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 July 11, 2011. 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], [13], and [31]. 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], [15], [23], and [24] 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 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], [10], [11], [25], [9], [12], [27], [28], [29], [30], [34], [37], [38], [39], [40], and [4], among many others; for analogous results on Post-LASSO see [4].

[1] It 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] [14] 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 [33] 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 [31] 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.

[3] This is in contrast to the variable selection method of [19] which relies on a *a priori* knowledge that allows one to order the instruments in terms of instrument strength.

[4] One 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 [21] (FULL)[5] which have been advocated as procedures that are robust to using many instruments (e.g. [23]). 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 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 [17]. We follow the identification strategy of [17] 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 [17]. 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,

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

[5] Note 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

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

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 $a \leqslant cb$ 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 $\to 1$ 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 $X_n \to_d Y$.

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

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

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

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

Given a k_d-vector of instruments $A(x_i)$, the standard IV estimator of α_0 is given by $\widehat{\alpha} = (\mathbb{E}_n[A(x_i)d_i'])^{-1}\mathbb{E}_n[A(x_i)y_i]$, where $\{(x_i, d_i, y_i), i = 1, ..., n\}$ is an i.i.d. sample from the IV model above. For a given $A(x_i)$, $\sqrt{n}(\widehat{\alpha} - \alpha_0) =_d N(0, Q_0^{-1}\Omega_0 Q_0^{-1\prime}) + o_P(1)$, where $Q_0 = \mathrm{E}[A(x_i)d_i']$ and $\Omega_0 = \sigma^2 E[A(x_i)A(x_i)']$ under the standard conditions. Setting

$$A(x_i) = D(x_i) = \mathrm{E}[d_i|x_i]$$

minimizes the limit variance which becomes $\Lambda^* = \sigma^2\{\mathrm{E}[D(x_i)D(x_i)']\}^{-1}$, the semi-parametric efficiency bound for estimating α_0; see [1], [13], and [31]. 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 $d_i = (d_1, ..., d_{k_e}, w_i')'$ 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

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

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

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

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

Suppose there is a very large list of instruments,

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

to be used in estimation of conditional expectations $D_l(x_i)$, $l = 1, ..., k_e$, 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. [31].

We mainly use the term "series instruments" and contrast our results with those in the seminal work of [31], though our results are not limited to canonical series regressors as in [31]. 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 [31] 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:

$$D_l(x_i) = f_i'\beta_{l0}, \quad l = 1, ..., k_e, \tag{2.2}$$

$$\max_{1\leqslant l\leqslant k_e} \|\beta_{l0}\|_0 = \max_{1\leqslant l\leqslant k_e} \sum_{j=1}^{p} 1\{\beta_{l0j} \neq 0\} \leqslant s \ll n. \tag{2.3}$$

This simple sparsity model substantially generalizes the classical parametric model of optimal instruments of [1] by letting the identities of the relevant instruments $T_l = \text{support}(\beta_{l0}) = \{j \in \{1, \ldots, p\} \ : \ |\beta_{l0j}| > 0\}$ 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:*

$$D_l(x_i) = f_i'\beta_{l0} + a_l(x_i), \qquad l = 1, ..., k_e, \tag{2.4}$$

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

This model generalizes the nonparametric model of the optimal instrument of [31] by letting the identities of the most important series terms

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

be unknown and potentially different for $l = 1, \ldots, k_e$. 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 [31] 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:

$$\frac{s^2(\log p)^2}{n} \to 0 \tag{2.6}$$

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 [31] and [32]. In order to explain this consider the case of $k_e = 1$ and the set $\{f_j(x), j \geqslant 1\}$ 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 $\mathrm{E}D_l^2(x_i) < \infty$ by assumption, we can represent D_l via a Fourier expansion, $D_l(x) = \sum_{j=1}^{\infty} \delta_j f_j(x)$, where $\{\delta_j, j \geqslant 1\}$ 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 $\sqrt{\mathbb{E}_n[a_l^{c2}(x_i)]} \lesssim_P \sqrt{\mathrm{E}[a_l^{c2}(x_i)]} \lesssim K^{\frac{-2a+1}{2}}$. 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 $T_l = \{1, ..., s\}$, 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 $j \leqslant s$ and

[6] The 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 $s+1 \leqslant j \leqslant p$, 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 $\|\beta_{l0}\|_0 \leqslant s$; moreover, the key growth condition (2.6) required for IV estimation holds if the smoothness index $a > 1$ is sufficiently high so that $n^{1/a}(\log p)^2/n \to 0$. 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 $j \leqslant M$ and $\delta_j \propto (j-M)^{-a}$ for $j > M$. Clearly in this case the standard series approximation based on the first $K \leqslant M$ 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 $j \leqslant M$ and $j > M+s$, and $\beta_{l0j} = \delta_j$ for $M+1 \leqslant j \leqslant M+s$. Hence $\|\beta_{l0}\|_0 \leqslant s$ and we have that $\sqrt{\mathbb{E}_n[a_l^2(x_i)]} \lesssim_P \sqrt{\mathrm{E}[a_l^2(x_i)]} \lesssim s^{\frac{-2a+1}{2}}(1+o(1)) \lesssim \sqrt{s/n} \lesssim n^{\frac{1-2a}{4a}}$. 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. □

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

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

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

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

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.

[7] This is not merely a finite sample phenomenon but is also accomodated 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 [37] is defined as a solution of the following optimization program:

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

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

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

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

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

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

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

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

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 [37] 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 $\widehat{\alpha}$ is simply the standard series estimator of the optimal instruments of [31] 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:

$$\max_{1 \leqslant l \leqslant k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n} \lesssim_P \sqrt{\frac{s \log p}{n}} \tag{2.12}$$

$$\max_{1 \leqslant l \leqslant k_e} \|\widehat{\beta}_l - \beta_{l0}\|_1 \lesssim_P \sqrt{\frac{s^2 \log p}{n}}. \tag{2.13}$$

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

$$\widehat{D}_l(x_i) = f_i'\widehat{\beta}_l, l = 1, ..., k_e, \tag{2.14}$$

and then set

$$\widehat{D}_i = (\widehat{D}_1(x_i), ..., \widehat{D}_l(x_i), w_i')'. \tag{2.15}$$

The resulting IV estimator takes the form

$$\widehat{\alpha}^* = \mathbb{E}_n[\widehat{D}_i d_i']^{-1}\mathbb{E}_n[\widehat{D}_i y_i]. \tag{2.16}$$

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:

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

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 [31] 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 $j \leqslant K$, 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:

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

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

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

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 algorithm. We establish the asymptotic validity of this algorithm in the subsequent sections. Let $K \geqslant 1$ denote a bounded number of iterations.

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

(1) For each $l = 1, \ldots, k_e$, specify penalty loadings according to (2.10). A simple default is to use the initial option. One can also use the refined option given an initial set of residual estimates $\widehat{v}_{il}$. Use the initial penalty loading in computing the LASSO estimator $\widehat{\beta}_{lL}$

[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.)

via (2.9). Then (i) compute residuals $\widehat{v}_{il} = d_{li} - f_i'\widehat{\beta}_{lL}$, $i = 1, ..., n$, or (ii) compute the Post-LASSO estimator $\widehat{\beta}_{lPL}$ via (2.11) and the residuals as $\widehat{v}_{il} = d_{li} - f_i'\widehat{\beta}_{lPL}$, $i = 1, ..., n$.

(2) For each $l = 1, \ldots, k_e$, update the penalty loadings according to the refined option in (2.10) and update the LASSO estimator $\widehat{\beta}_{lL}$ via (2.9). Then (i) compute a new set of residuals using the updated LASSO coefficients $\widehat{v}_{il} = d_{li} - f_i'\widehat{\beta}_{lL}$, $i = 1, ..., n$, or (ii) update the Post-LASSO estimator $\widehat{\beta}_{lPL}$ via (2.11) and compute residuals using the updated Post-LASSO coefficients as $\widehat{v}_{il} = d_{li} - f_i'\widehat{\beta}_{lPL}$, $i = 1, ..., n$.

