ECONOMETRICA

JOURNAL OF THE ECONOMETRIC SOCIETY

An International Society for the Advancement of Economic Theory in its Relation to Statistics and Mathematics

http://www.econometricsociety.org/

Econometrica, Vol. 80, No. 6 (November, 2012), 2369–2429

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

A. BELLONI
Duke University Fuqua School of Business, Durham, NC 27708, U.S.A.

D. CHEN
ETH Zurich, CH-8092, Zurich, Switzerland

V. CHERNOZHUKOV
Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

C. HANSEN
University of Chicago Booth School of Business, Chicago, IL 60637, U.S.A.

Econometrica, Vol. 80, No. 6 (November, 2012), 2369–2429

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

BY A. BELLONI, D. CHEN, V. CHERNOZHUKOV, AND C. HANSEN[1]

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. Our results apply even when p is much larger than the sample size, n. We show that the IV estimator based on using Lasso or post-Lasso in the first stage is root-n consistent and asymptotically normal when the first stage is approximately sparse, that is, when the conditional expectation of the endogenous variables given the instruments can be well-approximated by a relatively small set of variables whose identities may be unknown. We also show that the estimator is semi-parametrically efficient when the structural error is homoscedastic. Notably, our results allow for imperfect model selection, and do not rely upon the unrealistic "beta-min" conditions that are widely used to establish validity of inference following model selection (see also Belloni, Chernozhukov, and Hansen (2011b)). 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 outperforms an intuitive benchmark.

Optimal instruments are conditional expectations. In developing the IV results, we establish a series of new results for Lasso and post-Lasso estimators of nonparametric conditional expectation functions which are of independent theoretical and practical interest. We construct a modification of Lasso designed to deal with non-Gaussian, heteroscedastic disturbances that uses a data-weighted ℓ_1-penalty function. By innovatively using moderate deviation theory for self-normalized sums, we provide convergence rates for the resulting Lasso and post-Lasso estimators that are as sharp as the corresponding rates in the homoscedastic Gaussian case under the condition that $\log p = o(n^{1/3})$. We also provide a data-driven method for choosing the penalty level that must be specified in obtaining Lasso and post-Lasso estimates and establish its asymptotic validity under non-Gaussian, heteroscedastic disturbances.

KEYWORDS: Inference on a low-dimensional parameter after model selection, imperfect model selection, instrumental variables, Lasso, post-Lasso, data-driven penalty, heteroscedasticity, non-Gaussian errors, moderate deviations for self-normalized sums.

1. INTRODUCTION

INSTRUMENTAL VARIABLES (IV) TECHNIQUES are widely used in applied economic research. While these methods provide a useful tool for identifying

[1] This is a revision of an October 2010 ArXiv/CEMMAP paper with the same title. Preliminary results of this paper were first presented at Chernozhukov's invited Cowles Foundation lecture at the Northern American meetings of the Econometric Society in June of 2009. We thank seminar participants at Brown, Columbia, Hebrew University, Tel Aviv University, Harvard–MIT, the Dutch Econometric Study Group, Fuqua School of Business, NYU, and New Economic School, for helpful comments. We also thank Denis Chetverikov, JB Doyle, and Joonhwan Lee for thorough reading of this paper and very useful comments.

DOI: 10.3982/ECTA9626

structural effects of interest, their application often results in imprecise inference. One way to improve the precision of instrumental variables estimators is to use many instruments or to try to approximate the optimal instruments, as in Amemiya (1974), Chamberlain (1987), and Newey (1990). Estimation of optimal instruments is generally done nonparametrically and thus implicitly makes use of many constructed instruments such as polynomials. The promised improvement in efficiency is appealing, but IV estimators based on many instruments may have poor properties. See, for example, Bekker (1994), Chao and Swanson (2005), Hansen, Hausman, and Newey (2008), and Chao, Swanson, Hausman, Newey, and Woutersen (2012), which proposed solutions for this problem based on "many-instrument" asymptotics.[2]

In this paper, we contribute to the literature on IV estimation with many instruments by considering the use of Lasso and post-Lasso for estimating the first-stage regression of endogenous variables on the instruments. Lasso is a widely used method that acts both as an estimator of regression functions and as a model selection device. Lasso solves for regression coefficients by minimizing the sum of the usual least squares objective function and a penalty for model size through the sum of the absolute values of the coefficients. The resulting Lasso estimator selects instruments and estimates the first-stage regression coefficients via a shrinkage procedure. The post-Lasso estimator discards the Lasso coefficient estimates and uses the data-dependent set of instruments selected by Lasso to refit the first-stage regression via ordinary least squares (OLS) to alleviate Lasso's shrinkage bias. For theoretical and simulation evidence regarding Lasso's performance, see Bai and Ng (2008, 2009a), Bickel, Ritov, and Tsybakov (2009), Bunea, Tsybakov, and Wegkamp (2006, 2007a, 2007b), Candes and Tao (2007), Huang, Horowitz, and Wei (2010), Knight (2008), Koltchinskii (2009), Lounici (2008), Lounici et al. (2010), Meinshausen and Yu (2009), Rosenbaum and Tsybakov (2008), Tibshirani (1996), van de Geer (2008), Wainwright (2009), Zhang and Huang (2008), Belloni and Chernozhukov (2012), and Bühlmann and van de Geer (2011), among many others. See Belloni and Chernozhukov (2012) for analogous results on post-Lasso.

Using Lasso-based methods to form first-stage predictions in IV estimation provides a practical approach to obtaining the efficiency gains from using optimal instruments while dampening the problems associated with many instruments. We show that Lasso-based procedures produce first-stage predictions that provide good approximations to the optimal instruments even when the number of available instruments is much larger than the sample size when the first stage is approximately sparse—that is, when there exists a relatively small set of important instruments whose identities are unknown that well-approximate the conditional expectation of the endogenous variables given

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

the instruments. Under approximate sparsity, estimating the first-stage relationship using Lasso-based procedures produces IV estimators that are root-n consistent and asymptotically normal. The IV estimator with Lasso-based first stage also achieves the semiparametric efficiency bound under the additional condition that structural errors are homoscedastic. Our results allow imperfect model selection and do not impose "beta-min" conditions that restrict the minimum allowable magnitude of the coefficients on relevant regressors. We also provide a consistent asymptotic variance estimator. Thus, our results generalize the IV procedure of Newey (1990) and Hahn (2002) based on conventional series approximation of the optimal instruments. Our results also generalize Bickel, Ritov, and Tsybakov (2009) by providing inference and confidence sets for the second-stage IV estimator based on Lasso or post-Lasso estimates of the first-stage predictions. To our knowledge, our result is the first to verify root-n consistency and asymptotic normality of an estimator for a low-dimensional structural parameter in a high-dimensional setting without imposing the very restrictive "beta-min" condition.[3] (For similar results in the partially linear regression see Belloni, Chernozhukov, and Hansen (2011a, 2011b).) Our results also remain valid in the presence of heteroscedasticity and thus provide a useful complement to existing approaches in the many-instrument literature, which often rely on homoscedasticity and may be inconsistent in the presence of heteroscedasticity; see Chao et al. (2012) for a notable exception that allows for heteroscedasticity and gives additional discussion.

Instrument selection procedures complement existing/traditional methods that are meant to be robust to many instruments but are not a universal solution to the many-instruments problem. The good performance of instrument selection procedures relies on approximate sparsity. Unlike traditional IV methods, instrument selection procedures do not require the identity of these "important" variables to be known a priori, as the identity of these instruments will be estimated from the data. This flexibility comes with the cost that instrument selection tends not to work well when the first stage is not approximately sparse. When approximate sparsity breaks down, instrument selection procedures may select too few or no instruments or may select too many instruments. Two scenarios where this failure is likely to occur are the weak-instrument case (e.g., Staiger and Stock (1997), Andrews, Moreira, and Stock (2006), Andrews and Stock (2005), Moreira (2003), Kleibergen (2002, 2005)), and the many-weak-instrument case (e.g., Bekker (1994), Chao and Swanson

[3] The "beta-min" condition requires the relevant coefficients in the regression to be separated from zero by a factor that exceeds the potential estimation error. This condition implies that the identities of the relevant regressors may be perfectly determined. There is a large body of theoretical work that uses such a condition and thus implicitly assumes that the resulting post-model selection estimator is the same as the oracle estimator that knows the identities of the relevant regressors. See Bühlmann and van de Geer (2011) for the discussion of the "beta-min" condition and the theoretical role it plays in obtaining "oracle" results.

(2005), Hansen, Hausman, and Newey (2008), Chao et al. (2012)). We consider two modifications of our basic procedure aimed at alleviating these concerns. In Section 4, we present a sup-score testing procedure that is related to Anderson and Rubin (1949) and Staiger and Stock (1997) but is better suited to cases with very many instruments; and we consider a split-sample IV estimator in Section 5 that combines instrument selection via Lasso with the sample-splitting method of Angrist and Krueger (1995). While these two procedures are steps toward addressing weak identification concerns with very many instruments, further exploration of the interplay between weak-instrument or many-weak-instrument methods and variable selection would be an interesting avenue for additional research.

Our paper also contributes to the growing literature on Lasso-based methods by providing results for Lasso-based estimators of nonparametric conditional expectations. We consider a modified Lasso estimator with penalty weights designed to deal with non-Gaussianity and heteroscedastic errors. This new construction allows us to innovatively use the results of moderate deviation theory for self-normalized sums of Jing, Shao, and Wang (2003) to provide convergence rates for Lasso and post-Lasso. The derived convergence rates are as sharp as in the homoscedastic Gaussian case under the weak condition that the log of the number of regressors p is small relative to $n^{1/3}$, that is, $\log p = o(n^{1/3})$. Our construction generalizes the standard Lasso estimator of Tibshirani (1996) and allows us to generalize the Lasso results of Bickel, Ritov, and Tsybakov (2009) and post-Lasso results of Belloni and Chernozhukov (2012), both of which assume homoscedasticity and Gaussianity. The construction as well as theoretical results are important for applied economic analysis where researchers are concerned about heteroscedasticity and non-Gaussianity in their data. We also provide a data-driven method for choosing the penalty that must be specified to obtain Lasso and post-Lasso estimates, and we establish its asymptotic validity allowing for non-Gaussian, heteroscedastic disturbances. Ours is the first paper to provide such a data-driven penalty, which was previously not available even in the Gaussian case.[4] These results are of independent interest in a variety of theoretical and applied settings.

We illustrate the performance of Lasso-based IV through simulation experiments. In these experiments, we find that a feasible Lasso-based procedure that uses our data-driven penalty performs well across a range of simulation designs where sparsity is a reasonable approximation. In terms of estimation risk, it outperforms the estimator of Fuller (1977) (FULL),[5] which is robust

[4] One exception is the work of Belloni, Chernozhukov, and Wang (2011b), which considered square-root-Lasso estimators and showed that their use allows for pivotal penalty choices. Those results strongly rely on homoscedasticity.

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

to many instruments (e.g., Hansen, Hausman, and Newey (2008)), except in a design where sparsity breaks down and the sample size is large relative to the number of instruments. In terms of size of 5% level tests, the Lasso-based IV estimator performs comparably to or better than FULL in all cases we consider. Overall, the simulation results are in line with the theory and favorable to the proposed Lasso-based IV procedures.

Finally, we demonstrate the potential gains of the Lasso-based procedure in an application where there are many available instruments among which there is not a clear a priori way to decide which instruments to use. 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 Chen and Yeh (2010). We follow the identification strategy of Chen and Yeh (2010), who used the random assignment of judges to three-judge panels that are then assigned to eminent domain cases to justify using the demographic characteristics of the judges on the realized panels as instruments for their decision. This strategy produces a situation in which there are many potential instruments in that all possible sets of characteristics of the three-judge panel are valid instruments. We find that the Lasso-based estimates using the data-dependent penalty produce much larger first-stage Wald statistics and generally have smaller estimated second-stage standard errors than estimates obtained using the baseline instruments of Chen and Yeh (2010).

1.1. *Relationship to Econometric Literature on Variable Selection and Shrinkage*

The idea of instrument selection goes back to Kloek and Mennes (1960) and Amemiya (1966), who searched among principal components to approximate the optimal instruments. Related ideas appear in dynamic factor models as in Bai and Ng (2010), Kapetanios and Marcellino (2010), and Kapetanios, Khalaf, and Marcellino (2011). Factor analysis differs from our approach, though principal components, factors, ridge fits, and other functions of the instruments could be considered among the set of potential instruments to select from.[6]

There are several other papers that explore the use of modern variable selection methods in econometrics, including some papers that apply these procedures to IV estimation. Bai and Ng (2009b) considered an approach to instrument selection that is closely related to ours, based on boosting. The latter method is distinct from Lasso (cf. Bühlmann (2006)), but it also does not rely on knowing the identity of the most important instruments. They showed through simulation examples that instrument selection via boosting works well in the designs they considered, but did not provide formal results. Bai and Ng

[6] Approximate sparsity should be understood to be relative to a given structure defined by the set of instruments considered. Allowing for principal components or ridge fits among the potential regressors considerably expands the applicability of the approximately sparse framework.

(2009b) also expressly mentioned the idea of using the Lasso method for instrument selection, though they focused their analysis on the boosting method. Our paper complements their analysis by providing a formal set of conditions under which Lasso variable selection will provide good first-stage predictions, and providing theoretical estimation and inference results for the resulting IV estimator. One of our theoretical results for the IV estimator is also sufficiently general to cover the use of any other first-stage variable selection procedure, including boosting, that satisfies a set of provided rate conditions. Caner (2009) considered estimation by penalizing the generalized method of moments (GMM) criterion function by the ℓ_γ-norm of the coefficients for $0 < \gamma < 1$. The analysis of Caner (2009) assumed that the number of parameters p is fixed in relation to the sample size, and so it is complementary to our approach where we allow $p \to \infty$ as $n \to \infty$. Other uses of Lasso in econometrics include Bai and Ng (2008), Belloni, Chernozhukov, and Hansen (2011b), Brodie, Daubechies, Mol, Giannone, and Loris (2009), DeMiguel, Garlappi, Nogales, and Uppal (2009), Huang, Horowitz, and Wei (2010), Knight (2008), and others. An introductory treatment of this topic was given in Belloni and Chernozhukov (2011b), and Belloni, Chernozhukov, and Hansen (2011a) provided a review of Lasso targeted at economic applications.

Our paper is also related to other shrinkage-based approaches to dealing with many instruments. Chamberlain and Imbens (2004) considered IV estimation with many instruments using a shrinkage estimator based on putting a random coefficients structure over the first-stage coefficients in a homoscedastic setting. In a related approach, Okui (2011) considered the use of ridge regression for estimating the first-stage regression in a homoscedastic framework where the instruments may be ordered in terms of relevance. Okui (2011) derived the asymptotic distribution of the resulting IV estimator and provided a method for choosing the ridge regression smoothing parameter that minimizes the higher-order asymptotic mean squared error (MSE) of the IV estimator. These two approaches are related to the approach we pursue in this paper in that both use shrinkage in estimating the first stage, but differ in the shrinkage methods they use. Their results are also only supplied in the context of homoscedastic models. Donald and Newey (2001) considered a variable selection procedure that minimizes higher-order asymptotic MSE which relies on a priori knowledge that allows one to order the instruments in terms of instrument strength. Our use of Lasso as a variable selection technique does not require any a priori knowledge about the identity of the most relevant instruments, and so provides a useful complement to Donald and Newey (2001) and Okui (2011). Carrasco (2012) provided an interesting approach to IV estimation with many instruments based on directly regularizing the inverse that appears in the definition of the two-stage least squares (2SLS) estimator; see also Carrasco and Tchuente Nguembu (2012). Carrasco (2012) considered three regularization schemes, including Tikhonov regularization, which corresponds to ridge regression, and showed that the regularized estimators achieve the semiparametric efficiency bound under some conditions. Carrasco's (2012) approach implicitly uses ℓ_2-norm penalization and hence differs from and complements our approach. A valuable feature of Carrasco (2012) is the provision of a data-dependent method for choosing the regularization parameter based on minimizing higher-order asymptotic MSE, following Donald and Newey (2001) and Okui (2011). Finally, in work that is more recent than the present paper, Gautier and Tsybakov (2011) considered the important case where the structural equation in an instrumental variables model is itself very high-dimensional, and proposed a new estimation method related to the Dantzig selector and the square-root-Lasso. They also provided an interesting inference method that differs from the one we consider.

1.2. *Notation*

In what follows, we work with triangular array data $\{(z_{i,n}, i = 1, \dots, n), n = 1, 2, 3, \dots\}$ defined on some common probability space $(\Omega, \mathcal{A}, \mathrm{P})$. Each $z_{i,n} = (y_{i,n}', x_{i,n}', d_{i,n}')'$ is a vector, with components defined below in what follows, and these vectors are i.n.i.d.—independent across i, but not necessarily identically distributed. The law P_n of $\{z_{i,n}, i = 1, \dots, n\}$ can change with n, though we do not make explicit use of P_n. Thus, all parameters that characterize the distribution of $\{z_{i,n}, i = 1, \dots, n\}$ are implicitly indexed by the sample size n, but we omit the index n in what follows to simplify notation. We use triangular array asymptotics to better capture some finite-sample phenomena and to retain the robustness of conclusions to perturbations of the data-generating process. We also use the empirical process notation $\mathbb{E}_n[f] := \mathbb{E}_n[f(z_i)] := \sum_{i=1}^n f(z_i)/n$, and $\mathbb{G}_n(f) := \sum_{i=1}^n (f(z_i) - \mathrm{E}[f(z_i)])/\sqrt{n}$. Since we want to deal with i.n.i.d. data, we also introduce the average expectation operator: $\bar{\mathrm{E}}[f] := \mathrm{E}\mathbb{E}_n[f] = \mathrm{E}\mathbb{E}_n[f(z_i)] = \sum_{i=1}^n \mathrm{E}[f(z_i)]/n$. The ℓ_2-norm is denoted by $\|\cdot\|_2$, and the ℓ_0-norm, $\|\cdot\|_0$, denotes the number of nonzero components of a vector. We use $\|\cdot\|_\infty$ to denote the maximal element of a vector. The empirical $L^2(\mathbb{P}_n)$ norm of a random variable W_i is defined as $\|W_i\|_{2,n} := \sqrt{\mathbb{E}_n[W_i^2]}$. When the empirical $L^2(\mathbb{P}_n)$ norm is applied to regressors $f_1, \dots, f_p$ and a vector $\delta \in \mathbb{R}^p$, $\|f_i'\delta\|_{2,n}$, it is called the prediction norm. 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 denote $T^c := \{1, 2, \dots, p\} \setminus T$. We use the notation $(a)_+ = \max\{a, 0\}$, $a \vee b = \max\{a, b\}$, and $a \wedge b = \min\{a, b\}$. We also use the notation $a \lesssim b$ to denote $a \leq cb$ for some constant $c > 0$ that does not depend on n; and $a \lesssim_{\mathrm{P}} b$ to denote $a = O_{\mathrm{P}}(b)$. For an event E, we say that E w.p. → 1 when E occurs with probability approaching 1 as n grows. We say $X_n =_d Y_n + o_{\mathrm{P}}(1)$ to mean that X_n has the same distribution as Y_n up to a term $o_{\mathrm{P}}(1)$ that vanishes in probability.

2. SPARSE MODELS AND METHODS FOR OPTIMAL INSTRUMENTAL VARIABLES

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

2.1. *The IV Model and Statement of the Problem*

The model is $y_i = d_i'\alpha_0 + \epsilon_i$, where α_0 denotes the true value of a vector-valued parameter α. y_i is the response variable, and d_i is a finite k_d-vector of variables whose first k_e elements contain endogenous variables. The disturbance ϵ_i obeys, for all i (and n),

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

where x_i is a k_x-vector of instrumental variables.

As a motivation, suppose that the structural disturbance is conditionally homoscedastic, namely, for all i, $\mathrm{E}[\epsilon_i^2|x_i] = \sigma^2$. Given a k_d-vector of instruments $A(x_i)$, the standard IV estimator of α_0 is given by $\widehat{\alpha} = (\mathbb{E}_n[A(x_i)d_i'])^{-1} \times \mathbb{E}_n[A(x_i)y_i]$, where $\{(x_i, d_i, y_i), i = 1, \dots, n\}$ is an i.i.d. sample from the IV model above. For a given $A(x_i)$, $\sqrt{n}(\widehat{\alpha} - \alpha_0) =_d N(0, Q_0^{-1}\Omega_0 Q_0^{-1'}) + o_{\mathrm{P}}(1)$, where $Q_0 = \bar{\mathrm{E}}[A(x_i)d_i']$ and $\Omega_0 = \sigma^2\bar{\mathrm{E}}[A(x_i)A(x_i)']$ under standard conditions. Setting $A(x_i) = D(x_i) = \mathrm{E}[d_i|x_i]$ minimizes the asymptotic variance, which becomes

$$\Lambda^* = \sigma^2\big\{\bar{\mathrm{E}}\big[D(x_i)D(x_i)'\big]\big\}^{-1},$$

the semiparametric efficiency bound for estimating α_0; see Amemiya (1974), Chamberlain (1987), and Newey (1990). In practice, the optimal instrument $D(x_i)$ is an unknown function and has to be estimated. In what follows, we investigate the use of sparse methods—namely Lasso and post-Lasso—in estimating the optimal instruments. The resulting IV estimator is asymptotically as efficient as the infeasible optimal IV estimator above.

Note that if d_i contains exogenous components w_i, then $d_i = (d_{i1}, \dots, d_{ik_e}, w_i')'$, where the first k_e variables are endogenous. Since the rest of the components w_i are exogenous, they appear in $x_i = (w_i', \tilde{x}_i')'$. It follows that $D_i := D(x_i) := \mathrm{E}[d_i|x_i] = (\mathrm{E}[d_1|x_i], \dots, \mathrm{E}[d_{k_e}|x_i], w_i')'$; that is, the estimator of w_i is simply w_i. Therefore, we discuss estimation of the conditional expectation functions:

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

In what follows, we focus on the strong-instruments case, which translates into the assumption that $Q = \bar{\mathrm{E}}[D(x_i)D(x_i)']$ has eigenvalues bounded away from zero and from above. We also present an inference procedure that remains valid in the absence of strong instruments which is related to Anderson and Rubin (1949) and Staiger and Stock (1997) but allows for $p \gg n$.

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

Suppose there is a very large list of instruments,

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

to be used in estimation of conditional expectations $D_l(x_i), l = 1, \dots, k_e$, where the number of instruments p is possibly much larger than the sample size n.

