \sqrt{\mathbb{E}_n[v_{il}^4]} \lesssim_P 1$ by the bounded k_e , Markov inequality, and that $\mathbb{E}[v_{il}^4]$ is bounded uniformly in n by Condition RF. The second term is bounded by

$$\max_{1 \leq i \leq n, 1 \leq j \leq p} |f_{ij}^2| \frac{s \log p}{n} \rightarrow_P 0,$$

which converges to zero by Condition RF. □

APPENDIX E. PROOF OF LEMMA 1-4

E.1. Proof of Lemma 1. See [5] (Supplement). □

E.2. Proof of Lemma 2. See [5] (Supplement). □

E.3. Proof of Lemmas 3 and 4. To show part (1), we note that by simple union bounds and tail properties of Gaussian variable, we have that $\max_{ij} |f_{ij}^2| \lesssim_P \log p$, so we need $\log p \frac{s \log p}{n} \rightarrow 0$. Applying union bound and Bernstein inequality, it follows that this condition and that $(\log p)^2 = o(n)$, implied by this condition, suffice for $\max_j \mathbb{E}_n[f_{ij}^8] \lesssim_P 1$. Part (2) holds immediately. Parts (3) and (4) and Lemma 4 follow immediately from the definition of the conditionally bounded moments and since for any $m > 0$, $\mathbb{E}[|f_{ij}|^m]$ is bounded, uniformly in $1 \leq j \leq p$, uniformly in n , for both the Gaussian regressors of Lemma 1 and arbitrary bounded regressors of Lemma 2. \square

APPENDIX F. PROOF OF THEOREMS 4-6.

The proofs are original and they rely on the consistency of the sparsity-based estimators both with respect to the $L^2(\mathbb{P}_n)$ norm $\|\cdot\|_{2,n}$ and the ℓ_1 -norm $\|\cdot\|_1$. These proofs also exploit the use of moderate deviation theory for self-normalized sums.

Step 0. We have by Theorems 1 and 3 that the LASSO estimator with data-driven penalty loadings and by Theorems 2 and 3 the Post-LASSO estimator with data-driven penalty loadings obey:

$$\max_{1 \leq l \leq k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n} \lesssim_P \sqrt{\frac{s \log p}{n}} \rightarrow 0 \quad (\text{F.35})$$

$$\sqrt{\log p} \|\widehat{\beta}_l - \beta_{l0}\|_1 \lesssim_P \sqrt{\frac{s^2 \log^2 p}{n}} \rightarrow 0 \quad (\text{F.36})$$

In order to prove Theorem 5 we need also the condition

$$\max_{1 \leq l \leq k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n}^2 n^{2/q_e} \lesssim_P \frac{s \log p}{n} n^{2/q_e} \rightarrow 0, \quad (\text{F.37})$$

with the last statement holding by Condition SM. Note that Theorem 6 assumes (F.35) -(F.36) as high level conditions.

Step 1. We have that by $\mathbb{E}[\epsilon_i | D_i] = 0$

$$\begin{aligned} \sqrt{n}(\widehat{\alpha} - \alpha_0) &= \mathbb{E}_n[\widehat{D}_i d'_i]^{-1} \sqrt{n} \mathbb{E}_n[\widehat{D}_i \epsilon_i] \\ &= \{\mathbb{E}_n[\widehat{D}_i d'_i]\}^{-1} \mathbb{G}_n[\widehat{D}_i \epsilon_i] \\ &= \{\mathbb{E}[D_i d'_i] + o_P(1)\}^{-1} (\mathbb{G}_n[D_i \epsilon_i] + o_P(1)) \end{aligned}$$

where by Steps 2 and 3 below:

$$\mathbb{E}_n[\widehat{D}_i d'_i] = \mathbb{E}[D_i d'_i] + o_P(1) \quad (\text{F.38})$$

$$\mathbb{G}_n[\widehat{D}_i \epsilon_i] = \mathbb{G}_n[D_i \epsilon_i] + o_P(1) \quad (\text{F.39})$$

where $\mathbb{E}[D_i d'_i] = \mathbb{E}[D_i D'_i] = Q$ is bounded away from zero and bounded from above in the matrix sense, uniformly in n . Moreover, $\text{Var}(\mathbb{G}_n[D_i \epsilon_i]) = \Omega$ where $\Omega = \sigma^2 \mathbb{E}[D_i D'_i]$ under homoscedasticity and $\Omega = \mathbb{E}[\epsilon_i^2 D_i D'_i]$ under heteroscedasticity. In either case we have that Ω is bounded away from zero and from above in the matrix sense, uniformly in n , by the assumptions of the theorems. (Note that matrices Ω and Q are implicitly indexed by n , but we omit the index to simplify notations.) Therefore,

$$\sqrt{n}(\widehat{\alpha} - \alpha_0) = Q^{-1} \mathbb{G}_n[D_i \epsilon_i] + o_P(1),$$

and $Z_n = (Q^{-1} \Omega Q^{-1})^{-1/2} \sqrt{n}(\widehat{\alpha} - \alpha_0) = \mathbb{G}_n[z_{i,n}] + o_P(1)$, where $z_{i,n} = (Q^{-1} \Omega Q^{-1})^{-1/2} Q^{-1} D_i \epsilon_i$ are i.i.d. with mean zero and variance I . We have that for some small enough $\delta > 0$

$$\mathbb{E} \|z_{i,n}\|^{2+\delta} \lesssim \mathbb{E} \left[\|D_i\|^{2+\delta} |\epsilon_i|^{2+\delta} \right] \lesssim \sqrt{\mathbb{E} \|D_i\|^{4+2\delta}} \sqrt{\mathbb{E} |\epsilon_i|^{4+2\delta}} \lesssim 1,$$

by Condition SM.

This condition verifies the Lyapunov condition, and the application of the Lyapunov CLT for triangular arrays and the Cramer-Wold device implies that $Z_n \rightarrow_d N(0, I)$.

Step 2. To show (F.38), note that

$$\begin{aligned} \|\mathbb{E}_n[(\widehat{D}_i - D_i) d'_i]\| &\leq \mathbb{E}_n[\|\widehat{D}_i - D_i\| \|d_i\|] \leq \sqrt{\mathbb{E}_n[\|\widehat{D}_i - D_i\|^2] \mathbb{E}_n[\|d_i\|^2]} \\ &= \sqrt{\mathbb{E}_n \left[\sum_{l=1}^{k_e} \|\widehat{D}_{il} - D_{il}\|^2 \right] \mathbb{E}_n[\|d_i\|^2]} \\ &\leq \sqrt{k_e} \max_{1 \leq l \leq k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n} \cdot \|d_i\|_{2,n} \\ &\lesssim_P \max_{1 \leq l \leq k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n} = o_P(1). \end{aligned}$$

where $\|d_i\|_{2,n} \lesssim_P 1$ by $\mathbb{E} \|d_i\|^2 < \infty$ and Chebyshev, and the last assertion holds by Step 0.