(3) Repeat the previous step K times, where K is bounded. Then (i) compute the estimates of the optimal instrument, $\widehat{D}_{il} = f_i'\widehat{\beta}_{lL}$, for $i = 1, ..., n$ and each $l = 1, ..., k_e$; or (ii) compute the estimates of the optimal instrument, $\widehat{D}_{il} = f_i'\widehat{\beta}_{lPL}$, for $i = 1, ..., n$ and each $l = 1, ..., k_e$. 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 algorithm allows use of either LASSO or Post-LASSO at any step. Our results allow for LASSO and Post-LASSO to be mixed freely across the steps. Our preferred approach uses Post-LASSO at every stage.

3. Main Results on LASSO and Post-LASSO Estimators of Conditional Expectation Functions under Conditionally 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:

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

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

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

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 $\kappa_C(\mathbb{E}_n[f_i f_i']) \geqslant \kappa$ and $\widetilde{\kappa}_C(\mathbb{E}_n[f_i f_i']) \geqslant \kappa$ as $n \to \infty$.*

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 $p \geqslant n$, [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

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

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

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

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 $n \to \infty$, $\kappa' \leqslant \phi_{\min}(Cs)(\mathbb{E}_n[f_i f_i']) \leqslant \phi_{\max}(Cs)(\mathbb{E}_n[f_i f_i']) \leqslant \kappa''$.*

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. We refer to [4] for proofs; the first lemma builds upon results in [40] and the second builds upon results in [36].

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 $\mathrm{E}[f_i f_i']$ has $s \log n$-sparse eigenvalues bounded from above and away from zero uniformly in n. 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 $\|f_i\|_\infty \leqslant K_B$ a.s. Further suppose that the population Gram matrix $\mathrm{E}[f_i f_i']$ has $s \log n$-sparse eigenvalues bounded from above and away from zero uniformly in n. Then if $K_B^2 s \log^2(n) \log^2(s \log n) \log(p \vee n) = o(n)$, Conditions RE and SE hold.*

Recall that a standard assumption in econometric research is to assume that the the population Gram matrix $\mathrm{E}[f_i f_i']$ has eigenvalues bounded from above and below, see e.g. [32]. The lemmas above allow for this and even much more general behavior, requiring only that the sparse eigenvalues of the population Gram matrix $\mathrm{E}[f_i f_i']$ 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 $\tilde{d}_{il} := d_{il} - \mathrm{E}[d_{il}]$.

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

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

(ii) The moments $\mathrm{E}[\tilde{d}_{il}^8]$ *and* $\mathrm{E}[v_{il}^8]$ *are bounded uniformly in* $1 \leqslant l \leqslant k_e$ *and in* n*. (iii) The regressors* f_i *obey:* $\max_{1 \leqslant j \leqslant p} \mathbb{E}_n[f_{ij}^8] \lesssim_P 1$ *and* $\max_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p} |f_{ij}^2| \frac{s \log p}{n} \to_P 0$*. (iv) The moments* $\mathrm{E}[f_{ij}^2 v_{il}^2]$ *are bounded away from zero and from above uniformly in* $1 \leqslant j \leqslant p$*,* $1 \leqslant l \leqslant k_e$*, uniformly in* n*, and the moments* $\mathrm{E}[f_{ij}^6 \tilde{d}_{il}^6]$*,* $\mathrm{E}[f_{ij}^6 v_{il}^6]$*,* $\mathrm{E}[|f_{ij}|^3 |v_{il}|^3]$ *are bounded, uniformly in* $1 \leqslant j \leqslant p$*,* $1 \leqslant l \leqslant k_e$*, 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 $0 < B_1 < B_2 < \infty$:

$$B_1 \leqslant \mathrm{E}\Big[|g_i|^k \Big| x_i\Big] \leqslant B_2 \text{ with probability 1, for } k = 1, \ldots, K.$$

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* $s(\log p)^2/n \to 0$*. (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* $l = 1, \ldots, k_e$*, 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:

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

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,

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

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

$$\mathrm{P}\left(\lambda \geqslant cn \max_{1\leqslant l\leqslant k_e} \|S_l\|_\infty\right) \to 1, \tag{3.22}$$

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

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

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

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

$$\ell\widehat{\Upsilon}_l^0 \leqslant \widehat{\Upsilon}_l \leqslant u\widehat{\Upsilon}_l^0, \tag{3.24}$$

with $0 < \ell \leqslant 1 \leqslant u$ such that $\ell \to_P 1$ and $u \to_P u'$ with $u' \geqslant 1$.

Theorem 1 (Rates for LASSO under Non-Gaussian and Heteroscedastic Errors)**.** *Suppose that in the regression model (2.7) Conditions AS, RE, SE 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* $\widehat{D}_{il} = f_i'\widehat{\beta}_{lL}$, $l = 1, ..., k_e$*, satisfy*

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

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 AS, RE, 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* $\widehat{D}_{il} = f_i'\widehat{\beta}_{lPL}$*,* $l = 1, ..., k_e$*, satisfy*

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

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* $\widehat{\Upsilon}$ *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 $\mathrm{E}[\epsilon_i^2|x_i]$ 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 $\mathrm{E}[f_{ij}^2\epsilon_i^2] = 1$ for each $1 \leqslant j \leqslant p$ and for all n. (ii) $\mathrm{E}[\|D_i\|^q]$ and $\mathrm{E}[\|d_i\|^q]$ and $\mathrm{E}[|\epsilon_i|^{q_\epsilon}]$ are bounded uniformly in n, where $q_\epsilon > 4$ and $q > 4$. (iii) The moments $\mathrm{E}[\epsilon_i^4\|D_i\|^2]$ and $\mathrm{E}[|f_{ij}|^3|\epsilon_i|^3]$ are bounded uniformly in $1 \leqslant j \leqslant p$, uniformly in n. (iv) The following growth conditions hold:*

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

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 $\beta_{l0}, l = 1, ..., k_e$, 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 $s\log p/n \to 0$ 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 $\widehat{\alpha}$, based on either LASSO or Post-LASSO estimates of the optimal instrument, is root-n consistent, asymptotically normal, and achieves the efficiency bound:*

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

where $\Lambda^* := \sigma^2 Q^{-1}$ for $Q = \mathrm{E}[D(x_i)D(x_i)']$, 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 $\widehat{\Lambda}^* := \widehat{\sigma}^2 \widehat{Q}^{-1}$, where $\widehat{Q} = \mathbb{E}_n[\widehat{D}(x_i)\widehat{D}(x_i)']$ and $\widehat{\sigma}^2 = \mathbb{E}_n[(y_i - d_i'\widehat{\alpha})^2]$.

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 4 hold, except that now the structural errors ϵ_i can be heteroscedastic conditional on x_i. Then the IV estimator $\widehat{\alpha}$, based on either LASSO or Post-LASSO estimates of the optimal instrument, is root-n consistent and asymptotically normal:*

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

for $\Omega := \mathrm{E}[\epsilon_i^2 D(x_i)D(x_i)']$ and $Q := \mathrm{E}[D(x_i)D(x_i)']$, 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* $\widehat{\Omega} := \mathbb{E}_n[\widehat{\epsilon}_i^2 \widehat{D}(x_i)\widehat{D}(x_i)']$ *for* $\widehat{\epsilon}_i = y_i - d_i'\widehat{\alpha}$*, and* Q *replaced by* $\widehat{Q} := \mathbb{E}_n[\widehat{D}(x_i)\widehat{D}(x_i)']$.

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,* $\widehat{D}_{il} = f_i'\widehat{\beta}_l$*, are constructed using any estimator* $\widehat{\beta}_l$ *such that*

$$\max_{1\leqslant l\leqslant k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n} \lesssim_P \sqrt{\frac{s\log p}{n}} \tag{4.25}$$

$$\max_{1\leqslant l\leqslant k_e} \|\widehat{\beta}_l - \beta_{l0}\|_1 \lesssim_P \sqrt{\frac{s^2\log p}{n}}. \tag{4.26}$$

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, [12]
- $\sqrt{\text{LASSO}}$ and post-$\sqrt{\text{LASSO}}$, [7] and [6],
- thresholded LASSO and post-thresholded LASSO, [4]
- grouped LASSO and post-grouped LASSO, [25], [29]
- adaptive versions of the above, [25].

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).

[9] Note 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.