For example, high-dimensional instruments f_i could arise as any combination of the following two cases. First, the list of available instruments may simply be large, in which case $f_i = x_i$ as in, for example, Amemiya (1974) and Bekker (1994). Second, the list f_i could consist of a large number of series terms with respect to some elementary regressor vector x_i; for example, f_i could be composed of B-splines, dummies, polynomials, and various interactions as in Newey (1990) or Hahn (2002), among others. We term the first example the many-instrument case and the second example the many-series-instrument case and note that our formulation does not require us to distinguish between the two cases. We mainly use the term "series instruments" and contrast our results with those in the seminal work of Newey (1990) and Hahn (2002), though our results are not limited to canonical series regressors as in Newey (1990) and Hahn (2002). 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 Newey (1990) and Hahn (2002) to approximate the optimal instruments.

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

$$\begin{aligned} &D_l(x_i) = f_i'\beta_{l0}, \quad l = 1, \dots, k_e, \\ &\max_{1\le l\le k_e} \|\beta_{l0}\|_0 = \max_{1\le l\le k_e} \sum_{j=1}^{p} 1\{\beta_{l0j} \neq 0\} \le s \ll n. \end{aligned} \tag{2.1}$$

This simple sparsity model generalizes the classic parametric model of optimal instruments of Amemiya (1974) by letting the identities of the relevant instruments $T_l = \text{support}(\beta_{l0}) = \{j \in \{1, \dots, p\} : |\beta_{l0j}| > 0\}$ be unknown.

The model given by (2.1) is unrealistic in that it presumes exact sparsity. We make no formal use of this model, but instead use a much more general approximately sparse or nonparametric model:

CONDITION AS—Approximately Sparse Optimal Instrument: Each optimal instrument function $D_l(x_i)$ is well-approximated by a function of unknown $s \geq 1$ instruments:

$$D_l(x_i) = f_i'\beta_{l0} + a_l(x_i), \quad l = 1, \dots, k_e, \quad k_e \text{ fixed},$$
$$\max_{1 \leq l \leq k_e} \|\beta_{l0}\|_0 \leq s = o(n), \quad \max_{1 \leq l \leq k_e} \left[\mathbb{E}_n a_l(x_i)^2\right]^{1/2} \leq c_s \lesssim_P \sqrt{s/n}. \tag{2.2}$$

Condition AS is the key assumption. It requires that there are at most s terms for each endogenous variable that are able to approximate the conditional expectation function $D_l(x_i)$ up to approximation error $a_l(x_i)$, chosen to be no larger than the conjectured size $\sqrt{s/n}$ of the estimation error of the infeasible estimator that knows the identity of these important variables, the "oracle estimator." In other words, 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. Importantly, the assumption allows the identity

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

to be unknown and to differ for $l = 1, \dots, k_e$.

For a detailed motivation and discussion of this assumption, we refer the reader to Belloni, Chernozhukov, and Hansen (2011a). Condition AS generalizes the conventional series approximation of optimal instruments in Newey (1990, 1997) and Hahn (2002) by letting the identities of the most important s series terms T_l be unknown. The rate $\sqrt{s/n}$ generalizes the rate obtained with the optimal number s of series terms in Newey (1990) for estimating conditional expectations by not relying on knowledge of what s series terms to include. Knowing the identities of the most important series terms is unrealistic in many examples. 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 approximation could come from the combination of completely different bases, for example, 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 a key growth condition,

$$\frac{s^2 \log^2(p \vee n)}{n} \to 0,$$

holds along with other more technical conditions. The growth condition requires the optimal instruments to be sufficiently smooth so that a small (relative to n) number of series terms can be used to approximate them well. The use of a small set of instruments ensures that the impact of first-stage estimation on the IV estimator is asymptotically negligible. We can weaken this condition to $s\log(p \vee n) = o(n)$ by using the sample-splitting idea from the many-instruments literature.

2.3. *Lasso-Based Estimation Methods for Optimal Instruments and Other Conditional Expectation Functions*

Let us write the first-stage regression equations as

$$d_{il} = D_l(x_i) + v_{il}, \quad \mathrm{E}[v_{il}|x_i] = 0, \quad l = 1, \dots, k_e. \tag{2.3}$$

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

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

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

Consider the usual least squares criterion function:

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

The Lasso estimator is defined as a solution of the optimization program

$$\widehat{\beta}_{lL} \in \arg\min_{\beta \in \mathbb{R}^p} \widehat{Q}_l(\beta) + \frac{\lambda}{n}\|\widehat{\Upsilon}_l\beta\|_1, \tag{2.4}$$

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

Our analysis will first employ the following "ideal" penalty loadings:

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

The ideal option is not feasible but leads to rather sharp theoretical bounds on estimation risk. This option is not feasible since v_{il} is not observed. In practice, we estimate the ideal loadings by first using conservative penalty loadings and then plugging in the resulting estimated residuals in place of v_{il} to obtain the refined loadings. This procedure could be iterated via Algorithm A.1 stated in the Appendix.

The idea behind the ideal penalty loading is to introduce self-normalization of the first-order condition of the Lasso problem by using data-dependent penalty loadings. This self-normalization allows us to apply moderate deviation theory of Jing, Shao, and Wang (2003) for self-normalized sums to bound deviations of the maximal element of the score vector

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

which provides a representation of the estimation noise in the problem. Specifically, the use of self-normalized moderate deviation theory allows us to establish that

$$\mathrm{P}\Big(\sqrt{n}\max_{1\leq l\leq k_e}\|S_l\|_\infty\leq 2\Phi^{-1}\big(1-\gamma/(2k_e p)\big)\Big)\geq 1-\gamma+o(1), \tag{2.5}$$

from which we obtain sharp convergence results for the Lasso estimator under non-Gaussianity and heteroscedasticity. Without using these loadings, we may not be able to achieve the same sharp rate of convergence. It is important to emphasize that our construction of the penalty loadings for Lasso is new and differs from the canonical penalty loadings proposed in Tibshirani (1996) and Bickel, Ritov, and Tsybakov (2009). Finally, to ensure the good performance of the Lasso estimator, one needs to select the penalty level λ/n to dominate the noise for all k_e regression problems simultaneously; that is, the penalty level should satisfy

$$\mathrm{P}\Big(\lambda/n\geq c\max_{1\leq l\leq k_e}\|S_l\|_\infty\Big)\to 1 \tag{2.6}$$

for some constant $c>1$. The bound (2.5) suggests that this can be achieved by selecting

$$\begin{aligned}\lambda&=c2\sqrt{n}\Phi^{-1}\big(1-\gamma/(2k_e p)\big),\\ &\quad\text{with}\quad\gamma\to 0,\quad\log(1/\gamma)\lesssim\log(p\vee n),\end{aligned} \tag{2.7}$$

which implements (2.6). Our current recommendation is to set the confidence level $\gamma=0.1/\log(p\vee n)$ and the constant $c=1.1$.[7]

The post-Lasso estimator is defined as the ordinary least squares regression applied to the model $\widehat{I}_l\supseteq\widehat{T}_l$, where $\widehat{T}_l$ is the model selected by Lasso:

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

The set $\widehat{I}_l$ can contain additional variables not selected by Lasso, but we require the number of such variables to be similar to or smaller than the number selected by Lasso. The post-Lasso estimator $\widehat{\beta}_{l\mathrm{PL}}$ is

$$\widehat{\beta}_{l\mathrm{PL}}\in\arg\min_{\beta\in\mathbb{R}^p:\beta_{\widehat{I}_l^c}=0}\widehat{Q}_l(\beta),\quad l=1,\ldots,k_e. \tag{2.8}$$

[7] We note that there is not much room to change c. Theoretically, we require $c>1$, and finite-sample experiments show that increasing c away from $c=1$ worsens the performance. Hence a value slightly above unity, namely $c=1.1$, is our current recommendation. The simulation evidence suggests that setting c to any value near 1, including $c=1$, does not impact the result noticeably.

In words, this estimator is ordinary least squares (OLS) using only the instruments/regressors whose coefficients were estimated to be nonzero by Lasso and any additional variables the researcher feels are important despite having Lasso coefficient estimates of zero.

Lasso and post-Lasso are motivated by the desire to predict the target function well without overfitting. Clearly, the OLS estimator is not consistent for estimating the target function when $p > n$. Some approaches based on Bayesian Information Criterion (BIC)-penalization of model size are consistent but computationally infeasible. The Lasso estimator of Tibshirani (1996) 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 regression function $D_l(x_i)$. The estimator achieves these rates by adapting to the unknown smoothness or sparsity of $D_l(x_i)$. Nonetheless, the estimator has an important drawback: The regularization by the ℓ_1-norm employed in (2.4) naturally lets the Lasso estimator avoid overfitting the data, but it also shrinks the estimated coefficients toward zero, causing a potentially significant bias. The post-Lasso estimator is meant to remove some of this shrinkage bias. If model selection by Lasso works perfectly, that is, if it selects exactly the "relevant" instruments, then the resulting post-Lasso estimator is simply the standard OLS estimator using only the relevant variables. In cases where perfect selection does not occur, post-Lasso estimates of coefficients will still tend to be less biased than Lasso. We prove the post-Lasso estimator achieves the same rate of convergence as Lasso, which is a near-optimal rate, despite imperfect model selection by Lasso.

The introduction of self-normalization via the penalty loadings allows us to contribute to the broad Lasso literature cited in the Introduction by showing that, under possibly heteroscedastic and non-Gaussian errors, the Lasso and post-Lasso estimators obey the following near-oracle performance bounds:

$$(2.9) \qquad \max_{1\le l\le k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n} \lesssim_{\mathrm{P}} \sqrt{\frac{s\log(p\vee n)}{n}} \quad \text{and}$$

$$\max_{1\le l\le k_e} \|\widehat{\beta}_l - \beta_{l0}\|_1 \lesssim_{\mathrm{P}} \sqrt{\frac{s^2\log(p\vee n)}{n}}.$$

The performance bounds in (2.9) are called near-oracle because they coincide up to a $\sqrt{\log p}$ factor with the bounds achievable when the ideal series terms T_l for each of the k_e regression equations in (2.2) are known. Our results extend those of Bickel, Ritov, and Tsybakov (2009) for Lasso with Gaussian errors and those of Belloni and Chernozhukov (2012) for post-Lasso with Gaussian errors. Notably, these bounds are as sharp as the results for the Gaussian case under the weak condition $\log p = o(n^{1/3})$. They are also the first results in the literature that allow for data-driven choice of the penalty level.

It is also useful to contrast the rates given in (2.9) with the rates available for nonparametrically estimating conditional expectations in the series literature; see, for example, Newey (1997). Obtaining rates of convergence for series estimators relies on approximate sparsity just as our results do. Approximate sparsity in the series context is typically motivated by smoothness assumptions, but approximate sparsity is more general than typical smoothness assumptions.[8] The standard series approach postulates that the first K series terms are the most important for approximating the target regression function D_{il}. The Lasso approach postulates that s terms from a large number p of terms are important but does not require knowledge of the identity of these terms or the number of terms, s, needed to approximate the target function well enough that approximation errors are small relative to estimation error. Lasso methods estimate both the optimal number of series terms s as well as the identities of these terms, and thus automatically adapt to the unknown sparsity (or smoothness) of the true optimal instrument (conditional expectation). This behavior differs sharply from standard series procedures that do not adapt to the unknown sparsity of the target function unless the number of series terms is chosen by a model selection method. Lasso-based methods may also provide enhanced approximation of the optimal instrument by allowing selection of the most important terms from among a set of very many series terms, with total number of terms $p \gg K$ that can be much larger than the sample size.[9] For example, a standard series approach based on K terms will perform poorly when the terms $m+1, m+2, \dots, m+j$ are the most important for approximating the optimal instrument for any $K < m$. On the other hand, Lasso-based methods will find the important terms as long as $p > m+j$, which is much less stringent than what is required in usual series approaches since p can be very large. This point can also be made using the array asymptotics where the model changes with n in such a way that the important series terms are always missed by the first $K \to \infty$ terms. Of course, the additional flexibility allowed for by Lasso-based methods comes with a price, namely slowing the rate of convergence by $\sqrt{\log p}$ relative to the usual series rates.

2.4. *The Instrumental Variable Estimator Based on Lasso and Post-Lasso Constructed Optimal Instrument*

Given Condition AS, we take advantage of the approximate sparsity by using Lasso and post-Lasso methods to construct estimates of $D_l(x_i)$ of the form

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

[8] See, for example, Belloni, Chernozhukov, and Hansen (2011a, 2011b) for detailed discussion of approximate sparsity.

[9] We can allow for $p \gg n$ for series formed with orthonormal bases with bounded components, such as trigonometric bases, but further restrictions on the number of terms apply if bounds on components of the series are allowed to increase with the sample size. For example, if we work with B-spline series terms, we can only consider $p = o(n)$ terms.

and then set

$$\widehat{D}_i = \big(\widehat{D}_1(x_i), \dots, \widehat{D}_{k_e}(x_i), w_i'\big)'.$$

The resulting IV estimator takes the form

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

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 set of instruments to produce estimates of the optimal instruments $\widehat{D}_i$ such that the resulting IV estimator achieves the efficiency bound asymptotically:

$$\sqrt{n}(\widehat{\alpha} - \alpha_0) =_d N\big(0, \Lambda^*\big) + o_{\mathrm{P}}(1).$$

The estimator matches the performance of the classical/standard series-based IV estimator of Newey (1990) and has additional advantages mentioned in the previous subsection. We also show that the IV estimator with Lasso-based optimal instruments continues to be root-n consistent and asymptotically normal in the presence of heteroscedasticity:

$$\sqrt{n}(\widehat{\alpha} - \alpha_0) =_d N\big(0, Q^{-1}\Omega Q^{-1}\big) + o_{\mathrm{P}}(1), \tag{2.10}$$

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

$$\widehat{Q}^{-1}\widehat{\Omega}\widehat{Q}^{-1}, \quad \widehat{\Omega} := \mathbb{E}_n\big[\widehat{\epsilon}_i^2\widehat{D}(x_i)\widehat{D}(x_i)'\big], \quad \widehat{Q} := \mathbb{E}_n\big[\widehat{D}(x_i)\widehat{D}(x_i)'\big], \tag{2.11}$$

where $\widehat{\epsilon}_i := y_i - d_i'\widehat{\alpha}$, $i = 1, \dots, n$. Using (2.11), we can perform robust inference.

We note that our result (2.10) for the IV estimator does 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 (2.9).

We conclude by stressing that our result (2.10) does not rely on perfect model selection. Perfect model selection only occurs in extremely limited circumstances that are unlikely to occur in practice. We show that model selection mistakes do not affect the asymptotic distribution of the IV estimator $\widehat{\alpha}$ under mild regularity conditions. The intuition is that the model selection mistakes are sufficiently small to allow the Lasso or post-Lasso to estimate the first-stage predictions with a sufficient, near-oracle accuracy, which translates to the result above. Using analysis like that given in Belloni, Chernozhukov, and Hansen (2011b), the result (2.10) can be shown to hold uniformly over models with strong optimal instruments that are uniformly approximately sparse. We also offer an inference test procedure in Section 4.2 that remains valid in the absence of a strong optimal instrument, is robust to many weak instruments, and can be used even if $p \gg n$. This procedure could also be shown to be uniformly valid over a large class of models.

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

In this section, we present our main results on Lasso and post-Lasso estimators of conditional expectation functions under nonclassical assumptions and data-driven penalty choices. The problem we analyze in this section has many applications outside the IV framework of the present paper.

3.1. *Regularity Conditions for Estimating Conditional Expectations*

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

To define the restricted eigenvalue, first define the restricted set:

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

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

$$\kappa_C^2(M) := \min_{\delta \in \Delta_{C,T}, |T| \leq s} s\frac{\delta' M \delta}{\|\delta_T\|_1^2}. \tag{3.1}$$

This restricted eigenvalue can depend on n, but we suppress the dependence in our notation.

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

CONDITION RE: For any $C > 0$, there exists a finite constant $\kappa > 0$, which does not depend on n but may depend on C, such that the restricted eigenvalue obeys $\kappa_C(\mathbb{E}_n[f_i f_i']) \geq \kappa$ with probability approaching 1 as $n \to \infty$.

The restricted eigenvalue (3.1) is a variant of the restricted eigenvalues introduced in Bickel, Ritov, and Tsybakov (2009) to analyze the properties of Lasso in the classical Gaussian regression model. Even though the minimal eigenvalue of the empirical Gram matrix $\mathbb{E}_n[f_i f_i']$ is zero whenever $p \geq n$, Bickel, Ritov, and Tsybakov (2009) showed that its restricted eigenvalues can be bounded away from zero. Lemmas 1 and 2 below contain sufficient conditions for this.

Many other sufficient conditions are available from the literature; see Bickel, Ritov, and Tsybakov (2009). Consequently, we take restricted eigenvalues as primitive quantities and Condition RE as a primitive condition.

COMMENT 3.1—On Restricted Eigenvalues: To gain intuition about restricted eigenvalues, assume the exactly sparse model, in which there is no approximation error. In this model, the term δ stands for a generic deviation between an estimator and the true parameter vector β_0. Thus, the restricted eigenvalue represents a modulus of continuity between a penalty-related term and the prediction norm, which allows us to derive the rate of convergence. Indeed, the restricted eigenvalue bounds the minimum change in the prediction norm induced by a deviation δ within the restricted set $\Delta_{C,T}$ relative to the norm of δ_T, the deviation on the true support. Given a specific choice of the penalty level, the deviation of the estimator belongs to the restricted set, making the restricted eigenvalue relevant for deriving rates of convergence.

To define the sparse eigenvalues, let us define the m-sparse subset of a unit sphere as

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

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

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

To simplify asymptotic statements for post-Lasso, we use the following condition:

CONDITION SE: For any $C > 0$, there exist constants $0 < \kappa' < \kappa'' < \infty$, which do not depend on n but may depend on C, such that, with probability approaching 1, as $n \to \infty$, $\kappa' \le \phi_{\min}(Cs)(\mathbb{E}_n[f_i f_i']) \le \phi_{\max}(Cs)(\mathbb{E}_n[f_i f_i']) \le \kappa''$.

Condition SE requires only that certain “small” $m \times m$ submatrices of the large $p \times p$ empirical Gram matrix are well-behaved, which is a reasonable assumption and will be sufficient for the results that follow. Condition SE implies Condition RE by the argument given in Bickel, Ritov, and Tsybakov (2009). The following lemmas show that Conditions RE and SE are plausible for both many-instrument and many-series-instrument settings. We refer to Belloni and Chernozhukov (2012) for proofs; the first lemma builds upon results in Zhang and Huang (2008) and the second builds upon results in Rudelson and Vershynin (2008). The lemmas could also be derived from Rudelson and Zhou (2011).

LEMMA 1—Plausibility of Conditions RE and SE Under Many Gaussian Instruments: *Suppose f_i, $i = 1, \ldots, n$, are i.i.d. zero-mean Gaussian random vectors. Further suppose that the population Gram matrix* $\mathrm{E}[f_i f_i']$ *has* $s\log n$*-sparse eigenvalues bounded from above and away from zero uniformly in* n. *Then if* $s\log n = o(n/\log p)$, *Conditions* RE *and* SE *hold.*

LEMMA 2—Plausibility of Conditions RE and SE Under Many Series Instruments: *Suppose f_i, $i = 1, \ldots, n$, are i.i.d. bounded zero-mean random vectors with* $\|f_i\|_\infty \le K_B$ *a.s. Further suppose that the population Gram matrix* $\mathrm{E}[f_i f_i']$ *has* $s\log n$*-sparse eigenvalues bounded from above and away from zero uniformly in* n. *Then if* $K_B^2 s\log^2(n)\log^2(s\log n)\log(p\vee n) = o(n)$, *Conditions* RE *and* SE *hold.*

In the context of i.i.d. sampling, a standard assumption in econometric research is that the population Gram matrix $\mathrm{E}[f_i f_i']$ has eigenvalues bounded from above and below; see, for example, Newey (1997). The lemmas above allow for this and more general behavior, requiring only that the sparse eigenvalues of the population Gram matrix $\mathrm{E}[f_i f_i']$ are bounded from below and from above. The latter is important for allowing functions f_i to be formed as a combination of elements from different bases, for example, a combination of B-splines with polynomials. The lemmas above further show that the good behavior of the population sparse eigenvalues translates into good behavior of empirical sparse eigenvalues under some restrictions on the growth of s in relation to the sample size n. For example, if p grows polynomially with n and the components of technical regressors are uniformly bounded, Lemma 2 holds provided $s = o(n/\log^5 n)$.

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

CONDITION RF: (i) $\max_{l\le k_e, j\le p}\bar{\mathrm{E}}[\tilde{d}_{il}^2] + \bar{\mathrm{E}}[|f_{ij}^2\tilde{d}_{il}^2|] + 1/\bar{\mathrm{E}}[f_{ij}^2 v_{il}^2] \lesssim 1$, (ii) $\max_{l\le k_e, j\le p}\bar{\mathrm{E}}[|f_{ij}^3 v_{il}^3|] \lesssim K_n$, (iii) $K_n^2\log^3(p\vee n) = o(n)$ and $s\log(p\vee n) = o(n)$, (iv) $\max_{i\le n, j\le p} f_{ij}^2[s\log(p\vee n)]/n \to_{\mathrm{P}} 0$ and $\max_{l\le k_e, j\le p}|(\mathbb{E}_n - \bar{\mathrm{E}})[f_{ij}^2 v_{il}^2]| + |(\mathbb{E}_n - \bar{\mathrm{E}})[f_{ij}^2\tilde{d}_{il}^2]| \to_{\mathrm{P}} 0$.

We emphasize that the conditions given above are only one possible set of sufficient conditions, which are presented in a manner that reduces the complexity of the exposition.

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

$$B_1 \le \mathrm{E}\big[|g_i|^k|x_i\big] \le B_2 \quad \text{with probability } 1,$$
$$\text{for } k = 1, \ldots, K, i = 1, \ldots, n.$$

LEMMA 3—Plausibility of Condition RF: 1. *If the moments* $\bar{\mathrm{E}}[\tilde{d}_{il}^8]$ *and* $\bar{\mathrm{E}}[v_{il}^8]$ *are bounded uniformly in* $1 \leq l \leq k_e$ *and in* n, *the regressors* f_i *obey* $\max_{1\leq j\leq p} \mathbb{E}_n[f_{ij}^8] \lesssim_{\mathrm{P}} 1$ *and* $\max_{1\leq i\leq n, 1\leq j\leq p} f_{ij}^2 \frac{s\log(p\vee n)}{n} \to_{\mathrm{P}} 0$, *Conditions* RF(i)–(iii) *imply Condition* RF(iv). 2. *Suppose that* $\{(f_i, \tilde{d}_i, v_i), i = 1, \dots, n\}$ *are i.i.d. vectors, and that* $\tilde{d}_{il}$ *and* v_{il} *have uniformly bounded conditional moments of order* 4 *uniformly in* $l = 1, \dots, k_e$. (a) *If the regressors* f_i *are Gaussian as in Lemma* 1, *Condition* RF(iii) *holds, and* $s\log^2(p\vee n)/n \to 0$, *then Conditions* RF(i), (ii), *and* (iv) *hold*. (b) *If the regressors* f_i *have bounded entries as in Lemma* 2, *then Conditions* RF(i), (ii), *and* (iv) *hold under Condition* RF(iii).

