

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 limiting theory for the resulting IV estimators and provide conditions under which these estimators are oracle-efficient. Optimal instruments are conditional expectations; and in developing the IV results, we provide several new results for LASSO and post-LASSO estimation of conditional expectation functions, which can be of independent theoretical and practical interest. Specifically, we develop rate results for these estimators under non-Gaussian and heteroscedastic disturbances, obtaining rates of convergence that are as sharp as in the Gaussian case under the weak restriction that $(\log p)^3 = o(n)$ by using moderate deviation theory for self-normalized sums. Moreover, we propose and establish asymptotic validity of fully data-driven choices of penalty levels for LASSO. 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.

Key Words: Instrumental Variables, Optimal Instruments, LASSO, Post-LASSO, Sparsity, Eminent Domain

1. INTRODUCTION

Instrumental variables (IV) techniques are widely used in applied economic research. While these methods provide a useful tool for justifying the identification of structural effects of interest, their application often results in imprecise inference. One way to improve the precision of

Date: First version: June 2009, This VERSION OCTOBER 16, 2010.

instrumental variables estimators is to use many instruments or to try to approximate the optimal instrument as in [1], [15], and [35] which will generally be done nonparametrically and thus implicitly make use of many constructed instruments such as polynomials. The promised improvement in efficiency is appealing, but recent work in econometrics has also demonstrated that IV estimators that make use of many instruments may have very poor properties; see, for example, [3], [18], [26], and [28] which propose solutions for this problem based on “many-instrument” asymptotics.

In this paper, we contribute to the literature on IV estimation with many instruments by considering the use of LASSO-based methods for estimating the first-stage relationship between the instruments and the endogenous variables. LASSO is a widely-used shrinkage method in statistics that solves a penalized least-squares problem where, intuitively, the penalty function is large when the model includes many variables.¹ In this way, LASSO methods provide a way to construct parsimonious predictive models that both provide good predictions and rely on a relatively small number of the potentially many available regressors. For theoretical and simulation evidence regarding LASSO’s performance see, for example, [4], [8], [12], [13], [11], [14], [30], [31], [32], [33], [38], [40], [42], [43], and [44] among many others.

The use of LASSO 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 produces first-stage predictions that approximate the optimal instruments and performs well when the optimal instrument may be well-approximated using a small, but unknown, set of the available instruments even when number of potential instruments is allowed to be much larger than the sample size.²

Our paper also contributes to the growing literature on the theoretical properties of LASSO-based methods. We derive the asymptotic properties of the resulting IV estimator and provide

¹[16] consider an alternate shrinkage estimator in the context of IV estimation with many instruments. [2] consider IV estimation with many instruments based on principal components analysis and variable selection via boosting, and [36] provides results for Ridge regression.

²This is in contrast to the variable selection method of [22] which relies on a *a priori* knowledge that allows one to order the instruments in terms of instrument strength.

conditions under which the LASSO predictions approximate the optimal instruments. Under homoskedasticity, the resulting IV estimator achieves the semi-parametric efficiency bound; i.e. it is oracle efficient. In developing the results for the second stage IV estimator, we contribute to the LASSO literature by providing results for LASSO estimation of conditional expectations, of which the optimal instrument in instrumental variables estimation is an important special case in econometrics. We provide rates of convergence allowing for non-Gaussian, heteroscedastic disturbances which generalizes most LASSO results which assume both homoscedasticity and Gaussianity and is important for applied economic analysis where researchers are very concerned about potential heteroscedasticity and non-normality in their data. Using moderate deviation theory for self-normalized sums, we provide rates that are as sharp as in the Gaussian case under the weak condition that $(\log p)^3 = o(n)$. 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 allowing for non-Gaussian, heteroskedastic disturbances. Ours is the first paper to provide such a data-driven penalty selection result which was previously not available even in the Gaussian case.

We illustrate the performance of LASSO-based IV through a series of simulation experiments. In these experiments, we see that a feasible LASSO procedure that uses a data-dependent penalty performs very well across a wide range of simulation designs. In terms of estimation risk, it outperforms both LIML and its modification due to [24] (FULL)³ that have been advocated as procedures that are robust to using many instruments (e.g. [26]). In terms of inference based on 5% level tests, the LASSO-based IV performs comparably to LIML and FULL in the majority of cases, though it does have somewhat larger size distortions in the case of relatively weak instruments and strong endogeneity. Overall, the simulation results are very favorable to the proposed LASSO-based IV procedures.

In addition to the simulations, we consider an application in which there are many available instruments among which there is not a simple *a priori* ordering in terms of which instruments are expected to be strongest. 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

³Note that these procedures are only applicable when the number of instruments is less than or equal to the sample size.

and state-level GDP as in [20]. We follow the identification strategy of [20] who use the random assignment of judges to three judge panels that are then assigned to eminent domain cases to justify using the demographic characteristics of the judges on the realized panels as instruments for their decision. This strategy produces a situation in which there are many potential instruments in that all possible sets of characteristics of the three judge panel are valid instruments. The results of the application of LASSO-based IV in this example clearly demonstrate the potential gains that may be obtained through its use. Specifically, the LASSO-based estimates using the data-dependent penalty have substantially smaller estimated standard errors than estimates obtained using the baseline instruments of [20]. This improvement of precision clearly allows one to draw more precise conclusions about the effects of the judicial decisions on economic outcomes relative to the benchmark case.

Notation. In what follows, all true parameter values, we allow for the models to change with the sample size, i.e. allow for array asymptotics, so all parameters are implicitly indexed by the sample size n , but we omit the index to simplify notation. We need the array asymptotics to better capture some finite-sample phenomena using the asymptotic approximations. We use the notation $(a)_+ = \max\{a, 0\}$, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. The l_2 -norm is denoted by $\|\cdot\|$, and the l_0 -norm, $\|\cdot\|_0$, denotes the number of non-zero components of a vector. Given a vector $\delta \in \mathbb{R}^p$, and a set of indices $T \subset \{1, \dots, p\}$, 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 also use the following empirical process notation,

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

and

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

We use the notation $a \lesssim b$ to denote $a \leq cb$ for some constant $c > 0$ that does not depend on n ; and $a \lesssim_P b$ to denote $a = O_P(b)$. For an event E , we say that E wp $\rightarrow 1$ when E occurs with probability approaching one as n grows. We say $X_n =_d Y_n + o_P(1)$ to mean that X_n has the same distribution as Y_n up to a term $o_P(1)$ that vanishes in probability. Such statements are needed to accommodate statements for models that change with n . When Y_n is a fixed random vector, i.e. $Y_n = Y$, this notation is equivalent to $X_n \rightarrow_d Y$.

2. SPARSE MODELS AND METHODS FOR OPTIMAL INSTRUMENTAL VARIABLES

In this section of the paper, we present the model and methods 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_e -vector of endogenous variables. The disturbance ϵ_i obeys

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

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

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

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

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

minimizes the limit variance which becomes $\Lambda^* = \sigma^2 \{\mathbb{E}[D(x_i) D(x_i)']\}^{-1}$, the semi-parametric efficiency bound for estimating α_0 ; see [1], [15], and [35].

We would like to construct an IV estimator that is as efficient as the infeasible optimal IV estimator above. The optimal instrument $D(x_i)$ is unknown in practice and has to be estimated. In what follows, we investigate the use of sparse methods – namely LASSO and post-LASSO – for this purpose.

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

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

i.e. the estimator of w_i is simply w_i . Therefore, in what follows, we discuss modeling and estimating conditional expectation functions:

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

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

Suppose there is a very large list of technical instruments,

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

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

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

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

- **Many Instruments Case.** The list of available instruments is large, in which case, we have $f_i = x_i$. This case includes e.g. [3] as a special case.
- **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., B-splines, dummies, polynomials, and various interactions. This case includes e.g. [35] as a special case.

We mainly use the term “series instruments” and contrast our results with those in the seminal work of [35], though our results are not limited to canonical series regressors as in [35]. 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 [35] to approximate the optimal instruments.

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

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

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

This simple sparsity model substantially generalizes the classical model of the optimal instrument of [1] by letting the identities of the relevant instruments

$$T_l = \text{support}(\beta_{l0}) = \{j \in \{1, \dots, p\} : |\beta_{l0j}| > 0\}$$

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 an approximately sparse model:

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

$$D_l(x_i) = f'_i \beta_{l0} + a_{il}, \quad l = 1, \dots, k_e, \quad \max_{1 \leq l \leq k_e} [\mathbb{E}_n a_{il}^2]^{1/2} \leq c_s \lesssim_P \sqrt{s/n}, \quad (2.4)$$

$$\max_{1 \leq l \leq k_e} \|\beta_{l0}\|_0 \leq s \ll n, \quad (2.5)$$

This model generalizes [35] by letting the identities of the most important series terms $T_l = \text{support}(\beta_{l0})$ be unknown. 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 K of series terms in [35], $\sqrt{K/n}$, by not relying on knowledge of what K series terms to include. Knowing the identities of the most important series terms is unrealistic in many examples in practice. Indeed, the most important series terms need not be the first s terms, and the optimal number of series terms to consider is also unknown. Moreover, an optimal series approximation to the instrument could come from the combination of completely different bases e.g by using both polynomials and B-splines. LASSO and post-LASSO use the data to estimate the set of the most relevant series terms in a manner that allows the resulting IV estimator to achieve good performance if the following key growth condition holds: $\frac{s^2(\log p)^2}{n} \rightarrow 0$ along with other more technical condition. This condition requires the optimal instrument to be sufficiently smooth within the set of p considered instruments so that a small number of series terms can be used to well-approximate the target function, ensuring the impact of instrument estimation on the IV estimator is asymptotically negligible.

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

$$d_{il} = D_l(x_i) + v_{il}, \quad \mathbb{E}[v_{il}|x_i] = 0, \quad l = 1, \dots, k_e. \quad (2.6)$$

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

$$\widehat{D}_{li} := \widehat{D}_l(x_i) = f_i' \widehat{\beta}_l, \quad l = 1, \dots, k_e, \quad (2.7)$$

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

Consider the usual least squares criterion function:

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

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

$$\widehat{\beta}_{lL} \in \arg \min_{\beta \in \mathbb{R}^p} \widehat{Q}_l(\beta) + \frac{\lambda}{n} \|\widehat{\Upsilon}_l \beta\|_1 \quad (2.8)$$

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

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

$$\begin{aligned} \text{initial} \quad & \widehat{\gamma}_{lj} = \sqrt{\mathbb{E}_n[f_{ij}^2 (d_{il} - \mathbb{E}_n d_{il})^2]}, \quad \lambda = c2 \sqrt{2n \log(2pk_e)}, \\ \text{refined} \quad & \widehat{\gamma}_{lj} = \sqrt{\mathbb{E}_n[f_{ij}^2 \widehat{v}_{il}]}, \quad \lambda = c2 \sqrt{2n \log(2pk_e)}, \end{aligned} \quad (2.9)$$

where $c > 1$ is a constant. 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 $c = 1.1$.

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

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

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

$$\widehat{\beta}_{lPL} \in \arg \min_{\beta \in \mathbb{R}^p: \widehat{\beta}_{T_l^c} = 0} \widehat{Q}_l(\beta), \quad l = 1, \dots, k_e. \quad (2.10)$$

In words, this estimator is ordinary least squares (OLS) applied to the data after removing the instruments 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_0 in the setting with $p > n$. Some classical approaches based on BIC-penalization of model size are consistent but computationally infeasible. The LASSO estimator [40] resolves these difficulties by penalizing 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 $D_l(x_i)$. 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 (2.8) 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 we select 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 instrument of [35] 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 by showing that under possibly heteroscedastic and non-Gaussian reduced form errors the LASSO and Post-LASSO estimators obey the following near-oracle performance bounds:

$$\max_{1 \leq l \leq k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n} \lesssim_P \sqrt{\frac{s \log p}{n}} \quad (2.11)$$

$$\max_{1 \leq l \leq k_e} \|\widehat{\beta}_l - \beta_{l0}\|_1 \lesssim_P \sqrt{\frac{s^2 \log p}{n}}. \quad (2.12)$$

The performance bounds in (2.11) are called near-oracle because they coincide with the bounds achievable when the minimal true models T_l for each of the reduced form equations up to a $\log p$ factor. Our results extend those of [8] for LASSO with Gaussian errors and those of [4] for Post-LASSO with Gaussian errors. These results are notable because they are the first results in the literature that allow for data-driven choice of the penalty level. We achieve this through using moderate deviation theory for self-normalized sums. 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 (2.11).