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. 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:

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

where $\beta = 1$ is the parameter of interest, and $z_i = (z_{i1}, z_{i2}, ..., z_{i100})' \sim N(0, \Sigma_Z)$ is a 100 x 1 vector with $E[z_{ih}^2] = \sigma_z^2$ and $Corr(z_{ih}, z_{ij}) = .5^{|j-h|}$. 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 two different values for $Corr(e, v)$: .3 and .6. We also consider two values of σ_v^2 which are chosen to benchmark two different strengths of instruments. The two values of σ_v^2 are found as $\sigma_v^2 = \frac{n\Pi'\Sigma_Z\Pi}{F^*\Pi'\Pi}$ for F^*: 10 and 40.[10] Finally, we consider two different settings for the first-stage coefficients, Π. The first sets the S elements of Π equal to one and the remaining elements equal to zero. We report results for $S = 5$ and $S = 50$ to cover a sparse and less sparse case. We refer to this design as the "cut-off" design. The second model sets the coefficient on $z_{ih} = .7^{h-1}$ for $h = 1, ..., 100$. We refer to this design as the "exponential" design. In the exponential design, the model is not literally sparse although the majority of explanatory power is contained in the first few instruments, while the model is exactly sparse in the cut-off design. However, treating $\frac{s^2(\log p)^2}{n}$ as vanishingly small seems

[10] These values were chosen by roughly benchmarking to the first-stage F-statistics calculated using clustered standard errors and the instruments used by [17] from the emprical example in Section 6 which produces F's between 29 and 119. We took 40 as a rough intermediate value and then divided by four to obtain weaker and stronger identification.

like a poor approximation with $S = 50$ and the sample sizes considered. We expect LASSO to perform poorly in this case.[11]

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 [21] (FULL)[12] 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 [23] for example. We report results for these estimators in rows labeled 2SLS(100), LIML(100), and FULL(100) respectively.[13] For LASSO, we consider variable selection based on two different sets of instruments. In the first scenario, we use LASSO to select among the base 100 instruments and report results for the IV estimator based on the Post-LASSO (Post-LASSO) forecasts. In the second, we use LASSO to select among 120 instruments formed by augmenting the base 100 instruments by the first 20 principle components constructed from the sampled instruments in each replication. We then report results for the IV estimator based on the Post-LASSO (Post-LASSO-F) forecasts. In all cases, we use the refined data-dependent penalty loadings given in (2.10).[14] For each estimator, we report median bias (Med. Bias), median absolute deviation (MAD), and rejection frequencies for 5% level tests (rp(.05)). For computing rejection frequencies, we estimate conventional, homoscedastic 2SLS standard errors for 2SLS(100), Post-LASSO, and Post-LASSO-F and the many instrument robust standard

[11] In an online appendix, we provide simulation results for additional settings: $Corr(e, v) \in \{0, .3, .6\}$, $F^* \in \{2.5, 10, 40, 160\}$, and $S \in \{5, 25, 50, 100\}$. The settings reported in the main paper are sufficient to capture the main patterns in this larger set of simulations.

[12] [21] requires a user-specified parameter. We set this parameter equal to one which produces a higher-order unbiased estimator. See [22] for additional discussion.

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

[14] Specifically, we form an initial estimate of the first-stage residuals $\widehat{v}_i$ by regressing x_i on one element of z_i found by minimizing the LASSO objective function over β and λ subject to $\|\beta\|_0 = 1$. Using this initial set of residual estimates, we follow the algorithm in Section 2.5 with K = 1.

errors of [23] which rely on homoscedasticity for LIML(100) and FULL(100). We also report the number of cases in which LASSO selected no instruments in the column labeled N(0).[15]

Simulation results for Corr$(e, v) = .3$ and Corr$(e, v) = .6$ are respectively presented in Tables 1-2. When considering the results, it is useful to recall the importance of sparsity for the good performance of LASSO in selecting variables. Sparsity will provide a good approximation in scenarios in which there are a few important variables with coefficients that are large relative to estimation uncertainty. In our simulation design, these conditions should be met in the exponential and S = 5 cut-off cases with $F^* = 40$. Decreasing the value of the first-stage F-statistics while keeping the number of relevant instruments constant corresponds to decreasing the ratio of the coefficient to sampling error which should make it harder to detect which are the relevant instruments. Given a set of selected instruments, the performance of the resulting IV estimator should also suffer due to usual weak instrument concerns. In the scenario with $S = 50$, individual coefficients will also need to be small relative to sampling variation to keep the F^* values at the levels chosen. In this case, there are essentially no variables that are "individually relevant" meaning that LASSO as a variable selection device among the raw instruments should break down. For this reason, we also consider an estimation procedure in which one selects among a dictionary[16] of instrumental variables that includes sensible linear combinations of the baseline variables. Finally, note that increasing the sample size keeping F^* fixed implies a reduction in the size of the coefficients on individual instrumental variables.

A feature of the simulation results that immediately stands out is that LASSO often fails to select any instruments in cases in which there no individually relevant variables as intuition and theory would predict. With $S = 50$, LASSO on the original set of instruments results in failure to select any instruments in (essentially) all cases considered. With $S = 5$ or in the exponential design, we also see that LASSO selects no instruments in a non-negligible fraction of simulation replications when the instruments are relatively weak ($F^* = 10$). One possibility in a scenario such as this is that there is not a sparse representation in the set of instruments considered but there is a sparse representation in some other representation of the information

[15] In cases where LASSO selects no instruments, Med. Bias, and MAD use only the replications where LASSO selects a non-empty set of instruments, and we set the confidence interval eqaul to $(-\infty, \infty)$ and thus fail to reject.

[16] Loosely defined, a dictionary is a set of variables that contain redundant elements.

available in the instruments. To allow for this possibility, we consider LASSO among the set of raw instruments augmented with principle components of the instruments. We see that, despite a sparse representation in principle components not being exact in our designs, doing LASSO over this dictionary substantively reduces the number of cases in which LASSO fails to select any instruments, though the fraction remains sizeable when $S = 50$ and the instruments are weak. We feel that the ability of variable selection procedures to search over dictionaries comprised of elements from different bases is an important feature that makes sparsity more palatable and serves as a useful complement to approaches to dimension reduction based on factor analysis as in [2]. Of course, LASSO still fails to select instruments in many cases. As a practical issue, it seems likely that LASSO's selecting no instruments is an indication that instruments are (individually) weak and might be taken as an indication that weak or many instrument robust procedures are called for.[17]

Ignoring the $S = 50$ and $N = 100$ results where LASSO using either set of instruments selects no instruments the majority of the time, we see that Post-LASSO and Post-LASSO-F perform well in terms of estimator risk as measured by MAD. Not surprisingly, the LASSO-based estimators substantially outperform 2SLS, LIML, and FULL when $N = 100$. The LASSO-based estimators also perform systematically better than the remaining estimators with $N = 250$, though LIML, FULL, and the LASSO-based estimators are similar with $\mathrm{Corr}(e, v) = .6$ and $F^* = 40$ in the less-favorable cut-off designs. When $N = 500$, the LASSO-based estimators have smaller MAD's in the exponential design and slightly smaller MAD's with $S = 5$. With $S = 50$ and $N = 500$, Post-LASSO-F, LIML, and FULL all perform similarly. Looking at Median Bias, we see that the LASSO-based IV seems to do better than the other options when $N = 100$ but is outperformed by LIML and FULL in the larger sample sizes. This result is unsurprising given that LIML is approximately median unbiased and FULL is higher-order unbiased.

Finally, we see that LASSO does quite well in terms of rejection frequencies of 5% level tests. Unsurpisingly, there is no procedure that is uniformly dominant based on this metric, though 2SLS(100) performs very poorly across the board. Based on rejection frequencies, LASSO is competitive with inference based on FULL(100) or LIML(100) with many-instrument-robust

[17] The use of LASSO coupled with a many-instrument robust procedure such as LIML or FULL or a weak-instrument robust procedure as in [26], for example, is an interesting extension. We are investigating these possibilities as well as other approaches to instrument selection, e.g. based on [20], in ongoing research.

standard errors. In the $S = 5$ cut-off design, sizes of Post-LASSO-based tests are on par or better than those of LIML or FULL, though the Post-LASSO-F tests are somewhat more size-distorted across most cases. In the exponential design, tests using Post-LASSO and Post-LASSO-F do well when $\text{Corr}(e, v) = .3$ but suffer from modest size distortions when $corr(e, v) = .6$. In the $S = 50$ cut-off design, the Post-LASSO-F tests tend to be more size-distorted but generally on par with the LIML- and FULL-based tests. Whether one would be willing to trade this modest deterioration in testing performance for the potential improvements in MAD of the estimator will of course depend on the preferences of the researcher.