3.2. *Results on Lasso and Post-Lasso for Estimating Conditional Expectations*

We consider Lasso and post-Lasso estimators defined in equations (2.4) and (2.8) in the system of k_e nonparametric regression equations (2.3) with non-Gaussian and heteroscedastic errors. These results extend the previous results of Bickel, Ritov, and Tsybakov (2009) for Lasso and of Belloni and Chernozhukov (2012) for post-Lasso with Gaussian 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.

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

$$\ell\widehat{\Upsilon}_l^0 \leq \widehat{\Upsilon}_l \leq u\widehat{\Upsilon}_l^0, \tag{3.2}$$

with $0 < \ell \leq 1 \leq u$ such that $\ell \to_{\mathrm{P}} 1$ and $u \to_{\mathrm{P}} u'$ with $u' \geq 1$. The penalty loadings constructed by Algorithm A.1 satisfy this condition.

THEOREM 1—Rates for Lasso Under Non-Gaussian and Heteroscedastic Errors: *Suppose that in the regression model* (2.3), *Conditions* AS *and* RF *hold. Suppose the penalty level is specified as in* (2.7), *and consider any asymptotically valid penalty loadings* $\widehat{\Upsilon}$, *for example, penalty loadings constructed by Algorithm* A.1 *stated in Appendix* A. *Then, the Lasso estimator* $\widehat{\beta}_l = \widehat{\beta}_{l\mathrm{L}}$ *and the Lasso fit* $\widehat{D}_{il} = f_i'\widehat{\beta}_{l\mathrm{L}}$, $l = 1, \dots, k_e$, *satisfy*

$$\max_{1\leq l\leq k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n} \lesssim_{\mathrm{P}} \frac{1}{\kappa_{\bar{C}}}\sqrt{\frac{s\log(k_e p/\gamma)}{n}}$$

and

$$\max_{1\leq l\leq k_e} \|\widehat{\beta}_l - \beta_{l0}\|_1 \lesssim_{\mathrm{P}} \frac{1}{(\kappa_{2\bar{C}})^2}\sqrt{\frac{s^2\log(k_e p/\gamma)}{n}},$$

where $\bar{C} = \max_{1\leq l\leq k_e}\{\|\widehat{\Upsilon}_l^0\|_\infty \|(\widehat{\Upsilon}_l^0)^{-1}\|_\infty\}(uc+1)/(\ell c - 1)$ *and* $\kappa_{\bar{C}} = \kappa_{\bar{C}}(\mathbb{E}_n[f_i f_i'])$.

The theorem provides a rate result for the Lasso estimator constructed specifically to deal with non-Gaussian errors and heteroscedasticity. The rate result generalizes, and is as sharp as, the rate results of Bickel, Ritov, and Tsybakov (2009) obtained for the homoscedastic Gaussian case. This generalization is important for real applications where non-Gaussianity and heteroscedasticity are ubiquitous. Note that the obtained rate is near-optimal in the sense that if we happened to know the model T_ℓ, that is, if we knew the identities of the most important variables, we would only improve the rate by the $\log p$ factor. The theorem also shows that the data-driven penalty loadings defined in Algorithm A.1 are asymptotically valid.

The following theorem presents the properties of post-Lasso, which requires a mild assumption on the number of additional variables in the set $\widehat{I}_l$, $l = 1, \dots, k_e$. We assume that the size of these sets is not substantially larger than the model selected by Lasso, namely, a.s.

$$|\widehat{I}_l \setminus \widehat{T}_l| \lesssim 1 \vee |\widehat{T}_l|, \quad l = 1, \dots, k_e. \tag{3.3}$$

THEOREM 2—Rates for Post-Lasso Under Non-Gaussian and Heteroscedastic Errors: *Suppose that in the regression model* (2.3), *Conditions* AS *and* RF *hold. Suppose the penalty level for the Lasso estimator is specified as in* (2.7), *that Lasso's penalty loadings* $\widehat{\Upsilon}$ *are asymptotically valid, and the sets of additional variables obey* (3.3). *Then, the post-Lasso estimator* $\widehat{\beta}_l = \widehat{\beta}_{l\mathrm{PL}}$ *and the post-Lasso fit* $\widehat{D}_{il} = f_i'\widehat{\beta}_{l\mathrm{PL}}$, $l = 1, \dots, k_e$, *satisfy*

$$\max_{1\le l\le k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n} \lesssim_{\mathrm{P}} \frac{\mu}{\kappa_{\bar{C}}}\sqrt{\frac{s\log(k_e p/\gamma)}{n}}$$

and

$$\max_{1\le l\le k_e} \|\widehat{\beta}_l - \beta_{l0}\|_1 \lesssim_{\mathrm{P}} \frac{\mu^2}{(\kappa_{\bar{C}})^2}\sqrt{\frac{s^2\log(k_e p/\gamma)}{n}},$$

where $\mu^2 = \min_k\{\phi_{\max}(k)(\mathbb{E}_n[f_if_i'])/\phi_{\min}(k+s)(\mathbb{E}_n[f_if_i']) : k > 18\bar{C}^2 s\phi_{\max}(k) \times (\mathbb{E}_n[f_if_i'])/(\kappa_{\bar{C}})^2\}$ *for* $\bar{C}$ *defined in Theorem* 1.

The theorem provides a rate result for the post-Lasso estimator with non-Gaussian errors and heteroscedasticity. The rate result generalizes the results of Belloni and Chernozhukov (2012) obtained for the homoscedastic Gaussian case. The post-Lasso achieves the same near-optimal rate of convergence as Lasso. As stressed in the introductory sections, our analysis allows Lasso to make model selection mistakes, which is expected generically. We show that these model selection mistakes are small enough to allow the post-Lasso estimator to perform as well as Lasso.[10]

Rates of convergence in different norms can also be of interest in other applications. In particular, the ℓ_2-rate of convergence can be derived from the rate of convergence in the prediction norm and Condition SE using a sparsity result for Lasso established in Appendix D. Below we specialize the previous theorems to the important case where Condition SE holds.

COROLLARY 1—Rates for Lasso and Post-Lasso Under Condition SE: *Under the conditions of Theorem 2 and Condition SE, the Lasso and post-Lasso estimators satisfy*

$$\max_{1\le l\le k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n} \lesssim_{\mathrm{P}} \sqrt{\frac{s\log(p\vee n)}{n}},$$

$$\max_{1\le l\le k_e} \|\widehat{\beta}_l - \beta_{l0}\|_2 \lesssim_{\mathrm{P}} \sqrt{\frac{s\log(p\vee n)}{n}},$$

$$\max_{1\le l\le k_e} \|\widehat{\beta}_l - \beta_{l0}\|_1 \lesssim_{\mathrm{P}} \sqrt{\frac{s^2\log(p\vee n)}{n}}.$$

The rates of convergence in the prediction norm and ℓ_2-norm are faster than the rate of convergence in the ℓ_1-norm, which is typical of high-dimensional settings.

4. MAIN RESULTS ON IV ESTIMATION

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

4.1. *The IV Estimator With Lasso-Based Instruments*

We impose the following moment conditions on the instruments, the structural errors, and regressors.

CONDITION SM: (i) The eigenvalues of $Q = \bar{\mathrm{E}}[D(x_i)D(x_i)']$ are bounded uniformly from above and away from zero, uniformly in n. The conditional variance $\mathrm{E}[\epsilon_i^2|x_i]$ is bounded uniformly from above and away from zero, uniformly in i and n. Given this assumption, without loss of generality, we normalize the instruments so that $\bar{\mathrm{E}}[f_{ij}^2\epsilon_i^2] = 1$ for each $1 \le j \le p$ and for all n.

[10] Under further conditions stated in proofs, post-Lasso can sometimes achieve a faster rate of convergence. In special cases where perfect model selection is possible, post-Lasso becomes the so-called oracle estimator and can completely remove the $\log p$ factor.

(ii) For some $q > 2$ and $q_\epsilon > 2$, uniformly in n,

$$\max_{1\le j\le p} \bar{\mathrm{E}}\big[|f_{ij}\epsilon_i|^3\big] + \bar{\mathrm{E}}\big[\|D_i\|_2^q|\epsilon_i|^{2q}\big] + \bar{\mathrm{E}}\big[\|D_i\|_2^q\big] + \bar{\mathrm{E}}\big[|\epsilon_i|^{q_\epsilon}\big] + \bar{\mathrm{E}}\big[\|d_i\|_2^q\big] \lesssim 1.$$

(iii) In addition to $\log^3 p = o(n)$, the following growth conditions hold:

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

$$\text{(b)} \quad \frac{s^2\log^2(p\vee n)}{n} \to 0,$$

$$\text{(c)} \quad \max_{1\le j\le p} \mathbb{E}_n\big[f_{ij}^2\epsilon_i^2\big] \lesssim_{\mathrm{P}} 1.$$

COMMENT 4.1—On Condition SM: Condition SM(i) places restrictions on the variation of the structural errors (ϵ) and the optimal instruments ($D(x)$). The first condition about the variation in the optimal instrument guarantees that identification is strong; that is, it ensures that the conditional expectation of the endogenous variables given the instruments is a nontrivial function of the instruments. This assumption rules out non-identification, in which case $D(x)$ does not depend on x, and weak-identification, in which case $D(x)$ would be local to a constant function. We present an inference procedure that remains valid without this condition in Section 4.2. The remaining restriction in 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 β_{l0}, $l = 1, \dots, k_e$, via an isomorphic transformation. We use this normalization to simplify notation in the proofs, but do not use it in the construction of the estimators. Condition SM(ii) imposes some mild moment assumptions. Condition SM(iii) strengthens the growth requirement $s\log p/n \to 0$ needed for estimating conditional expectations. However, the restrictiveness of Condition SM(iii)(a) rapidly decreases as the number of bounded moments of the structural error increases. Condition SM(iii)(b) indirectly requires the optimal instruments in Condition AS to be smooth enough that the number of unknown series terms s needed to approximate them well is not too large. This condition ensures that the impact of the instrument estimation on the IV estimator is asymptotically negligible. This condition can be relaxed using the sample-splitting method.

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

LEMMA 4—Plausibility of Condition SM(iii): *Suppose that the structural disturbance ϵ_i has uniformly bounded conditional moments of order* 4 *uniformly in n and that $s^2 \log^2(p \vee n) = o(n)$. Then Condition* SM(iii) *holds if* (a) *the regressors f_i are Gaussian as in Lemma* 1, *or* (b) *the regressors f_i are arbitrary i.i.d. vectors with bounded entries as in Lemma* 2.

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. The result also provides a consistent estimator for the asymptotic variance of this estimator under heteroscedasticity.

THEOREM 3—Inference With Optimal IV Estimated by Lasso or Post-Lasso: *Suppose that data (y_i, x_i, d_i) are i.n.i.d. and obey the linear IV model described in Section* 2. *Suppose also that Conditions* AS, RF, SM, (2.7), *and* (3.2) *hold. To construct the estimate of the optimal instrument, suppose that Condition* RE *holds in the case of using Lasso or that Condition* SE *and* (3.3) *hold in the case of using post-Lasso. Then the IV estimator $\widehat{\alpha}$, based on either Lasso or post-Lasso estimates of the optimal instrument, is root-n consistent and asymptotically normal*:

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

for $\Omega := \bar{\mathrm{E}}[\epsilon_i^2 D(x_i)D(x_i)']$ and $Q := \bar{\mathrm{E}}[D(x_i)D(x_i)']$. Moreover, the result above continues to hold with Ω replaced by $\widehat{\Omega} := \mathbb{E}_n[\widehat{\epsilon}_i^2 \widehat{D}(x_i)\widehat{D}(x_i)']$ for $\widehat{\epsilon}_i = y_i - d_i'\widehat{\alpha}$, and Q replaced by $\widehat{Q} := \mathbb{E}_n[\widehat{D}(x_i)\widehat{D}(x_i)']$. In the case that the structural error ϵ_i is homoscedastic conditional on x_i, that is, $E[\epsilon_i^2|x_i] = \sigma^2$ a.s. for all $i = 1, \dots, n$, the IV estimator $\widehat{\alpha}$ based on either Lasso or post-Lasso estimates of the optimal instrument is root-n consistent, asymptotically normal, and achieves the efficiency bound $(\Lambda^*)^{-1/2}\sqrt{n}(\widehat{\alpha} - \alpha_0) \to_d N(0, I)$, where $\Lambda^* := \sigma^2 Q^{-1}$. The result above continues to hold with Λ^* replaced by $\widehat{\Lambda}^* := \widehat{\sigma}^2\widehat{Q}^{-1}$, where $\widehat{Q} := \mathbb{E}_n[\widehat{D}(x_i)\widehat{D}(x_i)']$ and $\widehat{\sigma}^2 := \mathbb{E}_n[(y_i - d_i'\widehat{\alpha})^2]$.

In the setting with homoscedastic structural errors, the estimator achieves the efficiency bound asymptotically. In the case of heteroscedastic structural errors, 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 final result of this section extends the previous result to any IV estimator with a generic sparse estimator of the optimal instruments.

THEOREM 4—Inference With IV Constructed by a Generic Sparsity-Based Procedure: *Suppose that Conditions* AS *and* SM *hold, and suppose that the fitted values of the optimal instrument, $\widehat{D}_{il} = f_i'\widehat{\beta}_l$, are constructed using any estimator*

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

$$\max_{1\le l\le k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n} \lesssim_{\mathrm{P}} \sqrt{\frac{s\log(p\vee n)}{n}} \quad \textit{and} \tag{4.1}$$

$$\max_{1\le l\le k_e} \|\widehat{\beta}_l - \beta_{l0}\|_1 \lesssim_{\mathrm{P}} \sqrt{\frac{s^2\log(p\vee n)}{n}}.$$

Then the conclusions reached in Theorem 3 *continue to apply*.

This result shows that the previous two theorems apply for any first-stage estimator that attains near-oracle performance given in (4.1). Examples of other sparse estimators covered by this theorem are Dantzig and Gauss–Dantzig (Candes and Tao (2007)), $\sqrt{\text{Lasso}}$ and post-$\sqrt{\text{Lasso}}$ (Belloni, Chernozhukov, and Wang (2011a, 2011b)), thresholded Lasso and post-thresholded Lasso (Belloni and Chernozhukov (2012)), group Lasso and post-group Lasso (Huang, Horowitz, and Wei (2010), Lounici et al. (2010)), adaptive versions of the above (Huang, Horowitz, and Wei (2010)), and boosting (Bühlmann (2006)). Verification of the near-oracle performance (4.1) can be done on a case by case basis using the best conditions in the literature.[11] Our results extend to Lasso-type estimators under alternative forms of regularity conditions that fall outside the framework of Conditions RE and RF; all that is required is the near-oracle performance of the kind (4.1).

4.2. *Inference When Instruments Are Weak*

When instruments are weak individually, Lasso may end up selecting no instruments or may produce unreliable estimates of the optimal instruments. To cover this case, we propose a method for inference based on inverting pointwise tests performed using a sup-score statistic defined below. The procedure is similar in spirit to Anderson and Rubin (1949) and Staiger and Stock (1997), but uses a different statistics that is better suited to cases with very many instruments. To describe the approach, we rewrite the main structural equation as

$$y_i = d_{ei}'\alpha_1 + w_i'\alpha_2 + \epsilon_i, \quad \mathrm{E}[\epsilon_i|x_i] = 0, \tag{4.2}$$

where y_i is the response variable, d_{ei} is a vector of endogenous variables, w_i is a k_w-vector of control variables, $x_i = (z_i', w_i')'$ is a vector of elementary instrumental variables, and ϵ_i is a disturbance such that $\epsilon_1, \dots, \epsilon_n$ are i.n.i.d.

[11] Post-ℓ_1-penalized procedures have only been analyzed for the case of Lasso and $\sqrt{\text{Lasso}}$; see Belloni and Chernozhukov (2012) and Belloni, Chernozhukov, and Wang (2011a). We expect that similar results carry over to other procedures listed above.

conditional on $X = [x_1', \dots, x_n']$. We partition $d_i = (d_{ei}{}', w_i')'$. The parameter of interest is $\alpha_1 \in \mathcal{A}_1 \subset \mathbb{R}^{k_e}$. We use $f_i = P(x_i)$, a vector which includes w_i, as technical instruments. In this subsection, we treat X as fixed; that is, we condition on X.

We would like to use a high-dimensional vector f_i of technical instruments for inference on α_1 that is robust to weak identification. To formulate a practical sup-score statistic, it is useful to partial-out the effect of w_i on the key variables. For an n-vector $\{u_i, i = 1, \dots, n\}$, define $\tilde{u}_i = u_i - w_i'\mathbb{E}_n[w_i w_i']^{-1}\mathbb{E}_n[w_i u_i]$, that is, the residuals left after regressing this vector on $\{w_i, i = 1, \dots, n\}$. Hence $\tilde{y}_i$, $\tilde{d}_{ei}$, and $\tilde{f}_{ij}$ are residuals obtained by partialling out controls w_i. Also, let

$$\tilde{f}_i = (\tilde{f}_{i1}, \dots, \tilde{f}_{ip})'. \tag{4.3}$$

In this formulation, we omit elements of w_i from $\tilde{f}_{ij}$ since they are eliminated by partialling out. We then normalize these technical instruments so that

$$\mathbb{E}_n\big[\tilde{f}_{ij}^2\big] = 1, \quad j = 1, \dots, p. \tag{4.4}$$

The sup-score statistic for testing the hypothesis $\alpha_1 = a$ takes the form

$$\Lambda_a = \max_{1\le j\le p}\big|n\mathbb{E}_n\big[(\tilde{y}_i - \tilde{d}_{ei}'a)\tilde{f}_{ij}\big]\big| / \sqrt{\mathbb{E}_n\big[(\tilde{y}_i - \tilde{d}_{ei}'a)^2\tilde{f}_{ij}^2\big]}. \tag{4.5}$$

As before, we apply self-normalized moderate deviation theory for self-normalized sums to obtain

$$\mathrm{P}\big(\Lambda_{\alpha_1} \le c\sqrt{n}\Phi^{-1}(1 - \gamma/2p)\big) \ge 1 - \gamma + o(1).$$

Therefore, we can employ $\Lambda(1-\gamma) := c\sqrt{n}\Phi^{-1}(1-\gamma/2p)$ for $c > 1$ as a critical value for testing $\alpha_1 = a$ using Λ_a as the test statistic. The asymptotic $(1 - \gamma)$-confidence region for α_1 is then given by $\mathcal{C} := \{a \in \mathcal{A}_1 : \Lambda_a \le \Lambda(1 - \gamma)\}$.

The construction of confidence regions above can be given the following *Inverse Lasso* interpretation. Let

$$\widehat{\beta}_a \in \arg\min_{\beta\in\mathbb{R}^p} \mathbb{E}_n\big[(\tilde{y}_i - \tilde{d}_{ei}'a) - \tilde{f}_{ij}'\beta\big]^2 + \frac{\lambda}{n}\sum_{j=1}^{p}\gamma_{aj}|\beta_j|,$$

$$\gamma_{aj} = \sqrt{\mathbb{E}_n\big[(\tilde{y}_i - \tilde{d}_{ei}'a)^2\tilde{f}_{ij}^2\big]}.$$

If $\lambda = 2\Lambda(1-\gamma)$, then $\mathcal{C}$ is equivalent to the region $\{a \in \mathbb{R}^{k_e} : \widehat{\beta}_a = 0\}$. In words, this confidence region collects all potential values of the structural parameter where the Lasso regression of the potential structural disturbance on the instruments yields zero coefficients on the instruments. This idea is akin to the

Inverse Quantile Regression and Inverse Least Squares ideas in Chernozhukov and Hansen (2008a, 2008b).

Below, we state the main regularity condition for the validity of inference using the sup-score statistic as well as the formal inference result.

CONDITION SM2: Suppose that, for each n, the linear model (4.2) holds with $\alpha_1 \in \mathcal{A}_1 \subset \mathbb{R}^{k_e}$ such that $\epsilon_1, \dots, \epsilon_n$ are i.n.i.d., X is fixed, and $\tilde{f}_1, \dots, \tilde{f}_n$ are p-vectors of technical instruments defined in (4.3) and (4.4). Suppose that (i) the dimension of w_i is k_w and $\|w_i\|_2 \leq \zeta_w$ such that $\sqrt{k_w}\zeta_w/\sqrt{n} \to 0$, (ii) the eigenvalues of $\mathbb{E}_n[w_i w_i']$ are bounded away from zero and eigenvalues of $\bar{\mathrm{E}}[\epsilon_i^2 w_i w_i']$ are bounded away from above, uniformly in n, (iii) $\max_{1\leq j\leq p} \bar{\mathrm{E}}[|\epsilon_i|^3|\tilde{f}_{ij}|^3]^{1/3}/\bar{\mathrm{E}}[\epsilon_i^2\tilde{f}_{ij}^2]^{1/2} \leq K_n$, and (iv) $K_n^2 \log(p \vee n) = o(n^{1/3})$.

THEOREM 5—Valid Inference Based on the Sup-Score Statistic: *Let* $\gamma \in (0,1)$ *be fixed or, more generally, such that* $\log(1/\gamma) \lesssim \log(p \vee n)$. *Under Condition* SM2, (i) *in large samples, the constructed confidence set* $\mathcal{C}$ *contains the true value* α_1 *with at least the prescribed probability, namely* $\mathrm{P}(\alpha_1 \in \mathcal{C}) \geq 1 - \gamma - o(1)$. (ii) *Moreover, the confidence set* $\mathcal{C}$ *necessarily excludes a sequence of parameter value* a, *namely* $\mathrm{P}(a \in \mathcal{C}) \to 0$, *if*

$$\max_{1\leq j\leq p} \frac{\sqrt{n/\log(p/\gamma)}|\mathbb{E}_n[(a-\alpha_1)'\tilde{d}_{ei}\tilde{f}_{ij}]|}{c\sqrt{\mathbb{E}_n[\epsilon_i^2\tilde{f}_{ij}^2]} + \sqrt{\mathbb{E}_n[\{(a-\alpha_1)'\tilde{d}_{ei}\}^2\tilde{f}_{ij}^2]}} \to_{\mathrm{P}} \infty.$$

The theorem shows that the confidence region $\mathcal{C}$ constructed above is valid in large samples and that the probability of including a false point a in $\mathcal{C}$ tends to zero as long as a is sufficiently distant from α_1 and instruments are not too weak. In particular, if there is a strong instrument, the confidence regions will eventually exclude points a that are further than $\sqrt{\log(p \vee n)/n}$ away from α_1. Moreover, if there are instruments whose correlation with the endogenous variable is of greater order than $\sqrt{\log(p \vee n)/n}$, then the confidence regions will asymptotically be bounded.

