ARTICLE IN PRESS

Journal of Econometrics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Inference for high-dimensional instrumental variables regression☆

David Gold [a], Johannes Lederer [a], Jing Tao [b,*]

[a] Department of Statistics, University of Washington, Seattle, WA, 98195, USA
[b] Department of Economics, University of Washington, Seattle, WA, 98195, USA

ARTICLE INFO

Article history:
Received 31 July 2018
Received in revised form 19 June 2019
Accepted 29 September 2019
Available online xxxx

JEL classification:
C14
C31
C36

Keywords:
High-dimensional inference
Instrumental variables
De-biasing

ABSTRACT

This paper concerns statistical inference for the components of a high-dimensional regression parameter despite possible endogeneity of each regressor. Given a first-stage linear model for the endogenous regressors and a second-stage linear model for the dependent variable, we develop a novel adaptation of the parametric one-step update to a generic second-stage estimator. We provide conditions under which the scaled update is asymptotically normal. We then introduce a two-stage Lasso procedure and show that the second-stage Lasso estimator satisfies the aforementioned conditions. Using these results, we construct asymptotically valid confidence intervals for the components of the second-stage regression coefficients. We complement our asymptotic theory with simulation studies, which demonstrate the performance of our method in finite samples.

1. Introduction

1.1. Overview

High-dimensional estimation has been extensively studied and is now ubiquitous in the data-intensive sciences (Bühlmann and van de Geer, 2011; Giraud, 2014; Hastie et al., 2015). High-dimensional inference, on the other hand, is much less developed. In particular, although considerable progress has been made for inference in standard high-dimensional regression (Javanmard and Montanari, 2014; van de Geer et al., 2014; Zhang and Zhang, 2014; Ning and Liu, 2017), much less is known for more complex models.

In this paper, we extend the study of high-dimensional inference to the *linear instrumental variables (IV) model*. To motivate the linear IV model, we consider the ordinary linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u},$$

where $\boldsymbol{y}$ is the vector of responses, $\boldsymbol{X}$ is the matrix of regressors, $\boldsymbol{\beta}$ is the regression vector, and $\boldsymbol{u}$ is a vector of random disturbances. Standard inference for $\boldsymbol{\beta}$ using ordinary least squares is valid only if $\mathrm{E}[\boldsymbol{u}|\boldsymbol{X}] = 0$. However, this assumption

☆ We are grateful to the editors and anonymous referees for their careful review and valuable suggestions. We sincerely thank Alex Belloni, Mehmet Caner, Denis Chetverikov, Yanqin Fan, Amit Gandhi, Eric Gautier, Mohamed Hebiri, Joseph Salmon, Alexandre Tsybakov, and Jon Wellner for their insightful comments and Donghui Mai for her help in setting up the cloud computing. We acknowledge support through Cloud Credits for Research by Amazon and through the University of Washington Royalty Research Fund.

* Corresponding author at: Department of Economics, University of Washington, Seattle, WA, 98195, USA.
E-mail addresses: ledererj@uw.edu (J. Lederer), jingtao@uw.edu (J. Tao).

https://doi.org/10.1016/j.jeconom.2019.09.009
0304-4076/

is easily violated in practice in view of selection biases, omitted variables, measurement errors, and many other challenges common to data collection. Hence, it is often more reasonable to allow for $\mathrm{E}[\boldsymbol{u}|\boldsymbol{X}] \neq 0$ and instead assume that $\boldsymbol{X}$ can be modeled based on observable variables $\boldsymbol{Z}$ that satisfy $\mathrm{E}[\boldsymbol{u}|\boldsymbol{Z}] = 0$. As standard in the econometric literature, we then call the regressors in $\boldsymbol{X}$ *endogenous*, because they can be correlated with $\boldsymbol{u}$, and we call the instrumental variables $\boldsymbol{Z}$ *exogenous*, because $\mathrm{E}[\boldsymbol{u}|\boldsymbol{Z}] = 0$.