The evidence from the simulations is supportive of the derived theory and favorable to LASSO-based IV methods. The LASSO-IV estimators appear clearly dominate on all metrics considered when $p = N$ and $S << N$. The LASSO-based IV estimators generally have relatively small median bias and estimator risk and do well in terms of testing properties, though they do not dominate LIML or FULL in these dimensions across all designs with $p < n$. In particular, we see that LIML and FULL become relatively more appealing as the sparsity assumption underlying the validity of LASSO variable selection becomes a less adequate approximation to the underlying structure of the model. This breakdown of sparsity is likely in situations with weak instruments, be they many or few, where none of the first-stage coefficients are well-separated from zero relative to sampling variation. Overall, the simulation results show that simple LASSO-based procedures are competitive to existing methods for many instruments and should usefully complememnt these methods in applied work.

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

$$y_{ct} = \alpha_c + \alpha_t + \gamma_c t + \beta \; Takings \; Law_{ct} + W_{ct}'\delta + \epsilon_{ct} \tag{6.27}$$

[18] See [17] for a detailed discussion of the economics of takings law (or eminent domain), relevant institutional features of the legal system, and a careful discussion of endogeneity concerns and the instrumental variables strategy in this context.

where y_{ct} is an economic outcome for circuit c at time t, $Takings\ Law_{ct}$ represents the number of pro-plaintiff apellate takings decisions in circuit c and year t; W_{ct} are judicial pool characteristics,[19] 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 $\gamma_c t$ 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: 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 repeat-sales residential housing prices. We also use state level GDP from the Bureau of Economic Analysis to form log(GDP). For simplicity and since all of the controls, instruments, and the endogenous variable all vary only at the circuit-year level, we use the within-circuit-year average of each of these variables as the dependent variables in our models. Due to the different coverage and time series lengths available for each of these series, the sample sizes and sets of available controls differ somewhat across the outcomes. These differences lead to different first-stages across the outcomes as well. The total sample sizes are 312 for FHFA and GDP which have identical first-stages. For Non-Metro and Case-Shiller the sample sizes are 110 and 183 respectively.

The analysis of the effects of takings law is complicated by the possible endogeneity between governmental takings and takings law decisions and economic variables. To address the potential endogeneity of takings law, we employ the instrumental variables strategy based on the identification argument of [16] and [17] 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

[19] The 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 contols available for FHFA house prices, non-metro house prices, Case-Shiller house prices, and GDP respectively.

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 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, there will clearly be some instruments which are more worthwhile 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 potential instruments for FHFA prices, non-metro prices, Case-Shiller, and GDP respectively that we select among using LASSO.

Table 3 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 [17].[20] We consider this the baseline. The second set of instruments are those selected through LASSO using the refined data-driven penalty.[21] The number of instruments selected by LASSO is reported in the row "S". In all cases, we use heteroscedasticity consistent standard error estimators. We use the Post-LASSO 2SLS estimator and report these results in the rows labeled "Post-LASSO". Finally, we report the value of a Wald test comparing

[20] [17] used two variables motivated on intuitive grouds, 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.

[21] LASSO selects the number of panels with at least one appointee whose law degree is from a public university (Public) cubed for GDP and FHFA. In the Case-Shiller data, LASSO selects Public and Public squared. For non-metro prices, LASSO selects Public interacted with the number of panels with at least one member who reports belonging to a mainline protestant religion, Public interacted with the number of panels with at least one appointee whose BA was obtained in-state (In-State), In-State interacted with the number of panels with at least one non-white appointee, and the interaction of the number of panels with at least one democrat appointee with the number of panels with at least one Jewish appointee.

the difference between the estimate of β obtained using the [17] instruments and the Post-LASSO estimate of β in the row labeled "Hausman."

The most interesting results from the standpoint of the present paper are found by comparing first-stage Wald-statistics and estimated standard errors across the instrument sets. The LASSO instruments are clearly much better first-stage predictors as measured by the first-stage Wald-statistic. The Wald-statistics increase in all cases which obviously corresponds to more significant first-stage relationships for FHFA prices, GDP, and the Case-Shiller prices compared to the benchmark of the [17] instruments. In the non-metro case, the p-value from the Wald test with the [17] instruments is larger than that of the LASSO-selected instruments. This improved first-stage prediction is associated with the resulting 2SLS estimator having smaller estimated standard errors than the benchmark case for non-metro prices, Case-Shiller prices, and GDP. The reduction in standard errors is sizable for both non-metro and Case-Shiller. Interestingly, the standard error estimate is somewhat larger in the FHFA case despite the improvement in first-stage prediction. Given that the Post-LASSO first-stage produces a larger first-stage Wald-statistic while choosing fewer instruments than the benchmark suggests that we might prefer the Post-LASSO results in any case. We also see that the t-statistics for testing the difference between the estimate using the [17] instruments and the Post-LASSO estimate is equal to zero are uniformly small.

Overall, we find evidence that the effect of takings law decisions on contemporaneous property price indices is small but positive in two of the price-indexes (non-metro and Case-Shiller) while one cannot reject the hypothesis of no-effect in the FHFA or GDP data at usual significance levels. The results are consistent with the developed asymptotic theory in that the 2SLS point-estimates based on the benchmark instruments are similar to the estimates based on the LASSO-selected instruments while LASSO produces a stronger first-stage relationship and the Post-LASSO estimates are more precise in three of the four cases. The example suggests 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 see that the LASSO-based results using the data-dependent penalty may potentially provide substantial reductions in estimated standard errors and consequently allow one to draw more precise conclusions about the effects of judicial decisions relative to a baseline obtained using the instruments of [17]. 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 [18].

Let $X_1, ..., X_n$ 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 $0 < \delta \leqslant 1$ set

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

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

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

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 $X_{1,n}, ..., X_{n,n}$ be the triangular array of i.i.d, zero-mean random variables. Suppose that*

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

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

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

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

$$\left| \frac{\mathrm{P}(|S_{n,n}/V_{n,n}| \geqslant x)}{2\bar{\Phi}(x)} - 1 \right| \leqslant \frac{A}{\ell_n^3} \to 0.$$

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. □

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 $l = 1, \ldots, k_e$, consider the following weighted restricted eigenvalue

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

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

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

for every $C > 0$ we have

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

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 $\lambda/n \geqslant c\|S_l\|_\infty$, and $\widehat{\Upsilon}_l$ satisfy (3.24) with $u \geqslant 1 \geqslant \ell > 1/c$ then*

$$\|f_i'(\widehat{\beta}_l - \beta_{l0})\|_{2,n} \leqslant \left(u + \frac{1}{c}\right)\frac{\lambda\sqrt{s}}{n\kappa_{c_0}^l} + 2c_s,$$

$$\|\widehat{\Upsilon}_l^0(\widehat{\beta}_l - \beta_{l0})\|_1 \leqslant 3c_0\frac{\sqrt{s}}{\kappa_{2c_0}^l}\left((u + [1/c])\frac{\lambda\sqrt{s}}{n\kappa_{c_0}^l} + 2c_s\right) + \frac{3c_0 n}{\lambda}c_s^2,$$

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

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

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

and we also have

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

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

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

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

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

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

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

Thus, relation (B.29) implies

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

and the result follows.