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 $D_l(x_i)$ of the form

$$\widehat{D}_l(x_i) = f_i' \widehat{\beta}_l, l = 1, \dots, k_e, \quad (2.13)$$

and then set

$$\widehat{D}_i = (\widehat{D}_1(x_i), \dots, \widehat{D}_l(x_i), w_i')'. \quad (2.14)$$

The resulting IV estimator takes the form

$$\widehat{\alpha}^* = \mathbb{E}_n[d_i \widehat{D}_i']^{-1} \mathbb{E}_n[\widehat{D}_i y_i]. \quad (2.15)$$

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

$$\sqrt{n}(\widehat{\alpha}^* - \alpha_0) =_d N(0, \Lambda^*) + o_P(1). \quad (2.16)$$

That is, the LASSO-based and post-LASSO based IV estimator asymptotically achieves oracle performance. Thus the estimator matches the performance of the series-based IV estimator of [35] with the following advantages:

- **Adaptivity to Unknown Smoothness/Sparsity.** The LASSO-based procedures adapt to the unknown smoothness/sparsity of the true optimal instrument D_i and automatically choose the optimal number of series terms. This is in contrast to the classical series procedure that does not adapt to the unknown smoothness and can fail if the incorrect number of terms is chosen. Note that both methods rely on the sufficient smoothness of the optimal instrument.
- **Enhanced Approximation of the Optimal Instrument.** The LASSO-based procedures can make the estimates of the optimal instrument more precise in finite samples since they may select among $p \gg K$ instruments to find the best small set to be used in the approximation of the optimal instrument. To illustrate this with an extreme example, suppose that the optimal instrument obeys $D_i = \sum_{j=1}^p \beta_{0j} f_j$, with $\beta_{0j} = 0$ for $j \leq K$, and $\beta_{0j} \propto j^{-a}$, then the canonical series procedure with K terms will fail to approximate the optimal instrument, but the LASSO-based procedure will deliver near oracle performance if the smoothness index a is not too low.⁴

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 heteroskedasticity:

$$\sqrt{n}(\hat{\alpha} - \alpha_0) =_d N(0, Q^{-1}\Omega Q^{-1}) + o_P(1), \quad (2.17)$$

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

$$\hat{Q}^{-1} \hat{\Omega} \hat{Q}^{-1}, \quad \hat{\Omega} := \{E_n[\epsilon_i^2 \hat{D}(x_i) \hat{D}(x_i)']\}, \quad \hat{Q} := E_n[\hat{D}(x_i) \hat{D}(x_i)']. \quad (2.18)$$

Using (2.18) permits us to perform robust inference.

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

Algorithm 2.1 (IV Estimation and Inference Using Post-LASSO).

⁴In order to accommodate examples such as this formally, we need to allow for triangular array asymptotics.

- (1) For each l , specify penalty loadings according to the initial option in (2.9) and compute the Post-LASSO estimator $\widehat{\beta}_{lPL}$ via (2.10) and the residuals $\widehat{v}_{il} = d_{li} - f'_i \widehat{\beta}_{lPL}$, $i = 1, \dots, n$.
- (2) Update the penalty loadings according to the refined option in (2.9) and update the Post-LASSO estimator $\widehat{\beta}_{lPL}$ via (2.10) and the residuals $\widehat{v}_{il} = d_{li} - f'_i \widehat{\beta}_{lPL}$, $i = 1, \dots, n$.
- (3) Repeat the previous step K times, where K is bounded. Compute the estimates of the optimal instrument $\widehat{D}_{li} = f'_i \widehat{\beta}_{lPL}$. Then compute the IV estimator defined in (2.15).
- (4) Compute the robust estimates (2.18) of the asymptotic variance matrix, and proceed to perform conventional inference using the normality result (2.17).

The first algorithm involves the use of Post-LASSO in the first two steps and is our preferred algorithm.

Algorithm 2.2 (IV Estimation and Inference Using LASSO).

- (1) For each l , specify penalty loadings according to the initial option in (2.9) and compute the LASSO estimator $\widehat{\beta}_{lL}$ via (2.8) and the residuals $\widehat{v}_{il} = d_{li} - f'_i \widehat{\beta}_{lL}$, $i = 1, \dots, n$.
- (2) Update the penalty loadings according to the refined option in (2.9) and update the LASSO estimator $\widehat{\beta}_{lL}$ via (2.8) and the residuals $\widehat{v}_{il} = d_{li} - f'_i \widehat{\beta}_{lL}$, $i = 1, \dots, n$.
- (3) Repeat the previous step K times, where K is bounded. Compute the estimates of the optimal instrument $\widehat{D}_{li} = f'_i \widehat{\beta}_{lL}$. Then compute the IV estimator defined in (2.15).
- (4) Compute the robust estimates (2.18) of the asymptotic variance matrix, and proceed to perform conventional inference using the normality result (2.17).

The second algorithm involves the use of LASSO in the first two steps. Note that our results allow for hybrids between this and the previous algorithm.

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

In this section we present our main results on LASSO and Post-LASSO estimators of the conditional expectation functions under non-classical assumptions and under 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 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, and 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:

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

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

$$\kappa_C^2(M) := \min_{\delta \in \Delta_C, |T| \leq s} s \frac{\delta' M \delta}{\|\delta_T\|_1^2} \text{ and } \tilde{\kappa}_C := \min_{\delta \in \Delta_C} \frac{\delta' M \delta}{\|\delta\|_2^2}. \quad (3.19)$$

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

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 $\kappa_C(\mathbb{E}_n[f_i f_i']) \geq \kappa$ and $\tilde{\kappa}_C(\mathbb{E}_n[f_i f_i']) \geq \kappa$ for all $n > n_0$.*

The restricted eigenvalue (3.19) 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 Gram matrix $\mathbb{E}_n[f_i f_i']$ is zero whenever $p \geq n$, [8] show that its restricted eigenvalues can in fact be bounded away from zero. Lemmas 1 and 2 below contain sufficient conditions for this. Many more sufficient conditions are available from the literature; see [8]. This makes conditions on restricted eigenvalues useful for many applications. 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

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

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

$$\phi_{\min}(m)(M) = \min_{\delta \in \Delta(m)} \delta' M \delta \quad \text{and} \quad \phi_{\max}(m)(M) = \max_{\delta \in \Delta(m)} \delta' M \delta. \quad (3.20)$$

We require the $s \log n$ -sparse eigenvalues to be well-behaved.

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

However, Condition SE requires that certain “small” $m \times m$ submatrices of the large $p \times p$ Gram matrix are well-behaved, which will be sufficient to make the results that follow work. Note that Condition SE implies Condition RE by the argument given in [8].

The following condition shows that the condition above is plausible for both many instruments and many series instrument settings.

Lemma 1 (Plausibility of RF 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 design matrix $\mathbb{E}[f_i f_i']$ has diagonal entries bounded above and away from zero and that its $s \log n$ -sparse eigenvalues are bounded from above and away from zero. Then if $s \log n = o(n/\log p)$, then Condition RE and SE holds with probability approaching one exponentially fast in n .*

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

The condition that the population covariance matrix $\mathbb{E}[f_i f_i']$ has eigenvalues bounded from below and from above is a standard assumption in econometric research. Our conditions allow for this and even more general behavior.

We also impose the following moment conditions on the reduced form errors and regressions. We use these conditions along with RE and SE to prove the general rate results for LASSO

and Post-LASSO in the next section. Therefore, we state these conditions separately from the conditions on the structural error. In the following, $\tilde{d}_{il} = d_{il} - \mathbb{E}[d_{il}]$.

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

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

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

We emphasize that the condition given above is only one possible set of sufficient conditions, which are presented in a manner that reduces the complexity of the exposition. The proofs contain a more refined set of conditions. The following lemma shows that the condition put forward above is quite plausible. It is satisfied for example if the regressors f_i are Gaussian or arbitrary with bounded entries and if the disturbances v_{il} and \tilde{d}_{il} have uniformly bounded conditional moments of order 6. Note that we say that a random variable g_i has uniformly bounded conditional moments of order K if for some positive constants $0 < B_1 < B_2 < \infty$:

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

Lemma 3 (Plausibility of RF). (1) *If the regressors f_i are Gaussian as in Lemma 1, then Conditions RF(ii) and (iii) hold under Condition RF (i) and under $s(\log p)^2/n \rightarrow 0$. (2) If the regressors f_i are arbitrary i.i.d. vectors with bounded entries as in Lemma 2, then Conditions RF(ii), (iii), and (iv) hold under Condition RF(i). Suppose that the disturbances v_{il} have uniformly bounded conditional moments of order 6 uniformly in l , then Condition RF(iv) holds if (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 (2.8) and (2.10) in the system of k_e non-parametric regression equations (2.6) with non-Gaussian and heteroscedastic errors. These results extend the previous results of [8] for LASSO and of [4]

for post-LASSO with classical i.i.d. errors. In addition, we account for the fact that we are simultaneously estimating k_e regressions and account for the dependence of our results on k_e .

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

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

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

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

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

plays a key role. We select the penalty level λ/n to dominate the noise for all k_e regression problems we are considering simultaneously, specifically so that

$$\mathbb{P}\left(\lambda \geq c' n \max_{1 \leq l \leq k_e} \|S_l\|_\infty\right) \rightarrow 1, \quad (3.21)$$

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

$$\lambda = c2\sqrt{2n \log(2pk_e)}, \quad (3.22)$$

with $c > c'$ implements (3.21).

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

$$\ell \hat{\Upsilon}_l^0 \leq \hat{\Upsilon}_l \leq u \hat{\Upsilon}_l^0,$$

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

Theorem 1 (Rates for LASSO under Non-Gaussian and Heteroscedastic Errors). *Suppose that in the regression model (2.6) Conditions RE and RF hold. Suppose the penalty level is specified as in (3.22), and consider any asymptotically valid penalty loadings $\hat{\Upsilon}$. Then, the LASSO estimator*

$\widehat{\beta}_l = \widehat{\beta}_{lL}$ and $\widehat{D}_{li} = f'_i \widehat{\beta}_{lL}$, $l = 1, \dots, k_e$, satisfy

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

The following theorem derives the properties of Post-LASSO.

Theorem 2 (Rates for Post-LASSO under Non-Gaussian and Heteroscedastic Errors). *Suppose that in the regression model (2.6) Conditions SE and RF hold. Suppose the penalty level for the LASSO estimator is specified as in (3.22), and that LASSO's penalty loadings $\widehat{\Upsilon}$ are asymptotically valid. Then, the Post-LASSO estimator $\widehat{\beta}_l = \widehat{\beta}_{lPL}$ and $\widehat{D}_{li} = f'_i \widehat{\beta}_{lPL}$, $l = 1, \dots, k_e$, satisfy*

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

Finally, we show that the data-driven penalty loadings that we have proposed in (2.9) obey the conditions put forward above. 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, note 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 5 or 6, the penalty loadings $\widehat{\Upsilon}$ specified in (2.9), obtained under the basic, the*

refined, and the K -step refined option based on residuals obtained from LASSO or Post-LASSO are asymptotically valid. (In particular, for the refined options $u' = 1$).

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

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

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

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

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

Condition SM(i) requires that structural errors are boundedly heteroscedastic. Given this we make a normalization assumption on the instruments. This entails no loss of generality, since this is equivalent to suitably rescaling the parameter space for coefficients $\beta_{l0}, l = 1, \dots, k_\epsilon$, via an isomorphic transformation. Moreover, we only need this normalization to simplify notations in the proofs, and we do not use it in the construction of the estimators. Condition SM(ii) imposes some mild moment assumptions. Condition (iv) strengthens the usual growth requirement $s \log p/n \rightarrow 0$ to somewhat stronger conditions, the restrictiveness of which rapidly decreases as the number of bounded moments of the structural error increases.

The following lemma shows that SM is quite plausible. For example, it is satisfied if the regressors f_i are Gaussian or arbitrary with bounded entries and if the structural disturbance ϵ_i have uniformly bounded conditional moments of order 3.