Inference for such models in low-dimensional settings, where the number of samples is much larger than both the number of regressors in $\boldsymbol{X}$ and the number of regressors in $\boldsymbol{Z}$, has been extensively studied and put to use in economic applications and beyond (Angrist and Pischke, 2009). In the era of Big Data, however, low-dimensional settings often do not apply, because one wants to allow for flexible parameter combinations, or because many variables are measured in the first place. We are thus interested in inference for double high-dimensional settings with the number of samples dominated by both the number of regressors in $\boldsymbol{X}$ and the number of regressors in $\boldsymbol{Z}$.

The method we propose is based on a novel adaptation of the parametric one-step update to a generic two-stage estimation procedure. In parametric models, the one-step update $\tilde{\boldsymbol{\beta}}$ to an initial estimator $\hat{\boldsymbol{\beta}}$ is one Newton–Raphson step in the direction of a solution to the empirical analogue of the score equations. This approach is similar to those of van de Geer et al. (2014) and Javanmard and Montanari (2014), who de-bias the Lasso, and Zhang and Zhang (2014), who use a low-dimensional projection technique, to obtain asymptotic pivots for the low-dimensional components of the high-dimensional linear regression models when endogeneity is absent. The present work extends this approach to two-stage estimation when both stages are high-dimensional and the regressors of interest may each be endogenous. To adapt the one-step update to handle endogeneity of $\boldsymbol{X}$, we (i) choose the update as a step towards the solution to the empirical analogue of a valid moment condition and (ii) apply the update to a generic second-stage estimator $\hat{\boldsymbol{\beta}}$ that depends on the predicted conditional means $\hat{\mathrm{E}}[\boldsymbol{X}|\boldsymbol{Z}]$. The resultant estimator decomposes into a main term and four remainder terms, which contrasts with the single remainder term in the case of the de-biased Lasso in the ordinary linear model.

We present high-level conditions under which the updated estimator yields asymptotic pivots for the components of $\boldsymbol{\beta}$, and we show, as an example, how these conditions may be satisfied by a two-stage Lasso estimation routine. We assume a sub-Gaussian regime throughout for the noise elements and instrumental variables in order to support flexibility of distributional assumptions. The main challenges of establishing the example are due to (i) the involved structure of the remainder terms, whose control requires a variety of concentration bounds and lead to extensive proofs and (ii) the estimation of the population precision matrix of the conditional means $\mathrm{E}[\boldsymbol{X}|\boldsymbol{Z}]$, since these are not observed directly.

1.2. Our contributions

Our primary contribution is to develop a method with which to conduct statistical inference for the components $\beta_j, j = 1, \ldots, p_{\boldsymbol{x}}$ of a high-dimensional regression vector $\boldsymbol{\beta}$ despite endogeneity of the respective regressors. We develop a novel adaptation of the one-step update and high-level conditions under which the updated estimator yields asymptotically Gaussian pivots for each β_j. Our rigorous demonstration of conditions under which such inference is possible in the doubly high-dimensional setting differentiates the present paper from similar works such as Belloni et al. (2011), who develop inferential methods for IV models with high-dimensional instruments and low-dimensional endogenous regressors, and (Zhu, 2018), who work under a doubly high-dimensional regime but focus primarily on bounding the error of a two-stage Lasso estimator such as that of Section 4.

A related contribution concerns sparse inverse covariance matrix estimation. The updated estimator $\tilde{\boldsymbol{\beta}}$ depends on an estimate of the inverse covariance matrix $\boldsymbol{\Theta}$ of the conditional means $\mathrm{E}[\boldsymbol{x}_i|\boldsymbol{z}_i]$. However, we do not observe these conditional means directly, and must base our estimate of $\boldsymbol{\Theta}$ on the predictions $\hat{\mathrm{E}}[\boldsymbol{x}_i|\boldsymbol{z}_i]$. For this, we use a modification of the CLIME estimator $\widehat{\boldsymbol{\Theta}}$ of Cai et al. (2011). Our paper is the first one to use such an estimator in the context of instrumental variable selection, and we account for the prediction step in deriving probabilistic guarantees for the estimator's performance.