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

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

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

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

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 $\|f_i'\delta_l\|_{2,n} \leqslant 2c_s$ by (B.29). Moreover,

$$\begin{aligned}
\|\widehat{\Upsilon}_l^0\delta_{lT_l^c}\|_1 &\leqslant_{(1)} c_0\|\widehat{\Upsilon}_l^0\delta_{lT_l}\|_1 + \tfrac{c}{\ell c-1}\tfrac{n}{\lambda}\|f_i'\delta_l\|_{2,n}(2c_s - \|f_i'\delta_l\|_{2,n}) \\
&\leqslant_{(2)} c_0\|\widehat{\Upsilon}_l^0\delta_{lT_l}\|_1 + \tfrac{c}{\ell c-1}\tfrac{n}{\lambda}c_s^2 \\
&\leqslant_{(3)} \tfrac{1}{2}\|\widehat{\Upsilon}_l^0\delta_{lT_l^c}\|_1 + \tfrac{c}{\ell c-1}\tfrac{n}{\lambda}c_s^2,
\end{aligned} \tag{B.31}$$

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

Thus,

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

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

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

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

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 $n \to \infty$ and $p \to \infty$*

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

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

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

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

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

$$\mathrm{P}\left(\max_{1\leqslant l\leqslant k_e}\sqrt{n}\|S_l\|_\infty \geqslant 2\sqrt{2\log(2pk_e/a)}\right) \leqslant \frac{a(1+o(1))}{\sqrt{2\log(2pk_e/a)}}, \tag{B.32}$$

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

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

To prove the bound, note that

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

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

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

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

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:

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

This follows from

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

Indeed, application of Lemma 5, gives us that

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

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

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

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

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

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

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 $\|D_i - f_i'\beta_{l0}\|_{2,n} \leqslant c_s \lesssim_P \sqrt{s/n}$ holding by Condition AS, the first result

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

[CHECK FROM HERE]

To prove the second result first assume $\|\widehat{\Upsilon}_l^0 \delta_{lT_l^c}\|_1 \leqslant 2c_0 \|\widehat{\Upsilon}_l^0 \delta_{lT_l}\|_1$. In this case the result follows from the definition of $\widetilde{\kappa}_{2c_0}(\mathbb{E}_n[f_i f_i'])$ which by Condition RE is bounded away from zero with probability approaching 1, thus

$$\|\widehat{\beta}_l - \beta_{l0}\|_2 \leqslant \|\widehat{\beta}_l - \beta_{l0}\|_{2,n}/\widetilde{\kappa}_{2c_0} \lesssim_P \sqrt{\frac{s\log p}{n}}.$$

On the other hand, consider the case that $\|\widehat{\Upsilon}_l^0 \delta_{lT_l^c}\|_1 > 2c_0 \|\widehat{\Upsilon}_l^0 \delta_{lT_l}\|_1$ which by (B.29) $\|f_i'\delta_l\|_{2,n} \leqslant 2c_s$ and by (B.31)

$$\|\widehat{\Upsilon}_l^0 \delta_{lT_l^c}\|_1 \leqslant \frac{2c}{\ell c - 1}\frac{n}{\lambda}c_s^2 \quad \text{so that} \quad \|\delta_{lT_l^c}\|_1 \leqslant \|(\widehat{\Upsilon}_l^0)^{-1}\|_\infty \frac{2c}{\ell c - 1}\frac{n}{\lambda}c_s^2. \qquad \text{(B.33)}$$

Let T_l^1 denote the m largest components of $\delta_{lT_l^c}$. Moreover, let $T_l^c = \cup_{k=1}^K T_l^k$ where $K = \lceil (p - |T_l|)/m \rceil$, $|T_l^k| \leqslant m$ and T_l^k corresponds to the m largest components of δ_l outside $T_l \cup (\cup_{d=1}^{k-1} T_l^d)$. This construction implies that

$$\|\delta_{lT_l^{k+1}}\| \leqslant \|\delta_{lT_l^k}\|_1/\sqrt{m}. \tag{B.34}$$

Indeed, consider the problem $\max\{\|v\|_2/\|u\|_1 : v, u \in \mathbb{R}^m, \max_i |v_i| \leqslant \min_i |u_i|\}$. Given a v and u we can always increase the objective function by using $\tilde{v} = \max_i |v_i|(1,\ldots,1)'$ and $\tilde{u}' = \min_i |u_i|(1,\ldots,1)'$ instead. Thus, the maximum is achieved at $v^* = u^* = (1,\ldots,1)'$, yielding $1/\sqrt{m}$.

We will bound $\|\delta_l\|_2 \leqslant \|\delta_{l(T_l\cup T_l^1)}\|_2 + \|\delta_{l(T_l\cup T_l^1)^c}\|_2$ as follows. First, by (B.34) we have

$$\|\delta_{l(T_l\cup T_l^1)^c}\|_2 \leqslant \sum_{k=2}^K \|\delta_{lT_l^k}\|_2 \leqslant \sum_{k=1}^{K-1} \frac{\|\delta_{lT_l^k}\|_1}{\sqrt{m}} \leqslant \frac{\|\delta_{lT_l^c}\|_1}{\sqrt{m}}.$$

Second,

$$\|\delta_{l(T_l\cup T_l^1)}\|_2 \leqslant \frac{\|\delta_{l(T_l\cup T_l^1)}\|_{2,n}}{\sqrt{\phi_{\min}(s+m)}} \leqslant \frac{\|\delta_l\|_{2,n} + \|\delta_{l(T_l\cup T_l^1)^c}\|_{2,n}}{\sqrt{\phi_{\min}(s+m)}} \leqslant \frac{2c_s + \|\delta_{l(T_l\cup T_l^1)^c}\|_{2,n}}{\sqrt{\phi_{\min}(s+m)}}$$

where we have

$$\|\delta_{l(T_l\cup T_l^1)^c}\|_{2,n} \leqslant \sum_{k=2}^K \|\delta_{lT_l^k}\|_{2,n} \leqslant \sqrt{\phi_{\max}(m)} \sum_{k=2}^K \|\delta_{lT_l^k}\|_2 \leqslant \sqrt{\phi_{\max}(m)} \frac{\|\delta_{lT_l^c}\|_1}{\sqrt{m}}.$$

The result follows by taking $m = s$, Condition SE, (B.33), and noting that $nc_s^2/\lambda \lesssim s/\sqrt{n}$ by Condition AS.

Finally, we obtain the third result follows by the second result in Lemma 6

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

which gives the result. □

[CHECK UNTIL HERE]

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 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, $\lambda/n \geqslant c\|S_l\|_\infty$, and $\widehat{\Upsilon}_l$ satisfies (3.24) with $u \geqslant 1 \geqslant \ell > 1/c$ in the first stage for LASSO for every $l = 1, \ldots, k_e$. Then we have*

$$\max_{1\leqslant l\leqslant k_e} \|f_i'(\widehat{\beta}_{lPL} - \beta_{l0})\|_{2,n} \lesssim_P \frac{\|\widehat{\Upsilon}^0_l\|_\infty}{\sqrt{\phi_{\min}(\widehat{m}_l + s)}} \left(\sqrt{k_e \wedge \log(sk_e)}\sqrt{\frac{s}{n}} + \sqrt{\frac{\widehat{m}_l \log(pk_e)}{n}} \right) + 2c_s +$$
$$+ \max\nolimits_{1\leqslant l\leqslant k_e} 1\{T_l \not\subseteq \widehat{T}_l\} \left(\frac{4u^2 s\lambda^2}{n^2(\kappa^l_{(u/\ell)})^2} + \frac{2uc_s\sqrt{s}\lambda}{n\kappa^l_{(u/\ell)}} \right)^{1/2}$$

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

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

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

Next note that the least squares criterion function satisfies

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

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

By Lemma 7, we have

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

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

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

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

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

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

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

solving which we obtain the stated result:

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

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

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

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

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

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

The second statement of the theorem follows from

$$\|\widehat{\Upsilon}_l\delta_l\|_1 \leqslant \|\widehat{\Upsilon}_l\|_\infty\|\delta_l\|_1 \leqslant \|\widehat{\Upsilon}_l\|_\infty\sqrt{\|\delta_l\|_0}\|\delta_l\|_2 \leqslant \|\widehat{\Upsilon}_l\|_\infty\sqrt{\|\delta_l\|_0}\|f_i'\delta_l\|_{2,n}/\sqrt{\phi_{\min}(\|\delta_l\|_0)},$$

and noting that $\|\delta_l\|_0 \leqslant \widehat{m}_l + s$ and $\|\widehat{\Upsilon}_l\|_\infty \leqslant \|\widehat{\Upsilon}_l^0\|_\infty + \|\widehat{\Upsilon}_l - \widehat{\Upsilon}_l^0\|_\infty$. □

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 $\widehat{T}_l$ denote the support selected by the LASSO estimator, $\widehat{m}_l = |\widehat{T}_l \setminus T_l|$, and assume that $\lambda/n \geqslant c \cdot \|S_l\|_\infty$ and $u \geqslant 1 \geqslant \ell > 1/c$ as in Lemma 6. Then, for $c_0 = (uc+1)/(\ell c - 1)$ we have*