5. FURTHER INFERENCE AND ESTIMATION RESULTS FOR THE IV MODEL

In this section, we provide further estimation and inference results. We develop an overidentification test that compares the IV-Lasso based estimates to estimates obtained using a baseline set of instruments. We also combine the IV selection using Lasso with a sample-splitting technique from the many-instruments literature which allows us to relax the growth requirement on the number of relevant instruments.

5.1. *A Specification Test for Validity of Instrumental Variables*

Here we develop a Hausman-style specification test for the validity of the instrumental variables. Let $A_i = A(x_i)$ be a baseline set of instruments, with $\dim(A_i) \geq \dim(\alpha) = k_\alpha$ bounded. Let $\tilde{\alpha}$ be the baseline instrumental variable estimator based on these instruments:

$$\tilde{\alpha} = \left(\mathbb{E}_n\left[d_i A_i'\right]\mathbb{E}_n\left[A_i A_i'\right]^{-1}\mathbb{E}_n\left[A_i d_i'\right]\right)^{-1}\mathbb{E}_n\left[d_i A_i'\right]\mathbb{E}_n\left[A_i A_i'\right]^{-1}\mathbb{E}_n[A_i y_i].$$

If the instrumental variable exclusion restriction is valid, then the unscaled difference between this estimator and the IV estimator $\widehat{\alpha}$ proposed in the previous sections should be small. If the exclusion restriction is not valid, the difference between $\tilde{\alpha}$ and $\widehat{\alpha}$ should be large. Therefore, we can reject the null hypothesis of instrument validity if the difference is large.

We formalize the test as follows. Suppose we care about $R'\alpha$ for some $k \times k_d$ matrix R of $\operatorname{rank}(R) = k$. For instance, we might care only about the first k components of α, in which case $R = [I_k \;\; 0]$ is a $k \times k_d$ matrix that selects the first k coefficients of α. Define the estimand for $\tilde{\alpha}$ as

$$\alpha = \left(\bar{\mathrm{E}}\left[d_i A_i'\right]\bar{\mathrm{E}}\left[A_i A_i'\right]^{-1}\bar{\mathrm{E}}\left[A_i d_i'\right]\right)^{-1}\bar{\mathrm{E}}\left[d_i A_i'\right]\bar{\mathrm{E}}\left[A_i A_i'\right]^{-1}\bar{\mathrm{E}}[A_i y_i],$$

and define the estimand of $\widehat{\alpha}$ as

$$\alpha_a = \bar{\mathrm{E}}\left[D(x_i)D(x_i)'\right]^{-1}\bar{\mathrm{E}}\left[D(x_i)y_i\right].$$

The null hypothesis H_0 is $R(\alpha - \alpha_a) = 0$ and the alternative H_a is $R(\alpha - \alpha_a) \neq 0$. We can form a test statistic

$$J = \sqrt{n}(\tilde{\alpha} - \widehat{\alpha})'R'\left(R\widehat{\Sigma}R'\right)^{-1}\sqrt{n}R(\tilde{\alpha} - \widehat{\alpha})$$

for a matrix $\widehat{\Sigma}$ defined below and reject H_0 if $J > c_\gamma$, where c_γ is the $(1 - \gamma)$-quantile of chi-squared random variable with k degrees of freedom. The justification for this test is provided by the following theorem, which builds upon the previous results coupled with conventional results for the baseline instrumental variable estimator.[12]

THEOREM 6—Specification Test: (a) *Suppose the conditions of Theorem 3 hold, that* $\bar{\mathrm{E}}[\|A_i\|_2^q]$ *is bounded uniformly in n for* $q > 4$, *and the eigenvalues of*

$$\Sigma := \bar{\mathrm{E}}\left[\epsilon_i^2\left(MA_i - Q^{-1}D(x_i)\right)\left(MA_i - Q^{-1}D(x_i)\right)'\right]$$

are bounded from above and below, uniformly in n, where

$$M = \left(\bar{\mathrm{E}}\left[d_i A_i'\right]\bar{\mathrm{E}}\left[A_i A_i'\right]^{-1}\bar{\mathrm{E}}\left[A_i d_i'\right]\right)^{-1}\bar{\mathrm{E}}\left[d_i A_i'\right]\bar{\mathrm{E}}\left[A_i A_i'\right]^{-1}.$$

[12] The proof of this result is provided in the Supplemental Material (Belloni, Chen, Chernozhukov, and Hansen (2012)).

Then $\sqrt{n}\widehat{\Sigma}^{-1/2}(\tilde{\alpha}-\widehat{\alpha})' \to_d N(0, I)$ *and* $J \to_d \chi^2(k)$, *where*

$$\widehat{\Sigma} = \mathbb{E}_n\big[\widehat{\epsilon}_i^2\big(\widehat{M}^{-1}A_i - \widehat{Q}^{-1}\widehat{D}(x_i)\big)\big(\widehat{M}^{-1}A_i - \widehat{Q}^{-1}\widehat{D}(x_i)\big)'\big]$$

for $\widehat{\epsilon}_i = y_i - d_i'\widehat{\alpha}$, $\widehat{Q} = \mathbb{E}_n[\widehat{D}(x_i)\widehat{D}(x_i)']$, *and*

$$\widehat{M} = \big(\mathbb{E}_n[d_iA_i']\mathbb{E}_n[A_iA_i']^{-1}\mathbb{E}_n[A_id_i']\big)^{-1}\mathbb{E}_n[d_iA_i']\mathbb{E}_n[A_iA_i']^{-1}.$$

(b) *Suppose the conditions of Theorem 3 hold with the exception that* $\mathrm{E}[A_i\epsilon_i] = 0$ *for all* $i = 1, \dots, n$ *and* n, *but* $\|\bar{\mathrm{E}}[D(x_i)\epsilon_i]\|_2$ *is bounded away from zero. Then* $J \to_{\mathrm{P}} \infty$.

5.2. *Split-Sample IV Estimator*

The rate condition $s^2\log^2(p \vee n) = o(n)$ can be substantive and cannot be substantially weakened for the full-sample IV estimator considered above. However, we can replace this condition with the weaker condition that

$$s\log(p \vee n) = o(n)$$

by employing a sample-splitting method from the many-instruments literature (Angrist and Krueger (1995)). Specifically, we consider dividing the sample randomly into (approximately) equal parts a and b, with sizes $n_a = \lceil n/2 \rceil$ and $n_b = n - n_a$. We use superscripts a and b for variables in the first and second subsample, respectively. The index i will enumerate observations in both samples, with ranges for the index given by $1 \le i \le n_a$ for sample a and $1 \le i \le n_b$ for sample b. We can use each of the subsamples to fit the first stage via Lasso or post-Lasso to obtain the first-stage estimates $\widehat{\beta}_l^k$, $k = a, b$, and $l = 1, \dots, k_e$. Then setting $\widehat{D}_{il}^a = f_i^{a\prime}\widehat{\beta}_l^b$, $1 \le i \le n_a$, $\widehat{D}_{il}^b = f_i^{b\prime}\widehat{\beta}_l^a$, $1 \le i \le n_b$, $\widehat{D}_i^k = (\widehat{D}_{i1}^k, \dots, \widehat{D}_{ik_e}^k, w_i^{k\prime})'$, $k = a, b$, we form the IV estimates in the two subsamples:

$$\widehat{\alpha}_a = \mathbb{E}_{n_a}\big[\widehat{D}_i^a d_i^{a\prime}\big]^{-1}\mathbb{E}_{n_a}\big[\widehat{D}_i^a y_i^a\big] \quad \text{and} \quad \widehat{\alpha}_b = \mathbb{E}_{n_b}\big[\widehat{D}_i^b d_i^{b\prime}\big]^{-1}\mathbb{E}_{n_b}\big[\widehat{D}_i^b y_i^b\big].$$

Then we combine the estimates into one:

$$\begin{aligned} \widehat{\alpha}_{ab} = {} & \big(n_a\mathbb{E}_{n_a}\big[\widehat{D}_i^a\widehat{D}_i^{a\prime}\big] + n_b\mathbb{E}_{n_b}\big[\widehat{D}_i^b\widehat{D}_i^{b\prime}\big]\big)^{-1} \\ & \times \big(n_a\mathbb{E}_{n_a}\big[\widehat{D}_i^a\widehat{D}_i^{a\prime}\big]\widehat{\alpha}_a + n_b\mathbb{E}_{n_b}\big[\widehat{D}_i^b\widehat{D}_i^{b\prime}\big]\widehat{\alpha}_b\big). \end{aligned} \tag{5.1}$$

The following result shows that the split-sample IV estimator $\widehat{\alpha}_{ab}$ has the same large sample properties as the estimator $\widehat{\alpha}$ of the previous section but requires a weaker growth condition.

THEOREM 7—Inference With a Split-Sample IV Based on Lasso or Post-Lasso: *Suppose that data* (y_i, x_i, d_i) *are i.n.i.d. and obey the linear IV model described in Section* 2. *Suppose also that Conditions* AS, RF, SM, (2.7), (3.2), *and* (3.3) *hold, except that instead of growth condition* $s^2\log^2(p \vee n) = o(n)$, *we now have a weaker growth condition* $s\log(p \vee n) = o(n)$. *Suppose also that Condition* SE *holds for* $M^k = \mathbb{E}_{n_k}[f_i^k f_i^{k\prime}]$ *for* $k = a, b$. *Let* $\widehat{D}_{il}^k = f_i^{k\prime}\widehat{\beta}_l^{k^c}$, *where* $\widehat{\beta}_l^{k^c}$ *is the Lasso or post-Lasso estimator applied to the subsample* $\{(d_{li}^{k^c}, f_i^{k^c}) : 1 \leq i \leq n_{k^c}\}$ *for* $k = a, b$, *and* $k^c = \{a, b\} \setminus k$. *Then the split-sample IV estimator based on equation* (5.1) *is* $\sqrt{n}$*-consistent and asymptotically normal, as* $n \to \infty$:

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

for $\Omega := \bar{\mathrm{E}}[\epsilon_i^2 D(x_i)D(x_i)']$ *and* $Q := \bar{\mathrm{E}}[D(x_i)D(x_i)']$. *Moreover, the result above continues to hold with* Ω *replaced by* $\widehat{\Omega} := \mathbb{E}_n[\widehat{\epsilon}_i^2\widehat{D}(x_i)\widehat{D}(x_i)']$ *for* $\widehat{\epsilon}_i = y_i - d_i'\widehat{\alpha}_{ab}$, *and* Q *replaced by* $\widehat{Q} := \mathbb{E}_n[\widehat{D}(x_i)\widehat{D}(x_i)']$.

6. SIMULATION EXPERIMENT

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

Our simulations are based on a simple instrumental variables model data-generating process (DGP):

$$y_i = \beta d_i + e_i,$$

$$d_i = z_i'\Pi + v_i,$$

$$(e_i, v_i) \sim N\left(0, \begin{pmatrix} \sigma_e^2 & \sigma_{ev} \\ \sigma_{ev} & \sigma_v^2 \end{pmatrix}\right) \text{ i.i.d.},$$

where $\beta = 1$ is the parameter of interest, and $z_i = (z_{i1}, z_{i2}, \ldots, z_{i100})' \sim N(0, \Sigma_Z)$ is a 100×1 vector with $E[z_{ih}^2] = \sigma_z^2$ and $\mathrm{Corr}(z_{ih}, z_{ij}) = 0.5^{|j-h|}$. In all simulations, we set $\sigma_e^2 = 1$ and $\sigma_z^2 = 1$. We also set $\mathrm{Corr}(e, v) = 0.6$.

For the other parameters, we consider various settings. We provide results for sample sizes, n, of 100 and 250. We set σ_v^2 so that the unconditional variance of the endogenous variable equals 1; that is, $\sigma_v^2 = 1 - \Pi'\Sigma_Z\Pi$. We use three different settings for the pattern of the first-stage coefficients, Π. In the first, we set $\Pi = C\widetilde{\Pi} = C(1, 0.7, 0.7^2, \ldots, 0.7^{98}, 0.7^{99})'$; we term this the "exponential" design. In the second and third case, we set $\Pi = C\widetilde{\Pi} = C(\iota_s, 0_{n-s})'$, where ι_s is a $1 \times s$ vector of ones and 0_{n-s} is a $1 \times n - s$ vector of zeros. We term this the "cut-off" design and consider two different values of s, $s = 5$ and $s = 50$. In the exponential design, the model is not literally sparse, although the majority of explanatory power is contained in the first few instruments. While the model is exactly sparse in the cut-off design, we expect Lasso to perform poorly with $s = 50$ since treating $\frac{s^2 \log^2 p}{n}$ as vanishingly small seems like a poor approximation given the sample sizes considered. We consider different values of the constant C that are chosen to generate target values for the concentration parameter, $\mu^2 = \frac{n\Pi'\Sigma_Z\Pi}{\sigma_v^2}$, which plays a key role in the behavior of IV estimators; see, for example, Stock, Wright, and Yogo (2002) or Hansen, Hausman, and Newey (2008).[13] Specifically, we choose C to solve $\mu^2 = \frac{nC^2\widetilde{\Pi}\Sigma_Z\widetilde{\Pi}}{1-C^2\widetilde{\Pi}\Sigma_Z\widetilde{\Pi}}$ for $\mu^2 = 30$ and for $\mu^2 = 180$. These values of the concentration parameter were chosen by using estimated values from the empirical example reported below as a benchmark.[14]

For each setting of the simulation parameter values, we report results from seven different procedures. A simple possibility, when presented with many instrumental variables (with $p < n$), 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, Bekker (1994). The estimator proposed in Fuller (1977) (FULL) is robust to many instruments (with $p < n$) as long as the presence of many instruments is accounted for when constructing standard errors for the estimators; see Hansen, Hausman, and Newey (2008), for example.[15] We report results for these estimators in rows labeled 2SLS(100) and FULL(100), respectively.[16] For our variable selection procedures, we use Lasso to select among the instruments using the refined data-dependent penalty loadings given in (A.1), described in Appendix A, and consider two post-model selection estimation procedures. The first, post-Lasso, runs 2SLS using the instruments selected by Lasso; and the second, post-Lasso-F, runs FULL using

[13] The concentration parameter is closely related to the first-stage Wald statistic and first-stage F-statistic for testing that the coefficients on the instruments are equal to 0. Under homoscedasticity, the first-stage Wald statistic is $W = \widehat{\Pi}'(Z'Z)\widehat{\Pi}/\widehat{\sigma}_v^2$ and the first-stage F-statistic is $W/\dim(Z)$.

[14] In the empirical example, first-stage Wald statistics based on the selected instruments range from between 44 and 243. In the cases with constant coefficients, our concentration parameter choices correspond naturally to "infeasible F-statistics" defined as μ^2/s of 6 and 36 with $s = 5$ and 0.6 and 3.6 with $s = 50$. In the Supplemental Material (Belloni et al. 2012), we provide additional simulation results. The results reported in the current section are sufficient to capture the key patterns.

[15] FULL requires a user-specified parameter. We set this parameter equal to 1, which produces a higher-order unbiased estimator. See Hahn, Hausman, and Kuersteiner (2004) for additional discussion. Limited information maximum likelihood (LIML) is another commonly proposed estimator that is robust to many instruments. In our designs, its performance was generally similar to that of FULL, and we report only FULL for brevity.

[16] With $n = 100$, estimates are based on a randomly selected 99 instruments.

the instruments selected by Lasso. In cases where no instruments are selected by Lasso, we use the point-estimate obtained by running 2SLS with the single instrument with the highest within sample correlation to the endogenous variable as the point estimate for post-Lasso and post-Lasso-F. In these cases, we use the sup-score test for performing inference.[17] We report inference results based on the weak-identification robust sup-score testing procedure in rows labeled "sup-Score."

The other two procedures, "post-Lasso (Ridge)" and "post-Lasso-F (Ridge)," use a combination of Ridge regression, Lasso, and sample-splitting. For these procedures, we randomly split the sample into two equal-sized parts. Call these subsamples "sample A" and "sample B." We then use leave-one-out cross-validation with only the data in sample A to select a ridge penalty parameter, and then estimate a set of ridge coefficients using this penalty and the data from sample A. We then use the data from sample B with these coefficients estimated using only data from sample A to form first-stage fitted values for sample B. Then, we take the full set of instruments augmented with the estimated fitted value just described and perform Lasso variable selection using only the data from sample B. We use the selected variables to run either 2SLS or Fuller in sample B to obtain estimates of β (and associated standard errors), say $\widehat{\beta}_{B,\text{2SLS}}$ ($s_{B,\text{2SLS}}$) and $\widehat{\beta}_{B,\text{Fuller}}$ ($s_{B,\text{Fuller}}$). We then repeat this exercise switching samples A and B to obtain estimates of β (and associated standard errors) from sample A, say $\widehat{\beta}_{A,\text{2SLS}}$ ($s_{A,\text{2SLS}}$) and $\widehat{\beta}_{A,\text{Fuller}}$ ($s_{A,\text{Fuller}}$). Post-Lasso (Ridge) is then $w_{A,\text{2SLS}}\widehat{\beta}_{A,\text{2SLS}} + (1 - w_{A,\text{2SLS}})\widehat{\beta}_{B,\text{2SLS}}$ for $w_{A,\text{2SLS}} = \frac{s^2_{B,\text{2SLS}}}{s^2_{A,\text{2SLS}}+s^2_{B,\text{2SLS}}}$, and post-Lasso-F (Ridge) is defined similarly. If instruments are selected in one subsample but not in the other, we put weight 1 on the estimator from the subsample where instruments were selected. If no instruments are selected in either subsample, we use the single instrument with the highest correlation to obtain the point-estimate and use the sup-score test for performing inference.

For each estimator, we report median bias (Med. Bias), median absolute deviation (MAD), and rejection frequencies for 5% level tests (rp(0.05)). For computing rejection frequencies, we estimate conventional, homoscedastic 2SLS standard errors for 2SLS(100) and post-Lasso and the many-instrument robust standard errors of Hansen, Hausman, and Newey (2008), which rely on homoscedasticity for FULL(100) and post-Lasso-F. We report the number of cases in which Lasso selected no instruments in the column labeled N(0).

We summarize the simulation results in Table I. It is apparent that the Lasso procedures are dominant when $n = 100$. In this case, the Lasso-based procedures outperform 2SLS(100) and FULL(100) on all dimensions considered. When the concentration parameter is 30 or $s = 50$, the instruments are relatively weak, and Lasso accordingly selects no instruments in many cases. In

[17] Inference based on the asymptotic approximation when Lasso selects instruments and based on the sup-Score test when Lasso fails to select instruments is our preferred procedure.

TABLE I

SIMULATION RESULTS[a]

	Exponential				$S = 5$				$S = 50$			
	Median				Median				Median			
Estimator	N(0)	Bias	MAD	rp(0.05)	N(0)	Bias	MAD	rp(0.05)	N(0)	Bias	MAD	rp(0.05)
A. Concentration Parameter = 30, $n = 100$												
2SLS(100)		0.524	0.524	1.000		0.520	0.520	1.000		0.528	0.528	0.998
FULL(100)		0.373	0.741	0.646		0.476	0.781	0.690		0.285	0.832	0.580
Post-LASSO	483	0.117	0.183	0.012	485	0.128	0.178	0.008	498	0.363	0.368	0.012
Post-LASSO-F	483	0.117	0.184	0.012	485	0.128	0.178	0.008	498	0.363	0.368	0.012
Post-LASSO (Ridge)	500	0.229	0.263	0.000	500	0.212	0.239	0.000	500	0.362	0.364	0.002
Post-LASSO-F (Ridge)	500	0.229	0.263	0.000	500	0.212	0.239	0.000	500	0.362	0.364	0.002
sup-Score				0.006				0.000				0.008
B. Concentration Parameter = 30, $n = 250$												
2SLS(100)		0.493	0.493	1.000		0.485	0.485	1.000		0.486	0.486	1.000
FULL(100)		0.028	0.286	0.076		0.023	0.272	0.056		0.046	0.252	0.072
Post-LASSO	396	0.106	0.163	0.044	423	0.105	0.165	0.042	499	0.358	0.359	0.008
Post-LASSO-F	396	0.107	0.164	0.048	423	0.105	0.166	0.044	499	0.358	0.359	0.008
Post-LASSO (Ridge)	500	0.191	0.223	0.004	500	0.196	0.217	0.006	500	0.353	0.355	0.000
Post-LASSO-F (Ridge)	500	0.191	0.223	0.004	500	0.196	0.217	0.006	500	0.353	0.355	0.000
sup-Score				0.002				0.010				0.006

(*Continues*)

TABLE I—*Continued*

	Exponential				$S = 5$				$S = 50$			
	Median				Median				Median			
Estimator	N(0)	Bias	MAD	rp(0.05)	N(0)	Bias	MAD	rp(0.05)	N(0)	Bias	MAD	rp(0.05)
C. Concentration Parameter = 180, $n = 100$												
2SLS(100)		0.353	0.353	0.952		0.354	0.354	0.958		0.350	0.350	0.948
FULL(100)		0.063	0.563	0.648		0.096	0.562	0.694		0.148	0.538	0.656
Post-LASSO	120	0.037	0.093	0.078	132	0.035	0.100	0.052	498	0.192	0.211	0.000
Post-LASSO-F	120	0.030	0.093	0.070	132	0.025	0.100	0.046	498	0.192	0.211	0.000
Post-LASSO (Ridge)	500	0.061	0.132	0.002	500	0.063	0.116	0.000	500	0.004	0.119	0.000
Post-LASSO-F (Ridge)	500	0.061	0.132	0.002	500	0.063	0.116	0.000	500	0.004	0.119	0.000
sup-Score				0.002				0.002				0.000
D. Concentration Parameter = 180, $n = 250$												
2SLS(100)		0.289	0.289	0.966		0.281	0.281	0.972		0.280	0.280	0.964
FULL(100)		0.008	0.082	0.058		0.007	0.081	0.044		0.008	0.083	0.048
Post-LASSO	0	0.032	0.073	0.054	0	0.019	0.067	0.060	411	0.233	0.237	0.044
Post-LASSO-F	0	0.024	0.069	0.038	0	0.014	0.068	0.046	411	0.235	0.236	0.040
Post-LASSO (Ridge)	211	0.062	0.095	0.098	225	0.058	0.084	0.082	295	−0.008	0.090	0.030
Post-LASSO-F (Ridge)	211	0.061	0.096	0.082	225	0.056	0.081	0.062	295	−0.004	0.090	0.032
sup-Score				0.012				0.012				0.012

[a] Results are based on 500 simulation replications and 100 instruments. Column labels indicate the structure of the first-stage coefficients as described in the text. 2SLS(100) and FULL(100) are respectively the 2SLS and Fuller(1) estimators using all 100 potential instruments. Post-LASSO and Post-LASSO-F respectively correspond to 2SLS and Fuller(1) using the instruments selected from LASSO variable selection among the 100 instruments, with inference based on the asymptotic normal approximation; in cases where no instruments are selected, the procedure switches to using the sup-Score test for inference. sup-Score provides the rejection frequency for a weak identification robust procedure that is suited to situations with more instruments than observations. Post-LASSO (Ridge) and Post-LASSO-F (Ridge) are defined as Post-LASSO and Post-LASSO-F but augment the instrument set with a fitted value obtained via ridge regression as described in the text. We report the number of replications in which LASSO selected no instruments (N(0)), median bias (Med. Bias), median absolute deviation (MAD), and rejection frequency for 5% level tests (rp(0.05)). In cases where LASSO selects no instruments, Med. Bias, and MAD are based on 2SLS using the single instrument with the largest sample correlation to the endogenous variable and rp(0.05) is based on the sup-Score test.