Lemma 4 (Plausibility of SM). *Suppose that the structural disturbance ϵ_i has uniformly bounded conditional moments of order 3 uniformly in n , then Condition SM(ii) 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 the introduction, and that the structural error ϵ_i is homoscedastic. 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 IV estimator is root- n consistent, asymptotically normal, and achieves the efficiency bound:*

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

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

The second result below describes the properties of the IV estimator with the IV constructed using LASSO or Post-LASSO in the setting of the standard model with heteroscedastic structural errors. In these settings 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 robust inference.

Theorem 5 (Robust Inference with IV Constructed by LASSO or Post-LASSO). *Suppose conditions of Theorem 1 hold, except that now that the structural errors ϵ_i can be heteroscedastic. Then the IV estimator is root- n consistent and asymptotically normal:*

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

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

The final third result extends the previous the first two results to any IV-estimator with a generic sparse estimator of the optimal instrument. This result shows that provided that the estimators obey the near-oracle performance given in (4.23)-(4.24), our results extend easily to other forms of sparse estimators such as

- Dantzig and post-Dantzig, [14]
- $\sqrt{\text{LASSO}}$ and post- $\sqrt{\text{LASSO}}$, [7] and [6],
- thresholded LASSO and post-thresholded LASSO, [4]
- grouped LASSO and post-grouped LASSO, [32]

Verification of the near-oracle performance follows on a case by case basis using the best current and future conditions in the literature.⁵ 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 (4.23)-(4.24).

Theorem 6 (Inference with IV Constructed by a Generic Sparsity-Based Procedure). *Suppose that conditions AS, RF, SM hold and suppose now that $\hat{D}_i = f_i'\hat{\beta}_l$ is constructed using any*

⁵Note 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.

estimator $\widehat{\beta}_l$ such that

$$\max_{1 \leq l \leq k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n} \lesssim_P \sqrt{\frac{s \log p}{n}} \quad (4.23)$$

$$\max_{1 \leq l \leq k_e} \|\widehat{\beta}_l - \beta_{l0}\|_1 \lesssim_P \sqrt{\frac{s^2 \log p}{n}}. \quad (4.24)$$

then the conclusions reached in Theorem 1 or Theorem 2 continue to apply in this case.

5. SIMULATION EXPERIMENT

The theoretic results presented in the previous sections suggest that using LASSO to aid in fitting the first-stage regression should result in IV estimators with good estimation and inference properties. In this section, we provide simulation evidence on estimation and inference properties of IV estimators using post-LASSO for the first-stage in the situation in which there are many possible instruments though only a small number are very informative about the endogenous regressor. We also compare the performance of LASSO-based estimators to other estimators that are available in the literature.

Our simulations are based on a simple instrumental variables model of the form

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

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

For the other parameters, we use a variety of different parameter settings. We provide simulation results for sample sizes, n , of 150, 750, and 1500. We consider three different values for $\text{Corr}(e, v)$: 0, .3, and .6. We also consider three values of σ_v^2 which are chosen to benchmark three different strengths of instruments. The three values of σ_v^2 are found as $\sigma_v^2 = \frac{n\Pi'\Sigma_Z\Pi}{F^*\Pi'\Pi}$ for three different values of F^* : 6.72, 26.87, and 107.48.⁶ Finally, we use two different settings for

⁶These values were chosen by taking the value of the first-stage F-statistic calculated using clustered standard errors and the instrument chosen by LASSO restricted to choose one instrument in the FHFA data used in Section 6 which yields an F of 26.87 and then multiplying and dividing this number by four to obtain weaker and stronger identification.

the first stage coefficients, Π . The first sets the first five elements of Π equal to one and the remaining elements equal to zero. We refer to this design as the “cut-off” design. The second model sets the coefficient on $z_{ih} = .7^{h-1}$ for $h = 1, \dots, 100$. We refer to this design as the “exponential” design. In the cut-off case, the first-stage has an exact sparse representation, while in the exponential design, the model is not literally sparse although the majority of explanatory power is contained in the first few instruments.

For each setting of the simulation parameter values, we report results from five different estimation procedures. A simple possibility when presented with many instrumental variables is to just estimate the model using 2SLS and all of the available instruments. It is well-known that this will result in poor-finite sample properties unless there are many more observations than instruments; see, for example, [3]. The limited information maximum likelihood estimator (LIML) and its modification by [24] (FULL)⁷ 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 [26] for example.⁸ We report results for these estimators in rows labeled 2SLS(100), LIML(100), and FULL(100) respectively. In addition, we report estimates based on the post-LASSO IV estimator that uses LASSO either to select one instrument (LASSO(1)) or the data-dependent LASSO penalty (LASSO(C)). For each estimator, we report root-mean-squared-error (RMSE), median bias (Med. Bias), mean absolute deviation (MAD), and rejection frequencies for 5% level tests (rp(.05)). For computing rejection frequencies, we estimate conventional 2SLS standard errors for 2SLS(100), LASSO(1), and LASSO(C), and the many instrument robust standard errors of [26] for LIML(100) and FULL(100).

Simulation results are presented in Tables 1-6. Tables 1-3 give results for the cut-off design with $n = 150$, $n = 750$, and $n = 1500$ respectively; and Tables 4-6 give results for the exponential design with $n = 150$, $n = 750$, and $n = 1500$ respectively. As expected, 2SLS(100) does extremely poorly along all dimensions except in the case with no correlation between e and v .

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

⁸Another reason that one may prefer LASSO that we do not pursue here is that LIML and Fuller lose their robustness to many instruments when data are not homoscedastic. In this case, robustness can be restored by using the continuous-updating estimator (CUE) as outlined in [28] which is computationally very burdensome relative to 2SLS.

In this case, there is no endogeneity bias and OLS would be the optimal estimator. 2SLS(100) performs well here since the many instrument bias moves 2SLS toward the OLS estimator and is thus favorable to the performance of the estimator; see, e.g. [3]. Of course, this bias works against 2SLS(100) once one moves away from the exogenous case, and we see that the performance of 2SLS(100) rapidly declines in the strength of the correlation of the errors.

There is interesting variation in the performance of the other estimators across the simulation parameters. LASSO(C) performs very well in terms of RMSE and MAD, having smaller RMSE and MAD than any of the other considered estimators in each design with $Corr(e, v) \neq 0$, and beating all estimators but 2SLS(100) when the two error terms are independent. The performance of LASSO(1) in terms of RMSE and MAD is more mixed. LASSO(1) outperforms LIML(100) and FULL(100) based on these metrics when the instruments are relatively weak ($F^* = 6.72$) or the sample is small ($n = 150$), but performs worse than these estimators when the instruments are strong ($F^* = 107.48$) or the sample is larger. This result can intuitively be associated to the “amount of information” available in the instruments relative to the sample size. With weak instruments, there is relatively little useful available for forming the first stage prediction, and LASSO(1) selects a relatively strong instrument and, by keeping the number of instruments small, keeps bias small as well. This same argument applies to LASSO(C) in this case as well, and it is interesting to note that LASSO(C) may be overly conservative in that it selects no instruments in many cases when instruments are weak.⁹ LIML(100) and FULL(100) use all of the instruments, the majority of which contribute little if anything other than noise, and so do not perform as well as LASSO(1) or LASSO(C). On the other hand, there is a substantial “amount of information” contained in several of the instruments when the first-stage relationship is stronger which cannot be exploited by LASSO(1) since it restricts the number of instruments to one. Using this information, as is done in LIML(100), FULL(100), and LASSO(C), thus reduces estimator risk as measured by RMSE or MAD. In other words, LASSO(C) appears to adapt to the “amount of information” available for forming the first-stage prediction and outperforms the less adaptive procedures.

⁹In these cases, the LASSO(C) results are based on the number of simulation replications in which LASSO(C) selected a non-empty set of instruments. The exact numbers of cases for each design may be found in Tables 1-6.

Considering next median bias, we see that LIML(100) and FULL(100) perform relatively well as theory would predict; see, for example, [25]. Given the performance in terms of RMSE and MAD discussed above, it must also be the case that the cost of this smaller bias is additional variability in the estimators which is pronounced in some designs. In most cases, the biases of LASSO(1) and LASSO(C) are also quite small, though the bias of these estimators does increase more rapidly than that of LIML(100) or FULL(100) as the strength of the correlation between e and v increases. When comparing between LASSO(1) and LASSO(C) in terms of bias, neither dominates the other and they are often close. It appears that LASSO(C) tends to do slightly better than LASSO(1) when the first-stage relationship as measured by F^* is large and that LASSO(1) tends to do slightly better than LASSO(C) when the first-stage relationship as measured by F^* is small, but the pattern is not striking.

Finally, we see that LASSO(C) does quite well in terms of rejection frequencies of 5% level tests. Once again, there is no procedure that is uniformly dominant based on this metric, though 2SLS(100) performs very poorly except when there is no endogeneity. LASSO(C) is very competitive with FULL(100) and LIML(100) using many-instrument-robust standard errors except when the correlation between the errors is .6 and the instruments are weak. In this case, FULL(100) and LIML(100) clearly dominate the LASSO procedures in terms of size of tests. Outside of this cell, LASSO(C) and LASSO(1) are competitive in all cases with test sizes between 1.2% and 9.2% for nominal 5% level tests and the large majority of sizes between 3.5% and 6.5%. LIML(100) and FULL(100) appear to control size more uniformly across the designs with sizes for FULL(100) and LIML(100) between 1.8% and 7.6% in every case except the exponential design with $n = 150$, $corr(e, v) = .6$, and weak instruments in which case the sizes are respectively 9.6% and 9.2% (compared to 12% and 12.2% for LASSO(1) and LASSO(C)). Whether one would be willing to trade the modest deterioration in testing performance using LASSO(C) versus FULL(100) or LIML(100) for the improvement in RMSE and MAD of the estimator will of course depend on the preferences of the researcher.¹⁰

¹⁰Another possibility that we do not consider is to use LASSO coupled with LIML or FULL and many-instrument robust standard errors.

Overall, the simulation results are quite favorable to post-LASSO-based IV methods and especially LASSO(C) (which uses a data-dependent penalty). Using the post-LASSO IV estimator provides one with an estimator that dominates the other estimators considered based on RMSE or MAD. It also produces an estimator with relatively small finite sample bias, though the higher-order unbiased FULL(100) estimator and the approximately median-unbiased LIML(100) do slightly better based on this metric. Finally, the LASSO-based procedures do a relatively good job in producing tests with size close to the nominal level that are not dominated, but do not dominate, FULL(100) and LIML(100) with many-instrument robust standard errors. While the results do not show that the LASSO-based procedures uniformly dominate the many-instrument robust procedures, they do show that the simple LASSO-based procedures are highly competitive, having uniformly lower risk (as measured by RMSE and MAD) and not being dominated in terms of testing performance. We also note that the design is favorable to LIML and FULL in that $p < n$ in all cases.

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 appellate court decisions regarding eminent domain on a variety of economic outcomes. The study of the economic consequences of the law of takings or eminent domain, when a government actor physically acquires the property rights of one or more individuals, is important for a variety of reasons. Takings are often justified based on “public use” arguments such as removing economic blight and/or promoting economic development through private commercial development. People worried about the government’s use of eminent domain, on the other hand, note potential undesirable redistribution of wealth from groups with little political power to those with more political power and distortions the exercise of eminent domain may induce in the efficient investment of capital. In particular, the government’s exercise of eminent domain could lead to underinvestment due to uncertainty induced by the possibility that the government may appropriate the investment; see [37]. On the flip side, scholars have raised arguments that the exercise of eminent domain and the associated “just compensation” could lead to overinvestment when property owners anticipate receiving higher compensation as a result of having their land condemned; see [9], [34], [29], and [41].

Despite the amount of theoretical and doctrinal writings that speculate about the effects of takings law on economic outcomes, little empirical evidence exists. One notable recent exception is [20] which provides a careful empirical analysis of the effect of appellate court decisions regarding takings law on economic outcomes. The results we provide in this section complement the analysis of [20] by taking their estimation strategy and augmenting it with the use of LASSO to choose instruments.

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

$$y_{ict} = \alpha_c + \alpha_t + \gamma_{ct} + \beta \text{ Takings Law}_{ct} + W'_{ct}\delta + \epsilon_{ict} \quad (6.25)$$

where y_{ict} is an economic outcome for area i in circuit c at time t , Takings Law_{ct} represents the number of pro-plaintiff appellate takings decisions in circuit c and year t ; W_{ct} are judicial pool characteristics,¹¹ a dummy for whether there were no cases in that circuit-year, and the number of takings appellate decisions; and α_c , α_t , and γ_{ct} are respectively circuit-specific effects, time-specific effects, and circuit-specific time trends. An appellate court decision is coded as pro-plaintiff if the court ruled that a taking was unlawful, thus overturning the government's seizure of the property in favor of the private owner. We construe pro-plaintiff decisions to indicate a regime that is more protective of individual property rights. The parameter of interest, β , thus represents the effect of an additional decision upholding individual property rights on an economic outcome.