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

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

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

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

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

where we used that

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

and similarly

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

Since $\lambda/c \geqslant n\|S_l\|_\infty$, and by Lemma 6, $\|\widehat{\beta}_l - \beta_{l0}\|_{2,n} \leqslant \left(u + \frac{1}{c}\right)\frac{\lambda\sqrt{s}}{n\kappa^l_{c_0}} + 2c_s$ we have

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

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

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

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 $\lambda/n \geqslant c\|S_l\|_\infty$, and let $\widehat{m}_l := |\widehat{T}_l \setminus T_l|$. Consider the set*

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

Then,

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

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

$$\widehat{m}_l \leqslant \phi_{\max}(\widehat{m}_l)\|(\widehat{\Upsilon}^0_l)^{-1}\|^2_\infty \left[\frac{2c_0}{\kappa^l_{c_0}} + \frac{6c_0 n c_s}{\lambda\sqrt{s}}\right]^2. \tag{C.35}$$

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

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

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

$$M \leqslant s\cdot 2\phi_{\max}(M)\|(\widehat{\Upsilon}^0_l)^{-1}\|^2_\infty \left[\frac{2c_0}{\kappa^l_{c_0}} + \frac{6c_0 n c_s}{\lambda\sqrt{s}}\right]^2$$

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

In turn, applying (C.35) once more with $\widehat{m} \leqslant (M \wedge n)$ we obtain

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

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

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

$$\widehat{m}_l \lesssim_P s \tag{C.36}$$

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

$$\max_{1\leqslant l \leqslant k_e} \|f_i'(\widehat{\beta}_{lPL} - \beta_{l0})\|_{2,n} \lesssim_P \sqrt{\frac{s \log p}{n}} + c_s + \max_{1\leqslant l \leqslant k_e} \frac{\lambda\sqrt{s}}{n\kappa^l_{(u/\ell)}}.$$

By the choice of $\lambda = 2c\sqrt{2n\log(2pk_e)}$, obtained in Lemma 7, and that $1/\kappa^l_{(u/\ell)} \lesssim_P 1$ for $l = 1, \ldots, k_e$ 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 $\lambda/n \geqslant c \max_{1\leqslant l \leqslant k_e} \|S_l\|_\infty$ 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.36). □

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 $\tilde{d}_{il} = d_{il} - \mathrm{E}[d_{il}]$. Here we consider the basic option, in which

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

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

$$\Delta_1 = \max_{1\leqslant l\leqslant k_e, 1\leqslant j\leqslant p} |\widehat{\gamma}_{jl}^2 - \tilde{\gamma}_{jl}^2| \to_P 0, \ \Delta_2 = \max_{1\leqslant l\leqslant k_e, 1\leqslant j\leqslant p} |\tilde{\gamma}_{jl}^2 - \gamma_{jl}^2| \to_P 0,$$

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

We have that

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

Indeed, we have for the first term that, $\max_{1\leqslant l\leqslant k_e, 1\leqslant j\leqslant p} |\mathbb{E}_n[f_{ij}^2\tilde{d}_{il}]| \leqslant \max_{1\leqslant j\leqslant p} \sqrt{\mathbb{E}_n[f_{ij}^4]} \max_{1\leqslant l\leqslant k_e} \sqrt{\mathbb{E}_n[\tilde{d}_{il}^2]} \lesssim_P 1$ by the assumption on the empirical moments of f_{ij} and the Markov inequality and by $\mathrm{Var}(d_{il})$ being uniformly bounded in n and l by assumption; also recall that k_e is fixed. Moreover, $\max_{1\leqslant l\leqslant k_e} |\mathbb{E}_n\tilde{d}_{il}| \lesssim_P \sqrt{k_e}/\sqrt{n}$ by the Chebyshev inequality and by $\mathrm{Var}(d_{il})$ being uniformly bounded by Condition RF. Likewise, the second term vanishes by a similar argument.

Furthermore,

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

Indeed, application of Lemma 5 gives us that

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

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

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

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

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

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

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

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

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

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

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

$$\|\widehat{D}_i - D_i\|_{2,n} \lesssim_P \sqrt{\frac{s \log p}{n}}. \tag{D.37}$$

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.37), are asymptotically valid. Thus by Theorem 4 and 5, the LASSO and Post-LASSO estimators based on the refined option also obey (D.37). 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 $\gamma_{jl}^{02} := \mathrm{E}[f_{ij}^2 v_{il}^2]$, which is bounded away from zero and from above by assumption. Hence it suffices to show that $\max_{1\leqslant j\leqslant p, 1\leqslant l\leqslant k_e} |\widehat{\gamma}_{jl}^2 - \gamma_{jl}^{02}| \to_P 0$. This in turn follows from

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

which we establish below.

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

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

The first term is bounded by

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

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

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

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 $\log p \frac{s\log p}{n} \to 0$. 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 $m > 0$, $\mathrm{E}[|f_{ij}|^m]$ is bounded, uniformly in $1 \leqslant j \leqslant p$, uniformly in n, for both the Gaussian regressors of Lemma 1 and arbitrary bounded regressors of Lemma 2. □

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:

$$\max_{1\leqslant l\leqslant k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n} \lesssim_P \sqrt{\frac{s\log p}{n}} \to 0 \tag{F.38}$$

$$\sqrt{\log p}\|\widehat{\beta}_l - \beta_{l0}\|_1 \lesssim_P \sqrt{\frac{s^2\log^2 p}{n}} \to 0 \tag{F.39}$$

In order to prove Theorem 5 we need also the condition

$$\max_{1\leqslant l\leqslant k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n}^2 n^{2/q_\epsilon} \lesssim_P \frac{s\log p}{n} n^{2/q_\epsilon} \to 0, \tag{F.40}$$

with the last statement holding by Condition SM. Note that Theorem 6 assumes (F.38) -(F.39) as high level conditions.

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

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

where by Steps 2 and 3 below:

$$\mathbb{E}_n[\widehat{D}_i d_i'] = \mathrm{E}[D_i d_i'] + o_P(1) \tag{F.41}$$

$$\mathbb{G}_n[\widehat{D}_i\epsilon_i] = \mathbb{G}_n[D_i\epsilon_i] + o_P(1) \tag{F.42}$$

where $\mathrm{E}[D_i d_i'] = \mathrm{E}[D_i D_i'] = Q$ is bounded away from zero and bounded from above in the matrix sense, uniformly in n. Moreover, $\mathrm{Var}(\mathbb{G}_n[D_i\epsilon_i]) = \Omega$ where $\Omega = \sigma^2\mathrm{E}[D_i D_i']$ under homoscedasticity and $\Omega = \mathrm{E}[\epsilon_i^2 D_i D_i']$ 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 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 $\delta > 0$

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

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 $Z_n \to_d N(0,I)$.

Step 2. To show (F.41), note that

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

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

Moreover,

$$\mathbb{E}_n[D_iD_i'] - \mathrm{E}[D_iD_i'] \to_P 0$$

by Rosenthal's [35] inequality using that $\mathrm{E}\|D_i\|^q$ for $q > 2$ is bounded uniformly in n.