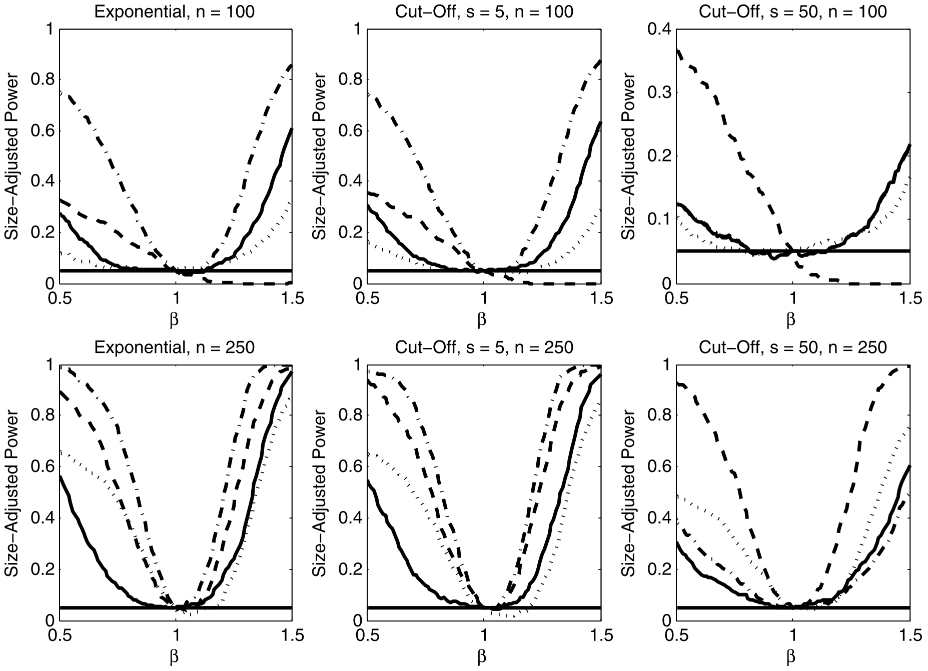

FIGURE 1.—Size-adjusted power curves for post-Lasso-F (dot–dash), post-Lasso-F (Ridge) (dotted), FULL(100) (dashed), and sup-Score (solid) from the simulation example with concentration parameter of 180 for $n = 100$ and $n = 250$.

these cases, inference switches to the robust sup-score procedure, which controls size. With a concentration parameter of 180, the instruments are relatively more informative and sparsity provides a good approximation in the exponential design and $s = 5$ cut-off design. In these cases, Lasso selects instruments in the majority of replications and the procedure has good risk and inference properties relative to the other procedures considered. In the $n = 100$ case, the simple Lasso procedures also clearly dominate Lasso augmented with Ridge, as this procedure often results in no instruments being selected and relatively low power; see Figure 1. We also see that the sup-score procedure controls size across the designs considered.

In the $n = 250$ case, the conventional many-instrument asymptotic sequence, which has p proportional to n but $p/n < 1$, provides a reasonable approximation to the DGP, and one would expect FULL to perform well. In this case, 2SLS(100) is clearly dominated by the other procedures. However, there is no obvious ranking between FULL(100) and the Lasso-based procedures. With $s = 50$, sparsity is a poor approximation in that there is signal in the combination of the 50 relevant instruments, but no small set of instruments has much explanatory power. In this setting, FULL(100) has lower estimation risk than the Lasso procedure, which is not effectively able to capture the diffuse signal though both inference procedures have size close to the prescribed level. Lasso augmented with the Ridge fit also does relatively well in this setting, being roughly on par with FULL(100). In the exponential and cut-off with $s = 5$ designs, sparsity is a much better approximation. In these cases, the simple Lasso-based estimators have smaller risk than FULL(100) or Lasso with Ridge, and produce tests that have size close to the nominal 5% level. Finally, we see that the sup-score procedure continues to control size with $n = 250$.

Given that the sup-score procedure uniformly controls size across the designs considered but is actually substantially undersized, it is worth presenting additional results regarding power. We plot size-adjusted power curves for the sup-score test, post-Lasso-F, post-Lasso-F (Ridge), and FULL(100) across the different designs in the $\mu^2 = 180$ cases in Figure 1. We focus on $\mu^2 = 180$ since we expect it is when identification is relatively strong that differences in power curves will be most pronounced. From these curves, it is apparent that the robustness of the sup-score test comes with a substantial loss of power in cases where identification is strong. Exploring other procedures that are robust to weak identification, allow for $p \gg n$, and do not suffer from such power losses may be interesting for future research.

6.1. *Conclusions From Simulation Experiments*

The evidence from the simulations is supportive of the derived theory and favorable to Lasso-based IV methods. The Lasso-IV estimators clearly dominate on all metrics considered when $p = n$ and $s \ll n$. The Lasso-based IV estimators generally have relatively small median bias and estimator risk and do well in terms of testing properties, though they do not dominate FULL in these dimensions across all designs with $p < n$. The simulation results verify that FULL becomes more appealing as the sparsity assumption breaks down. This breakdown of sparsity is likely in situations with weak instruments, be they many or few, where none of the first-stage coefficients are well-separated from zero relative to sampling variation. Overall, the simulation results show that simple Lasso-based procedures can usefully complement other many-instrument methods.

7. THE IMPACT OF EMINENT DOMAIN ON ECONOMIC OUTCOMES

As an example of the potential application of Lasso to select instruments, we consider IV estimation of the effects of federal appellate court decisions regarding eminent domain on a variety of economic outcomes.[18] To try to un-

[18] See Chen and Yeh (2010) for a detailed discussion of the economics of takings law (or eminent domain), relevant institutional features of the legal system, and a careful discussion of endogeneity concerns and the instrumental variables strategy in this context.

cover the relationship between takings law and economic outcomes, we estimate structural models of the form

$$y_{ct} = \alpha_c + \alpha_t + \gamma_c t + \beta \text{Takings Law}_{ct} + W_{ct}'\delta + \epsilon_{ct},$$

where y_{ct} is an economic outcome for 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,[19] a dummy for whether there were no cases in that circuit-year, and the number of takings appellate decisions; and α_c, α_t, and $\gamma_c t$ are respectively circuit-specific effects, time-specific effects, and circuit-specific time trends. An appellate court decision is coded as pro-plaintiff if the court ruled that a taking was unlawful, thus overturning the government's seizure of the property in favor of the private owner. We construe pro-plaintiff decisions to indicate a regime that is more protective of individual property rights. The parameter of interest, β, thus represents the effect of an additional decision upholding individual property rights on an economic outcome.

We provide results using four different economic outcomes: the log of three home price indices and log(GDP). The three different home price indices we consider are the quarterly, weighted, repeat-sales FHFA/OFHEO house price index that tracks single-family house prices at the state level for metro (FHFA) and non-metro (Non-Metro) areas and the Case–Shiller home price index (Case–Shiller) by month for 20 metropolitan areas based on repeat-sales residential housing prices. We also use state-level GDP from the Bureau of Economic Analysis to form log(GDP). For simplicity and since all of the controls, instruments, and the endogenous variable vary only at the circuit-year level, we use the within-circuit-year average of each of these variables as the dependent variables in our models. Due to the different coverage and time series lengths available for each of these series, the sample sizes and sets of available controls differ somewhat across the outcomes. These differences lead to different first stages across the outcomes as well. The total sample sizes are 312 for FHFA and GDP, which have identical first stages. For Non-Metro and Case–Shiller, the sample sizes are 110 and 183, respectively.

The analysis of the effects of takings law is complicated by the possible endogeneity between governmental takings and takings law decisions and economic variables. To address the potential endogeneity of takings law, we employ an instrumental variables strategy based on the identification argument of Chen and Sethi (2010) and Chen and Yeh (2010) that relies on the random assignment of judges to federal appellate panels. Since judges are randomly assigned to three-judge panels to decide appellate cases, the exact identity of the judges

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

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 are many potential characteristics of three-judge panels that may be used as instruments. While the basic identification argument suggests any set of characteristics of the three-judge panel will be uncorrelated with the structural unobservable, there will clearly be some instruments that are more worthwhile than others in obtaining precise second-stage estimates. For simplicity, we consider only the following demographics: gender, race, religion, political affiliation, whether the judge's bachelor's degree was obtained in-state, whether the bachelor's degree is from a public university, whether the JD was obtained from a public university, and whether the judge was elevated from a district court, along with various interactions. In total, we have 138, 143, 147, and 138 potential instruments for FHFA prices, non-metro prices, Case–Shiller, and GDP, respectively, that we select among using Lasso.[20]

Table II contains estimation results for β. We report OLS estimates and results based on three different sets of instruments. The first set of instruments, used in the rows labeled 2SLS, are the instruments adopted in Chen and Yeh (2010).[21] We consider this the baseline. The second set of instruments are those selected through Lasso using the refined data-driven penalty.[22] The number of instruments selected by Lasso is reported in the row "S." We use the post-Lasso 2SLS estimator and report these results in the rows labeled "post-Lasso." The third set of instruments is simply the union of the first two instrument sets. Results for this set of instruments are in the rows labeled "post-Lasso+." In this case, "S" is the total number of instruments used. In all cases, we use heteroscedasticity consistent standard error estimators. Finally, we report the value of the test statistic discussed in Section 5.1, comparing es-

[20] Given the sample sizes and numbers of variables, estimators using all the instruments without shrinkage are only defined in the GDP and FHFA data. For these outcomes, the Fuller (1977) point-estimate (standard error) is -0.0020 (3.123) for FHFA and 0.0120 (0.1758) for GDP.

[21] Chen and Yeh (2010) used two variables motivated on intuitive grounds, whether a panel was assigned an appointee who did not report a religious affiliation and whether a panel was assigned an appointee who earned his or her first law degree from a public university, as instruments.

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

TABLE II

EFFECT OF FEDERAL APPELLATE TAKINGS LAW DECISIONS ON ECONOMIC OUTCOMES[a]

	Home Prices			GDP
	log(FHFA)	log(Non-Metro)	log(Case–Shiller)	log(GDP)
Sample Size	312	110	183	312
OLS	0.0114	0.0108	0.0152	0.0099
s.e.	0.0132	0.0066	0.0132	0.0048
2SLS	0.0262	0.0480	0.0604	0.0165
s.e.	0.0441	0.0212	0.0296	0.0162
FS-W	28.0859	82.9647	67.7452	28.0859
Post-LASSO	0.0369	0.0357	0.0631	0.0133
s.e.	0.0465	0.0132	0.0249	0.0161
FS-W	44.5337	243.1946	89.5950	44.5337
S	1	4	2	1
Post-LASSO+	0.0314	0.0348	0.0628	0.0144
s.e.	0.0366	0.0127	0.0245	0.0131
FS-W	73.3010	260.9823	105.3206	73.3010
S	3	6	3	3
Spec. Test	−0.2064	0.5753	−0.0985	0.1754

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

timates using the first and second sets of instruments in the row labeled "Spec. Test."

The most interesting results from the standpoint of the present paper are found by comparing first-stage Wald statistics and estimated standard errors across the instrument sets. The Lasso instruments are clearly much better first-stage predictors as measured by the first-stage Wald statistic compared to the Chen and Yeh (2010) benchmark. Given the degrees of freedom, this increase obviously corresponds to Lasso-based IV providing a stronger first-stage relationship for FHFA prices, GDP, and the Case–Shiller prices. In the non-metro case, the p-value from the Wald test with the baseline instruments of Chen and Yeh (2010) is larger than that of the Lasso-selected instruments. This improved first-stage prediction is associated with the resulting 2SLS estimator having smaller estimated standard errors than the benchmark case for non-metro prices, Case–Shiller prices, and GDP. The reduction in standard errors is sizable for both non-metro and Case–Shiller. The standard error estimate is somewhat larger in the FHFA case despite the improvement in first-stage prediction. Given that the post-Lasso first-stage produces a larger first-stage Wald statistic while choosing fewer instruments than the benchmark suggests that we might prefer the post-Lasso results in any case. We also see that the test statistics for testing that the difference between the estimate using the Chen and Yeh (2010) instruments and the post-Lasso estimate are uniformly small. Given the small differences between estimates using the first two sets of instruments, it is unsurprising that the results using the union of the two instrument sets are similar to those already discussed.

The results are also economically interesting. The point-estimates for the effect of an additional pro-plaintiff decision, a decision in favor of individual property holders, are positive, suggesting that these decisions are associated with increases in property prices and GDP. These point-estimates are all small, and it is hard to draw any conclusion about the likely effect on GDP or the FHFA index given their estimated standard errors. On the other hand, confidence intervals for non-metro and Case–Shiller constructed at usual confidence levels exclude zero. Overall, the results do suggest that the causal effect of decisions reinforcing individual property rights is an increase in the value of holding property, at least in the short term. The results are also consistent with the developed asymptotic theory in that the 2SLS point-estimates based on the benchmark instruments are similar to the estimates based on the Lasso-selected instruments, while Lasso produces a stronger first-stage relationship and the post-Lasso estimates are more precise in three of the four cases. The example suggests that there is the potential for Lasso to be fruitfully employed to choose instruments in economic applications.

APPENDIX A: IMPLEMENTATION ALGORITHMS

It is useful to organize the precise implementation details into the following algorithm. Feasible options for setting the penalty level and the loadings for $j = 1, \dots, p$, and $l = 1, \dots, k_e$ are

$$\text{(A.1)} \qquad \text{initial} \quad \widehat{\gamma}_{lj} = \sqrt{\mathbb{E}_n\big[f_{ij}^2(d_{il} - \bar{d}_l)^2\big]}, \quad \lambda = 2c\sqrt{n}\Phi^{-1}\big(1 - \gamma/(2k_e p)\big),$$

$$\text{refined} \quad \widehat{\gamma}_{lj} = \sqrt{\mathbb{E}_n\big[f_{ij}^2\widehat{v}_{il}^2\big]}, \quad \lambda = 2c\sqrt{n}\Phi^{-1}\big(1 - \gamma/(2k_e p)\big),$$

where $c > 1$ is a constant, $\gamma \in (0, 1)$, $\bar{d}_l := \mathbb{E}_n[d_{il}]$, and $\widehat{v}_{il}$ is an estimate of v_{il}. Let $K \geq 1$ denote a bounded number of iterations. We used $c = 1.1$, $\gamma = 0.1/\log(p \vee n)$, and $K = 15$ in the simulations. In what follows, Lasso/post-Lasso estimator indicates that the practitioner can apply either the Lasso or post-Lasso estimator. Our preferred approach uses post-Lasso at every stage.

ALGORITHM A.1—Lasso/Post-Lasso Estimators: (a) For each $l = 1, \dots, k_e$, specify penalty loadings according to the initial option in (A.1). Use these penalty loadings in computing the Lasso/post-Lasso estimator $\widehat{\beta}_l$ via equations (2.4) or (2.8). Then compute residuals $\widehat{v}_{il} = d_{li} - f_i'\widehat{\beta}_l$, $i = 1, \dots, n$. (b) For each $l = 1, \dots, k_e$, update the penalty loadings according to the refined option in (A.1) and update the Lasso/post-Lasso estimator $\widehat{\beta}_l$. Then compute a new set of residuals using the updated Lasso/post-Lasso coefficients $\widehat{v}_{il} = d_{li} - f_i'\widehat{\beta}_l$, $i = 1, \dots, n$. (c) Repeat the previous step K times.

If Algorithm A.1 selected no instruments other than intercepts, or, more generally, if $\mathbb{E}_n[\widehat{D}_{il}\widehat{D}_{il}']$ is near-singular, proceed to Algorithm A.3; otherwise, we recommend the following algorithm.

ALGORITHM A.2—IV Inference Using Estimates of Optimal Instrument: (a) Compute the estimates of the optimal instrument, $\widehat{D}_{il} = f_i'\widehat{\beta}_l$, for $i = 1, \dots, n$ and each $l = 1, \dots, k_e$, where $\widehat{\beta}_l$ is computed by Algorithm A.1. Compute the IV estimator $\widehat{\alpha} = \mathbb{E}_n[\widehat{D}_i d_i']^{-1}\mathbb{E}_n[\widehat{D}_i y_i]$. (b) Compute estimates of the asymptotic variance matrix $\widehat{Q}^{-1}\widehat{\Omega}\widehat{Q}^{-1}$, where $\widehat{\Omega} := \mathbb{E}_n[\widehat{\epsilon}_i^2\widehat{D}_i\widehat{D}_i']$ for $\widehat{\epsilon}_i = y_i - d_i'\widehat{\alpha}$, and $\widehat{Q} := \mathbb{E}_n[\widehat{D}_i\widehat{D}_i']$. (c) Proceed to perform conventional inference using the normality result (2.10).

The following algorithm is only invoked if the weak-instruments problem has been diagnosed, for example, using the methods of Stock and Yogo (2005). In the algorithm below, $\mathcal{A}_1$ is the parameter space, and $\mathcal{G}_1 \subset \mathcal{A}_1$ is a grid of potential values for α_1. Choose the confidence level $1 - \gamma$ of the interval, and set $\Lambda(1 - \gamma) = c\sqrt{n}\Phi^{-1}(1 - \gamma/2p)$.

ALGORITHM A.3—IV Inference Robust to Weak Identification: (a) Set $\mathcal{C} = \emptyset$. (b) For each $a \in \mathcal{G}_1$, compute Λ_a as in (4.5). If $\Lambda_a \leq \Lambda(1 - \gamma)$, add a to $\mathcal{C}$. (c) Report $\mathcal{C}$.

APPENDIX B: TOOLS

The following useful lemma is a consequence of moderate deviations theorems for self-normalized sums in Jing, Shao, and Wang (2003) and de la Peña, Lai, and Shao (2009).

We use the following result—Theorem 7.4 in de la Peña, Lai, and Shao (2009). Let $X_1, \dots, X_n$ be independent, zero-mean variables, and $S_n = \sum_{i=1}^n X_i$,

$V_n^2 = \sum_{i=1}^n X_i^2$. For $0 < \mu \le 1$, set $B_n^2 = \sum_{i=1}^n \mathrm{E} X_i^2$, $L_{n,\mu} = \sum_{i=1}^n \mathrm{E}|X_i|^{2+\mu}$, $d_{n,\mu} = B_n / L_{n,\mu}^{1/(2+\mu)}$. Then uniformly in $0 \le x \le d_{n,\mu}$,

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

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

where the terms $O(1)$ are bounded in absolute value by a universal constant A, $\bar{\Phi} := 1 - \Phi$, and Φ is the cumulative distribution function of a standard Gaussian random variable.

LEMMA 5—Moderate Deviation Inequality for Maximum of a Vector: *Suppose that for each j*

$$\mathcal{S}_j = \frac{\sum_{i=1}^n U_{ij}}{\sqrt{\sum_{i=1}^n U_{ij}^2}},$$

where U_{ij} are independent variables across i with mean zero. We have that

$$\mathrm{P}\Bigl(\max_{1\le j\le p} |\mathcal{S}_j| > \Phi^{-1}(1-\gamma/2p)\Bigr) \le \gamma\left(1 + \frac{A}{\ell_n^3}\right),$$

where A is an absolute constant, provided that, for $\ell_n > 0$,

$$0 \le \Phi^{-1}\bigl(1-\gamma/(2p)\bigr) \le \frac{n^{1/6}}{\ell_n} \min_{1\le j\le p} M[U_j] - 1,$$

$$M[U_j] := \frac{\left(\frac{1}{n}\sum_{i=1}^n \mathrm{E} U_{ij}^2\right)^{1/2}}{\left(\frac{1}{n}\sum_{i=1}^n \mathrm{E}|U_{ij}^3|\right)^{1/3}}.$$

PROOF: *Step 1.* We first note the following simple consequence of the result of Theorem 7.4 in de la Peña, Lai, and Shao (2009). Let $X_{1,n}, \dots, X_{n,n}$ be the triangular array of i.n.i.d., zero-mean random variables. Suppose that

$$n^{1/6}M_n/\ell_n \geq 1, \quad M_n := \frac{\left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}X_{i,n}^2\right)^{1/2}}{\left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}|X_{i,n}|^3\right)^{1/3}}.$$

Then uniformly on $0 \leq x \leq n^{1/6}M_n/\ell_n - 1$, the quantities $S_{n,n} = \sum_{i=1}^{n} X_{i,n}$ and $V_{n,n}^2 = \sum_{i=1}^{n} X_{i,n}^2$ obey

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

This corollary follows by the application of the quoted theorem to the case with $\mu = 1$. The calculated error bound follows from the triangular inequalities and conditions on ℓ_n and M_n.

Step 2. It follows that

$$\begin{aligned}
&\mathrm{P}\Big(\max_{1\leq j\leq p} |\mathcal{S}_j| > \Phi^{-1}(1-\gamma/2p)\Big) \\
&\quad \leq_{(1)} p \max_{1\leq j\leq p} \mathrm{P}\big(|\mathcal{S}_j| > \Phi^{-1}(1-\gamma/2p)\big) \\
&\quad =_{(2)} p\mathrm{P}\big(|\mathcal{S}_{j_n}| > \Phi^{-1}(1-\gamma/2p)\big) \\
&\quad \leq_{(3)} p2\bar{\Phi}\big(\Phi^{-1}(1-\gamma/2p)\big)\left(1 + \frac{A}{\ell_n^3}\right) \\
&\quad \leq 2p\gamma/(2p)\left(1 + \frac{A}{\ell_n^3}\right) \\
&\quad \leq \gamma\left(1 + \frac{A}{\ell_n^3}\right),
\end{aligned}$$

on the set $0 \leq \Phi^{-1}(1-\gamma/(2p)) \leq \frac{n^{1/6}}{\ell_n}M_{j_n} - 1$, where inequality (1) follows by the union bound, equality (2) is the maximum taken over finite set, so the maximum is attained at some $j_n \in \{1, \dots, p\}$, and the last inequality follows by the application of Step 1, by setting $X_{i,n} = U_{i,j_n}$. *Q.E.D.*

APPENDIX C: PROOF OF THEOREM 1

The proof of Theorem 1 has four steps. The most important steps are Steps 1–3. One half of Step 1 for bounding the $\|\cdot\|_{2,n}$-rate follows the strategy of Bickel, Ritov, and Tsybakov (2009), but accommodates data-driven penalty loadings. The other half of Step 1 for bounding the $\|\cdot\|_1$-rate is new for the nonparametric case. Step 2 innovatively uses the moderate deviation theory for self-normalized sums, which allows us to obtain sharp results for non-Gaussian and heteroscedastic errors as well as handle data-driven penalty loadings. Step 3 relates the ideal penalty loadings and the feasible penalty loadings. Step 4 puts the results together to reach the conclusions.