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

¹¹The judicial pool characteristics are the probability of a panel being assigned with each set of characteristics defined by the instruments. There are 144 total probability controls. The list is available upon request.

influence judicial decisions related to takings. [20] provide additional discussion of potential sources of endogeneity that motivate the use of an instrumental variables strategy.

To address the potential endogeneity of *Takings Law*, we employ the instrumental variables based on the identification argument of [19] and [20] that relies on the random assignment of judges to appellate panels that decide federal appellate cases. Since judges are randomly assigned to three judge panels to decide appellate cases, the exact identity of the judges and, more importantly, their demographics are randomly assigned conditional on the distribution of characteristics of federal circuit court judges in a given circuit-year. Thus, once the distribution of characteristics is controlled for, the realized characteristics of the randomly assigned three judge panel should be unrelated to other factors besides judicial decisions that may be related to economic outcomes. There is also substantial evidence that the demographic characteristics of judges are related to their judicial decision making.¹² Thus, the characteristics of judges on panels deciding eminent domain cases should provide valid instruments for learning about the effect of appellate court decisions regarding takings law on economic outcomes.

This analysis is complicated by the number of potential characteristics of three judge panels that may be used as instruments. While the identification argument outlined above suggests any set of characteristics of the three judge panel will be uncorrelated with the structural unobservable conditional on the set of controls in the model, there will clearly be some instruments which are more worthwhile than others in obtaining precise second-stage estimates. For simplicity, we consider only the following demographics: gender, race, religion, political affiliation, education,¹³ and whether the judge was elevated from a district court. We also only consider interactions between gender and race, gender and religion, race and religion, gender and whether the law degree is from a public university, race and whether the law degree is from a public university, and religion and whether a law degree is from a public university. Imposing this set of restrictions gives a total of 92 potential instruments that we select among using LASSO. The exact description of the instrument set is available upon request.

¹²See, e.g., [10], [17], [23], [19], and [20].

¹³The education variables are whether the judge's bachelor was obtained in-state, from a public university, whether the JD was obtained from a public university, whether the judge has an LLM, and whether a judge has an SJD.

We provide results using four different economic outcomes as dependent variables: the log of three home-price-indices and $\log(\text{GDP})$. 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. The total sample sizes are 5304, 1920, and 4320 for FHFA, Non-Metro, and Case-Shiller respectively. We also use state level GDP from the Bureau of Economic Analysis to form $\log(\text{GDP})$. The GDP regressions are based on 1326 observations.

Table 7 contains second-stage estimation results. We report results using three different sets of instruments. In the first results, labeled 2SLS, we use the instruments used in [20] and consider this the baseline. [20] used two variables, whether a panel was assigned an appointee who did not report a public religious affiliation and whether a panel was assigned an appointee who earned their first law degree from a public university, as instruments. The choice of these instruments is motivated on intuitive grounds. [20] note that judges who are strongly affiliated with a religious group are more likely to vote anti-government (pro-plaintiff) as many religious groups are populated by people who believe in smaller government and more private agency. They also argue that judges who attended public institutions to obtain their law degrees are more likely to hold populist positions that would also lead them to vote pro-plaintiff. Regardless of the intuitive justification, [20] find that these two variables do predict the number of pro-plaintiff decisions in first-stage regressions. The second and third sets of results are based on instruments selected through LASSO.¹⁴ Estimates based on LASSO(2) use instruments selected by LASSO where the LASSO penalty was chosen so that LASSO would select only two instruments. LASSO(C) uses a data-dependent penalty to select the instruments as discussed above. For these results, the number of instruments selected by LASSO in each case is also reported in the row “S”. In all cases, estimated standard errors are clustered at the circuit level and critical values are adjusted for the small number of clusters as in [27]. In all cases we use the Post-LASSO 2SLS estimator.

¹⁴We do not report the exact instruments selected by LASSO for brevity. The identity of the selected instruments is available from the authors upon request. It is interesting to note that in no case were the same two instruments as used in [20] jointly selected.

There are a number of key patterns that emerge when looking at the results in Table 7. Consistent with the view that any of the potential instruments would be valid in the sense of being unrelated to the structural error, we see that the point estimates produced using any set of instruments are broadly consistent with each other for all of the dependent variables considered. For the log-price-indices, the point estimates are all small and positive, suggesting that one more pro-plaintiff decision in a year is associated with a small increase in house prices.¹⁵ The estimated effects for log(GDP) are all very near zero. We also see that using LASSO(2) as opposed to the original instruments from [20] leads to increases in the estimated precision of the effect of takings law in three price-indices but interestingly results in lower precision in the log(GDP) regression.

The most interesting results are found by comparing LASSO(C) to 2SLS. The point estimates are similar across the two sets of instruments. However, we see that LASSO selects more than two instruments in every case and that estimated standard errors are substantially smaller, ranging between .18 and .76 times the original standard error. The reduction in standard errors obviously produces more precise inference. The resulting changes in 95% confidence intervals, for example, are substantial, changing them from (-0.0269,0.0677) to (-0.0025,0.0543), from (0.0012,0.1224) to (0.0078,0.0300), from (0.0126,0.0532) to (0.0188,0.0496), and from (-0.0546,0.0576) to (-0.0317,0.0486) for FHFA, Non-Metro, Case-Shiller, and log(GDP) respectively. In each case, the LASSO(C) interval is a strict subset of the original 2SLS interval. For FHFA, the LASSO(C) interval is essentially always positive while the 2SLS interval has substantial length below zero, and for Case-Shiller, the LASSO(C) interval safely excludes zero while the lower endpoint of the 2SLS interval is barely greater than zero. The largest change comes in the Non-Metro result where the 2SLS result does not exclude rather large positive effects while the LASSO(C) interval has an upper endpoint of only .03. The GDP results are both reasonably tight around zero though one cannot exclude small positive or negative effects.

In summary, we find evidence that the effect of takings law decisions on contemporaneous property prices is small but positive while there is little evidence of any appreciable effect on GDP. The results are consistent with the theory in that the 2SLS point-estimates based on each set of instruments are similar while the post-LASSO estimates are considerably more precise. In

¹⁵Note that the average number of pro-plaintiff decisions in a circuit-year is around .15 in the sample.

this example, we also see that the potential precision gains based on LASSO variable selection can be large, producing standard errors that are 18% of the original standard errors in one case. Overall, the findings suggest that there is the potential for LASSO to be fruitfully employed to choose instruments in economic applications.

7. CONCLUSION

In this paper, we have considered the use of LASSO and post-LASSO methods for forming first-stage predictions in a linear instrumental variables model with potentially many instruments. We note that two leading cases where this might arise are when a researcher has a small set of many-valued, possibly continuous, instruments and wishes to nonparametrically estimate the optimal instrument or when the set of potential basic instrument 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 simulation, we see that a feasible post-LASSO procedure that uses a data-dependent penalty performs very well across the range of simulation designs we consider. It performs as well as or better than recently advocated many-instrument robust procedures in the majority of designs, though it does do somewhat worse than these procedures in terms of size of 5% level tests in the case of relatively weak instruments and strong endogeneity. This performance suggests that it may be useful to use LASSO-based instrument selection in conjunction with the many instrument robust procedures, and exploring this may be an interesting avenue for future research.

In the empirical example, we look at the effect of judicial decisions at the federal circuit court level regarding the government's exercise of eminent domain on house prices and state-level GDP as in [20]. We use the random assignment of judges to three judge panels who decide the

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

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

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

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

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

For $0 < \delta \leq 1$ set

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

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

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

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

where the terms $O(1)$ are bounded in absolute value by an absolute constant A .

Application of this result gives the following lemma:

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

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

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

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

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

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

obey

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

Proof. This follows by the application of the quoted theorem to the i.i.d. case with $\delta = 3$ and $d_{n,3} = n^{1/6}M_n$. The calculated error bound follows from the triangular inequalities and conditions on ℓ_n and M_n . \square

APPENDIX B. PROOF OF THEOREM 1

The proof is delicate, because it relies on primitive assumptions and allows for data-driven choices of the penalty loading. We split the proof in several steps.

Step 1. Consider the following quantity,

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

This quantity controls the modulus of continuity between the prediction norm and the l_1 -norm within a restricted region.

The main result of this step is the following lemma:

Lemma 6. *If $\lambda/n \geq c\|S_l\|_\infty$, then*

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

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

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

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

and we also have

$$\left| \widehat{Q}_l(\widehat{\beta}_l) - \widehat{Q}_l(\beta_{l0}) - \|f'_i \delta_l\|_{2,n}^2 \right| \leq \|S_l\|_\infty \|\widehat{\Upsilon}_l^0 \delta_l\|_1 + 2c_s \|f'_i \delta_l\|_{2,n} \quad (\text{B.26})$$

so that from $\lambda \geq cn \|S_l\|_\infty$

$$\begin{aligned} \|f'_i \delta_l\|_{2,n}^2 &\leq \frac{\lambda}{n} \left(\|\widehat{\Upsilon}_l \delta_{lT_l}\|_1 - \|\widehat{\Upsilon}_l \delta_{lT_l^c}\|_1 \right) + \|S_l\|_\infty \|\widehat{\Upsilon}_l^0 \delta_l\|_1 + 2c_s \|f'_i \delta_l\|_{2,n} \\ &\leq \left(u + \frac{1}{c} \right) \frac{\lambda}{n} \|\widehat{\Upsilon}_l^0 \delta_{lT_l}\|_1 - \left(\ell - \frac{1}{c} \right) \frac{\lambda}{n} \|\widehat{\Upsilon}_l^0 \delta_{lT_l^c}\|_1 + 2c_s \|f'_i \delta_l\|_{2,n}. \end{aligned} \quad (\text{B.27})$$

To show the first statement we can assume $\|f'_i \delta_l\|_{2,n} \geq 2c_s$, otherwise we are done. This condition together with relation (B.27) implies that for $c_0 = (uc + 1)/(\ell c - 1)$ we have

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

Therefore, by definition of the restricted eigenvalue, we have

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

Thus, relation (B.27) implies

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

and the result follows.

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

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

In this case, by definition of the restricted eigenvalue, we have

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

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

On the other hand, consider the case that

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

which would already imply $\|f'_i \delta_l\|_{2,n} \leq 2c_s$ by (B.27). Moreover, (B.27) implies that

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

Thus,

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

and the result follows from $c/(\ell c - 1) \leq c_0$. \square

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

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

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

Lemma 7. *For $c > c'$ and $\lambda = c2\sqrt{n2\log(2pk_e)}$, we have that as $n \rightarrow \infty$ and $p \rightarrow \infty$*

$$\mathbb{P}\left(c' \max_{1 \leq l \leq k_e} n \|S_l\|_\infty > \lambda\right) \leq \frac{(1 + o(1))}{(2pk_e)^{c/c' - 1}} = o(1),$$

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

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

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

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

$$\mathbb{P}\left(\max_{1 \leq l \leq k_e} \sqrt{n} \|S_l\|_\infty \geq 2\sqrt{2\log(2pk_e/a)}\right) \leq a(1 + o(1)), \quad (\text{B.28})$$

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

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

To prove the bound, note that

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

uniformly over the region specified above. 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 $\Phi(t) \leq \phi(t)/t$. Finally, boundedness of M_{jl} from below is immediate from Condition RF. \square

Step 3. The main result of this step is the following: Let

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

where the entries of Υ_l^0 are bounded away from zero and from above uniformly in n by Condition RF. Then we have

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

since

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

Indeed, application of Lemma 5, gives us that

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

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

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

Note that under our assumption on moments in Condition RF and Lyapunov moment inequality, the term $\min_{1 \leq j \leq p, 1 \leq l \leq k_e} W_{jl}^2$ is bounded away from zero, so the growth conditions holds for some

$a \rightarrow 0$ and $b_n \rightarrow \infty$ under our condition $\log p = o(n^{1/3})$. Moreover,

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

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

Step 5. Combining the results of all the steps above, we have that given λ specified in the statement of the theorem and the penalty loadings specified in the statement of the theorem, and using the bound $c_s \lesssim_P \sqrt{s/n}$, we obtain the conclusion that

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

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

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

Finally, we also obtain

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

which gives the result. □

APPENDIX C. PROOF OF THEOREM 2

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 \hat{T}_l denote the selected support by $\hat{\beta}_l = \hat{\beta}_{lL}$, $\hat{m}_l = |\hat{T}_l \setminus T_l|$, $\hat{\beta}_{lPL}$ be the Post-LASSO estimator, and $\lambda/n > c\|S_l\|_\infty$ in the first stage*

for LASSO for every $l = 1, \dots, k_e$. Then we have

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

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

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

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

Next note that the least squares criterion function satisfies

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

By taking the maximum over $l = 1, \dots, k_e$ on each side we need to bound the quantities $\max_{1 \leq l \leq k_e} \|S_{T_l^c}\|_\infty$ and $\max_{1 \leq l \leq k_e} \|S_{T_l}\|$.