Moreover,

$$\mathbb{E}_n[D_i D'_i] - \mathbb{E}[D_i D'_i] \rightarrow_P 0$$

by Rosenthal's [42] inequality using that $\mathbb{E} \|D_i\|^q$ for $q > 2$ is bounded uniformly in n .

Step 3. To show (F.39), note that

$$\begin{aligned}
& \max_{1 \leq l \leq k_e} |\mathbb{G}_n[(\widehat{D}_{il} - D_{il})\epsilon_i]| \\
&= \max_{1 \leq l \leq k_e} |\mathbb{G}_n\{f'_i(\widehat{\beta}_l - \beta_{l0})\epsilon_i\} + \mathbb{G}_n\{a_{il}\epsilon_i\}| \\
&= \max_{1 \leq l \leq k_e} \left| \sum_{j=1}^p \mathbb{G}_n\{f_{ij}\epsilon_i\}'(\widehat{\beta}_{lj} - \beta_{lj0}) + \mathbb{G}_n\{a_{il}\epsilon_i\} \right| \\
&\leq \max_{1 \leq j \leq p} \left| \frac{\mathbb{G}_n[f_{ij}\epsilon_i]}{\sqrt{\mathbb{E}_n[f_{ij}^2\epsilon_i^2]}} \right| \max_{1 \leq j \leq p} \sqrt{\mathbb{E}_n[f_{ij}^2\epsilon_i^2]} \max_{1 \leq l \leq k_e} \|\widehat{\beta}_l - \beta_{l0}\|_1 + \max_{1 \leq l \leq k_e} |\mathbb{G}_n\{a_{il}\epsilon_i\}|.
\end{aligned}$$

Next we note that for each $l = 1, \dots, k_e$

$$|\mathbb{G}_n\{a_{il}\epsilon_i\}| \lesssim_P [\mathbb{E}_n a_{il}^2]^{1/2} \lesssim_P \sqrt{s/n} \rightarrow 0,$$

by the Condition AS on $[\mathbb{E}_n a_{il}^2]^{1/2}$ and by Chebyshev inequality, since in the homoscedastic case of Theorem 4:

$$\text{Var} [\mathbb{G}_n\{a_{il}\epsilon_i\} | x_1, \dots, x_n] \leq \sigma \mathbb{E}_n a_{il}^2,$$

and in the boundedly heteroscedastic case of Theorem 5:

$$\text{Var} [\mathbb{G}_n\{a_{il}\epsilon_i\} | x_1, \dots, x_n] \lesssim \mathbb{E}_n a_{il}^2.$$

Next we have the following maximal inequality for self-normalized sums:

$$\max_{1 \leq j \leq p} \left| \frac{\mathbb{G}_n[f_{ij}\epsilon_i]}{\sqrt{\mathbb{E}_n[f_{ij}^2\epsilon_i^2]}} \right| \lesssim_P \sqrt{\log p}$$

provided that p obeys the growth condition $\log p = o(n^{1/3})$. To prove this, note that

$$\begin{aligned}
& \mathbb{P} \left(\max_{1 \leq j \leq p} \left| \frac{\mathbb{G}_n[f_{ij}\epsilon_i]}{\sqrt{\mathbb{E}_n[f_{ij}^2\epsilon_i^2]}} \right| \geq \sqrt{2 \log(2p/a)} \right) \\
&\leq_{(1)} p \max_{1 \leq j \leq p} \mathbb{P} \left(\frac{|\mathbb{G}_n(f_{ij}\epsilon_i)|}{\sqrt{\mathbb{E}_n[f_{ij}^2\epsilon_i^2]}} > \sqrt{2 \log(2p/a)} \right) \\
&\leq_{(2)} p 2\bar{\Phi}(\sqrt{2 \log(2p/a)})(1 + o(1)) \leq_{(3)} a(1 + o(1)),
\end{aligned}$$

uniformly for all $0 \leq a \leq 1$ and p such that

$$2 \log(2p/a) \leq \frac{n^{1/3}}{b_n} \min_{1 \leq j \leq p} M_{j0}^2, \quad M_{j0} := \frac{\mathbb{E}[f_{ij}^2\epsilon_i^2]^{1/2}}{\mathbb{E}[|f_{ij}|^3|\epsilon_i|^3]^{1/3}}. \quad (\text{F.40})$$

The bound (1) follows by the union bound; (2) follows by the moderate deviation theory for self-normalized sums, specifically Lemma 5; and (3) by $\bar{\Phi}(t) \leq \phi(t)/t$. Finally, by Condition SM $\min_{1 \leq j \leq p} M_{j0}^2$ is bounded away from zero, so the condition (F.40) is satisfied asymptotically for some $b_n \rightarrow \infty$ and some $a \rightarrow 0$ provided $\log p = o(n^{1/3})$.

Finally, we have that

$$\max_{1 \leq j \leq p} \mathbb{E}_n[f_{ij}^2 \epsilon_i^2] \leq \max_{1 \leq j \leq p} \sqrt{\mathbb{E}_n[f_{ij}^4]} \sqrt{\mathbb{E}_n[\epsilon_i^4]} \lesssim_P 1,$$

since $\max_{1 \leq j \leq p} \sqrt{\mathbb{E}_n[f_{ij}^4]} \lesssim_P 1$ by assumption and $\mathbb{E}_n[\epsilon_i^4] \lesssim_P 1$ by $\mathbb{E}[|\epsilon|^{q_\epsilon}]$ uniformly bounded in n for $q_\epsilon > 4$ and Markov inequality.

Thus, combining bounds above with bounds in (F.35)-(F.36)

$$\max_{1 \leq l \leq k_e} |\mathbb{G}_n[(\hat{D}_{il} - D_{il})\epsilon_i]| \lesssim_P \sqrt{\frac{s^2 \log^2 p}{n}} + \sqrt{\frac{s}{n}} \rightarrow 0,$$

where the conclusion follows by the assumed growth condition on the sparsity index s in Condition SM.

Step 4. This step establishes consistency of the variance estimator in the homoscedastic case of Theorem 4.

Since σ^2 and $Q = \mathbb{E}[D_i D_i']$ are bounded away from zero and from above uniformly in n , it suffices to show $\hat{\sigma}^2 - \sigma^2 \rightarrow_P 0$ and $\mathbb{E}_n[\hat{D}_i \hat{D}_i'] - \mathbb{E}[D_i D_i'] \rightarrow_P 0$.