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

$$\kappa_C^l = \min_{\delta\in\mathbb{R}^p:\|\widehat{\Upsilon}_l^0\delta_{T_l^c}\|_1\leq C\|\widehat{\Upsilon}_l^0\delta_{T_l}\|_1,\|\delta\|_2\neq 0} \frac{\sqrt{s}\|f_i'\delta\|_{2,n}}{\|\widehat{\Upsilon}_l^0\delta_{T_l}\|_1}.$$

This quantity controls the modulus of continuity between the prediction norm $\|f_i'\delta\|_{2,n}$ and the ℓ_1-norm $\|\delta\|_1$ within a restricted region that depends on $l = 1, \dots, k_e$. Note that if $a = \min_{1\leq l\leq k_e}\min_{1\leq j\leq p}\widehat{\Upsilon}_{lj}^0 \leq \max_{1\leq l\leq k_e}\|\widehat{\Upsilon}_l^0\|_\infty = b$, for every $C > 0$, because $\{\delta \in \mathbb{R}^p : \|\widehat{\Upsilon}_l^0\delta_{T_l^c}\|_1 \leq C\|\widehat{\Upsilon}_l^0\delta_{T_l}\|_1\} \subseteq \{\delta \in \mathbb{R}^p : a\|\delta_{T_l^c}\|_1 \leq bC\|\delta_{T_l}\|_1\}$ and $\|\widehat{\Upsilon}_l^0\delta_{T_l}\|_1 \leq b\|\delta_{T_l}\|_1$, we have

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

where the latter is the restricted eigenvalue defined in (3.1). If $C = c_0 = (uc + 1)/(\ell c - 1)$, we have $\min_{1\leq l\leq k_e}\kappa_{c_0}^l \geq (1/b)\kappa_{\bar{C}}(\mathbb{E}_n[f_if_i'])$. By Condition RF and by Step 3 of Appendix C below, we have a bounded away from zero and b bounded from above with probability approaching 1 as n increases.

The main result of this step is the following lemma.

LEMMA 6: *Under Condition* AS, *if* $\lambda/n \geq c\|S_l\|_\infty$, *and* $\widehat{\Upsilon}_l$ *satisfies* (3.2) *with* $u \geq 1 \geq \ell > 1/c$, *then*

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

$$\big\|\widehat{\Upsilon}_l^0(\widehat{\beta}_l - \beta_{l0})\big\|_1 \leq 3c_0\frac{\sqrt{s}}{\kappa_{2c_0}^l}\left(\big(u + [1/c]\big)\frac{\lambda\sqrt{s}}{n\kappa_{c_0}^l} + 2c_s\right) + \frac{3c_0n}{\lambda}c_s^2,$$

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

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

$$\widehat{Q}_l(\widehat{\beta}_l) - \widehat{Q}_l(\beta_{l0}) \leq \frac{\lambda}{n}\big(\|\widehat{\Upsilon}_l\beta_{l0}\|_1 - \|\widehat{\Upsilon}_l\widehat{\beta}_l\|_1\big). \tag{C.1}$$

Expanding the quadratic function $\widehat{Q}_l$, and using that $S_l = 2(\widehat{\Upsilon}_l^0)^{-1}\mathbb{E}_n[v_{il}f_i]$, we have

$$\text{(C.2)} \qquad \begin{aligned} \big|\widehat{Q}_l(\widehat{\beta}_l) - \widehat{Q}_l(\beta_{l0}) - \|f_i'\delta_l\|_{2,n}^2\big| &= \big|2\mathbb{E}_n\big[v_{il}f_i'\delta_l\big] + 2\mathbb{E}_n\big[a_{il}f_i'\delta_l\big]\big| \\ &\le \|S_l\|_\infty \big\|\widehat{\Upsilon}_l^0\delta_l\big\|_1 + 2c_s\big\|f_i'\delta_l\big\|_{2,n}. \end{aligned}$$

So combining (C.1) and (C.2) with $\lambda/n \ge c\|S_l\|_\infty$ and the conditions imposed on $\widehat{\Upsilon}_l$ in the statement of the theorem,

$$\text{(C.3)} \qquad \begin{aligned} \|f_i'\delta_l\|_{2,n}^2 &\le \frac{\lambda}{n}\big(\|\widehat{\Upsilon}_l\delta_{lT_l}\|_1 - \|\widehat{\Upsilon}_l\delta_{lT_l^c}\|_1\big) + \|S_l\|_\infty\big\|\widehat{\Upsilon}_l^0\delta_l\big\|_1 \\ &\quad + 2c_s\big\|f_i'\delta_l\big\|_{2,n} \\ &\le \left(u + \frac{1}{c}\right)\frac{\lambda}{n}\big\|\widehat{\Upsilon}_l^0\delta_{lT_l}\big\|_1 - \left(\ell - \frac{1}{c}\right)\frac{\lambda}{n}\big\|\widehat{\Upsilon}_l^0\delta_{lT_l^c}\big\|_1 \\ &\quad + 2c_s\big\|f_i'\delta_l\big\|_{2,n}. \end{aligned}$$

To show the first statement of the lemma, we can assume $\|f_i'\delta_l\|_{2,n} \ge 2c_s$; otherwise we are done. This condition together with relation (C.3) implies that, for $c_0 = (uc+1)/(\ell c - 1)$, we have $\|\widehat{\Upsilon}_l^0\delta_{lT_l^c}\|_1 \le c_0\|\widehat{\Upsilon}_l^0\delta_{lT_l}\|_1$. Therefore, by definition of $\kappa_{c_0}^l$, we have $\|\widehat{\Upsilon}_l^0\delta_{lT_l}\|_1 \le \sqrt{s}\|f_i'\delta_l\|_{2,n}/\kappa_{c_0}^l$. Thus, relation (C.3) implies $\|f_i'\delta_l\|_{2,n}^2 \le (u + \frac{1}{c})\frac{\lambda\sqrt{s}}{n\kappa_{c_0}^l}\|f_i'\delta_l\|_{2,n} + 2c_s\|f_i'\delta_l\|_{2,n}$ and the result follows.

To establish the second statement of the lemma, we consider two cases. First, assume $\|\widehat{\Upsilon}_l^0\delta_{lT_l^c}\|_1 \le 2c_0\|\widehat{\Upsilon}_l^0\delta_{lT_l}\|_1$. In this case, by definition of $\kappa_{2c_0}^l$, we have

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

and the result follows by applying the first bound to $\|f_i'\delta_l\|_{2,n}$. On the other hand, consider the case that

$$\text{(C.4)} \qquad \big\|\widehat{\Upsilon}_l^0\delta_{lT_l^c}\big\|_1 > 2c_0\big\|\widehat{\Upsilon}_l^0\delta_{lT_l}\big\|_1,$$

which would already imply $\|f_i'\delta_l\|_{2,n} \le 2c_s$ by (C.3). Moreover,

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

where (1) holds by (C.3), (2) holds since $\|f_i'\delta_l\|_{2,n}(2c_s - \|f_i'\delta_l\|_{2,n}) \leq \max_{x\geq 0} x(2c_s - x) \leq c_s^2$, and (3) follows from (C.4). Thus,

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

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

Step 2. In this step, we prove a lemma about the quantiles of the maximum of the scores $S_l = 2\mathbb{E}_n[(\widehat{\Upsilon}_l^0)^{-1}f_i v_{il}]$, and use it to pin down the level of the penalty. For $\lambda = c2\sqrt{n}\Phi^{-1}(1-\gamma/(2k_e p))$, we have that as $\gamma \to 0$ and $n\to\infty$, $\mathrm{P}(c\max_{1\leq l\leq k_e} n\|S_l\|_\infty > \lambda) = o(1)$, provided that, for some $b_n \to \infty$,

$$2\Phi^{-1}\big(1-\gamma/(2k_e p)\big) \leq \frac{n^{1/6}}{b_n}\min_{1\leq j\leq p, 1\leq l\leq k_e} M_{jl}, \quad M_{jl} := \frac{\bar{\mathrm{E}}[f_{ij}^2 v_{il}^2]^{1/2}}{\bar{\mathrm{E}}[|f_{ij}|^3|v_{il}|^3]^{1/3}}.$$

Note that the last condition is satisfied under our conditions for large n for some $b_n\to\infty$, since k_e is fixed, $\log(1/\gamma) \lesssim \log(p\vee n)$, $K_n^2\log^3(p\vee n) = o(n)$, and $\min_{1\leq j\leq p,1\leq l\leq k_e} M_{jl} \gtrsim 1/K_n^{1/3}$. This result follows from the bounds on moderate deviations of a maximum of a vector provided in Lemma 5, by $\bar{\Phi}(t) \leq \phi(t)/t$, $\max_{j\leq p, l\leq k_e} 1/M_{jl} \lesssim K_n^{1/3}$, and $K_n^{2/3}\log(p\vee n) = o(n^{1/3})$ holding by Condition RF.

Step 3. The main result of this step is the following: Define the expected "ideal" penalty loadings $\Upsilon_l^0 := \mathrm{diag}(\sqrt{\bar{\mathrm{E}}[f_{i1}^2 v_{il}^2]}, \dots, \sqrt{\bar{\mathrm{E}}[f_{ip}^2 v_{il}^2]})$, where the entries of Υ_l^0 are bounded away from zero and from above uniformly in n by Condition RF. Then the empirical "ideal" loadings converge to the expected "ideal" loadings: $\max_{1\leq l\leq k_e}\|\widehat{\Upsilon}_l^0 - \Upsilon_l^0\|_\infty \to_{\mathrm{P}} 0$. This is assumed in Condition RF.

Step 4. Combining the results of all the steps above, given that $\lambda = 2c\sqrt{n}\Phi^{-1}(1-\gamma/(2pk_e)) \lesssim c\sqrt{n\log(pk_e/\gamma)}$, k_e fixed, and asymptotic valid penalty loadings $\widehat{\Upsilon}_l$, and using the bound $c_s \lesssim_{\mathrm{P}} \sqrt{s/n}$ from Condition AS, we obtain the conclusion that

$$\|f_i'(\widehat{\beta}_l - \beta_{l0})\|_{2,n} \lesssim_{\mathrm{P}} \frac{1}{\kappa_{c_0}^l}\sqrt{\frac{s\log(k_e p/\gamma)}{n}} + \sqrt{\frac{s}{n}},$$

which gives, by the triangular inequality and by $\|D_{il} - f_i'\beta_{l0}\|_{2,n} \leq c_s \lesssim_{\mathrm{P}} \sqrt{s/n}$ holding by Condition AS,

$$\|\widehat{D}_{il} - D_{il}\|_{2,n} \lesssim_{\mathrm{P}} \frac{1}{\kappa_{c_0}^l}\sqrt{\frac{s\log(k_e p/\gamma)}{n}}.$$

The first result follows since $\kappa_{\bar{C}} \lesssim_{\mathrm{P}} \kappa_{c_0}^l$ by Step 1.

To derive the ℓ_1-rate, we apply the second result in Lemma 6 as follows:

$$\begin{aligned}\|\widehat{\beta}_l-\beta_{l0}\|_1 &\leq \big\|(\widehat{Y}_l^0)^{-1}\big\|_\infty\big\|\widehat{Y}_l^0(\widehat{\beta}_l-\beta_{l0})\big\|_1\\ &\lesssim_{\mathrm{P}} \big\|(\widehat{Y}_l^0)^{-1}\big\|_\infty\left(\frac{\sqrt{s}}{\kappa^l_{2c_0}}\left(\frac{1}{\kappa^l_{c_0}}\sqrt{\frac{s\log(k_e p/\gamma)}{n}}+\sqrt{\frac{s}{n}}\right)\right.\\ &\quad\left.+\frac{1}{\sqrt{\log(p/\gamma)}}\frac{s}{\sqrt{n}}\right)\\ &\lesssim_{\mathrm{P}} \frac{1}{(\kappa^l_{2c_0})^2}\sqrt{\frac{s^2\log(k_e p/\gamma)}{n}}.\end{aligned}$$

That yields the result since $\kappa_{2\bar{C}}\lesssim_{\mathrm{P}}\kappa^l_{2c_0}$ by Step 1. *Q.E.D.*

APPENDIX D: PROOF OF THEOREM 2

The proof proceeds in three steps. The general strategy of Step 1 follows Belloni and Chernozhukov (2011a, 2012), but a major difference is the use of moderate deviation theory for self-normalized sums, which allows us to obtain the results for non-Gaussian and heteroscedastic errors as well as handle data-driven penalty loadings. The sparsity proofs are motivated by Belloni and Chernozhukov (2012) but adjusted for the data-driven penalty loadings that contain self-normalizing factors.

Step 1. Here we derive a general performance bound for post-Lasso, that actually contains more information than the statement of the theorem. This lemma will be invoked in Step 3 below.

Let $F=[f_1;\dots;f_n]'$ denote an n by p matrix and, for a set of indices $S\subset\{1,\dots,p\}$, we define $\mathcal{P}_S=F[S](F[S]'F[S])^{-1}F[S]'$ as the projection matrix on the columns associated with the indices in S.

LEMMA 7—Performance of the Post-Lasso: *Under Conditions* AS *and* RF, *let $\widehat{T}_l$ denote the support selected by $\widehat{\beta}_l=\widehat{\beta}_{l\mathrm{L}}$, $\widehat{T}_l\subseteq\widehat{I}_l$, $\widetilde{m}_l=|\widehat{I}_l\setminus T_l|$, and let $\widehat{\beta}_{l\mathrm{PL}}$ be the post-Lasso estimator based on $\widehat{I}_l$, $l=1,\dots,k_e$. Then we have*

$$\begin{aligned}\max_{l\le k_e}\big\|D_{il}-f_i'\widehat{\beta}_{l\mathrm{PL}}\big\|_{2,n} &\lesssim_{\mathrm{P}}\sqrt{\frac{s}{n}}\sqrt{\frac{k_e\wedge\log(sk_e)}{\phi_{\min}(s)}}+\max_{l\le k_e}\frac{\sqrt{\widetilde{m}_l\log(pk_e)}}{\sqrt{n\phi_{\min}(\widetilde{m}_l)}}\\ &\quad+\big\|(D_l-\mathcal{P}_{\widehat{I}_l}D_l)/\sqrt{n}\big\|_2,\end{aligned}$$

$$\begin{aligned}\max_{1\le l\le k_e}\big\|\widehat{Y}_l(\widehat{\beta}_{l\mathrm{PL}}-\beta_{l0})\big\|_1 &\leq \max_{1\le l\le k_e}\frac{(\|\widehat{Y}_l^0\|_\infty+\|\widehat{Y}_l-\widehat{Y}_l^0\|_\infty)\sqrt{\widetilde{m}_l+s}}{\sqrt{\phi_{\min}(\widetilde{m}_l+s)}}\\ &\quad\times\big\|f_i'(\widehat{\beta}_{l\mathrm{PL}}-\beta_{l0})\big\|_{2,n}.\end{aligned}$$

If, in addition, $\lambda/n \geq c\|S_l\|_\infty$, and $\widehat{Y}_l$ satisfies (3.2) *with $u \geq 1 \geq \ell > 1/c$ in the first stage for Lasso for every $l = 1, \ldots, k_e$, then we have*

$$\max_{l\leq k_e}\big\|(D_l - \mathcal{P}_{\widehat{I}_l}D_l)/\sqrt{n}\big\|_2 \leq \max_{l\leq k_e}\left(u+\frac{1}{c}\right)\frac{\lambda\sqrt{s}}{n\kappa^l_{c_0}} + 3c_s.$$

PROOF: We have that $D_l - F\widehat{\beta}_{l\mathrm{PL}} = (I - \mathcal{P}_{\widehat{I}_l})D_l - \mathcal{P}_{\widehat{I}_l}v_l$, where I is the identity operator. Therefore, for every $l = 1, \ldots, k_e$, we have

$$\text{(D.1)} \qquad \|D_l - F\widehat{\beta}_{l\mathrm{PL}}\|_2 \leq \big\|(I - \mathcal{P}_{\widehat{I}_l})D_l\big\|_2 + \|\mathcal{P}_{T_l}v_l\|_2 + \|\mathcal{P}_{\widehat{I}_l\setminus T_l}v_l\|_2.$$

Since $\|F[\widehat{I}_l \setminus T_l]/\sqrt{n}(F[\widehat{I}_l \setminus T_l]'F[\widehat{I}_l \setminus T_l]/n)^{-1}\| \leq \sqrt{1/\phi_{\min}(\widetilde{m}_l)}$, $\widetilde{m}_l = |\widehat{I}_l \setminus T_l|$, the last term in (D.1) satisfies

$$\begin{aligned}\|\mathcal{P}_{\widehat{I}_l\setminus T_l}v_l\|_2 &\leq \sqrt{1/\phi_{\min}(\widetilde{m}_l)}\big\|F[\widehat{I}_l \setminus T_l]'v_l/\sqrt{n}\big\|_2 \\ &\leq \sqrt{\widetilde{m}_l/\phi_{\min}(\widetilde{m}_l)}\big\|F'v_l/\sqrt{n}\big\|_\infty.\end{aligned}$$

Under Condition RF, by Lemma 5 we have

$$\max_{l=1,\ldots,k_e}\big\|F'v_l/\sqrt{n}\big\|_\infty \lesssim_{\mathrm{P}} \sqrt{\log(pk_e)}\max_{l\leq k_e, j\leq p}\sqrt{\mathbb{E}_n\big[f_{ij}^2v_{il}^2\big]}.$$

Note that Condition RF also implies $\max_{l\leq k_e, j\leq p}\sqrt{\mathbb{E}_n[f_{ij}^2v_{il}^2]} \lesssim_{\mathrm{P}} 1$ since

$$\max_{l\leq k_e, j\leq p}\big|(\mathbb{E}_n - \bar{\mathrm{E}})\big[f_{ij}^2v_{il}^2\big]\big| \to_{\mathrm{P}} 0$$

and

$$\max_{l\leq k_e, j\leq p}\bar{\mathrm{E}}\big[f_{ij}^2v_{il}^2\big] \leq \max_{l\leq k_e, j\leq p}\bar{\mathrm{E}}\big[f_{ij}^2\tilde{d}_{il}^2\big] \lesssim 1.$$

We bound the second term in (D.1) in two ways. First, proceeding as above, we have

$$\max_{l=1,\ldots,k_e}\|\mathcal{P}_{T_l}v_l\|_2 \lesssim_{\mathrm{P}} \sqrt{\log(k_e s)}\sqrt{s/\phi_{\min}(s)}\max_{l\leq k_e, j\leq p}\sqrt{\mathbb{E}_n\big[f_{ij}^2v_{il}^2\big]}.$$

Second, since $\mathrm{E}[\|F[T_l]'v_l\|_2^2] = \mathrm{E}[\sum_{j\in T_l}(\sum_{i=1}^n f_{ij}v_{il})^2] = \sum_{j\in T_l}\sum_{i=1}^n \mathrm{E}[f_{ij}^2v_{il}^2]$, we have

$$\max_{l=1,\ldots,k_e}\|\mathcal{P}_{T_l}v_l\|_2 \lesssim_{\mathrm{P}} \sqrt{sk_e/\phi_{\min}(s)}\max_{l\leq k_e, j\leq p}\sqrt{\bar{\mathrm{E}}\big[f_{ij}^2v_{il}^2\big]}.$$

These relations yield the first result.

Letting $\delta_l = \widehat{\beta}_{l\mathrm{PL}} - \beta_{l0}$, the statement regarding the ℓ_1-norm of the theorem follows from

$$\begin{aligned}\|\widehat{Y}_l\delta_l\|_1 &\le \|\widehat{Y}_l\|_\infty\|\delta_l\|_1 \\ &\le \|\widehat{Y}_l\|_\infty\sqrt{\|\delta_l\|_0}\|\delta_l\|_2 \\ &\le \|\widehat{Y}_l\|_\infty\sqrt{\|\delta_l\|_0}\big\|f_i'\delta_l\big\|_{2,n}\Big/\sqrt{\phi_{\min}\big(\|\delta_l\|_0\big)},\end{aligned}$$

and noting that $\|\delta_l\|_0 \le \widetilde{m}_l + s$ and $\|\widehat{Y}_l\|_\infty \le \|\widehat{Y}_l^0\|_\infty + \|\widehat{Y}_l - \widehat{Y}_l^0\|_\infty$.

The last statement follows from noting that the Lasso solution provides an upper bound to the approximation of the best model based on $\widehat{I}_l$, since $\widehat{T}_l \subseteq \widehat{I}_l$, and the application of Lemma 6. *Q.E.D.*

COMMENT D.1—Comparison Between Lasso and Post-Lasso Performance: Under mild conditions on the empirical Gram matrix and on the number of additional variables, Lemma 10 below derives sparsity bounds on the model selected by Lasso, which establishes that

$$|\widehat{T}_l \setminus T_l| = \widehat{m}_l \lesssim_{\mathrm{P}} s.$$

Under this condition, we have that the rate of post-Lasso is no worse than Lasso's rate. This occurs despite the fact that Lasso may, in general, fail to correctly select the oracle model T_l as a subset, that is, $T_l \not\subseteq \widehat{T}_l$. However, if the oracle model has well-separated coefficients and the approximation error does not dominate the estimation error, then the post-Lasso rate improves upon Lasso's rate. Specifically, this occurs if Condition AS holds, $\widehat{m}_l = o_{\mathrm{P}}(s)$ and $T_l \subseteq \widehat{T}_l$ w.p. → 1, or if $T = \widehat{T}$ w.p. → 1 as under the conditions of Wainwright (2009). In such cases, the rates found for Lasso are sharp, and they cannot be faster than $\sqrt{s\log p/n}$. Thus, the improvement in the rate of convergence of post-Lasso over Lasso is strict in these cases. Note that, as shown in the proof of Lemma 8, a higher penalty level will tend to reduce $\widehat{m}_l$ but will increase the likelihood of $T_l \not\subseteq \widehat{T}_l$. On the other hand, a lower penalty level will decrease the likelihood of $T_l \not\subseteq \widehat{T}_l$ (bias) but will tend to increase $\widehat{m}_l$ (variance). The impact in the estimation of this trade-off is captured by the last term of the bound in Lemma 7.