Step 3. To show (F.42), note that

$$\begin{array}{l} \max_{1\leqslant l\leqslant k_e}|\mathbb{G}_n[(\widehat{D}_{il} - D_{il})\epsilon_i]| \\ = \max_{1\leqslant l\leqslant k_e}|\mathbb{G}_n\{f_i'(\widehat{\beta}_l - \beta_{l0})\epsilon_i\} + \mathbb{G}_n\{a_{il}\epsilon_i\}| \\ = \max_{1\leqslant l\leqslant k_e}|\sum_{j=1}^{p}\mathbb{G}_n\{f_{ij}\epsilon_i\}'(\widehat{\beta}_{lj} - \beta_{l0j}) + \mathbb{G}_n\{a_{il}\epsilon_i\}| \\ \leqslant \max_{1\leqslant j\leqslant p}\left|\frac{\mathbb{G}_n[f_{ij}\epsilon_i]}{\sqrt{\mathbb{E}_n[f_{ij}^2\epsilon_i^2]}}\right|\max_{1\leqslant j\leqslant p}\sqrt{\mathbb{E}_n[f_{ij}^2\epsilon_i^2]}\max_{1\leqslant l\leqslant k_e}\|\widehat{\beta}_l - \beta_{l0}\|_1 + \max_{1\leqslant l\leqslant k_e}|\mathbb{G}_n\{a_{il}\epsilon_i\}|. \end{array}$$

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

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

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:

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

and in the boundedly heteroscedastic case of Theorem 5:

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

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

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

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

$$\begin{aligned}
&\mathrm{P}\left(\max_{1\leqslant j\leqslant p}\left|\frac{\mathbb{G}_n[f_{ij}\epsilon_i]}{\sqrt{\mathbb{E}_n[f_{ij}^2\epsilon_i^2]}}\right| \geqslant \sqrt{2\log(2p/a)}\right)\\
&\leqslant_{(1)} p\max_{1\leqslant j\leqslant p}\mathrm{P}\left(\frac{|\mathbb{G}_n(f_{ij}\epsilon_i)|}{\sqrt{\mathbb{E}_n[f_{ij}^2\epsilon_i^2]}} > \sqrt{2\log(2p/a)}\right)\\
&\leqslant_{(2)} p2\bar{\Phi}(\sqrt{2\log(2p/a)})(1+o(1)) \leqslant_{(3)} a(1+o(1)),
\end{aligned}$$

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

$$2\log(2p/a) \leqslant \frac{n^{1/3}}{b_n}\min_{1\leqslant j\leqslant p} M_{j0}^2, \quad M_{j0} := \frac{\mathrm{E}[f_{ij}^2\epsilon_i^2]^{1/2}}{\mathrm{E}[|f_{ij}|^3|\epsilon_i|^3]^{1/3}}. \tag{F.43}$$

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 $\bar{\Phi}(t) \leqslant \phi(t)/t$. Finally, by Condition SM $\min_{1\leqslant j\leqslant p} M_{j0}^2$ is bounded away from zero, so the condition (F.43) is satisfied asymptotically for some $b_n \to \infty$ and some $a \to 0$ provided $\log p = o(n^{1/3})$.

Finally, we have that

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

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

Thus, combining bounds above with bounds in (F.38)-(F.39)

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

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 $Q = \mathrm{E}[D_iD_i']$ are bounded away from zero and from above uniformly in n, it suffices to show $\widehat{\sigma}^2 - \sigma^2 \to_P 0$ and $\mathbb{E}_n[\widehat{D}_i\widehat{D}_i'] - \mathrm{E}[D_iD_i'] \to_P 0$.

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

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

which is bounded up to a constant by

$$\sqrt{k_e}\max_{1\leqslant l\leqslant k_e}\|\widehat{D}_{il} - D_{il}\|_{2,n}\|D_i\|_{2,n} + k_e\max_{1\leqslant l\leqslant k_e}\|\widehat{D}_{il} - D_{il}\|_{2,n}^2 \to_P 0$$

by (F.38) and by $\|D_i\|_{2,n} \lesssim_P 1$ holding by Markov inequality. Moreover, $\mathbb{E}_n[D_iD_i'] - \mathrm{E}[D_iD_i'] \to_P 0$ by Step 2.

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

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

First, we note

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

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

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

which is bounded up to a constant by

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

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

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

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

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. Least squares after model selection in high-dimensional sparse 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] F. Bunea, A. Tsybakov, and M. H. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[10] 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.

[11] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.

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

[13] G. Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34:305–334, 1987.

[14] G. Chamberlain and G. Imbens. Random effects estimators with many instrumental variables. *Econometrica*, 72:295–306, 2004.

[15] J. Chao and N. Swanson. Consistent estimation with a large number of weak instruments. *Econometrica*, 73:1673–1692, 2005.

[16] D. L. Chen and J. Sethi. Does forbidding sexual harassment exacerbate gender inequality. unpublished manuscript, 2010.

[17] D. L. Chen and S. Yeh. The economic impacts of eminent domain. unpublished manuscript, 2010.

[18] 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.

[19] Stephen G. Donald and Whitney K. Newey. Choosing the number of instruments. *Econometrica*, 69(5):1161–1191, 2001.

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

[21] Wayne A. Fuller. Some properties of a modification of the limited information estimator. *Econometrica*, 45:939–954, 1977.

[22] 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.

[23] Christian Hansen, Jerry Hausman, and Whitney K. Newey. Estimation with many instrumental variables. *Journal of Business and Economic Statistics*, 26:398–422, 2008.

[24] J. Hausman, W. Newey, T. Woutersen, J. Chao, and N. Swanson. Instrumental variable estimation with heteroskedasticity and many instruments. mimeo, 2009.

[25] Jian Huang, Joel L. Horowitz, and Fengrong Wei. Variable selection in nonparametric additive models. *Ann. Statist.*, 38(4):2282–2313, 2010.

[26] Frank Kleibergen. Testing parameters in gmm without assuming that they are identified. *Econometrica*, 73:1103–1124, 2005.

[27] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincar Probab. Statist.*, 45(1):7–57, 2009.

[28] K. Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electron. J. Statist.*, 2:90–102, 2008.

[29] 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.

[30] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):2246–2270, 2009.

[31] Whitney K. Newey. Efficient instrumental variables estimation of nonlinear models. *Econometrica*, 58:809–837, 1990.

[32] Whitney K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79:147–168, 1997.

[33] R. Okui. Instrumental variable estimation in the presence of many moment conditions. *forthcoming Journal of Econometrics*, 2010.

[34] M. Rosenbaum and A. B. Tsybakov. Sparse recovery under matrix uncertainty. *arXiv:0812.2818v1 [math.ST]*, 2008.

[35] 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.

[36] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61:10251045, 2008.

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

[38] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.

[39] 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.

[40] 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. Simulation Results. Corr(e,v) = .3

		Exponential				S = 5				S = 50		
		Median				Median				Median		
Estimator	N(0)	Bias	MAD	rp(.05)	N(0)	Bias	MAD	rp(.05)	N(0)	Bias	MAD	rp(.05)
						A. F* = 10, N = 100						
2SLS(100)		0.168	0.168	0.758		0.110	0.110	0.728		0.024	0.024	0.256
LIML(100)		0.113	0.527	0.122		0.095	0.373	0.124		0.004	0.079	0.062
FULL(100)		0.113	0.526	0.122		0.095	0.373	0.124		0.004	0.079	0.062
Post-LASSO	77	0.042	0.085	0.056	232	0.026	0.059	0.048	500	*	*	0.000
Post-LASSO-F	13	0.068	0.094	0.092	22	0.040	0.060	0.090	468	0.002	0.016	0.008
						B. F* = 10, N = 250						
2SLS(100)		0.104	0.104	0.774		0.071	0.071	0.718		0.015	0.015	0.248
LIML(100)		0.020	0.126	0.034		0.006	0.071	0.038		0.000	0.011	0.032
FULL(100)		0.021	0.123	0.034		0.007	0.070	0.038		0.000	0.011	0.032
Post-LASSO	90	0.024	0.054	0.036	253	0.019	0.036	0.046	500	*	*	0.000
Post-LASSO-F	34	0.043	0.055	0.058	88	0.022	0.036	0.076	92	0.002	0.009	0.030
						C. F* = 10, N = 500						
2SLS(100)		0.074	0.074	0.758		0.050	0.050	0.750		0.011	0.012	0.262
LIML(100)		0.003	0.082	0.040		0.004	0.048	0.056		0.000	0.007	0.066
FULL(100)		0.005	0.079	0.036		0.005	0.047	0.050		0.000	0.007	0.066
Post-LASSO	105	0.015	0.040	0.038	306	0.014	0.024	0.030	500	*	*	0.000
Post-LASSO-F	52	0.027	0.039	0.060	136	0.016	0.026	0.070	0	0.003	0.007	0.066
						D. F* = 40, N = 100						
2SLS(100)		0.193	0.193	0.530		0.110	0.110	0.432		0.015	0.018	0.114
LIML(100)		0.070	0.718	0.082		0.098	0.367	0.060		0.012	0.049	0.034
FULL(100)		0.070	0.718	0.082		0.098	0.367	0.060		0.012	0.049	0.034
Post-LASSO	0	0.025	0.088	0.066	0	0.012	0.051	0.046	500	*	*	0.000
Post-LASSO-F	0	0.038	0.090	0.056	0	0.023	0.052	0.056	400	0.004	0.016	0.010
						E. F* = 40, N = 250						
2SLS(100)		0.115	0.115	0.476		0.070	0.070	0.416		0.010	0.012	0.106
LIML(100)		-0.008	0.082	0.046		0.000	0.046	0.044		0.001	0.010	0.048
FULL(100)		-0.007	0.081	0.044		0.001	0.046	0.044		0.001	0.010	0.048
Post-LASSO	0	0.012	0.054	0.060	0	0.004	0.033	0.052	500	*	*	0.000
Post-LASSO-F	0	0.021	0.053	0.064	0	0.011	0.033	0.056	0	0.002	0.010	0.056
						F. F* = 40, N = 500						
2SLS(100)		0.085	0.085	0.532		0.048	0.048	0.426		0.006	0.008	0.108
LIML(100)		0.003	0.049	0.044		0.001	0.029	0.034		0.000	0.007	0.066
FULL(100)		0.004	0.049	0.044		0.002	0.029	0.036		0.000	0.007	0.064
Post-LASSO	0	0.007	0.037	0.050	0	0.006	0.023	0.036	500	*	*	0.000
Post-LASSO-F	0	0.013	0.038	0.070	0	0.010	0.023	0.046	0	0.002	0.007	0.058