Indeed, $\hat{\sigma}^2 = \mathbb{E}_n[(\epsilon_i - d_i'(\hat{\alpha} - \alpha))^2] = \mathbb{E}_n[\epsilon_i^2] + 2\mathbb{E}_n[\epsilon_i d_i'(\alpha - \hat{\alpha})] + \mathbb{E}_n[(d_i'(\alpha - \hat{\alpha}))^2]$ so that $\mathbb{E}_n[\epsilon_i^2] - \sigma^2 \rightarrow_P 0$ by Chebyshev inequality since $\mathbb{E}|\epsilon_i|^4$ is bounded uniformly in n , and the remaining terms converge to zero in probability since $\hat{\alpha} - \alpha \rightarrow_P 0$ by Step 3, $\|\mathbb{E}_n[d_i \epsilon_i]\| \lesssim_P 1$ by Markov and since $\mathbb{E}\|d_i \epsilon_i\| \leq \sqrt{\mathbb{E}\|d_i\|^2} \sqrt{\mathbb{E}|\epsilon_i|^2}$ is uniformly bounded in n by Condition SM, and $\mathbb{E}_n\|d_i\|^2 \lesssim_P 1$ by Markov and $\mathbb{E}\|d_i\|^2$ bounded uniformly in n by Condition SM. Next, note that

$$\|\mathbb{E}_n[\hat{D}_i \hat{D}_i'] - \mathbb{E}_n[D_i D_i']\| = \|\mathbb{E}_n[D_i(\hat{D}_i - D_i)' + (\hat{D}_i - D_i)D_i'] + \mathbb{E}_n[(\hat{D}_i - D_i)(\hat{D}_i - D_i)']\|$$

which is bounded up to a constant by

$$\sqrt{k_e} \max_{1 \leq l \leq k_e} \|\hat{D}_{il} - D_{il}\|_{2,n} \|D_i\|_{2,n} + k_e \max_{1 \leq l \leq k_e} \|\hat{D}_{il} - D_{il}\|_{2,n}^2 \rightarrow_P 0$$

by (F.35) and by $\|D_i\|_{2,n} \lesssim_P 1$ holding by Markov inequality. Moreover, $\mathbb{E}_n[D_i D_i'] - \mathbb{E}[D_i D_i'] \rightarrow_P 0$ by Step 2.

Step 5. This step establishes consistency of the variance estimator in the boundedly heteroscedastic case of Theorem 5.

Recall that $\widehat{\Omega} := \mathbb{E}_n[\widehat{\epsilon}_i^2 \widehat{D}_i \widehat{D}_i']$ and $\Omega := \mathbb{E}[\epsilon_i^2 D_i D_i']$, where the latter is bounded away from zero and from above uniformly in n . Also, $Q = \mathbb{E}[D_i D_i']$ is bounded away from zero and from above uniformly in n . Therefore, it suffices to show $\widehat{\Omega} - \Omega \rightarrow_P 0$ and that $\mathbb{E}_n[\widehat{D}_i \widehat{D}_i'] - \mathbb{E}[D_i D_i'] \rightarrow_P 0$. The latter has been show in the previous step, and we only need to show the former.

First, we note

$$\begin{aligned} \|\mathbb{E}_n[(\widehat{\epsilon}_i^2 - \epsilon_i^2) \widehat{D}_i \widehat{D}_i']\| &\leq \|\mathbb{E}_n[\{d_i'(\widehat{\alpha} - \alpha_0)\}^2 \widehat{D}_i \widehat{D}_i']\| + 2\|\mathbb{E}_n[\epsilon_i d_i'(\widehat{\alpha} - \alpha_0) \widehat{D}_i \widehat{D}_i']\| \\ &\lesssim_P \max_{i \leq n} \|d_i\|^2 n^{-1} \|\mathbb{E}_n[\widehat{D}_i \widehat{D}_i']\| + \max_{i \leq n} |\epsilon_i| \|d_i\| n^{-1/2} \cdot \|\mathbb{E}_n[\widehat{D}_i \widehat{D}_i']\| \rightarrow_P 0, \end{aligned}$$

since $\|\widehat{\alpha} - \alpha\|^2 \lesssim 1/n$, $\|\mathbb{E}_n \widehat{D}_i \widehat{D}_i'\| \lesssim_P 1$ by Step 4, and $\max_{i \leq n} \|d_i\|^2 n^{-1} \rightarrow_P 0$ by $\mathbb{E}_n[\|d_i\|^2 - \mathbb{E}\|d_i\|^2] \rightarrow_P 0$ occurring by the Rosenthal inequality and by $\mathbb{E}\|d_i\|^q$ uniformly bounded in n for $q > 2$, and $\max_{i \leq n} [|\epsilon_i| \|d_i\|] n^{-1/2} \rightarrow_P 0$ by $\mathbb{E}_n[\|d_i\|^2 |\epsilon_i|^2 - \mathbb{E}[\|d_i\|^2 |\epsilon_i|^2]] \rightarrow_P 0$ holding by the Rosenthal inequality and by $\mathbb{E}[\|d_i\|^{2+\delta} |\epsilon_i|^{2+\delta}] \leq \sqrt{\mathbb{E}[\|d_i\|^{4+2\delta}]} \sqrt{\mathbb{E}[|\epsilon_i|^{4+\delta}]}$ uniformly bounded in n by assumption, for small enough $\delta > 0$. Next we note that

$$\|\mathbb{E}_n[\epsilon_i^2 \widehat{D}_i \widehat{D}_i'] - \mathbb{E}_n[\epsilon_i^2 D_i D_i']\| = \|\mathbb{E}_n[\epsilon_i^2 D_i (\widehat{D}_i - D_i)'] + \epsilon_i^2 (\widehat{D}_i - D_i) D_i'\| + \|\mathbb{E}_n[\epsilon_i^2 (\widehat{D}_i - D_i) (\widehat{D}_i - D_i)']\|$$

which is bounded up to a constant by

$$\sqrt{k_e} \max_{1 \leq l \leq k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n} \|\epsilon_i^2\|_{D_i} + k_e \max_{1 \leq l \leq k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n}^2 \max_{i \leq n} \epsilon_i^2 \rightarrow_P 0.$$