Step 2. In this step, we provide a sparsity bound for Lasso, which is important for establishing various rate results and fundamental to the analysis of post-Lasso. It relies on the following lemmas.

LEMMA 8—Empirical Pre-Sparsity for Lasso: *Let $\widehat{T}_l$ denote the support selected by the Lasso estimator, $\widehat{m}_l = |\widehat{T}_l \setminus T_l|$, and assume that $\lambda/n \ge c\|S_l\|_\infty$ and*

$u \geq 1 \geq \ell > 1/c$ as in Lemma 6. Then, for $c_0 = (uc+1)/(\ell c - 1)$, we have

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

PROOF: We have from the optimality conditions that the Lasso estimator $\widehat{\beta}_l = \widehat{\beta}_{l\mathrm{L}}$ satisfies

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

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

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

where we used that

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

and similarly,

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

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

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

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

LEMMA 9—Sublinearity of Maximal Sparse Eigenvalues: *Let M be a semidefinite positive matrix. For any integer $k \geq 0$ and constant $\ell \geq 1$, we have $\phi_{\max}(\lceil \ell k \rceil)(M) \leq \lceil \ell \rceil \phi_{\max}(k)(M)$.*

PROOF: Denote $\phi_M(k) = \phi_{\max}(k)(M)$, and let $\bar{\alpha}$ achieve $\phi_M(\ell k)$. Moreover let $\sum_{i=1}^{\lceil \ell \rceil} \alpha_i = \bar{\alpha}$ such that $\sum_{i=1}^{\lceil \ell \rceil} \|\alpha_i\|_0 = \|\bar{\alpha}\|_0$. We can choose α_i's such that $\|\alpha_i\|_0 \leq k$ since $\lceil \ell \rceil k \geq \ell k$. Since M is positive semidefinite, for any i, j we have $\alpha_i' M\alpha_i + \alpha_j' M\alpha_j \geq 2|\alpha_i' M\alpha_j|$. Therefore,

$$\begin{aligned}
\phi_M(\ell k) = \bar{\alpha}' M \bar{\alpha} &= \sum_{i=1}^{\lceil \ell \rceil} \alpha_i' M \alpha_i + \sum_{i=1}^{\lceil \ell \rceil} \sum_{j \neq i} \alpha_i' M \alpha_j \\
&\leq \sum_{i=1}^{\lceil \ell \rceil} \big\{\alpha_i' M \alpha_i + \big(\lceil \ell \rceil - 1\big)\alpha_i' M \alpha_i\big\} \leq \lceil \ell \rceil \sum_{i=1}^{\lceil \ell \rceil} \|\alpha_i\|^2 \phi_M\big(\|\alpha_i\|_0\big) \\
&\leq \lceil \ell \rceil \max_{i=1,\dots,\lceil \ell \rceil} \phi_M\big(\|\alpha_i\|_0\big) \leq \lceil \ell \rceil \phi_M(k),
\end{aligned}$$

where we used that $\sum_{i=1}^{\lceil \ell \rceil} \|\alpha_i\|_2^2 = 1$. *Q.E.D.*

LEMMA 10—Sparsity Bound for Lasso Under Data-Driven Penalty: *Consider the Lasso estimator $\widehat{\beta}_l = \widehat{\beta}_{l\mathrm{L}}$ with $\lambda/n \geq c\|S_l\|_\infty$, and let $\widehat{m}_l = |\widehat{T}_l \setminus T_l|$. Consider the set*

$$\mathcal{M} = \left\{m \in \mathbb{N} : m > s2\phi_{\max}(m)\big\|\big(\widehat{Y}_l^0\big)^{-1}\big\|_\infty^2 \left[\frac{2c_0}{\kappa^l_{c_0}} + \frac{6c_0 nc_s}{\lambda\sqrt{s}}\right]^2\right\}.$$

Then,

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

COMMENT D.2—Sparsity Bound: Provided that the regularization event $\lambda/n \geq c\|S_l\|_\infty$ occurs, Lemma 10 bounds the number of components $\widehat{m}_l$ incorrectly selected by Lasso. Essentially, the bound depends on s and on the ratio between the maximum sparse eigenvalues and the restricted eigenvalues. Thus, the empirical Gram matrix can impact the sparsity bound substantially. However, under Condition SE, the ratio mentioned is bounded from above uniformly in n. As expected, the bound improves and the regularization event is more likely to occur if a larger value of the penalty parameter λ is used.

PROOF OF LEMMA 10: Rewriting the conclusion in Lemma 8, we have

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

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

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

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

$$M \leq s2\phi_{\max}(M)\big\|\big(\widehat{\Upsilon}_l^0\big)^{-1}\big\|_\infty^2\left[\frac{2c_0}{\kappa_{c_0}^l}+\frac{6c_0nc_s}{\lambda\sqrt{s}}\right]^2,$$

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

In turn, applying (D.2) once more with $\widehat{m}_l \leq (M \wedge n)$, we obtain

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

The result follows by minimizing the bound over $M \in \mathcal{M}$. *Q.E.D.*

Step 3. Next we combine the previous steps to establish Theorem 2. As in Step 3 of Appendix C, recall that $\max_{1\leq l\leq k_e}\|\widehat{\Upsilon}_l^0 - \Upsilon_l^0\|_\infty \to_{\mathrm{P}} 0$.

Let $\bar{k}$ be the integer that achieves the minimum in the definition of μ^2. Since $c_s \lesssim_{\mathrm{P}} \sqrt{s/n}$ leads to $nc_s/[\lambda\sqrt{s}] \to_{\mathrm{P}} 0$, we have that $\bar{k} \in \mathcal{M}$ with high probability as $n \to \infty$. Moreover, as long as $\lambda/n \geq c\max_{1\leq l\leq k_e}\|S_l\|_\infty$, $\ell \to_{\mathrm{P}} 1$, and $c > 1$, by Lemma 10 we have, for every $l = 1, \dots, k_e$, that

$$\widehat{m}_l \lesssim_{\mathrm{P}} s\mu^2\phi_{\min}(\bar{k}+s)/\kappa_{\bar{C}}^2 \lesssim_{\mathrm{P}} s\mu^2\phi_{\min}(\widehat{m}_l+s)/\kappa_{\bar{C}}^2, \tag{D.3}$$

since $\bar{k} \in \mathcal{M}$ implies $\bar{k} \geq \widehat{m}_l$.

By the choice of $\lambda = 2c\sqrt{n}\Phi^{-1}(1-\gamma/(2pk_e))$ in (2.7), since $\gamma \to 0$, the event $\lambda/n \geq c\max_{1\leq l\leq k_e}\|S_l\|_\infty$ holds with probability approaching 1. Therefore, by the first and last results in Lemma 7, we have

$$\max_{1\leq l\leq k_e}\|D_{il} - f_i'\widehat{\beta}_{l\mathrm{PL}}\|_{2,n} \lesssim_{\mathrm{P}} \frac{\mu}{\kappa_{\bar{C}}}\sqrt{\frac{s\log p}{n}} + c_s + \max_{1\leq l\leq k_e}\frac{\lambda\sqrt{s}}{n\kappa_{c_0}^l}.$$

Because $\max_{1\leq l\leq k_e} 1/\kappa_{c_0}^l \leq \max_{1\leq l\leq k_e}\|\widehat{\Gamma}_l^0\|_\infty/\kappa_{\bar{C}} \lesssim_{\mathrm{P}} 1/\kappa_{\bar{C}}$ by Step 1 of Theorem 1, we have

$$\max_{1\leq l\leq k_e}\|D_{il} - f_i'\widehat{\beta}_{l\mathrm{PL}}\|_{2,n} \lesssim_{\mathrm{P}} \frac{\mu}{\kappa_{\bar{C}}}\sqrt{\frac{s\log(k_e p/\gamma)}{n}}, \tag{D.4}$$

since $k_e \lesssim p$ and $c_s \lesssim_{\mathrm{P}} \sqrt{s/n}$. That establishes the first inequality of Theorem 2.

To establish the second inequality of Theorem 2, since $\|\widehat{\beta}_{l\mathrm{PL}} - \beta_{l0}\|_0 \leq \widehat{m}_l + s$, we have

$$\begin{aligned}\|\widehat{\beta}_{l\mathrm{PL}} - \beta_{l0}\|_1 &\leq \sqrt{\|\widehat{\beta}_{l\mathrm{PL}} - \beta_{l0}\|_0}\|\widehat{\beta}_{l\mathrm{PL}} - \beta_{l0}\|_2 \\ &\leq \sqrt{\widehat{m}_l + s}\frac{\|f_i'(\widehat{\beta}_{l\mathrm{PL}} - \beta_{l0})\|_{2,n}}{\sqrt{\phi_{\min}(\widehat{m}_l + s)}}.\end{aligned}$$

The sparsity bound (D.3), the prediction norm bound (D.4), and the relation $\|D_{il} - f_i'\widehat{\beta}_{l\mathrm{PL}}\|_{2,n} \leq c_s + \|f_i'(\widehat{\beta}_{l\mathrm{PL}} - \beta_{l0})\|_{2,n}$ yield the result with the relation above. *Q.E.D.*

LEMMA 11—Asymptotic Validity of the Data-Driven Penalty Loadings: *Under the conditions of Theorem 1 and Condition RF or the conditions of Theorem 2 and Condition SE, the penalty loadings $\widehat{\Upsilon}$ constructed by the K-step Algorithm A.1 are asymptotically valid. In particular, for $K \geq 2$, we have $u' = 1$.*

For proof of Lemma 11, see the Supplemental Material (Belloni et al. (2012)).

APPENDIX E: PROOFS OF LEMMAS 1–4

For proof of Lemma 1, see Belloni and Chernozhukov (2011a, Supplement). For proof of Lemma 2, see Belloni and Chernozhukov (2012). For proofs of Lemmas 3 and 4, see the Supplemental Material (Belloni et al. (2012)).

APPENDIX F: PROOFS OF THEOREMS 3–7

F.1. *Proofs of Theorems 3 and 4*

The proofs are original and they rely on the consistency of the sparsity-based estimators with respect to both the $L^2(\mathbb{P}_n)$ norm $\|\cdot\|_{2,n}$ and the ℓ_1-norm

$\|\cdot\|_1$. These proofs also exploit the use of moderate deviation theory for self-normalized sums.

Step 0. Using data-driven penalty satisfying (2.7) and (3.2), we have, by Theorem 1 and Condition RE that the Lasso estimator, and by Theorem 2 and Condition SE that the post-Lasso estimator, obey

$$\max_{1\le l\le k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n} \lesssim_{\mathrm{P}} \sqrt{\frac{s\log(p\vee n)}{n}} \to 0, \tag{F.1}$$

$$\sqrt{\log p}\|\widehat{\beta}_l - \beta_{l0}\|_1 \lesssim_{\mathrm{P}} \sqrt{\frac{s^2\log^2(p\vee n)}{n}} \to 0. \tag{F.2}$$

To prove Theorem 3, we need also the condition

$$\max_{1\le l\le k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n}^2 n^{2/q_\epsilon} \lesssim_{\mathrm{P}} \frac{s\log(p\vee n)}{n} n^{2/q_\epsilon} \to 0,$$

with the last statement holding by Condition SM. Note that Theorem 4 assumes (F.1) and (F.2) as high-level conditions.

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

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

where, by Steps 2 and 3 below,

$$\mathbb{E}_n\bigl[\widehat{D}_i d_i'\bigr] = \bar{\mathrm{E}}\bigl[D_i d_i'\bigr] + o_{\mathrm{P}}(1), \tag{F.3}$$

$$\sqrt{n}\mathbb{E}_n[\widehat{D}_i\epsilon_i] = \mathbb{G}_n[D_i\epsilon_i] + o_{\mathrm{P}}(1), \tag{F.4}$$

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

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

and $Z_n = (Q^{-1}\Omega Q^{-1})^{-1/2}\sqrt{n}(\widehat{\alpha} - \alpha_0) = \mathbb{G}_n[z_{i,n}] + o_{\mathrm{P}}(1)$, where $z_{i,n} = (Q^{-1}\Omega Q^{-1})^{-1/2}Q^{-1}D_i\epsilon_i$ are i.n.i.d. with mean zero and variance I. We have that, for some small enough $\delta > 0$, $\bar{\mathrm{E}}\|z_{i,n}\|_2^{2+\delta} \lesssim \bar{\mathrm{E}}[\|D_i\|_2^{2+\delta}|\epsilon_i|^{2+\delta}] \lesssim 1$, by Condition SM. This condition verifies the Lyapunov condition, and the application of the Lyapunov CLT for i.n.i.d. triangular arrays and the Cramer–Wold device implies that $Z_n \to_d N(0, I)$.

Step 2. To show (F.3), note that

$$\begin{aligned}
\big\|\mathbb{E}_n\big[(\widehat{D}_i - D_i)d_i'\big]\big\| &\le \mathbb{E}_n\big[\|\widehat{D}_i - D_i\|_2\|d_i\|_2\big] \\
&\le \sqrt{\mathbb{E}_n\big[\|\widehat{D}_i - D_i\|_2^2\big]\mathbb{E}_n\big[\|d_i\|_2^2\big]} \\
&= \sqrt{\mathbb{E}_n\left[\sum_{l=1}^{k_e}|\widehat{D}_{il} - D_{il}|^2\right]\mathbb{E}_n\big[\|d_i\|_2^2\big]} \\
&\le \sqrt{k_e}\max_{1\le l\le k_e}\|\widehat{D}_{il} - D_{il}\|_{2,n}\sqrt{\mathbb{E}_n\big[\|d_i\|_2^2\big]} \\
&\lesssim_{\mathrm{P}} \max_{1\le l\le k_e}\|\widehat{D}_{il} - D_{il}\|_{2,n} = o_{\mathrm{P}}(1),
\end{aligned}$$

where $\sqrt{\mathbb{E}_n[\|d_i\|_2^2]} \lesssim_{\mathrm{P}} 1$ by $\bar{\mathrm{E}}\|d_i\|_2^2 \lesssim 1$ and Chebyshev inequality, and the last assertion holds by Step 0.

Moreover, $\mathbb{E}_n[D_iD_i'] - \bar{\mathrm{E}}[D_iD_i'] \to_{\mathrm{P}} 0$ by von Bahr–Essen inequality (von Bahr and Esseen (1965)), using that $\bar{\mathrm{E}}[\|D_i\|_2^q]$ for a fixed $q > 2$ is bounded uniformly in n by Condition SM.

Step 3. To show (F.4), let $a_{il} := a_l(x_i)$, note that $\mathrm{E}[f_{ij}\epsilon_i] = 0$, $\mathrm{E}[\epsilon_i|D_{il}] = 0$, and $\mathrm{E}[\epsilon_{il}|a_{il}] = 0$, and

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

Next we note that, for each $l = 1, \dots, k_e$, $|\mathbb{G}_n\{a_{il}\epsilon_i\}| \lesssim_{\mathrm{P}} [\mathbb{E}_n a_{il}^2]^{1/2} \lesssim_{\mathrm{P}} \sqrt{s/n} \to 0$, by Condition AS on $[\mathbb{E}_n a_{il}^2]^{1/2}$ and by Chebyshev inequality, since in the homoscedastic case of Theorem 3, $\mathrm{Var}[\mathbb{G}_n\{a_{il}\epsilon_i\}|x_1, \dots, x_n] \le \sigma^2\mathbb{E}_n a_{il}^2$, and in the bounded heteroscedastic case of Theorem 3, $\mathrm{Var}[\mathbb{G}_n\{a_{il}\epsilon_i\}|x_1, \dots, x_n] \lesssim$

$\mathbb{E}_n a_{il}^2$. Next we can bound $\max_{1\le j\le p} |\mathbb{G}_n[f_{ij}\epsilon_i]/\sqrt{\mathbb{E}_n[f_{ij}^2\epsilon_i^2]}| \lesssim_{\mathrm{P}} \sqrt{\log p}$ provided that p obeys the growth condition $\log p = o(n^{1/3})$, and

$$\text{(F.5)}\qquad \min_{1\le j\le p} M_{j0} := \frac{\bar{\mathrm{E}}[f_{ij}^2\epsilon_i^2]^{1/2}}{\bar{\mathrm{E}}[|f_{ij}|^3|\epsilon_i|^3]^{1/3}} \gtrsim 1.$$

This result follows by the bound on moderate deviations of a maximum of a self-normalized vector stated in Lemma 5, and by (F.5) holding by Condition SM. Finally, $\max_{1\le j\le p} \mathbb{E}_n[f_{ij}^2\epsilon_i^2] \lesssim_{\mathrm{P}} 1$, by Condition SM. Thus, combining bounds above with bounds in (F.1) and (F.2),

$$\max_{1\le l\le k_e} \bigl|\sqrt{n}\mathbb{E}_n\bigl[(\widehat{D}_{il} - D_{il})\epsilon_i\bigr]\bigr| \lesssim_{\mathrm{P}} \sqrt{\frac{s^2\log^2(p\vee n)}{n}} + \sqrt{\frac{s}{n}} \to 0,$$

where the conclusion holds by Condition SM(iii).

Step 4. This step establishes consistency of the variance estimator in the homoscedastic case of Theorem 3.

Since σ^2 and $Q = \bar{\mathrm{E}}[D_iD_i']$ are bounded away from zero and from above uniformly in n, it suffices to show $\widehat{\sigma}^2 - \sigma^2 \to_{\mathrm{P}} 0$ and $\mathbb{E}_n[\widehat{D}_i\widehat{D}_i'] - \bar{\mathrm{E}}[D_iD_i'] \to_{\mathrm{P}} 0$. Indeed, $\widehat{\sigma}^2 = \mathbb{E}_n[(\epsilon_i - d_i'(\widehat{\alpha} - \alpha_0))^2] = \mathbb{E}_n[\epsilon_i^2] + 2\mathbb{E}_n[\epsilon_i d_i'(\alpha_0 - \widehat{\alpha})] + \mathbb{E}_n[(d_i'(\alpha_0 - \widehat{\alpha}))^2]$, so that $\mathbb{E}_n[\epsilon_i^2] - \sigma^2 \to_{\mathrm{P}} 0$ by Chebyshev inequality since $\bar{\mathrm{E}}[|\epsilon_i|^4]$ is bounded uniformly in n, and the remaining terms converge to zero in probability since $\widehat{\alpha} - \alpha_0 \to_{\mathrm{P}} 0$ by Step 3, $\|\mathbb{E}_n[d_i\epsilon_i]\|_2 \lesssim_{\mathrm{P}} 1$ by Markov, and since $\bar{\mathrm{E}}\|d_i\epsilon_i\|_2 \le \sqrt{\bar{\mathrm{E}}\|d_i\|_2^2}\sqrt{\bar{\mathrm{E}}|\epsilon_i|^2}$ is uniformly bounded in n by Condition SM, and $\mathbb{E}_n\|d_i\|_2^2 \lesssim_{\mathrm{P}} 1$ by Markov, and $\bar{\mathrm{E}}\|d_i\|_2^2$ bounded uniformly in n by Condition SM. Next, note that

$$\begin{aligned}\bigl\|\mathbb{E}_n\bigl[\widehat{D}_i\widehat{D}_i'\bigr] - \mathbb{E}_n\bigl[D_iD_i'\bigr]\bigr\| = \bigl\|&\mathbb{E}_n\bigl[D_i(\widehat{D}_i - D_i)' + (\widehat{D}_i - D_i)D_i'\bigr] \\ &+ \mathbb{E}_n\bigl[(\widehat{D}_i - D_i)(\widehat{D}_i - D_i)'\bigr]\bigr\|,\end{aligned}$$

which is bounded up to a constant by

$$\sqrt{k_e}\max_{1\le l\le k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n}\bigl\|\|D_i\|_2\bigr\|_{2,n} + k_e\max_{1\le l\le k_e} \|\widehat{D}_{il} - D_{il}\|_{2,n}^2 \to_{\mathrm{P}} 0$$

by (F.1) and by $\|\|D_i\|_2\|_{2,n} \lesssim_{\mathrm{P}} 1$ holding by Markov inequality. Moreover, $\mathbb{E}_n[D_iD_i'] - \bar{\mathrm{E}}[D_iD_i'] \to_{\mathrm{P}} 0$ by Step 2.

Step 5. This step establishes consistency of the variance estimator in the boundedly heteroscedastic case of Theorem 3.

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

In what follows, we shall repeatedly use the following elementary inequality: for arbitrary nonnegative random variables $W_1, \dots, W_n$ and $q > 1$,

$$\text{(F.6)} \qquad \max_{i \le n} W_i \lesssim n^{1/q} \quad \text{if} \quad \bar{\mathrm{E}}\big[W_i^q\big] \lesssim 1,$$

which follows by Markov inequality from $\mathrm{E}[\max_{i\le n} W_i] \le n^{1/q}\mathrm{E}(\frac{1}{n}\sum_{i=1}^n W_i^q)^{1/q} \le n^{1/q}(\bar{\mathrm{E}}[W_i^q])^{1/q}$, which follows from the trivial bound $\max_{i\le n}|w_i| \le \sum_{i=1}^n |w_i|$ and Jensen's inequality.

First, we note

$$\begin{aligned} \big\|\mathbb{E}_n\big[\big(\widehat{\epsilon}_i^2 - \epsilon_i^2\big)\widehat{D}_i\widehat{D}_i'\big]\big\| &\le \big\|\mathbb{E}_n\big[\big\{d_i'(\widehat{\alpha} - \alpha_0)\big\}^2\widehat{D}_i\widehat{D}_i'\big]\big\| \\ &\quad + 2\big\|\mathbb{E}_n\big[\epsilon_i d_i'(\widehat{\alpha} - \alpha_0)\widehat{D}_i\widehat{D}_i'\big]\big\| \\ &\lesssim_{\mathrm{P}} \max_{i\le n}\|d_i\|_2^2 n^{-1}\big\|\mathbb{E}_n\big[\widehat{D}_i\widehat{D}_i'\big]\big\| \\ &\quad + \max_{i\le n}|\epsilon_i|\|d_i\|_2 n^{-1/2}\big\|\mathbb{E}_n\big[\widehat{D}_i\widehat{D}_i'\big]\big\| \\ &\to_{\mathrm{P}} 0, \end{aligned}$$

since $\|\widehat{\alpha} - \alpha_0\|_2^2 \lesssim_{\mathrm{P}} 1/n$, $\|\mathbb{E}_n\widehat{D}_i\widehat{D}_i'\| \lesssim_{\mathrm{P}} 1$ by Step 4, and $\max_{i\le n}\|d_i\|_2^2 n^{-1} \to_{\mathrm{P}} 0$ (by $\max_{i\le n}\|d_i\|_2 \lesssim_{\mathrm{P}} n^{1/q}$ for $q > 2$, holding by $\bar{\mathrm{E}}[\|d_i\|_2^q] \lesssim 1$ and inequality (F.6)) and $\max_{i\le n}[\|d_i\|_2|\epsilon_i|]n^{-1/2} \to_{\mathrm{P}} 0$ (by $\max_{i\le n}[\|d_i\|_2|\epsilon_i|] \lesssim_{\mathrm{P}} n^{1/q}$ for $q > 2$ holding by $\bar{\mathrm{E}}[(\|d_i\|_2|\epsilon_i|^2)^q] \lesssim 1$ and inequality (F.6)).