Note: Results are based on 500 simulation replications and 100 instruments. Column labels indicate the structure of the first-stage coefficients as described in the text. 2SLS(100), LIML(100), and FULL(100) are respectively the 2SLS, LIML, and Fuller(1) estimators using all 100 potential instruments. Post-LASSO and Post-LASSO-F respectively correspond to Post-LASSO-IV based on LASSO variable selection among the 100 instruments and and LASSO variable selection among the 120 instruments formed by augmenting the original 100 instruments with the first 20 principal components. We report the number of replications in which LASSO selected no instruments (N(0)), median bias (Med. Bias), mean absolute deviation (MAD), and rejection frequency for 5% level tests (rp(.05)). In cases where LASSO selects no instruments, Med. Bias, and MAD use only the replications where LASSO selects a non-empty set of instruments, and we set the confidence interval eqaul to (-∞,∞) and thus fail to reject.

Table 2. Simulation Results. Corr(e,v) = .6

	Exponential				S = 5				S = 50			
		Median				Median				Median		
Estimator	N(0)	Bias	MAD	rp(.05)	N(0)	Bias	MAD	rp(.05)	N(0)	Bias	MAD	rp(.05)
A. F* = 10, N = 100												
2SLS(100)		0.335	0.335	0.998		0.221	0.221	1.000		0.049	0.049	0.702
LIML(100)		0.298	0.588	0.210		0.191	0.397	0.194		0.016	0.080	0.098
FULL(100)		0.298	0.588	0.210		0.191	0.396	0.194		0.016	0.080	0.098
Post-LASSO	93	0.060	0.094	0.114	215	0.040	0.058	0.070	500	*	*	0.000
Post-LASSO-F	12	0.129	0.144	0.244	34	0.079	0.081	0.224	477	-0.005	0.014	0.002
B. F* = 10, N = 250												
2SLS(100)		0.208	0.208	1.000		0.139	0.139	0.998		0.030	0.030	0.712
LIML(100)		0.015	0.119	0.076		0.001	0.057	0.070		0.000	0.010	0.058
FULL(100)		0.019	0.114	0.076		0.004	0.056	0.072		0.000	0.010	0.060
Post-LASSO	78	0.052	0.069	0.144	249	0.025	0.037	0.050	500	*	*	0.000
Post-LASSO-F	29	0.086	0.090	0.282	78	0.048	0.049	0.170	94	0.004	0.011	0.046
C. F* = 10, N = 500												
2SLS(100)		0.150	0.150	1.000		0.096	0.096	0.996		0.020	0.020	0.656
LIML(100)		0.004	0.077	0.068		-0.001	0.037	0.060		-0.001	0.007	0.058
FULL(100)		0.008	0.073	0.070		0.002	0.038	0.060		-0.001	0.007	0.054
Post-LASSO	106	0.038	0.048	0.132	324	0.019	0.023	0.044	500	*	*	0.000
Post-LASSO-F	46	0.061	0.064	0.236	118	0.029	0.031	0.154	1	0.003	0.007	0.088
D. F* = 40, N = 100												
2SLS(100)		0.381	0.381	0.988		0.221	0.221	0.948		0.027	0.027	0.264
LIML(100)		0.158	0.607	0.140		0.054	0.295	0.144		0.002	0.041	0.058
FULL(100)		0.158	0.607	0.140		0.054	0.295	0.144		0.002	0.041	0.058
Post-LASSO	0	0.052	0.094	0.096	0	0.014	0.055	0.074	500	*	*	0.000
Post-LASSO-F	0	0.082	0.101	0.136	0	0.041	0.060	0.116	391	0.002	0.017	0.014
E. F* = 40, N = 250												
2SLS(100)		0.244	0.244	0.994		0.138	0.138	0.964		0.017	0.017	0.278
LIML(100)		-0.003	0.072	0.044		-0.001	0.036	0.024		-0.001	0.010	0.048
FULL(100)		-0.001	0.072	0.044		0.001	0.036	0.022		-0.001	0.010	0.048
Post-LASSO	0	0.031	0.061	0.098	0	0.012	0.033	0.056	500	*	*	0.000
Post-LASSO-F	0	0.052	0.065	0.134	0	0.026	0.038	0.082	0	0.003	0.009	0.048
F. F* = 40, N = 500												
2SLS(100)		0.172	0.172	0.988		0.097	0.097	0.950		0.012	0.012	0.248
LIML(100)		0.008	0.044	0.064		-0.001	0.026	0.040		0.000	0.006	0.046
FULL(100)		0.010	0.044	0.064		0.000	0.026	0.040		0.000	0.006	0.042
Post-LASSO	0	0.020	0.042	0.078	0	0.008	0.023	0.064	497	-0.005	0.008	0.000
Post-LASSO-F	0	0.033	0.044	0.156	0	0.016	0.025	0.084	0	0.002	0.007	0.056

Note: Results are based on 500 simulation replications and 100 instruments. Column labels indicate the structure of the first-stage coefficients as described in the text. 2SLS(100), LIML(100), and FULL(100) are respectively the 2SLS, LIML, and Fuller(1) estimators using all 100 potential instruments. Post-LASSO and Post-LASSO-F respectively correspond to Post-LASSO-IV based on LASSO variable selection among the 100 instruments and and LASSO variable selection among the 120 instruments formed by augmenting the original 100 instruments with the first 20 principal components. We report the number of replications in which LASSO selected no instruments (N(0)), median bias (Med. Bias), mean absolute deviation (MAD), and rejection frequency for 5% level tests (rp(.05)). In cases where LASSO selects no instruments, Med. Bias, and MAD use only the replications where LASSO selects a non-empty set of instruments, and we set the confidence interval eqaul to (-∞,∞) and thus fail to reject.

Table 3: Effect of Federal Appellate Takings Law Decisions on Economic Outcomes

	Home Prices			GDP
	log(FHFA)	log(Non-Metro)	log(Case-Shiller)	log(GDP)
Sample Size	312	110	183	312
OLS	0.0114	0.0108	0.0152	0.0099
s.e.	0.0132	0.0066	0.0132	0.0048
2SLS	0.0262	0.0480	0.0604	0.0165
s.e.	0.0441	0.0212	0.0296	0.0162
FS-W	28.0859	82.9647	67.7452	28.0859
Post-LASSO	0.0369	0.0357	0.0631	0.0133
s.e.	0.0465	0.0132	0.0249	0.0161
FS-W	44.5337	243.1946	89.5950	44.5337
S	1	4	2	1
Hausman Test	-0.2064	0.5753	-0.0985	0.1754

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(FHFA), log(Non-Metro), and log(Case-Shiller) are within-circuit averages of log-house-price-indexes, and log(GDP) is the within-circuit average of 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-W is the value of the first-stage Wald statistic using the selected instrument. S is the number of instruments chosen by LASSO. Hausman test is the value of a Hausman test statistic comparing the 2SLS estimate of the effect of takings law decisions using the Chen and Yeh (2010) instruments to the estimated effect using the LASSO-selected instruments.