The latter occurs because $\|\epsilon_i^2\|_{D_i} = \sqrt{\mathbb{E}_n[\epsilon_i^4 \|D_i\|^2]} \lesssim_P 1$ by $\mathbb{E}[\epsilon_i^4 \|D_i\|^2]$ uniformly bounded in n by Condition SM and by Markov inequality, and

$$\max_{1 \leq l \leq k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n}^2 \max_{i \leq n} \epsilon_i^2 \lesssim_P \frac{s \log p}{n} n^{2/q_e} \rightarrow 0,$$

where the latter step holds by Step 0 and by $\max_{i \leq n} \epsilon_i^2 \lesssim_P n^{2/q_e}$ by $\mathbb{E}_n[|\epsilon_i|^{q_e}] \lesssim_P 1$ holding by Markov and by $\mathbb{E}[|\epsilon_i|^{q_e}]$ bounded uniformly in n . Finally, $\mathbb{E}_n[\epsilon_i^2 D_i D_i'] - \mathbb{E}[\epsilon_i^2 D_i D_i'] \rightarrow_P 0$ by the Rosenthal's inequality and by $\mathbb{E}[|\epsilon_i|^{2+\delta} \|D_i\|^{2+\delta}]$ bounded uniformly in n for small enough $\delta > 0$, as shown in the proof of Step 1. We conclude that $\mathbb{E}_n[\widehat{\epsilon}_i^2 \widehat{D}_i \widehat{D}_i'] - \mathbb{E}[\epsilon_i^2 D_i D_i'] \rightarrow_P 0$. \square

REFERENCES

- [1] Takeshi Amemiya. The non-linear two-stage least squares estimator. *Journal of Econometrics*, 2:105–110, 1974.
- [2] J. Bai and S. Ng. Selecting instrumental variables in a data rich environment. *Journal of Time Series Econometrics*, 1(1), 2009.
- [3] Paul A. Bekker. Alternative approximations to the distributions of instrumental variables estimators. *Econometrica*, 63:657–681, 1994.
- [4] A. Belloni and V. Chernozhukov. Post- ℓ_1 -penalized estimators in high-dimensional linear regression models. *arXiv:[math.ST]*, 2009.
- [5] A. Belloni and V. Chernozhukov. ℓ_1 -penalized quantile regression for high dimensional sparse models. *accepted at the Annals of Statistics*, 2010.
- [6] A. Belloni, V. Chernozhukov, and L. Wang. Square-root-lasso: Pivotal recovery of nonparametric regression functions via conic programming. *Duke and MIT Working Paper*, 2010.
- [7] A. Belloni, V. Chernozhukov, and L. Wang. Square-root-lasso: Pivotal recovery of sparse signals via conic programming. *arXiv:[math.ST]*, 2010.
- [8] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.
- [9] L. Blume, D. L. Rubinfeld, and P. Shapiro. The taking of land: When should compensation be paid? *Quarterly Journal of Economics*, 100:71–92, 1984.
- [10] C. L. Boyd, L. Epstein, and A. D. Martin. Untangling the causal effects of sex on judging. forthcoming *American Journal of Political Science*, 2010.
- [11] F. Bunea, A. Tsybakov, and M. H. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.
- [12] F. Bunea, A. B. Tsybakov, , and M. H. Wegkamp. Aggregation and sparsity via ℓ_1 penalized least squares. In *Proceedings of 19th Annual Conference on Learning Theory (COLT 2006) (G. Lugosi and H. U. Simon, eds.)*, pages 379–391, 2006.
- [13] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.
- [14] E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n . *Ann. Statist.*, 35(6):2313–2351, 2007.
- [15] G. Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34:305–334, 1987.
- [16] G. Chamberlain and G. Imbens. Random effects estimators with many instrumental variables. *Econometrica*, 72:295–306, 2004.
- [17] T. Chang and A. Schoar. Judge specific differences in chapter 11 and firm outcomes. Working Paper, MIT Sloan School of Management, NBER, and CEPR, 2007.

- [18] J. Chao and N. Swanson. Consistent estimation with a large number of weak instruments. *Econometrica*, 73:1673–1692, 2005.
- [19] D. L. Chen and J. Sethi. Does forbidding sexual harassment exacerbate gender inequality. unpublished manuscript, 2010.
- [20] D. L. Chen and S. Yeh. The economic impacts of eminent domain. unpublished manuscript, 2010.
- [21] Victor H. de la Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes*. Probability and its Applications (New York). Springer-Verlag, Berlin, 2009. Limit theory and statistical applications.
- [22] Stephen G. Donald and Whitney K. Newey. Choosing the number of instruments. *Econometrica*, 69(5):1161–1191, 2001.
- [23] L. M. Ellman, C. R. Sunstein, and D. Schkade. Ideological voting on federal courts of appeals: A preliminary investigation. AEI-Brookings Joint Center for Regulatory Studies Working Paper No. 03-9, 2003.
- [24] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(5):849–911, 2008.
- [25] Wayne A. Fuller. Some properties of a modification of the limited information estimator. *Econometrica*, 45:939–954, 1977.
- [26] Jinyong Hahn, Jerry A. Hausman, and Guido M. Kuersteiner. Estimation with weak instruments: Accuracy of higher-order bias and mse approximations. *Econometrics Journal*, 7(1):272–306, 2004.
- [27] Christian Hansen, Jerry Hausman, and Whitney K. Newey. Estimation with many instrumental variables. *Journal of Business and Economic Statistics*, 26:398–422, 2008.
- [28] J. Hausman, W. Newey, T. Woutersen, J. Chao, and N. Swanson. Instrumental variable estimation with heteroskedasticity and many instruments. mimeo, 2009.
- [29] Jian Huang, Joel L. Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *Ann. Statist.*, 38(4):2282–2313, 2010.
- [30] R. Innes. Takings, compensation, and equal treatment for owners of developed and undeveloped property. *Journal of Law and Economics*, 40:403–432, 1997.
- [31] Frank Kleibergen. Testing parameters in gmm without assuming that they are identified. *Econometrica*, 73:1103–1124, 2005.
- [32] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincaré Probab. Statist.*, 45(1):7–57, 2009.
- [33] K. Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electron. J. Statist.*, 2:90–102, 2008.
- [34] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. *arXiv:0903.1468v1 [stat.ML]*, 2010.
- [35] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):2246–2270, 2009.
- [36] T. J. Miceli and K. Segerson. Regulatory takings: When should compensation be paid? *Journal of Legal Studies*, 23:749–776, 1994.