Next we note that

$$\begin{aligned} \big\|\mathbb{E}_n\big[\epsilon_i^2\widehat{D}_i\widehat{D}_i'\big] - \mathbb{E}_n\big[\epsilon_i^2 D_iD_i'\big]\big\| &= \big\|\mathbb{E}_n\big[\epsilon_i^2 D_i(\widehat{D}_i - D_i)' + \epsilon_i^2(\widehat{D}_i - D_i)D_i'\big] \\ &\quad + \mathbb{E}_n\big[\epsilon_i^2(\widehat{D}_i - D_i)(\widehat{D}_i - D_i)'\big]\big\|, \end{aligned}$$

which is bounded up to a constant by

$$\begin{aligned} &\sqrt{k_e}\max_{1\le l\le k_e}\|\widehat{D}_{il} - D_{il}\|_{2,n}\big\|\epsilon_i^2\|D_i\|_2\big\|_{2,n} \\ &\quad + k_e\max_{1\le l\le k_e}\|\widehat{D}_{il} - D_{il}\|_{2,n}^2\max_{i\le n}\epsilon_i^2 \to_{\mathrm{P}} 0. \end{aligned}$$

The latter occurs because $\|\epsilon_i^2\|D_i\|_2\|_{2,n} = \sqrt{\mathbb{E}_n[\epsilon_i^4\|D_i\|_2^2]} \lesssim_{\mathrm{P}} 1$ by $\bar{\mathrm{E}}[\epsilon_i^4\|D_i\|_2^2]$ uniformly bounded in n by Condition SM and by Markov inequality, and

$$\max_{1\le l\le k_e}\|\widehat{D}_{il} - D_{il}\|_{2,n}^2\max_{i\le n}\epsilon_i^2 \lesssim_{\mathrm{P}} \frac{s\log(p \vee n)}{n}n^{2/q_\epsilon} \to 0,$$

where the latter step holds by Step 0 and by $\max_{i\le n}\epsilon_i^2 \lesssim_{\mathrm{P}} n^{2/q_\epsilon}$ holding by $\bar{\mathrm{E}}[\epsilon_i^{q_\epsilon}] \lesssim 1$ and inequality (F.6). Finally, $\mathbb{E}_n[\epsilon_i^2 D_iD_i'] - \bar{\mathrm{E}}[\epsilon_i^2 D_iD_i'] \to_{\mathrm{P}} 0$

by the von Bahr–Essen inequality (von Bahr and Esseen (1965)) and by $\bar{\mathrm{E}}[|\epsilon_i|^{2+\mu}\|D_i\|_2^{2+\mu}]$ bounded uniformly in n for small enough $\mu > 0$ by Condition SM.

We conclude that $\mathbb{E}_n[\widehat{\epsilon}_i^2\widehat{D}_i\widehat{D}_i'] - \bar{\mathrm{E}}[\epsilon_i^2 D_i D_i'] \to_{\mathrm{P}} 0$. *Q.E.D.*

F.2. *Proof of Theorem 5*

Step 1. To establish claim (i), using the properties of projection, we note that

$$n\mathbb{E}_n[\tilde{\epsilon}_i\tilde{f}_{ij}] = n\mathbb{E}_n[\epsilon_i\tilde{f}_{ij}]. \tag{F.7}$$

Since for $\widehat{\mu}_\epsilon = (\mathbb{E}_n[w_i w_i'])^{-1}\mathbb{E}_n[w_i\epsilon_i]$, we have $\|\widehat{\mu}_\epsilon\|_2 \le \|\mathbb{E}_n[w_i w_i']^{-1}\|\|\mathbb{E}_n[w_i\epsilon_i]\|_2$, where $\|\mathbb{E}_n[w_i w_i']^{-1}\|$ is bounded by Condition SM2(ii) and $\|\mathbb{E}_n[w_i\epsilon_i]\|_2$ is of stochastic order $\sqrt{k_w/n}$ by Chebyshev inequality and Condition SM2(ii). Hence $\|\widehat{\mu}_\epsilon\|_2 \lesssim_{\mathrm{P}} \sqrt{k_w/n}$. Since $\|w_i\|_2 \le \zeta_w$ by Condition SM2(i), we conclude that $\max_{i\le n}|w_i'\widehat{\mu}_\epsilon| \lesssim_{\mathrm{P}} \zeta_w\sqrt{k_w}/\sqrt{n} \to 0$. Hence, uniformly in $j \in \{1, \dots, p\}$,

$$\begin{aligned}\Big|\sqrt{\mathbb{E}_n\big[\tilde{\epsilon}_i^2\tilde{f}_{ij}^2\big]} - \sqrt{\mathbb{E}_n\big[\epsilon_i^2\tilde{f}_{ij}^2\big]}\Big| &\overset{\text{(a)}}{\le} \sqrt{\mathbb{E}_n\big[\big(w_i'\widehat{\mu}_\epsilon\big)^2\tilde{f}_{ij}^2\big]} \\ &\overset{\text{(b)}}{=} o_{\mathrm{P}}(1)\sqrt{\mathbb{E}_n\big[\tilde{f}_{ij}^2\big]} \overset{\text{(c)}}{=} o_{\mathrm{P}}(1),\end{aligned} \tag{F.8}$$

where (a) is by the triangular inequality and the decomposition $\tilde{\epsilon}_i = \epsilon_i - w_i'\widehat{\mu}_\epsilon$, (b) is by the Holder inequality, and (c) is by the normalization $\sqrt{\mathbb{E}_n[\tilde{f}_{ij}^2]} = 1$ for each j. Hence, for $c > 1$, by (F.7) and (F.8) w.p. → 1,

$$\Lambda_{\alpha_1} \le c\bar{\Lambda}_{\alpha_1}, \quad \bar{\Lambda}_{\alpha_1} := \max_{1\le j\le p} n\big|\mathbb{E}_n[\epsilon_i\tilde{f}_{ij}]\big|/\sqrt{\mathbb{E}_n\big[\epsilon_i^2\tilde{f}_{ij}^2\big]}.$$

Since $\bar{\Lambda}_{\alpha_1}$ is a maximum of self-normalized sum of i.n.i.d. terms conditional on X, application of Condition SM2(iii)–(iv) and the moderate deviation bound from Lemma 5 for the self-normalized sum with $U_{ij} = \epsilon_i\tilde{f}_{ij}$, conditional on X, implies that $\mathrm{P}(c\bar{\Lambda}_{\alpha_1} \le \Lambda(1-\gamma)) \ge 1 - \gamma - o(1)$. This verifies claim (i).

Step 2. To show claim (ii) we note that, using triangular and other elementary inequalities,

$$\begin{aligned}\Lambda_a &= \max_{1\le j\le p}\left|\frac{n\mathbb{E}_n[(\tilde{\epsilon}_i - (a-\alpha_1)'\tilde{d}_{ei})\tilde{f}_{ij}]}{\sqrt{\mathbb{E}_n[(\tilde{\epsilon}_i - (a-\alpha_1)'\tilde{d}_{ei})^2\tilde{f}_{ij}^2]}}\right| \\ &\ge \max_{1\le j\le p}\left|\frac{n|\mathbb{E}_n[(a-\alpha_1)'\tilde{d}_{ei}\tilde{f}_{ij}]|}{\sqrt{\mathbb{E}_n[\tilde{\epsilon}_i^2\tilde{f}_{ij}^2]} + \sqrt{\mathbb{E}_n[\{(a-\alpha_1)'\tilde{d}_{ei}\}^2\tilde{f}_{ij}^2]}}\right| - \Lambda_{\alpha_1}.\end{aligned}$$

The first term on the right side is bounded below by, w.p. → 1,

$$\max_{1\leq j\leq p} \frac{n|\mathbb{E}_n[(a-\alpha_1)'\tilde{d}_{ei}\tilde{f}_{ij}]|}{c\sqrt{\mathbb{E}_n[\epsilon_i^2\tilde{f}_{ij}^2]}+\sqrt{\mathbb{E}_n[\{(a-\alpha_1)'\tilde{d}_{ei}\}^2\tilde{f}_{ij}^2]}},$$

by Step 1 for some $c > 1$, and $\Lambda_{\alpha_1} \lesssim_{\mathrm{P}} \sqrt{n\log(p/\gamma)}$ also by Step 1. Hence for any constant C, by the last condition in the statement of the theorem, with probability converging to 1, $\Lambda_a - C\sqrt{n\log(p/\gamma)} \to +\infty$, so that claim (ii) immediately follows, since $\Lambda(1-\gamma) \lesssim \sqrt{n\log(p/\gamma)}$. *Q.E.D.*

F.3. *Proof of Theorems 6 and 7*

See the Supplemental Material (Belloni et al. (2012)).

REFERENCES

AMEMIYA, T. (1966): "On the Use of Principal Components of Independent Variables in Two-Stage Least-Squares Estimation," *International Economic Review*, 7, 283–303. [2373]

——— (1974): "The Non-Linear Two-Stage Least Squares Estimator," *Journal of Econometrics*, 2, 105–110. [2370,2376,2377]

ANDERSON, T. W., AND H. RUBIN (1949): "Estimation of the Parameters of Single Equation in a Complete System of Stochastic Equations," *The Annals of Mathematical Statistics*, 20, 46–63. [2372,2377,2392]

ANDREWS, D. W. K., AND J. H. STOCK (2005): "Inference With Weak Instruments," Discussion Paper 1530, Cowles Foundation. [2371]

ANDREWS, D. W. K., M. J. MOREIRA, AND J. H. STOCK (2006): "Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression," *Econometrica*, 74 (3), 715–752. [2371]

ANGRIST, J. D., AND A. B. KRUEGER (1995): "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business & Economic Statistics*, 13 (2), 225–235. [2372, 2396]

BAI, J., AND S. NG (2008): "Forecasting Economic Time Series Using Targeted Predictors," *Journal of Econometrics*, 146, 304–317. [2370,2374]

——— (2009a): "Boosting Diffusion Indices," *Journal of Applied Econometrics*, 24, 607–629. [2370]

——— (2009b): "Selecting Instrumental Variables in a Data Rich Environment," *Journal of Time Series Econometrics*, 1 (1), Article 4. [2373,2374]

——— (2010): "Instrumental Variable Estimation in a Data Rich Environment," *Econometric Theory*, 26, 1577–1606. [2373]

BEKKER, P. A. (1994): "Alternative Approximations to the Distributions of Instrumental Variables Estimators," *Econometrica*, 63, 657–681. [2370,2371,2377,2398]

BELLONI, A., AND V. CHERNOZHUKOV (2011a): "ℓ_1-Penalized Quantile Regression for High Dimensional Sparse Models," *The Annals of Statistics*, 39 (1), 82–130. [2414,2420]

——— (2011b): "High-Dimensional Sparse Econometric Models, an Introduction," in *Inverse Problems and High-Dimensional Estimation*. Lecture Notes in Statistics. Berlin: Springer, 121–156. [2374]

——— (2012): "Least Squares After Model Selection in High-Dimensional Sparse Models," *Bernoulli* (forthcoming). Available at arXiv:1001.0188. [2370,2372,2381,2385,2387,2388,2392, 2414,2420]

BELLONI, A., D. CHEN, V. CHERNOZHUKOV, AND C. HANSEN (2012): "Supplement to 'Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain'," *Econometrica Supplemental Material*, 80, http://www.econometricsociety.org/ecta/Supmat/9626_Proofs.pdf; http://www.econometricsociety.org/ecta/Supmat/9626_data_and_programs-1.zip; http://www.econometricsociety.org/ecta/Supmat/9626_data_and_programs-2.zip. [2395,2398,2420,2426]

BELLONI, A., V. CHERNOZHUKOV, AND C. HANSEN (2011a): "Estimation and Inference Methods for High-Dimensional Sparse Econometric Models," in *Advances in Economics and Econometrics, 10th World Congress of Econometric Society*. Cambridge: Cambridge University Press. [2371,2374,2378,2382]

——— (2011b): "Inference on Treatment Effects After Selection Amongst High-Dimensional Controls With an Application to Abortion on Crime," available at arXiv: 1201.0224. [2369,2371,2374,2382,2383]

BELLONI, A., V. CHERNOZHUKOV, AND L. WANG (2011a): "Pivotal Estimation of Nonparametric Functions via Square-Root Lasso," available at arXiv:1105.1475. [2392]

——— (2011b): "Square-Root-LASSO: Pivotal Recovery of Sparse Signals via Conic Programming," *Biometrika*, 98, 791–806. [2372,2392]

BICKEL, P. J., Y. RITOV, AND A. B. TSYBAKOV (2009): "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37 (4), 1705–1732. [2370-2372,2380,2381,2384,2385, 2387,2388,2410]

BRODIE, J., I. DAUBECHIES, C. D. MOL, D. GIANNONE, AND I. LORIS (2009): "Sparse and Stable Markowitz Portfolios," *PNAS*, 106 (30), 12267–12272. [2374]

BÜHLMANN, P. (2006): "Boosting for High-Dimensional Linear Models," *The Annals of Statistics*, 34 (2), 559–583. [2373,2392]

BÜHLMANN, P., AND S. VAN DE GEER (2011): *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin: Springer. [2370,2371]

BUNEA, F., A. B. TSYBAKOV, AND M. H. WEGKAMP (2006): "Aggregation and Sparsity via ℓ_1 Penalized Least Squares," in *Proceedings of 19th Annual Conference on Learning Theory (COLT 2006)*, ed. by G. Lugosi and H. U. Simon. Berlin: Springer, 379–391. [2370]

——— (2007a): "Sparsity Oracle Inequalities for the Lasso," *Electronic Journal of Statistics*, 1, 169–194. [2370]

——— (2007b): "Aggregation for Gaussian Regression," *The Annals of Statistics*, 35 (4), 1674–1697. [2370]

CANDES, E., AND T. TAO (2007): "The Dantzig Selector: Statistical Estimation When p Is Much Larger Than n," *The Annals of Statistics*, 35 (6), 2313–2351. [2370,2392]

CANER, M. (2009): "LASSO-Type GMM Estimator," *Econometric Theory*, 25, 270–290. [2374]

CARRASCO, M. (2012): "A Regularization Approach to the Many Instruments Problem," *Journal of Econometrics*, 170 (2), 383–398. [2374,2375]

CARRASCO, M., AND G. TCHUENTE NGUEMBU (2012): "Regularized LIML With Many Instruments," Discussion Paper, University of Montreal. [2374]

CHAMBERLAIN, G. (1987): "Asymptotic Efficiency in Estimation With Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334. [2370,2376]

CHAMBERLAIN, G., AND G. IMBENS (2004): "Random Effects Estimators With Many Instrumental Variables," *Econometrica*, 72, 295–306. [2374]

CHAO, J., AND N. SWANSON (2005): "Consistent Estimation With a Large Number of Weak Instruments," *Econometrica*, 73, 1673–1692. [2370,2372]

CHAO, J., N. SWANSON, J. HAUSMAN, W. NEWEY, AND T. WOUTERSEN (2012): "Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression With Many Instruments," *Econometric Theory*, 28 (1), 42–86. [2370-2372]

CHEN, D. L., AND J. SETHI (2010): "Does Forbidding Sexual Harassment Exacerbate Gender Inequality," Unpublished Manuscript. [2404]

CHEN, D. L., AND S. YEH (2010): "The Economic Impacts of Eminent Domain," Unpublished Manuscript. [2373,2403-2407]

CHERNOZHUKOV, V., AND C. HANSEN (2008a): "Instrumental Variable Quantile Regression: A Robust Inference Approach," *Journal of Econometrics*, 142, 379–398. [2394]

——— (2008b): "The Reduced Form: A Simple Approach to Inference With Weak Instruments," *Economics Letters*, 100, 68–71. [2394]

DE LA PEÑA, V. H., T. L. LAI, AND Q.-M. SHAO (2009): *Self-Normalized Processes: Limit Theory and Statistical Applications*. Probability and Its Applications (New York). Berlin: Springer. [2408,2409]

DEMIGUEL, V., L. GARLAPPI, F. NOGALES, AND R. UPPAL (2009): "A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms," *Management Science*, 55 (5), 798–812. [2374]

DONALD, S. G., AND W. K. NEWEY (2001): "Choosing the Number of Instruments," *Econometrica*, 69 (5), 1161–1191. [2374,2375]

FULLER, W. A. (1977): "Some Properties of a Modification of the Limited Information Estimator," *Econometrica*, 45, 939–953. [2372,2398,2405]

GAUTIER, E., AND A. B. TSYBAKOV (2011): "High-Dimensional Instrumental Variables Regression and Confidence Sets," available at arXiv:1105.2454. [2375]

HAHN, J. (2002): "Optimal Inference With Many Instruments," *Econometric Theory*, 18, 140–168. [2371,2377,2378]

HAHN, J., J. A. HAUSMAN, AND G. M. KUERSTEINER (2004): "Estimation With Weak Instruments: Accuracy of Higher-Order Bias and MSE Approximations," *Econometrics Journal*, 7 (1), 272–306. [2398]

HANSEN, C., J. HAUSMAN, AND W. K. NEWEY (2008): "Estimation With Many Instrumental Variables," *Journal of Business & Economic Statistics*, 26, 398–422. [2370,2372,2373,2398,2399]

HUANG, J., J. L. HOROWITZ, AND F. WEI (2010): "Variable Selection in Nonparametric Additive Models," *The Annals of Statistics*, 38 (4), 2282–2313. [2370,2374,2392]

JING, B.-Y., Q.-M. SHAO, AND Q. WANG (2003): "Self-Normalized Cramér-Type Large Deviations for Independent Random Variables," *The Annals of Probability*, 31 (4), 2167–2215. [2372,2379,2408]

KAPETANIOS, G., AND M. MARCELLINO (2010): "Factor-GMM Estimation With Large Sets of Possibly Weak Instruments," *Computational Statistics & Data Analysis*, 54 (11), 2655–2675. [2373]

KAPETANIOS, G., L. KHALAF, AND M. MARCELLINO (2011): "Factor Based Identification-Robust Inference in IV Regressions," Working Paper. [2373]

KLEIBERGEN, F. (2002): "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70, 1781–1803. [2371]

——— (2005): "Testing Parameters in GMM Without Assuming That They Are Identified," *Econometrica*, 73, 1103–1123. [2371]

KLOEK, T., AND L. MENNES (1960): "Simultaneous Equations Estimation Based on Principal Components of Predetermined Variables," *Econometrica*, 28, 45–61. [2373]

KNIGHT, K. (2008): "Shrinkage Estimation for Nearly Singular Designs," *Econometric Theory*, 24, 323–337. [2370,2374]

KOLTCHINSKII, V. (2009): "Sparsity in Penalized Empirical Risk Minimization," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 45 (1), 7–57. [2370]

LOUNICI, K. (2008): "Sup-Norm Convergence Rate and Sign Concentration Property of Lasso and Dantzig Estimators," *Electronical Journal of Statistic*, 2, 90–102. [2370]

LOUNICI, K., M. PONTIL, A. B. TSYBAKOV, AND S. VAN DE GEER (2010): "Taking Advantage of Sparsity in Multi-Task Learning," available at arXiv:0903.1468v1. [2370,2392]

MEINSHAUSEN, N., AND B. YU (2009): "Lasso-Type Recovery of Sparse Representations for High-Dimensional Data," *The Annals of Statistics*, 37 (1), 246–270. [2370]

MOREIRA, M. J. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71, 1027–1048. [2371]

NEWEY, W. K. (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models," *Econometrica*, 58, 809–837. [2370,2371,2376-2378,2383]

——— (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168. [2378,2382,2386]

OKUI, R. (2011): "Instrumental Variable Estimation in the Presence of Many Moment Conditions," *Journal of Econometrics*, 165, 70–86. [2374,2375]

ROSENBAUM, M., AND A. B. TSYBAKOV (2008): "Sparse Recovery Under Matrix Uncertainty," available at arXiv:0812.2818. [2370]

RUDELSON, M., AND R. VERSHYNIN (2008): "On Sparse Reconstruction From Fourier and Gaussian Measurements," *Communications on Pure and Applied Mathematics*, 61, 1025–1045. [2385]

RUDELSON, M., AND S. ZHOU (2011): "Reconstruction From Anisotropic Random Measurements," available at arXiv:1106.1151. [2385]

STAIGER, D., AND J. H. STOCK (1997): "Instrumental Variables Regression With Weak Instruments," *Econometrica*, 65, 557–586. [2371,2372,2377,2392]

STOCK, J. H., AND M. YOGO (2005): "Testing for Weak Instruments in Linear IV Regression," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. by D. W. K. Andrews and J. H. Stock. Cambridge: Cambridge University Press, Chap. 5, 80–108. [2408]

STOCK, J. H., J. H. WRIGHT, AND M. YOGO (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business & Economic Statistics*, 20 (4), 518–529. [2398]

TIBSHIRANI, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288. [2370,2372,2380,2381]

VAN DE GEER, S. A. (2008): "High-Dimensional Generalized Linear Models and the Lasso," *The Annals of Statistics*, 36 (2), 614–645. [2370]

VON BAHR, B., AND C.-G. ESSEEN (1965): "Inequalities for the rth Absolute Moment of a Sum of Random Variables, $1 \leq r \leq 2$," *The Annals of Mathematical Statistics*, 36, 299–303. [2422,2425]

WAINWRIGHT, M. (2009): "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. [2370,2416]

ZHANG, C.-H., AND J. HUANG (2008): "The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression," *The Annals of Statistics*, 36 (4), 1567–1594. [2370,2385]

Duke University Fuqua School of Business, Durham, NC 27708, U.S.A.; abn5@duke.edu,

D-GESS, ETH Zurich, IFW E 48.3, Haldeneggsteig 4, CH-8092, Zurich, Switzerland; chendan@ethz.ch,

Dept. of Economics, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.; vchern@mit.edu,

and

University of Chicago Booth School of Business, Chicago, IL 60637, U.S.A.; chansen1@chicagobooth.edu.

Manuscript received October, 2010; final revision received June, 2012.