By Lemma 7, we have

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

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

Next, note that for any $j \in T_l$ we have $E[S_{lj}^2] \lesssim 1/n$, so that

$$E\left[\max_{1 \leq l \leq k_e} \|S_{T_l}\| \right] \leq \sqrt{\sum_{l=1}^{k_e} E[\|S_{T_l}\|^2]} \lesssim \sqrt{k_e s/n}.$$

Thus, by Chebyshev inequality, we have $\max_{1 \leq l \leq k_e} \|S_{T_l}\| \lesssim_P \sqrt{k_e s/n}$. On the other hand, $\max_{1 \leq l \leq k_e} \|S_l\| \leq \sqrt{s} \max_{1 \leq l \leq k_e} \|S_{T_l}\|_\infty \lesssim_P \sqrt{\log(sk_e)}/\sqrt{n}$ by Lemma 7.

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

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

solving which we obtain the stated result:

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

Next we bound the goodness of fit terms $B_{l,n}$ and $C_{l,n}$. If $T_l \subseteq \hat{T}_l$ we directly have $C_{l,n} \leq 0$. Otherwise, the definition of $\hat{\beta}_{lPL}$ implies that

$$\hat{Q}_l(\hat{\beta}_{lPL}) - \hat{Q}_l(\beta_{l0}) \leq \hat{Q}_l(\hat{\beta}_{lL}) - \hat{Q}_l(\beta_{l0}).$$

Thus, letting $\delta_{lL} := \hat{\beta}_{lL} - \beta_{l0}$, by definition of LASSO and that $\ell \hat{\Upsilon}_l^0 \leq \hat{\Upsilon}_l \leq u \hat{\Upsilon}_l^0$, we have

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

If $\|\hat{\Upsilon}_l^0 \delta_{lLT_l^c}\|_1 > (u/\ell) \|\hat{\Upsilon}_l^0 \delta_{lLT_l}\|_1$, we have $\hat{Q}_l(\hat{\beta}_{lL}) - \hat{Q}_l(\beta_{l0}) \leq 0$. Otherwise, $\|\hat{\Upsilon}_l^0 \delta_{lLT_l^c}\|_1 \leq (u/\ell) \|\hat{\Upsilon}_l^0 \delta_{lLT_l}\|_1$ and we have $\|\hat{\Upsilon}_l^0 \delta_{lLT_l}\|_1 \leq \sqrt{s} \|f'_i \delta_{lL}\|_{2,n} / \kappa_{(u/\ell)}^l$ by definition of the restricted eigenvalue. Then, if $\lambda/n \geq c \|S_l\|_\infty$, we have by (B.27) that

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

and the result follows. \square

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

Lemma 9 (Empirical pre-sparsity for LASSO). *In either the parametric model or the nonparametric model, let $\hat{m}_l = |\hat{T}_l \setminus T_l|$ and $\lambda/n \geq c \cdot \|S_l\|_\infty$. We have*

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

Proof of Lemma 9. We have from the optimality conditions that

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

Therefore, noting that $\|\widehat{\Upsilon}_l^{-1}\widehat{\Upsilon}_l^0\|_\infty \leq 1/\ell$, we have for $R = (a_{l1}, \dots, a_{ln})'$

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

where we used that

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

and similarly $\|(X'R)_{\widehat{T}_l \setminus T_l}\| \leq n\sqrt{\phi_{\max}(\widehat{m}_l)}c_s$.

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

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

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

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

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

Lemma 11 (Sparsity bound for LASSO under data-driven penalty). *Consider the LASSO estimator with $\lambda/n \geq c\|S_l\|_\infty$, and let $\widehat{m}_l := |\widehat{T}_l \setminus T_l|$. Consider the set $\mathcal{M} = \{m \in \mathbb{N} : m > s \cdot 2\phi_{\max}(m)\frac{\|(\widehat{\Upsilon}_l^0)^{-1}\|_\infty^2}{\ell^2} \left[\frac{2c_0}{\kappa_{c_0}^l} + \frac{6c_0nc_s}{\lambda\sqrt{s}} \right]^2\}$. Then,*

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

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

$$\widehat{m}_l \leq \phi_{\max}(\widehat{m}_l)\frac{\|(\widehat{\Upsilon}_l^0)^{-1}\|_\infty^2}{\ell^2} \left[\frac{2c_0}{\kappa_{c_0}^l} + \frac{6c_0nc_s}{\lambda\sqrt{s}} \right]^2. \quad (\text{C.29})$$

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

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

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

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

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

In turn, applying (C.29) once more with $\widehat{m} \leq (M \wedge n)$ we obtain

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

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 C, recall that $\max_{1 \leq l \leq k_e} \|\widehat{\Upsilon}_l^0 - \Upsilon_l^0\|_\infty \rightarrow_P 0$. Moreover, under conditions RE and SE, as long as $\lambda/n > c \max_{1 \leq l \leq k_e} \|S_l\|_\infty$, by Lemma 11 we have for every $l = 1, \dots, k_e$ that

$$\widehat{m}_l \lesssim_P s$$

since $\kappa_{c_0}^l$ is bounded away from zero, and that $c_s \lesssim_P \sqrt{s/n}$ leads to $nc_s/[\lambda\sqrt{s}] \lesssim_P 1$. Therefore, by Lemma 8 we have

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

By the choice of $\lambda = 2c\sqrt{2n \log(2pk_e)}$, obtained in Lemma 7, we have

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

since the event $\lambda/n > c' \max_{1 \leq l \leq k_e} \|S_l\|_\infty$ holds with probability $1 - o(1)$. 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. □

APPENDIX D. PROOF OF THEOREM 3

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

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

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

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

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

We have that

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

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

Furthermore,

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

Indeed, application of Lemma 5 gives us that

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

uniformly in $0 < a < 1$ and p and k_e on the region,

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

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

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

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

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

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

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

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

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

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

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 the condition above, are asymptotically valid. Thus by Theorem 4 and 5, the LASSO and Post-LASSO estimators based on the refined option also obey the condition above. This, establishes that we can iterate on the refined option a bounded number of times, without affecting the validity of the approach.

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

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

which we establish below.

Now note that we have proven $\Delta_2 \rightarrow_P 0$ in the proof of Theorem 4. As for Δ_1 we note that

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

The first term is bounded by

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

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

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

which converges to by assumption on the empirical maximum of the regressors appearing in Condition RF. \square

APPENDIX E. PROOF OF LEMMA 1-4

E.1. **Proof of Lemma 1.** See [5] (Supplement). \square

E.2. **Proof of Lemma 2.** See [5] (Supplement) . \square

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

APPENDIX F. PROOF OF THEOREM 4-6.

Step 0. We have by Theorem 1 and 3 and Condition SM that the LASSO estimator with data-driven penalty loadings and by Theorem 2 and 3 the Post-LASSO estimator with data-driven

penalty loadings obey:

$$\max_{1 \leq l \leq k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n} \lesssim_P \sqrt{\frac{s \log p}{n}} \rightarrow 0 \quad (\text{F.30})$$

$$\sqrt{\log p} \|\widehat{\beta}_l - \beta_{l0}\|_1 \lesssim_P \sqrt{\frac{s^2 \log^2 p}{n}} \rightarrow 0 \quad (\text{F.31})$$

In order to prove Theorem 5 we need also the condition

$$\max_{1 \leq l \leq k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n}^2 n^{2/q_e} \lesssim_P \frac{s \log p}{n} n^{2/q_e} \rightarrow 0. \quad (\text{F.32})$$

Note that Theorem 6 assumes (F.30) -(F.32) as high level conditions.

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

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

where by Steps 2 and 3 below:

$$a = \mathbb{E}_n[d_i \widehat{D}'_i] = E[d_i D'_i] + o_P(1) \quad (\text{F.33})$$

$$b = \mathbb{G}_n[\widehat{D}_i \epsilon_i] = \mathbb{G}_n[D_i \epsilon_i] + o_P(1) \quad (\text{F.34})$$

where $E[d_i D'_i] = E[D_i D'_i] = Q$ is bounded away from zero and bounded from above in the matrix sense, uniformly in n . Moreover, $Var(b) = \Omega$ where $\Omega = \sigma^2 E[D_i D'_i]$ under homoscedasticity and $\Omega = E[\epsilon_i^2 D_i D'_i]$ under heteroscedasticity. In either case we have that Ω is bounded away from zero and from above in the matrix sense, uniformly in n . (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 $\delta > 0$

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

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

Step 2. To show (F.33), note that

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

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

Step 3. To show (F.34), note that

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

Next we note that for each l

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

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

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

and in the boundedly heteroscedastic case of Theorem 5:

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

Next we note that by the maximal inequality for self-normalized sums:

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

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

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

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

$$2 \log(2p/a) \leq \frac{n^{1/3}}{b_n} \min_{1 \leq j \leq p} M_{j0}^2, \quad M_{j0} := \frac{\mathbb{E}[f_{ij}^2\epsilon_i^2]^{1/2}}{\mathbb{E}[|f_{ij}|^3|\epsilon_i|^3]^{1/3}}. \quad (\text{F.35})$$

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 $\Phi(t) \leq \phi(t)/t$. Finally, by Condition SM $\min_{1 \leq j \leq p} M_{j0}^2$ is bounded away from zero, so the condition (F.35) is satisfied asymptotically for some $b_n \rightarrow \infty$ and some $a \rightarrow 0$ provided $\log p = o(n^{1/3})$.

Finally, we have that

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

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

Thus, combining bounds above with bounds in (F.30)

$$\max_{1 \leq l \leq k_e} |b_l| \lesssim_P \sqrt{\frac{s^2 \log^2 p}{n}} + \sqrt{\frac{s}{n}} \rightarrow 0,$$

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

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

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

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

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

which is bounded up to a constant by

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

by (F.30) and by $\|D_i\|_{2,n} \lesssim_P 1$ holding by Markov inequality. Moreover, $\mathbb{E}_n[D_i D'_i] - \mathbb{E}[D_i D'_i] \rightarrow_P 0$ by Rosenthal's [39] inequality using that $\mathbb{E}\|D_i\|^q$ for $q > 2$ is bounded uniformly in n .

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

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

First, we note

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

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

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

which is bounded up to a constant by

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

The latter occurs because $\|\epsilon_i^2 D_i\|_{2,n} \leq \|\epsilon_i^2\|_{2,n} \|D_i\|_{2,n} \lesssim_P 1$ by $E[|\epsilon_i|^q]$ and $E[\|D_i\|^4]$ uniformly bounded in n for $q > 4$ and by Markov inequality, and

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

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

REFERENCES

- [1] Takeshi Amemiya. The non-linear two-stage least squares estimator. *Journal of Econometrics*, 2:105–110, 1974.
- [2] J. Bai and S. Ng. Selecting instrumental variables in a data rich environment. *Journal of Time Series Econometrics*, 1(1), 2009.
- [3] Paul A. Bekker. Alternative approximations to the distributions of instrumental variables estimators. *Econometrica*, 63:657–681, 1994.
- [4] A. Belloni and V. Chernozhukov. Post- ℓ_1 -penalized estimators in high-dimensional linear regression models. *arXiv:[math.ST]*, 2009.
- [5] A. Belloni and V. Chernozhukov. ℓ_1 -penalized quantile regression for high dimensional sparse models. *accepted at the Annals of Statistics*, 2010.
- [6] A. Belloni, V. Chernozhukov, and L. Wang. Square-root-lasso: Pivotal recovery of nonparametric regression functions via conic programming. *Duke and MIT Working Paper*, 2010.
- [7] A. Belloni, V. Chernozhukov, and L. Wang. Square-root-lasso: Pivotal recovery of sparse signals via conic programming. *arXiv:[math.ST]*, 2010.
- [8] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.
- [9] L. Blume, D. L. Rubinfeld, and P. Shapiro. The taking of land: When should compensation be paid? *Quarterly Journal of Economics*, 100:71–92, 1984.
- [10] C. L. Boyd, L. Epstein, and A. D. Martin. Untangling the causal effects of sex on judging. forthcoming *American Journal of Political Science*, 2010.