- [37] Whitney K. Newey. Efficient instrumental variables estimation of nonlinear models. *Econometrica*, 58:809–837, 1990.
- [38] Whitney K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79:147–168, 1997.
- [39] R. Okui. Instrumental variable estimation in the presence of many moment conditions. *forthcoming Journal of Econometrics*, 2010.
- [40] T. J. Riddiough. The economic consequences of regulatory taking risk on land value and development activity. *Journal of Urban Economics*, 41:56–77, 1997.
- [41] M. Rosenbaum and A. B. Tsybakov. Sparse recovery under matrix uncertainty. *arXiv:0812.2818v1 [math.ST]*, 2008.
- [42] Haskell P. Rosenthal. On the subspaces of L^p ($p > 2$) spanned by sequences of independent random variables. *Israel J. Math.*, 8:273–303, 1970.
- [43] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58:267–288, 1996.
- [44] G. K. Turnbull. Land development under the threat of taking. *Southern Economic Journal*, 69:468–501, 2002.
- [45] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.
- [46] M. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1 -constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, May 2009.
- [47] C.-H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594, 2008.

Table 1: 2SLS Simulation Results. Cut-Off Design. N = 100

Estimator	Corr(e,v) = 0					Corr(e,v) = .3					Corr(e,v) = .6				
	RMSE	Med. Bias	MAD	rp(.05)	rp(.05)	RMSE	Med. Bias	MAD	rp(.05)	rp(.05)	RMSE	Med. Bias	MAD	rp(.05)	rp(.05)
F* = 10															
2SLS(100)	0.045	0.000	0.032	0.062	0.117	0.109	0.109	0.109	0.704	0.220	0.216	0.216	0.216	1.000	0.205
LIML(100)	19.174	0.013	0.385	0.016	6.334	0.080	0.341	0.341	0.116	2.078	0.161	0.336	0.336	0.205	0.205
FULL(100)	3.656	0.013	0.385	0.016	2.253	0.080	0.341	0.341	0.114	1.931	0.161	0.336	0.336	0.205	0.205
LASSO	0.081	0.001	0.058	0.030	0.078	0.011	0.052	0.052	0.028	0.080	0.038	0.060	0.060	0.042	0.042
Post-LASSO	0.079	-0.002	0.057	0.026	0.075	0.009	0.047	0.047	0.034	0.076	0.038	0.057	0.057	0.048	0.048
F* = 40															
2SLS(100)	0.064	-0.004	0.046	0.054	0.132	0.114	0.114	0.114	0.428	0.230	0.222	0.222	0.222	0.948	0.136
LIML(100)	12.643	-0.011	0.374	0.018	21.973	0.052	0.367	0.367	0.106	90.998	0.099	0.388	0.388	0.136	0.136
FULL(100)	5.520	-0.011	0.374	0.018	3.859	0.052	0.367	0.367	0.106	3.539	0.099	0.388	0.388	0.136	0.136
LASSO	0.082	-0.001	0.052	0.066	0.084	0.016	0.057	0.057	0.070	0.083	0.026	0.051	0.051	0.086	0.086
Post-LASSO	0.081	0.000	0.049	0.064	0.082	0.014	0.055	0.055	0.068	0.082	0.024	0.048	0.048	0.086	0.086
F* = 160															
2SLS(100)	0.075	-0.008	0.053	0.054	0.103	0.072	0.077	0.077	0.182	0.166	0.141	0.141	0.141	0.526	0.088
LIML(100)	4.165	0.015	0.253	0.034	22.001	0.059	0.253	0.253	0.024	1.911	0.017	0.194	0.194	0.088	0.088
FULL(100)	3.854	0.015	0.253	0.034	4.889	0.059	0.253	0.253	0.024	1.904	0.017	0.194	0.194	0.088	0.088
LASSO	0.077	-0.009	0.054	0.048	0.078	0.007	0.052	0.052	0.050	0.081	0.004	0.055	0.055	0.066	0.066
Post-LASSO	0.077	-0.010	0.055	0.050	0.077	0.008	0.052	0.052	0.044	0.081	0.005	0.054	0.054	0.066	0.066

Note: Results are based on 500 simulation replications and 100 instruments. The first five first-stage coefficients were set equal to one and the remaining 95 to zero in this design. Corr(e,v) is the correlation between first-stage and structural errors. F* measures the strength of the instruments as outlined in the text. 2SLS(100), LIML(100), and FULL(100) are respectively the 2SLS, LIML, and Fuller(1) estimator using all 100 potential instruments. Many-instrument robust standard errors are computed for LIML(100) and FULL(100) to obtain testing rejection frequencies. LASSO and Post-LASSO respectively correspond to IV using LASSO or Post-LASSO with the refined data-driven penalty. We report root-mean-square-error (RMSE), median bias (Med. Bias), mean absolute deviation (MAD), and rejection frequency for 5% level tests (rp(.05)). In the weak instrument design (F* = 10), LASSO chose no instruments in 205, 210, and 221 simulation replications for Corr(e,v) = 0, .3, and .6 respectively. In these cases, RMSE, Med. Bias, and MAD use only the replications where LASSO selects a non-empty set of instruments, and we set the confidence interval equal to $(-\infty, \infty)$ and thus fail to reject.

Table 2: 2SLS Simulation Results. Cut-Off Design. N = 250

Estimator	Corr(e, v) = 0					Corr(e, v) = .3					Corr(e, v) = .6					
	RMSE	Med. Bias	MAD	rp(.05)	RMSE	Med. Bias	MAD	rp(.05)	RMSE	Med. Bias	MAD	rp(.05)	RMSE	Med. Bias	MAD	rp(.05)
	$F^* = 10$															
2SLS(100)	0.028	0.001	0.020	0.040	0.073	0.069	0.069	0.708	0.141	0.140	0.140	0.140	0.141	0.140	0.140	0.998
LIML(100)	0.226	0.001	0.071	0.040	0.266	0.000	0.066	0.046	0.118	-0.001	0.061	0.066	0.118	-0.001	0.061	0.066
FULL(100)	0.163	0.001	0.070	0.038	0.156	0.001	0.065	0.046	0.110	0.002	0.058	0.066	0.110	0.002	0.058	0.066
LASSO	0.044	-0.001	0.030	0.014	0.046	0.017	0.031	0.018	0.053	0.031	0.037	0.052	0.044	-0.001	0.037	0.052
Post-LASSO	0.043	-0.003	0.026	0.022	0.044	0.016	0.029	0.018	0.050	0.027	0.035	0.052	0.043	-0.003	0.035	0.052
	$F^* = 40$															
2SLS(100)	0.038	0.000	0.028	0.030	0.080	0.070	0.070	0.440	0.142	0.137	0.137	0.137	0.142	0.137	0.137	0.956
LIML(100)	0.064	0.001	0.043	0.026	0.067	0.003	0.043	0.038	0.064	-0.002	0.044	0.040	0.064	-0.002	0.044	0.040
FULL(100)	0.064	0.001	0.043	0.026	0.067	0.003	0.042	0.038	0.063	-0.001	0.044	0.040	0.063	-0.001	0.044	0.040
LASSO	0.047	0.000	0.032	0.034	0.048	0.006	0.032	0.046	0.049	0.012	0.034	0.062	0.047	0.000	0.034	0.062
Post-LASSO	0.046	0.001	0.032	0.032	0.047	0.006	0.032	0.048	0.049	0.008	0.034	0.068	0.046	0.001	0.034	0.068
	$F^* = 160$															
2SLS(100)	0.046	0.002	0.030	0.048	0.063	0.043	0.046	0.164	0.104	0.096	0.096	0.096	0.104	0.096	0.096	0.564
LIML(100)	0.054	0.001	0.035	0.048	0.053	-0.005	0.033	0.048	0.054	0.002	0.034	0.070	0.054	0.002	0.034	0.070
FULL(100)	0.054	0.001	0.035	0.048	0.053	-0.004	0.033	0.048	0.054	0.003	0.035	0.068	0.054	0.003	0.035	0.068
LASSO	0.049	0.003	0.032	0.046	0.049	0.002	0.034	0.054	0.050	0.007	0.035	0.060	0.049	0.003	0.035	0.060
Post-LASSO	0.049	0.003	0.032	0.038	0.048	0.003	0.033	0.058	0.050	0.006	0.034	0.060	0.049	0.003	0.034	0.060