Another contribution is to show that the updated second-stage Lasso estimator studied in Section 4 satisfies the high-level conditions in Section 3 and therefore supports inference for the $\tilde{\beta}_j$. To show as much, we develop probabilistic bounds for the second-stage ℓ_1 estimation error, and we use these bounds to show asymptotic negligibility of the four remainder terms described in the previous section. We also demonstrate the feasibility of the compatibility condition in the second-stage regression, thereby justifying the practical use of the second-stage rates.

A majority of the proofs factor into deterministic and stochastic components. This also allows future analysts easily to combine the generic bounds contained in Section B of the Appendix with concentration results for specific error and design matrix distribution regimes and thereby derive the growth conditions required for good asymptotic behavior of the updated second-stage estimator under a variety of models.

1.3. Related work

Our work is related to the recent research on inference for high-dimensional linear instrumental variables models such as Belloni et al. (2012, 2018), Fan and Liao (2014), Gautier and Tsybakov (2014), Cheng and Liao (2015) and Neykov et al. (2015). Belloni et al. (2012) use the Lasso to obtain representations of the optimal IVs of Amemiya (1974, 1977) and Hansen (1982) for models in which the conditional mean of the response is linear in a small and fixed number

Please cite this article as: D. Gold, J. Lederer and J. Tao, Inference for high-dimensional instrumental variables regression. Journal of Econometrics (2020), https://doi.org/10.1016/j.jeconom.2019.09.009.

of endogenous variables and show that the second-stage estimator is $\sqrt{n}$-consistent. Following the line of the seminal work by Belloni et al. (2012), some recent papers propose different novel procedures to select many instruments when the number of second-stage regressors remains to be fixed or low-dimensional (for instance, Hansen and Kozbur, 2014; Caner and Fan, 2015; Fan and Zhong, 2018, among others). Cheng and Liao (2015) propose a Lasso procedure to select valid and relevant moments for the GMM estimation when the number of moments increases with sample size. However, and in contrast with the present paper, the dimensions of moments and parameters of interest are both smaller than the sample size. Compared to Belloni et al. (2012) and Cheng and Liao (2015), we allow both the number of IVs and the number of endogenous regressors to be bigger than the sample size. Gautier and Tsybakov (2014, 2018) construct robust confidence sets based on their Self-Tuned Instrumental Variables (STIV) estimator and confidence bands after bias correction (requiring a type of strong instruments condition).

Much of the present work is devoted to solving the inference problem for parameters of interest in a high-dimensional linear IV model with homoscedasticity by accounting for the prediction error when the first- and second-stage regression models are both high-dimensional. This contrasts with the methods of Fan and Liao (2014), who do not account for the need to predict the optimal instruments (Amemiya, 1974, 1977; Newey, 1990; Imbens et al., 2003). To our knowledge, such analysis under an ℓ_1-regularized estimation procedure is new in the literature.

Recent work by Lin et al. (2015) and Zhu (2018) also propose estimation methods for linear IV models when the regressors of both stages are high-dimensional but do not rigorously develop asymptotic methods for inference. We notice that Neykov et al. (2015) also provide an inferential method for high-dimensional linear IV models by using a Dantzig selector. However, their method requires that the number of instruments equals the number of endogenous regressors and they do not account for the first-stage of estimation. Another related work is concurrently developed by Belloni et al. (2018), who propose regularized estimation of nuisance parameters that appear in carefully constructed empirical orthogonality conditions. Our one-step update approach can be interpreted as an iteration of the Newton–Raphson method, while Belloni et al. (2018) build on the idea of Neyman orthogonality. Even though the two papers take different approaches, both show that $\sqrt{n}$-consistent estimators for low-dimensional parameters can be constructed in high-dimensional IV models.