- [11] F. Bunea, A. Tsybakov, and M. H. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.
- [12] F. Bunea, A. B. Tsybakov, , and M. H. Wegkamp. Aggregation and sparsity via ℓ_1 penalized least squares. In *Proceedings of 19th Annual Conference on Learning Theory (COLT 2006)* (G. Lugosi and H. U. Simon, eds.), pages 379–391, 2006.
- [13] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.
- [14] E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n . *Ann. Statist.*, 35(6):2313–2351, 2007.
- [15] G. Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34:305–334, 1987.
- [16] G. Chamberlain and G. Imbens. Random effects estimators with many instrumental variables. *Econometrica*, 72:295–306, 2004.
- [17] T. Chang and A. Schoar. Judge specific differences in chapter 11 and firm outcomes. Working Paper, MIT Sloan School of Management, NBER, and CEPR, 2007.
- [18] J. Chao and N. Swanson. Consistent estimation with a large number of weak instruments. *Econometrica*, 73:1673–1692, 2005.
- [19] D. L. Chen and J. Sethi. Does forbidding sexual harassment exacerbate gender inequality. unpublished manuscript, 2010.
- [20] D. L. Chen and S. Yeh. The economic impacts of eminent domain. unpublished manuscript, 2010.
- [21] Victor H. de la Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes*. Probability and its Applications (New York). Springer-Verlag, Berlin, 2009. Limit theory and statistical applications.
- [22] Stephen G. Donald and Whitney K. Newey. Choosing the number of instruments. *Econometrica*, 69(5):1161–1191, 2001.
- [23] L. M. Ellman, C. R. Sunstein, and D. Schkade. Ideological voting on federal courts of appeals: A preliminary investigation. AEI-Brookings Joint Center for Regulatory Studies Working Paper No. 03-9, 2003.
- [24] Wayne A. Fuller. Some properties of a modification of the limited information estimator. *Econometrica*, 45:939–954, 1977.
- [25] 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.
- [26] Christian Hansen, Jerry Hausman, and Whitney K. Newey. Estimation with many instrumental variables. *Journal of Business and Economic Statistics*, 26:398–422, 2008.
- [27] Christian B. Hansen. Asymptotic properties of a robust variance matrix estimator for panel data when t is large. *Journal of Econometrics*, 141:597–620, 2007.
- [28] J. Hausman, W. Newey, T. Woutersen, J. Chao, and N. Swanson. Instrumental variable estimation with heteroskedasticity and many instruments. mimeo, 2009.

- [29] R. Innes. Takings, compensation, and equal treatment for owners of developed and undeveloped property. *Journal of Law and Economics*, 40:403–432, 1997.
- [30] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. H. Poincar Probab. Statist.*, 45(1):7–57, 2009.
- [31] K. Lounici. Sup-norm convergence rate and sign concentration property of lasso and dantzig estimators. *Electron. J. Statist.*, 2:90–102, 2008.
- [32] 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.
- [33] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):2246–2270, 2009.
- [34] T. J. Miceli and K. Segerson. Regulatory takings: When should compensation be paid? *Journal of Legal Studies*, 23:749–776, 1994.
- [35] Whitney K. Newey. Efficient instrumental variables estimation of nonlinear models. *Econometrica*, 58:809–837, 1990.
- [36] R. Okui. Instrumental variable estimation in the presence of many moment conditions. *forthcoming Journal of Econometrics*, 2010.
- [37] T. J. Riddiough. The economic consequences of regulatory taking risk on land value and development activity. *Journal of Urban Economics*, 41:56–77, 1997.
- [38] M. Rosenbaum and A. B. Tsybakov. Sparse recovery under matrix uncertainty. *arXiv:0812.2818v1 [math.ST]*, 2008.
- [39] 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.
- [40] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58:267–288, 1996.
- [41] G. K. Turnbull. Land development under the threat of taking. *Southern Economic Journal*, 69:468–501, 2002.
- [42] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.
- [43] 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.
- [44] C.-H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594, 2008.

Table 1: 2SLS Simulation Results. Cut-Off Design. N = 150

| Estimator | Corr(e, v) = 0 | | | | | Corr(e, v) = .3 | | | | | Corr(e, v) = .6 | | | | | |
|----------------|--------------------|-----------|-------|---------|--------|---------------------|-------|---------|-------|-----------|---------------------|---------|------|-----------|-----|---------|
| | RMSE | Med. Bias | MAD | rp(.05) | RMSE | Med. Bias | MAD | rp(.05) | RMSE | Med. Bias | MAD | rp(.05) | RMSE | Med. Bias | MAD | rp(.05) |
| 2SLS(100) | 0.031 | 0.000 | 0.020 | 0.050 | 0.088 | 0.084 | 0.084 | 0.758 | 0.169 | 0.168 | 0.168 | 1.000 | | | | |
| LIML(100) | 0.970 | 0.007 | 0.114 | 0.026 | 21.706 | 0.030 | 0.131 | 0.074 | 0.881 | 0.012 | 0.114 | 0.076 | | | | |
| FULL(100) | 0.364 | 0.007 | 0.113 | 0.026 | 0.391 | 0.030 | 0.130 | 0.072 | 0.341 | 0.014 | 0.113 | 0.078 | | | | |
| LASSO(1) | 0.079 | -0.004 | 0.049 | 0.028 | 0.084 | 0.018 | 0.056 | 0.064 | 0.079 | 0.034 | 0.055 | 0.090 | | | | |
| LASSO(C) | 0.069 | 0.002 | 0.042 | 0.036 | 0.072 | 0.026 | 0.049 | 0.065 | 0.072 | 0.045 | 0.054 | 0.108 | | | | |
| $F^* = 26.87$ | | | | | | | | | | | | | | | | |
| 2SLS(100) | 0.049 | 0.003 | 0.032 | 0.048 | 0.103 | 0.091 | 0.091 | 0.490 | 0.192 | 0.186 | 0.186 | 0.984 | | | | |
| LIML(100) | 0.153 | -0.003 | 0.082 | 0.042 | 0.152 | -0.013 | 0.079 | 0.044 | 0.119 | -0.001 | 0.068 | 0.046 | | | | |
| FULL(100) | 0.151 | -0.003 | 0.082 | 0.042 | 0.149 | -0.012 | 0.078 | 0.044 | 0.117 | 0.000 | 0.067 | 0.046 | | | | |
| LASSO(1) | 0.079 | -0.004 | 0.053 | 0.038 | 0.085 | 0.014 | 0.056 | 0.052 | 0.085 | 0.028 | 0.063 | 0.072 | | | | |
| LASSO(C) | 0.065 | 0.003 | 0.045 | 0.048 | 0.068 | 0.007 | 0.044 | 0.062 | 0.065 | 0.019 | 0.045 | 0.054 | | | | |
| $F^* = 107.48$ | | | | | | | | | | | | | | | | |
| 2SLS(100) | 0.059 | 0.001 | 0.039 | 0.064 | 0.091 | 0.070 | 0.070 | 0.230 | 0.150 | 0.136 | 0.136 | 0.712 | | | | |
| LIML(100) | 0.083 | 0.000 | 0.054 | 0.046 | 0.083 | -0.002 | 0.050 | 0.066 | 0.070 | -0.002 | 0.045 | 0.026 | | | | |
| FULL(100) | 0.083 | 0.000 | 0.054 | 0.044 | 0.083 | -0.002 | 0.050 | 0.066 | 0.070 | -0.001 | 0.045 | 0.026 | | | | |
| LASSO(1) | 0.082 | -0.001 | 0.057 | 0.038 | 0.085 | 0.014 | 0.055 | 0.054 | 0.082 | 0.010 | 0.055 | 0.046 | | | | |
| LASSO(C) | 0.065 | 0.005 | 0.044 | 0.068 | 0.066 | 0.006 | 0.043 | 0.068 | 0.061 | 0.012 | 0.042 | 0.040 | | | | |

Note: Results are based on 500 simulation replications and 100 instruments. The first five first-stage coefficients were set equal to one and the remaining 95 to zero in this design. Corr(e, v) is the correlation between first-stage and structural errors. F^* measures the strength of the instruments as outlined in the text. 2SLS(100), LIML(100), and FULL(100) are respectively the 2SLS, LIML, and Fuller(1) estimator using all 100 potential instruments. Many-instrument robust standard errors are computed for LIML(100) and FULL(100) to obtain testing rejection frequencies. LASSO(1) reports results for the 2SLS estimator using LASSO restricted to choose only one instrument. LASSO(C) reports results for the 2SLS estimator using LASSO with a data-dependent penalty. For each estimation procedure, we report root-mean-square-error (RMSE), median bias (Med. Bias), mean absolute deviation (MAD), and rejection frequency for 5% level tests (rp(.05)). For LASSO(C) in the weak instrument design ($F^* = 6.72$), LASSO chose no instruments in 112, 103, and 121 simulation replications for Corr(e, v) = 0, .3, and .6 respectively. Results for LASSO(C) in these cases are reported for the subset of simulation draws where the set of instruments selected by LASSO was non-empty.

Table 2: 2SLS Simulation Results. Cut-Off Design. N = 750

| Estimator | Corr(ϵ, v) = 0 | | | | Corr(ϵ, v) = .3 | | | | Corr(ϵ, v) = .6 | | | |
|----------------|---------------------------|-----------|-------|---------|----------------------------|-----------|-------|---------|----------------------------|-----------|-------|---------|
| | RMSE | Med. Bias | MAD | rp(.05) | RMSE | Med. Bias | MAD | rp(.05) | RMSE | Med. Bias | MAD | rp(.05) |
| 2SLS(100) | 0.014 | 0.001 | 0.010 | 0.036 | 0.039 | 0.037 | 0.037 | 0.752 | 0.075 | 0.074 | 0.074 | 0.998 |
| LIML(100) | 0.215 | 0.006 | 0.041 | 0.030 | 0.220 | 0.006 | 0.040 | 0.046 | 0.648 | 0.001 | 0.031 | 0.052 |
| FULL(100) | 0.090 | 0.006 | 0.040 | 0.030 | 0.080 | 0.007 | 0.039 | 0.046 | 0.076 | 0.004 | 0.031 | 0.056 |
| LASSO(1) | 0.035 | 0.001 | 0.023 | 0.026 | 0.035 | 0.010 | 0.023 | 0.054 | 0.037 | 0.015 | 0.025 | 0.112 |
| LASSO(C) | 0.031 | 0.000 | 0.019 | 0.035 | 0.030 | 0.010 | 0.021 | 0.054 | 0.032 | 0.016 | 0.022 | 0.119 |
| $F^* = 26.87$ | | | | | | | | | | | | |
| 2SLS(100) | 0.021 | 0.000 | 0.014 | 0.040 | 0.048 | 0.043 | 0.043 | 0.544 | 0.087 | 0.086 | 0.086 | 0.990 |
| LIML(100) | 0.040 | -0.001 | 0.027 | 0.036 | 0.039 | 0.001 | 0.025 | 0.042 | 0.036 | 0.004 | 0.025 | 0.046 |
| FULL(100) | 0.040 | -0.001 | 0.027 | 0.034 | 0.038 | 0.002 | 0.025 | 0.040 | 0.036 | 0.005 | 0.025 | 0.050 |
| LASSO(1) | 0.037 | 0.000 | 0.024 | 0.054 | 0.036 | 0.003 | 0.024 | 0.046 | 0.039 | 0.014 | 0.027 | 0.106 |
| LASSO(C) | 0.028 | -0.001 | 0.017 | 0.044 | 0.027 | 0.005 | 0.019 | 0.032 | 0.029 | 0.010 | 0.021 | 0.076 |
| $F^* = 107.48$ | | | | | | | | | | | | |
| 2SLS(100) | 0.027 | -0.002 | 0.018 | 0.048 | 0.041 | 0.032 | 0.032 | 0.242 | 0.065 | 0.062 | 0.062 | 0.682 |
| LIML(100) | 0.032 | -0.002 | 0.020 | 0.052 | 0.032 | 0.002 | 0.021 | 0.062 | 0.032 | -0.001 | 0.021 | 0.066 |
| FULL(100) | 0.032 | -0.002 | 0.020 | 0.052 | 0.032 | 0.003 | 0.021 | 0.066 | 0.032 | -0.001 | 0.020 | 0.064 |
| LASSO(1) | 0.038 | -0.003 | 0.025 | 0.052 | 0.037 | 0.007 | 0.026 | 0.046 | 0.038 | 0.009 | 0.026 | 0.058 |
| LASSO(C) | 0.028 | -0.001 | 0.019 | 0.036 | 0.029 | 0.004 | 0.019 | 0.056 | 0.030 | 0.001 | 0.019 | 0.072 |