Note: Results are based on 500 simulation replications and 100 instruments. The first five first-stage coefficients were set equal to one and the remaining 95 to zero in this design. $\text{Corr}(e, v)$ is the correlation between first-stage and structural errors. F^* measures the strength of the instruments as outlined in the text. 2SLS(100), LIML(100), and FULL(100) are respectively the 2SLS, LIML, and Fuller(1) estimator using all 100 potential instruments. Many-instrument robust standard errors are computed for LIML(100) and FULL(100) to obtain testing rejection frequencies. LASSO and Post-LASSO respectively correspond to IV using LASSO or Post-LASSO with the refined data-driven penalty. We report root-mean-square-error (RMSE), median bias (Med. Bias), mean absolute deviation (MAD), and rejection frequency for 5% level tests (rp(.05)). In the weak instrument design ($F^* = 10$), LASSO chose no instruments in 250, 258, and 277 simulation replications for $\text{Corr}(e, v) = 0, .3$, and $.6$ respectively. In these cases, RMSE, Med. Bias, and MAD use only the replications where LASSO selects a non-empty set of instruments, and we set the confidence interval equal to $(-\infty, \infty)$ and thus fail to reject.

Table 3: 2SLS Simulation Results. Cut-Off Design. N = 500

Estimator	Corr(e,v) = 0					Corr(e,v) = .3					Corr(e,v) = .6					
	RMSE	Med. Bias	MAD	rp(.05)	RMSE	Med. Bias	MAD	rp(.05)	RMSE	Med. Bias	MAD	rp(.05)	RMSE	Med. Bias	MAD	rp(.05)
F* = 10																
2SLS(100)	0.020	-0.001	0.013	0.034	0.052	0.048	0.048	0.720	0.099	0.098	0.098	0.098	1.000			
LIML(100)	0.420	-0.004	0.046	0.036	0.206	-0.001	0.043	0.038	0.180	0.004	0.038	0.048				
FULL(100)	0.098	-0.004	0.045	0.036	0.080	0.000	0.043	0.038	0.078	0.006	0.037	0.048				
LASSO	0.033	-0.002	0.022	0.016	0.036	0.013	0.023	0.030	0.038	0.029	0.031	0.060				
Post-LASSO	0.032	-0.001	0.021	0.018	0.035	0.011	0.021	0.026	0.036	0.027	0.030	0.048				
F* = 40																
2SLS(100)	0.028	0.000	0.019	0.048	0.053	0.045	0.045	0.368	0.102	0.099	0.099	0.968				
LIML(100)	0.046	0.001	0.031	0.038	0.046	-0.002	0.027	0.056	0.042	-0.001	0.027	0.066				
FULL(100)	0.045	0.000	0.030	0.034	0.045	-0.002	0.027	0.056	0.042	0.001	0.027	0.066				
LASSO	0.035	0.000	0.024	0.034	0.034	0.001	0.023	0.052	0.035	0.011	0.025	0.078				
Post-LASSO	0.035	0.000	0.024	0.038	0.034	0.000	0.024	0.050	0.035	0.008	0.025	0.062				
F* = 160																
2SLS(100)	0.031	0.001	0.021	0.038	0.044	0.032	0.035	0.124	0.073	0.067	0.067	0.568				
LIML(100)	0.035	0.001	0.023	0.032	0.036	-0.002	0.024	0.050	0.035	0.000	0.024	0.032				
FULL(100)	0.035	0.001	0.023	0.032	0.036	-0.001	0.023	0.050	0.035	0.000	0.024	0.032				
LASSO	0.032	0.003	0.021	0.036	0.034	0.002	0.022	0.060	0.034	0.003	0.023	0.046				
Post-LASSO	0.032	0.003	0.020	0.034	0.034	0.002	0.022	0.058	0.034	0.003	0.023	0.048				

Note: Results are based on 500 simulation replications and 100 instruments. The first five first-stage coefficients were set equal to one and the remaining 95 to zero in this design. Corr(e,v) is the correlation between first-stage and structural errors. F* measures the strength of the instruments as outlined in the text. 2SLS(100), LIML(100), and FULL(100) are respectively the 2SLS, LIML, and Fuller(1) estimator using all 100 potential instruments. Many-instrument robust standard errors are computed for LIML(100) and FULL(100) to obtain testing rejection frequencies. LASSO and Post-LASSO respectively correspond to IV using LASSO or Post-LASSO with the refined data-driven penalty. We report root-mean-square-error (RMSE), median bias (Med. Bias), mean absolute deviation (MAD), and rejection frequency for 5% level tests (rp(.05)). In the weak instrument design (F* = 10), LASSO chose no instruments in 315, 317, and 311 simulation replications for Corr(e,v) = 0, .3, and .6 respectively. In these cases, RMSE, Med. Bias, and MAD use only the replications where LASSO selects a non-empty set of instruments, and we set the confidence interval equal to $(-\infty, \infty)$ and thus fail to reject.