In a very recent work Caner and Kock (2018b) introduce a de-sparsified, ℓ_1-penalized two-stage GMM estimator. They develop estimation error bounds and inferential procedures based on this estimator in the doubly high-dimensional setting and extend our work in that they (i) allow conditional heteroskedasticity in the second-stage noise elements, (ii) specify random components in terms of moment conditions rather than sub-Gaussianity, and (iii) do not require ℓ_0-sparsity of the first-stage regression vectors. Also noteworthy is that the authors do not predict the conditional means ZA for the weighting scheme of the second-stage. This leads to expressions for the scale factor of the asymptotic pivots in their Theorem 2.(i) that differ those of our Theorem 3.4. The scale factors of the latter such pivots is identical to the asymptotic variance of optimal estimators in low-dimensional IV models with homoscedastic structural errors (Chamberlain, 1987; Newey, 1990).[1]

1.4. Organization

The rest of the paper is organized as follows. We introduce our model and a generic two-stage estimation procedure in Section 2. In Section 3, we propose the one-step update inference procedure and demonstrate conditions under which the update yields asymptotically Gaussian pivots. In Section 4, we introduce a two-stage Lasso estimator of the parameter $\boldsymbol{\beta}$ and show that it is suitable for use with the inference procedure developed in Section 3. Finally, in Section 5, we present the results of numerical studies that demonstrate the relevance of our theoretical results to finite samples. All proofs are contained in Appendix.

1.5. Basic notation and preliminaries

We adopt the following general notational conventions. For $p \in \mathbb{N}$, we let $[p] := \{1, \dots, p\}$. We typically use bold and non-bold lowercase letters to denote vectors and scalars, respectively. We use bold uppercase letters to denote matrices. We typically denote the components of a vector (matrix) by the non-bold (lowercase) counterpart of the letter that denotes the vector (matrix). If $\boldsymbol{M} \in \mathbb{R}^{n\times p}$, with components m_{ij}, we use a superscript to refer to columns $\boldsymbol{m}^j = (m_{1j}, \dots, m_{nj})^\top$ and a subscript to refer to rows $\boldsymbol{m}_i = (m_{i1}, \dots, m_{ip})$. We let $\|\cdot\|_q$ and $\langle\cdot,\cdot\rangle$ denote the usual ℓ_q norm and inner product over Euclidean spaces, respectively.

For $\boldsymbol{m} \in \mathbb{R}^p$, we let $\operatorname{supp}(\boldsymbol{m}) := \{j \in [p] : m_j \neq 0\}$, $\|\boldsymbol{m}\|_0 = |\operatorname{supp}(\boldsymbol{m})|$, and $\|\boldsymbol{m}\|_\infty = \max_{j\in[p]}\{m_j\}$. For matrices $\boldsymbol{M} \in \mathbb{R}^{n\times p}$, we let $\|\boldsymbol{M}\|_\infty = \max_{i,j\in[n]\times[p]} |m_{ij}|$, $\|\boldsymbol{M}\|_{L_1} = \max_{j\in[p]} \|\boldsymbol{m}^j\|_1$, and $\|\boldsymbol{M}\|_{L_2} = \max_{j\in[p]} \|\boldsymbol{m}^j\|_2$. For matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2 \in \mathbb{R}^{n\times p}$, we write $\boldsymbol{M}_1 \succ \boldsymbol{M}_2$ if $\boldsymbol{M}_1 - \boldsymbol{M}_2$ is positive-definite.