Note: Results are based on 500 simulation replications and 100 instruments. The first five first-stage coefficients were set equal to one and the remaining 95 to zero in this design. Corr(ϵ, v) is the correlation between first-stage and structural errors. F^* measures the strength of the instruments as outlined in the text. 2SLS(100), LIML(100), and FULL(100) are respectively the 2SLS, LIML, and Fuller(1) estimator using all 100 potential instruments. Many-instrument robust standard errors are computed for LIML(100) and FULL(100) to obtain testing rejection frequencies. LASSO(1) reports results for the 2SLS estimator using LASSO restricted to choose only one instrument. LASSO(C) reports results for the 2SLS estimator using LASSO with a data-dependent penalty. For each estimation procedure, we report root-mean-square-error (RMSE), median bias (Med. Bias), mean absolute deviation (MAD), and rejection frequency for 5% level tests (rp(.05)). For LASSO(C) in the weak instrument design ($F^* = 6.72$), LASSO chose no instruments in 45, 36, and 39 simulation replications for Corr(ϵ, v) = 0, .3, and .6 respectively. Results for LASSO(C) in these cases are reported for the subset of simulation draws where the set of instruments selected by LASSO was non-empty.

Table 3: 2SLS Simulation Results. Cut-Off Design. N = 1500

| Estimator | Corr(ϵ, v) = 0 | | | | | Corr(ϵ, v) = .3 | | | | | Corr(ϵ, v) = .6 | | | | | |
|----------------|---------------------------|-----------|-------|---------|-------|----------------------------|-------|---------|-------|-----------|----------------------------|---------|------|-----------|-----|---------|
| | RMSE | Med. Bias | MAD | rp(.05) | RMSE | Med. Bias | MAD | rp(.05) | RMSE | Med. Bias | MAD | rp(.05) | RMSE | Med. Bias | MAD | rp(.05) |
| 2SLS(100) | 0.009 | 0.001 | 0.006 | 0.036 | 0.028 | 0.027 | 0.027 | 0.766 | 0.052 | 0.051 | 0.051 | 1.000 | | | | |
| LIML(100) | 1.570 | 0.002 | 0.026 | 0.032 | 4.065 | 0.000 | 0.027 | 0.042 | 0.168 | 0.001 | 0.024 | 0.064 | | | | |
| FULL(100) | 0.062 | 0.002 | 0.025 | 0.026 | 0.060 | 0.001 | 0.027 | 0.040 | 0.045 | 0.003 | 0.023 | 0.068 | | | | |
| LASSO(1) | 0.024 | 0.002 | 0.015 | 0.032 | 0.027 | 0.009 | 0.018 | 0.058 | 0.026 | 0.013 | 0.019 | 0.126 | | | | |
| LASSO(C) | 0.021 | 0.001 | 0.014 | 0.030 | 0.022 | 0.008 | 0.014 | 0.078 | 0.022 | 0.013 | 0.017 | 0.139 | | | | |
| $F^* = 26.87$ | | | | | | | | | | | | | | | | |
| 2SLS(100) | 0.016 | -0.002 | 0.010 | 0.066 | 0.033 | 0.030 | 0.030 | 0.548 | 0.061 | 0.060 | 0.060 | 0.990 | | | | |
| LIML(100) | 0.029 | -0.002 | 0.018 | 0.058 | 0.027 | 0.002 | 0.016 | 0.050 | 0.026 | 0.003 | 0.017 | 0.044 | | | | |
| FULL(100) | 0.029 | -0.002 | 0.018 | 0.056 | 0.027 | 0.002 | 0.015 | 0.046 | 0.026 | 0.004 | 0.017 | 0.044 | | | | |
| LASSO(1) | 0.027 | -0.001 | 0.018 | 0.044 | 0.025 | 0.004 | 0.016 | 0.046 | 0.028 | 0.011 | 0.020 | 0.088 | | | | |
| LASSO(C) | 0.021 | -0.001 | 0.013 | 0.076 | 0.020 | 0.002 | 0.013 | 0.058 | 0.021 | 0.008 | 0.015 | 0.080 | | | | |
| $F^* = 107.48$ | | | | | | | | | | | | | | | | |
| 2SLS(100) | 0.019 | 0.001 | 0.012 | 0.076 | 0.028 | 0.022 | 0.023 | 0.224 | 0.047 | 0.044 | 0.044 | 0.708 | | | | |
| LIML(100) | 0.023 | 0.000 | 0.015 | 0.072 | 0.022 | 0.001 | 0.014 | 0.052 | 0.020 | 0.000 | 0.012 | 0.044 | | | | |
| FULL(100) | 0.023 | 0.000 | 0.015 | 0.072 | 0.022 | 0.001 | 0.014 | 0.052 | 0.020 | 0.001 | 0.013 | 0.042 | | | | |
| LASSO(1) | 0.027 | 0.000 | 0.020 | 0.044 | 0.027 | 0.003 | 0.018 | 0.048 | 0.025 | 0.006 | 0.018 | 0.046 | | | | |
| LASSO(C) | 0.021 | 0.001 | 0.014 | 0.060 | 0.020 | 0.002 | 0.013 | 0.062 | 0.019 | 0.002 | 0.013 | 0.040 | | | | |

Note: Results are based on 500 simulation replications and 100 instruments. The first five first-stage coefficients were set equal to one and the remaining 95 to zero in this design. Corr(ϵ, v) is the correlation between first-stage and structural errors. F^* measures the strength of the instruments as outlined in the text. 2SLS(100), LIML(100), and FULL(100) are respectively the 2SLS, LIML, and Fuller(1) estimator using all 100 potential instruments. Many-instrument robust standard errors are computed for LIML(100) and FULL(100) to obtain testing rejection frequencies. LASSO(1) reports results for the 2SLS estimator using LASSO restricted to choose only one instrument. LASSO(C) reports results for the 2SLS estimator using LASSO with a data-dependent penalty. For each estimation procedure, we report root-mean-square-error (RMSE), median bias (Med. Bias), mean absolute deviation (MAD), and rejection frequency for 5% level tests (rp(.05)). For LASSO(C) in the weak instrument design ($F^* = 6.72$), LASSO chose no instruments in 36, 49, and 41 simulation replications for Corr(ϵ, v) = 0, .3, and .6 respectively. Results for LASSO(C) in these cases are reported for the subset of simulation draws where the set of instruments selected by LASSO was non-empty.

Table 4: 2SLS Simulation Results. Exponential Design. N = 150

| Estimator | Corr(e,v) = 0 | | | | Corr(e,v) = .3 | | | | Corr(e,v) = .6 | | | |
|-------------|---------------|-----------|-------|---------|----------------|-----------|-------|---------|----------------|-----------|-------|---------|
| | RMSE | Med. Bias | MAD | rp(.05) | RMSE | Med. Bias | MAD | rp(.05) | RMSE | Med. Bias | MAD | rp(.05) |
| F* = 6.72 | | | | | | | | | | | | |
| 2SLS(100) | 0.045 | 0.003 | 0.032 | 0.054 | 0.131 | 0.122 | 0.122 | 0.824 | 0.245 | 0.244 | 0.244 | 1.000 |
| LIML(100) | 3.589 | -0.008 | 0.227 | 0.036 | 3.752 | 0.018 | 0.257 | 0.050 | 22.419 | 0.042 | 0.210 | 0.092 |
| FULL(100) | 0.681 | -0.008 | 0.223 | 0.036 | 0.686 | 0.018 | 0.250 | 0.050 | 0.589 | 0.044 | 0.207 | 0.096 |
| LASSO(1) | 0.120 | 0.003 | 0.078 | 0.026 | 0.132 | 0.035 | 0.080 | 0.068 | 0.140 | 0.063 | 0.089 | 0.120 |
| LASSO(C) | 0.111 | 0.021 | 0.067 | 0.041 | 0.104 | 0.045 | 0.070 | 0.085 | 0.112 | 0.072 | 0.081 | 0.122 |
| F* = 26.87 | | | | | | | | | | | | |
| 2SLS(100) | 0.073 | 0.008 | 0.051 | 0.064 | 0.170 | 0.153 | 0.153 | 0.596 | 0.321 | 0.316 | 0.316 | 1.000 |
| LIML(100) | 0.374 | 0.012 | 0.153 | 0.042 | 0.585 | -0.006 | 0.148 | 0.032 | 0.275 | 0.012 | 0.124 | 0.054 |
| FULL(100) | 0.343 | 0.012 | 0.152 | 0.042 | 0.426 | -0.005 | 0.147 | 0.032 | 0.259 | 0.014 | 0.124 | 0.054 |
| LASSO(1) | 0.129 | 0.009 | 0.089 | 0.042 | 0.130 | 0.014 | 0.090 | 0.050 | 0.130 | 0.043 | 0.100 | 0.066 |
| LASSO(C) | 0.113 | 0.006 | 0.079 | 0.046 | 0.113 | 0.018 | 0.069 | 0.056 | 0.117 | 0.038 | 0.082 | 0.078 |
| F* = 107.48 | | | | | | | | | | | | |
| 2SLS(100) | 0.091 | 0.000 | 0.060 | 0.052 | 0.157 | 0.125 | 0.127 | 0.290 | 0.278 | 0.259 | 0.259 | 0.830 |
| LIML(100) | 0.144 | 0.007 | 0.099 | 0.050 | 0.154 | 0.000 | 0.101 | 0.078 | 0.140 | 0.004 | 0.096 | 0.044 |
| FULL(100) | 0.143 | 0.007 | 0.098 | 0.050 | 0.154 | 0.000 | 0.101 | 0.078 | 0.140 | 0.005 | 0.096 | 0.044 |
| LASSO(1) | 0.131 | -0.001 | 0.088 | 0.040 | 0.141 | 0.014 | 0.095 | 0.050 | 0.135 | 0.032 | 0.091 | 0.060 |
| LASSO(C) | 0.106 | -0.005 | 0.070 | 0.036 | 0.111 | 0.004 | 0.074 | 0.070 | 0.109 | 0.019 | 0.076 | 0.060 |

Note: Results are based on 500 simulation replications and 100 instruments. The first first-stage coefficients were set equal to $(.7)^{j-1}$ for $j=1, \dots, 100$ denoting the associated instrument. $\text{Corr}(e,v)$ is the correlation between first-stage and structural errors. F^* measures the strength of the instruments as outlined in the text. 2SLS(100), LIML(100), and FULL(100) are respectively the 2SLS, LIML, and Fuller(1) estimator using all 100 potential instruments. Many-instrument robust standard errors are computed for LIML(100) and FULL(100) to obtain testing rejection frequencies. LASSO(1) reports results for the 2SLS estimator using LASSO restricted to choose only one instrument. LASSO(C) reports results for the 2SLS estimator using LASSO with a data-dependent penalty. For each estimation procedure, we report root-mean-square-error (RMSE), median bias (Med. Bias), mean absolute deviation (MAD), and rejection frequency for 5% level tests (rp(.05)). For LASSO(C) in the weak instrument design ($F^* = 6.72$), LASSO chose no instruments in 233, 230, and 221 simulation replications for $\text{Corr}(e,v) = 0, .3,$ and $.6$ respectively. Results for LASSO(C) in these cases are reported for the subset of simulation draws where the set of instruments selected by LASSO was non-empty.