Table 4: 2SLS Simulation Results. Exponential Design. N = 100

Estimator	Corr(e,v) = 0					Corr(e,v) = .3					Corr(e,v) = .6					
	RMSE	Med. Bias	MAD	rp(.05)	RMSE	Med. Bias	MAD	rp(.05)	RMSE	Med. Bias	MAD	rp(.05)	RMSE	Med. Bias	MAD	rp(.05)
2SLS(100)	0.068	0.001	0.044	0.058	0.178	0.171	0.171	0.774	0.334	0.329	0.329	1.000	0.068	0.001	0.044	0.058
LIML(100)	6.511	-0.080	0.562	0.024	39.280	0.171	0.611	0.114	28.813	0.191	0.511	0.197	6.511	-0.080	0.562	0.024
FULL(100)	4.764	-0.080	0.562	0.024	5.432	0.171	0.611	0.114	3.664	0.191	0.511	0.197	4.764	-0.080	0.562	0.024
LASSO	0.138	0.018	0.087	0.034	0.147	0.034	0.091	0.070	0.136	0.068	0.094	0.094	0.138	0.018	0.087	0.034
Post-LASSO	0.134	0.018	0.088	0.038	0.144	0.038	0.094	0.076	0.137	0.075	0.097	0.112	0.134	0.018	0.088	0.038
F* = 40																
2SLS(100)	0.097	-0.006	0.064	0.048	0.211	0.190	0.190	0.540	0.389	0.377	0.377	0.972	0.097	-0.006	0.064	0.048
LIML(100)	79.644	-0.047	0.547	0.036	32.712	0.072	0.584	0.082	11.476	0.190	0.621	0.160	79.644	-0.047	0.547	0.036
FULL(100)	14.149	-0.047	0.547	0.036	5.058	0.072	0.584	0.082	8.515	0.190	0.621	0.160	14.149	-0.047	0.547	0.036
LASSO	0.134	-0.005	0.088	0.048	0.138	0.025	0.091	0.062	0.140	0.026	0.087	0.062	0.134	-0.005	0.088	0.048
Post-LASSO	0.131	-0.007	0.087	0.052	0.138	0.026	0.089	0.066	0.138	0.032	0.092	0.084	0.131	-0.007	0.087	0.052
F* = 160																
2SLS(100)	0.118	-0.007	0.083	0.040	0.187	0.146	0.149	0.234	0.306	0.287	0.287	0.692	0.118	-0.007	0.083	0.040
LIML(100)	26.697	-0.019	0.467	0.020	64.775	-0.005	0.517	0.054	8.391	0.057	0.399	0.092	26.697	-0.019	0.467	0.020
FULL(100)	8.935	-0.019	0.467	0.020	6.538	-0.005	0.516	0.054	8.086	0.057	0.399	0.092	8.935	-0.019	0.467	0.020
LASSO	0.128	-0.001	0.089	0.042	0.134	0.013	0.091	0.046	0.133	0.031	0.092	0.064	0.128	-0.001	0.089	0.042
Post-LASSO	0.127	-0.001	0.089	0.044	0.134	0.021	0.089	0.056	0.132	0.039	0.092	0.072	0.127	-0.001	0.089	0.044

Note: Results are based on 500 simulation replications and 100 instruments. The first five first-stage coefficients were set equal to one and the remaining 95 to zero in this design. Corr(e,v) is the correlation between first-stage and structural errors. F* measures the strength of the instruments as outlined in the text. 2SLS(100), LIML(100), and FULL(100) are respectively the 2SLS, LIML, and Fuller(1) estimator using all 100 potential instruments. Many-instrument robust standard errors are computed for LIML(100) and FULL(100) to obtain testing rejection frequencies. LASSO and Post-LASSO respectively correspond to IV using LASSO or Post-LASSO with the refined data-driven penalty. We report root-mean-square-error (RMSE), median bias (Med. Bias), mean absolute deviation (MAD), and rejection frequency for 5% level tests (rp(.05)). In the weak instrument design (F* = 10), LASSO chose no instruments in 84, 68, and 82 simulation replications for Corr(e,v) = 0, .3, and .6 respectively. In these cases, RMSE, Med. Bias, and MAD use only the replications where LASSO selects a non-empty set of instruments, and we set the confidence interval equal to $(-\infty, \infty)$ and thus fail to reject.

Table 5: 2SLS Simulation Results. Exponential Design. N = 250

Estimator	Corr(e,v) = 0					Corr(e,v) = .3					Corr(e,v) = .6				
	RMSE	Med. Bias	MAD	rp(.05)		RMSE	Med. Bias	MAD	rp(.05)		RMSE	Med. Bias	MAD	rp(.05)	
2SLS(100)	0.041	-0.003	0.028	0.050		0.113	0.107	0.107	0.784		0.211	0.209	0.209	1.000	
LIML(100)	277.357	-0.015	0.134	0.030		137.311	0.016	0.124	0.054		3.472	-0.003	0.111	0.062	
FULL(100)	0.353	-0.012	0.131	0.030		0.321	0.017	0.121	0.054		0.265	0.005	0.109	0.062	
LASSO	0.086	0.001	0.057	0.046		0.086	0.023	0.054	0.050		0.089	0.039	0.058	0.098	
Post-LASSO	0.084	0.002	0.052	0.040		0.083	0.026	0.051	0.058		0.090	0.046	0.064	0.110	
F* = 10															
2SLS(100)	0.064	0.004	0.045	0.046		0.137	0.124	0.124	0.522		0.247	0.241	0.241	0.998	
LIML(100)	0.165	0.009	0.084	0.050		0.133	0.002	0.079	0.056		0.113	0.004	0.080	0.042	
FULL(100)	0.154	0.010	0.084	0.048		0.132	0.003	0.078	0.056		0.112	0.007	0.079	0.042	
LASSO	0.083	-0.002	0.059	0.038		0.086	0.018	0.060	0.064		0.083	0.029	0.054	0.082	
Post-LASSO	0.081	0.001	0.058	0.038		0.085	0.017	0.060	0.060		0.083	0.030	0.058	0.084	
F* = 160															
2SLS(100)	0.075	-0.001	0.052	0.050		0.117	0.087	0.090	0.224		0.192	0.182	0.182	0.764	
LIML(100)	0.095	-0.006	0.065	0.044		0.095	-0.002	0.060	0.068		0.083	0.007	0.053	0.026	
FULL(100)	0.095	-0.006	0.065	0.044		0.095	-0.002	0.060	0.066		0.083	0.008	0.053	0.026	
LASSO	0.084	-0.005	0.058	0.056		0.085	0.007	0.057	0.054		0.077	0.019	0.053	0.044	
Post-LASSO	0.083	-0.007	0.058	0.056		0.084	0.011	0.057	0.058		0.077	0.023	0.050	0.052	

Note: Results are based on 500 simulation replications and 100 instruments. The first five first-stage coefficients were set equal to one and the remaining 95 to zero in this design. Corr(e,v) is the correlation between first-stage and structural errors. F* measures the strength of the instruments as outlined in the text. 2SLS(100), LIML(100), and FULL(100) are respectively the 2SLS, LIML, and Fuller(1) estimator using all 100 potential instruments. Many-instrument robust standard errors are computed for LIML(100) and FULL(100) to obtain testing rejection frequencies. LASSO and Post-LASSO respectively correspond to IV using LASSO or Post-LASSO with the refined data-driven penalty. We report root-mean-square-error (RMSE), median bias (Med. Bias), mean absolute deviation (MAD), and rejection frequency for 5% level tests (rp(.05)). In the weak instrument design (F* = 10), LASSO chose no instruments in 78, 93, and 103 simulation replications for Corr(e,v) = 0, .3, and .6 respectively. In these cases, RMSE, Med. Bias, and MAD use only the replications where LASSO selects a non-empty set of instruments, and we set the confidence interval equal to $(-\infty, \infty)$ and thus fail to reject.

Table 6: 2SLS Simulation Results. Exponential Design. N = 500

Estimator	Corr(e, v) = 0				Corr(e, v) = .3				Corr(e, v) = .6			
	RMSE	Med. Bias	MAD	rp(.05)	RMSE	Med. Bias	MAD	rp(.05)	RMSE	Med. Bias	MAD	rp(.05)
2SLS(100)	0.030	-0.001	0.020	0.060	0.082	0.077	0.077	0.784	0.152	0.148	0.148	1.000
LIML(100)	0.370	-0.012	0.090	0.048	2.605	0.005	0.087	0.064	3.157	0.012	0.070	0.052
FULL(100)	0.182	-0.011	0.088	0.046	0.188	0.007	0.085	0.064	0.147	0.016	0.068	0.052
LASSO	0.057	0.001	0.037	0.022	0.060	0.011	0.041	0.052	0.061	0.035	0.046	0.098
Post-LASSO	0.056	0.002	0.037	0.020	0.059	0.013	0.041	0.052	0.061	0.036	0.047	0.122
F* = 10												
2SLS(100)	0.046	0.001	0.031	0.062	0.097	0.088	0.088	0.548	0.174	0.169	0.169	0.986
LIML(100)	0.087	0.000	0.055	0.060	0.081	0.003	0.050	0.042	0.082	0.002	0.055	0.050
FULL(100)	0.087	-0.001	0.055	0.056	0.080	0.004	0.049	0.042	0.081	0.004	0.054	0.054
LASSO	0.058	0.000	0.043	0.034	0.061	0.012	0.041	0.064	0.061	0.017	0.041	0.074
Post-LASSO	0.057	0.001	0.044	0.034	0.060	0.016	0.040	0.062	0.061	0.018	0.041	0.080
F* = 160												
2SLS(100)	0.051	0.000	0.035	0.040	0.085	0.066	0.068	0.290	0.136	0.127	0.127	0.700
LIML(100)	0.063	0.001	0.042	0.032	0.065	0.006	0.045	0.054	0.063	0.004	0.046	0.044
FULL(100)	0.063	0.001	0.042	0.032	0.065	0.007	0.045	0.054	0.063	0.005	0.046	0.044
LASSO	0.055	-0.001	0.037	0.040	0.059	0.008	0.041	0.062	0.059	0.018	0.043	0.060
Post-LASSO	0.055	-0.002	0.037	0.042	0.059	0.009	0.040	0.066	0.059	0.020	0.043	0.062

Note: Results are based on 500 simulation replications and 100 instruments. The first five first-stage coefficients were set equal to one and the remaining 95 to zero in this design. Corr(e, v) is the correlation between first-stage and structural errors. F* measures the strength of the instruments as outlined in the text. 2SLS(100), LIML(100), and FULL(100) are respectively the 2SLS, LIML, and Fuller(1) estimator using all 100 potential instruments. Many-instrument robust standard errors are computed for LIML(100) and FULL(100) to obtain testing rejection frequencies. LASSO and Post-LASSO respectively correspond to IV using LASSO or Post-LASSO with the refined data-driven penalty. We report root-mean-square-error (RMSE), median bias (Med. Bias), mean absolute deviation (MAD), and rejection frequency for 5% level tests (rp(.05)). In the weak instrument design (F* = 10), LASSO chose no instruments in 112, 125, and 106 simulation replications for Corr(e, v) = 0, .3, and .6 respectively. In these cases, RMSE, Med. Bias, and MAD use only the replications where LASSO selects a non-empty set of instruments, and we set the confidence interval equal to $(-\infty, \infty)$ and thus fail to reject.

Table 7: Effect of Federal Appellate Takings Law Decisions on Economic Outcomes

	Home Prices			GDP
	log(FHFA)	log(Non-Metro)	log(Case-Shiller)	log(GDP)
OLS	0.0017	0.0087	0.0178	0.0020
s.e.	0.0106	0.0110	0.0080	0.0058
2SLS	-0.0180	0.0720	0.0480	-0.0261
s.e.	0.0337	0.0325	0.0164	0.0225
FS-F	13.9258	13.3136	32.4474	13.2854
Post-LASSO	-0.0090	0.0507	0.0415	-0.0232
s.e.	0.0328	0.0171	0.0132	0.0202
FS-F	23.1442	21.3627	23.0862	22.0799
S	2	4	4	2

Note: This table reports the estimated effect of an additional pro-plaintiff takings decision, a decision that goes against the government and leaves the property in the hands of the private owner, on various economic outcomes using two-stage least squares (2SLS). The characteristics of randomly assigned judges serving on the panel that decides the case are used as instruments for the decision variable. All estimates include circuit effects, circuit-specific time trends, time effects, controls for the number of cases in each circuit-year, and controls for the demographics of judges available within each circuit-year. Each column corresponds to a different dependent variable. $\log(\text{FHFA})$, $\log(\text{Non-Metro})$, and $\log(\text{Case-Shiller})$ are log-house-price-indexes, and $\log(\text{GDP})$ is the log of state-level GDP. OLS are ordinary least squares estimates. 2SLS is the 2SLS estimator with the original instruments in Chen and Yeh (2010). Post-LASSO provides 2SLS estimates obtained using instruments selected by LASSO with the refined data-dependent penalty choice. Rows labeled s.e. provide the estimated standard errors of the associated estimator. All standard errors are computed with clustering at the circuit-year level. FS-F is the value of the first-stage F-statistic using the selected instrument. S is the number of instruments chosen by LASSO.