Table 5: 2SLS Simulation Results. Exponential Design. N = 750

| Estimator | Corr(ϵ, ν) = 0 | | | | | Corr(ϵ, ν) = .3 | | | | | Corr(ϵ, ν) = .6 | | | | |
|----------------|-----------------------------|-----------|-------|---------|---------|------------------------------|-----------|-------|---------|---------|------------------------------|-----------|-------|---------|---------|
| | RMSE | Med. Bias | MAD | rp(.05) | rp(.05) | RMSE | Med. Bias | MAD | rp(.05) | rp(.05) | RMSE | Med. Bias | MAD | rp(.05) | rp(.05) |
| 2SLS(100) | 0.020 | -0.001 | 0.014 | 0.036 | 0.057 | 0.054 | 0.054 | 0.054 | 0.776 | 0.109 | 0.108 | 0.108 | 0.108 | 1.000 | 1.000 |
| LIML(100) | 1.410 | -0.003 | 0.072 | 0.028 | 2.111 | 0.002 | 0.081 | 0.062 | 0.062 | 8.000 | 0.008 | 0.067 | 0.067 | 0.070 | 0.070 |
| FULL(100) | 0.171 | -0.003 | 0.069 | 0.018 | 0.157 | 0.004 | 0.078 | 0.060 | 0.060 | 0.132 | 0.013 | 0.063 | 0.063 | 0.074 | 0.074 |
| LASSO(1) | 0.051 | 0.002 | 0.033 | 0.012 | 0.053 | 0.009 | 0.035 | 0.054 | 0.054 | 0.060 | 0.029 | 0.039 | 0.039 | 0.146 | 0.146 |
| LASSO(C) | 0.045 | 0.000 | 0.031 | 0.020 | 0.046 | 0.012 | 0.032 | 0.064 | 0.064 | 0.054 | 0.034 | 0.040 | 0.040 | 0.189 | 0.189 |
| $F^* = 26.87$ | | | | | | | | | | | | | | | |
| 2SLS(100) | 0.034 | 0.002 | 0.022 | 0.050 | 0.077 | 0.071 | 0.071 | 0.071 | 0.624 | 0.140 | 0.136 | 0.136 | 0.136 | 1.000 | 1.000 |
| LIML(100) | 0.082 | 0.003 | 0.052 | 0.052 | 0.070 | 0.002 | 0.048 | 0.056 | 0.056 | 0.062 | 0.001 | 0.040 | 0.040 | 0.054 | 0.054 |
| FULL(100) | 0.081 | 0.003 | 0.051 | 0.048 | 0.069 | 0.004 | 0.047 | 0.054 | 0.054 | 0.061 | 0.003 | 0.040 | 0.040 | 0.056 | 0.056 |
| LASSO(1) | 0.060 | 0.004 | 0.040 | 0.044 | 0.061 | 0.011 | 0.043 | 0.062 | 0.062 | 0.054 | 0.016 | 0.037 | 0.037 | 0.062 | 0.062 |
| LASSO(C) | 0.049 | 0.002 | 0.031 | 0.042 | 0.048 | 0.011 | 0.032 | 0.064 | 0.064 | 0.047 | 0.013 | 0.032 | 0.032 | 0.072 | 0.072 |
| $F^* = 107.48$ | | | | | | | | | | | | | | | |
| 2SLS(100) | 0.040 | -0.001 | 0.029 | 0.036 | 0.071 | 0.057 | 0.057 | 0.057 | 0.278 | 0.124 | 0.117 | 0.117 | 0.117 | 0.846 | 0.846 |
| LIML(100) | 0.052 | 0.000 | 0.038 | 0.028 | 0.055 | 0.000 | 0.036 | 0.060 | 0.060 | 0.053 | 0.001 | 0.037 | 0.037 | 0.068 | 0.068 |
| FULL(100) | 0.052 | 0.000 | 0.038 | 0.028 | 0.054 | 0.001 | 0.036 | 0.056 | 0.056 | 0.052 | 0.002 | 0.037 | 0.037 | 0.066 | 0.066 |
| LASSO(1) | 0.059 | 0.001 | 0.039 | 0.052 | 0.061 | 0.006 | 0.042 | 0.062 | 0.062 | 0.058 | 0.011 | 0.039 | 0.039 | 0.048 | 0.048 |
| LASSO(C) | 0.046 | 0.001 | 0.029 | 0.050 | 0.049 | 0.003 | 0.033 | 0.060 | 0.060 | 0.048 | 0.011 | 0.034 | 0.034 | 0.076 | 0.076 |

Note: Results are based on 500 simulation replications and 100 instruments. The first first-stage coefficients were set equal to $(.7)^{j-1}$ for $j=1, \dots, 100$ denoting the associated instrument. $\text{Corr}(\epsilon, \nu)$ is the correlation between first-stage and structural errors. F^* measures the strength of the instruments as outlined in the text. 2SLS(100), LIML(100), and FULL(100) are respectively the 2SLS, LIML, and Fuller(1) estimator using all 100 potential instruments. Many-instrument robust standard errors are computed for LIML(100) and FULL(100) to obtain testing rejection frequencies. LASSO(1) reports results for the 2SLS estimator using LASSO restricted to choose only one instrument. LASSO(C) reports results for the 2SLS estimator using LASSO with a data-dependent penalty. For each estimation procedure, we report root-mean-square-error (RMSE), median bias (Med. Bias), mean absolute deviation (MAD), and rejection frequency for 5% level tests (rp(.05)). For LASSO(C) in the weak instrument design ($F^* = 6.72$), LASSO chose no instruments in 151, 127, and 150 simulation replications for $\text{Corr}(\epsilon, \nu) = 0, .3,$ and $.6$ respectively. Results for LASSO(C) in these cases are reported for the subset of simulation draws where the set of instruments selected by LASSO was non-empty.

Table 6: 2SLS Simulation Results. Exponential Design. N = 1500

| Estimator | Corr(e,v) = 0 | | | | Corr(e,v) = .3 | | | | Corr(e,v) = .6 | | | |
|------------|---------------|-----------|-------|---------|----------------|-----------|-------|---------|----------------|-----------|-------|---------|
| | RMSE | Med. Bias | MAD | rp(.05) | RMSE | Med. Bias | MAD | rp(.05) | RMSE | Med. Bias | MAD | rp(.05) |
| 2SLS(100) | 0.014 | 0.001 | 0.010 | 0.060 | 0.042 | 0.040 | 0.040 | 0.842 | 0.077 | 0.077 | 0.077 | 1.000 |
| LIML(100) | 0.182 | 0.005 | 0.052 | 0.046 | 0.585 | 0.010 | 0.049 | 0.052 | 1.036 | 0.002 | 0.049 | 0.070 |
| FULL(100) | 0.105 | 0.005 | 0.050 | 0.042 | 0.098 | 0.011 | 0.047 | 0.046 | 0.087 | 0.006 | 0.045 | 0.070 |
| LASSO(1) | 0.040 | 0.001 | 0.025 | 0.036 | 0.040 | 0.014 | 0.027 | 0.064 | 0.048 | 0.025 | 0.030 | 0.166 |
| LASSO(C) | 0.034 | 0.002 | 0.020 | 0.045 | 0.034 | 0.016 | 0.025 | 0.068 | 0.038 | 0.026 | 0.029 | 0.189 |
| F* = 6.72 | | | | | | | | | | | | |
| 2SLS(100) | 0.023 | -0.002 | 0.015 | 0.048 | 0.054 | 0.049 | 0.049 | 0.602 | 0.101 | 0.099 | 0.099 | 0.996 |
| LIML(100) | 0.056 | -0.006 | 0.034 | 0.046 | 0.057 | -0.005 | 0.033 | 0.038 | 0.049 | 0.001 | 0.030 | 0.056 |
| FULL(100) | 0.055 | -0.006 | 0.033 | 0.044 | 0.055 | -0.004 | 0.032 | 0.038 | 0.048 | 0.003 | 0.029 | 0.056 |
| LASSO(1) | 0.043 | -0.003 | 0.027 | 0.062 | 0.041 | 0.005 | 0.024 | 0.058 | 0.042 | 0.016 | 0.029 | 0.082 |
| LASSO(C) | 0.035 | -0.003 | 0.022 | 0.066 | 0.034 | 0.002 | 0.023 | 0.048 | 0.035 | 0.013 | 0.025 | 0.092 |
| F* = 26.87 | | | | | | | | | | | | |
| 2SLS(100) | 0.031 | -0.002 | 0.021 | 0.066 | 0.049 | 0.039 | 0.040 | 0.284 | 0.086 | 0.082 | 0.082 | 0.838 |
| LIML(100) | 0.041 | -0.001 | 0.027 | 0.070 | 0.037 | -0.001 | 0.023 | 0.048 | 0.037 | 0.000 | 0.025 | 0.058 |
| FULL(100) | 0.040 | -0.001 | 0.026 | 0.068 | 0.037 | -0.001 | 0.023 | 0.050 | 0.037 | 0.001 | 0.025 | 0.060 |
| LASSO(1) | 0.044 | -0.001 | 0.030 | 0.068 | 0.042 | 0.002 | 0.029 | 0.054 | 0.043 | 0.012 | 0.028 | 0.064 |
| LASSO(C) | 0.035 | 0.000 | 0.025 | 0.054 | 0.033 | 0.003 | 0.021 | 0.060 | 0.036 | 0.008 | 0.025 | 0.088 |

Note: Results are based on 500 simulation replications and 100 instruments. The first first-stage coefficients were set equal to $(.7)^{j-1}$ for $j=1, \dots, 100$ denoting the associated instrument. $Corr(e,v)$ is the correlation between first-stage and structural errors. F^* measures the strength of the instruments as outlined in the text. 2SLS(100), LIML(100), and FULL(100) are respectively the 2SLS, LIML, and Fuller(1) estimator using all 100 potential instruments. Many-instrument robust standard errors are computed for LIML(100) and FULL(100) to obtain testing rejection frequencies. LASSO(1) reports results for the 2SLS estimator using LASSO restricted to choose only one instrument. LASSO(C) reports results for the 2SLS estimator using LASSO with a data-dependent penalty. For each estimation procedure, we report root-mean-square-error (RMSE), median bias (Med. Bias), mean absolute deviation (MAD), and rejection frequency for 5% level tests (rp(.05)). For LASSO(C) in the weak instrument design ($F^* = 6.72$), LASSO chose no instruments in 145, 146, and 151 simulation replications for $Corr(e,v) = 0, .3$, and $.6$ respectively. Results for LASSO(C) in these cases are reported for the subset of simulation draws where the set of instruments selected by LASSO was non-empty.

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

| | Home Prices | | | GDP |
|----------|-------------|-----------|--------------|----------|
| | FHFA | Non-Metro | Case-Shiller | log(GDP) |
| 2SLS | 0.0204 | 0.0618 | 0.0329 | 0.0015 |
| s.e. | 0.0215 | 0.0272 | 0.0091 | 0.0255 |
| p-value | 0.3631 | 0.0464 | 0.0047 | 0.9542 |
| LASSO(2) | -0.0001 | 0.0288 | 0.0496 | -0.0004 |
| s.e. | 0.0207 | 0.0141 | 0.0091 | 0.0292 |
| p-value | 0.9962 | 0.0684 | 0.0003 | 0.9893 |
| LASSO(C) | 0.0259 | 0.0189 | 0.0342 | 0.0088 |
| s.e. | 0.0129 | 0.0050 | 0.0069 | 0.0184 |
| p-value | 0.0699 | 0.0036 | 0.0006 | 0.6418 |
| S | 13 | 14 | 24 | 9 |

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, and controls for the demographics of judges available within the circuit. Each column corresponds to a different dependent variable. FHFA, Non-Metro, and Case-Shiller are house price indexes, and log(GDP) is the log of state-level GDP. 2SLS is uses the 2SLS estimator with the original instruments in Chen and Yeh (2010). LASSO(2) reports 2SLS estimates obtained using instruments selected by LASSO restricted to choose two instruments. LASSO(C) provides 2SLS estimates obtained using instruments selected by LASSO with a data-dependent penalty choice. Rows labeled s.e. provide the estimated standard errors of the associated estimator, and rows labeled p-value provide the corresponding p-values for the null hypothesis that the coefficient is zero. All standard errors are computed with clustering at the circuit level. Case-Shiller and Non-Metro estimates are based on 11 circuits, and FHFA and log(GDP) estimates are based on 12 circuits. The corresponding t-critical values for t-distributions with 10 and 11 degrees of freedom are respectively 2.228 and 2.201 for 5% level two-sided tests. S is the number of instruments chosen with LASSO using the data-dependent penalty.