For quantities x indexed by $i \in [n]$, we let $\mathbb{E}_n[x_i] = n^{-1}\sum_{i=1}^n x_i$. If X_n is a sequence of random variables, we write $X_n \rightsquigarrow X$ if X_n converges weakly to X. For $a, b \in \mathbb{R}$, we let $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. We write $a_n \lesssim b_n$ if $a_n \leq C_n b_n$ for a C_n that is of constant order. We say that a sequence of events $\mathcal{E} \equiv \mathcal{E}_n$ occurs with probability approaching one if $\lim_{n\to\infty} \mathrm{P}\,\mathcal{E}_n = 1$.

[1] In Caner and Kock (2018a), the same authors develop asymptotic theory to support both estimation and inference for the conservative Lasso under a wide variety (including heteroscedasticity and non-sub-Gaussianity) of error regimes, thereby laying foundations for future work considering two-stage estimation.

Please cite this article as: D. Gold, J. Lederer and J. Tao, Inference for high-dimensional instrumental variables regression. Journal of Econometrics (2020), https://doi.org/10.1016/j.jeconom.2019.09.009.

We recall the following definitions for sub-Gaussian and sub-exponential norms.

Definition 1.1 (*Sub-Gaussian and Sub-exponential Norms*)**.** For $q \geq 1$ and a random variable X, we write

$$\|X\|_{\psi_q} := \inf\{t \in (0,\infty) : \mathrm{E}[\exp(|X|^q/t^q) - 1] \leq 1\}.$$

if the infimum exists. The *sub-Gaussian norm* of a random variable X is given by $\|X\|_{\psi_2}$; the *sub-exponential norm* of a random variable X is given by $\|X\|_{\psi_1}$. The corresponding norms for a random p-vector $\boldsymbol{X}$ are given by

$$\|\boldsymbol{X}\|_{\psi_q} := \sup_{\boldsymbol{x}\in\mathbb{R}^p : \|\boldsymbol{x}\|_2=1} \|\langle \boldsymbol{X}, \boldsymbol{x}\rangle\|_{\psi_q}.$$

2. Two-stage estimation

To contend with endogeneity, the method of instrumental variables isolates variation in the endogenous regressors induced by the instrumental variables. In Section 2.1, we posit the two-stage linear IV model to describe this relationship. In Section 2.2, we discuss a generic two-stage estimation routine that respects the structure of the model.

2.1. Model

Our model of interest is

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + u_i, \tag{1}$$

$$x_{ij} = \boldsymbol{z}_i^\top \boldsymbol{\alpha}^j + v_{ij}, \tag{2}$$

where: i ranges from 1 to n (unless stated otherwise); j ranges from 1 to $p_{\boldsymbol{x}}$ (unless stated otherwise); the vectors $\boldsymbol{x}_i \in \mathbb{R}^{p_{\boldsymbol{x}}}$ consist of the *second-stage regressors* $x_{i1}, \ldots, x_{ip_{\boldsymbol{x}}}$; the vector $\boldsymbol{\beta} \in \mathbb{R}^{p_{\boldsymbol{x}}}$ is the parameter of interest; the vectors $\boldsymbol{z}_i \in \mathbb{R}^{p_{\boldsymbol{z}}}$ consist of the *first-stage regressors* $z_{i1}, \ldots, z_{ip_{\boldsymbol{z}}}$; the quantities u_i and $\boldsymbol{v}_i := (v_{i1}, \ldots, v_{ip_{\boldsymbol{x}}})^\top$ are random noise elements that satisfy

$$\mathrm{E}[u_i|\boldsymbol{z}_i] = 0, \qquad \mathrm{E}[\boldsymbol{v}_i|\boldsymbol{z}_i] = \boldsymbol{0}, \tag{3}$$

and the vectors $\boldsymbol{\alpha}^j$ are regression parameters up to which the respective conditional means $d_{ij} := \mathrm{E}[x_{ij}|\boldsymbol{z}_i] = \boldsymbol{z}_i^\top \boldsymbol{\alpha}^j$ are specified. We call the models of (2) and (1) the first-stage and second-stage models, respectively. Note that the setup in (3) is similar to the one in Belloni et al. (2012) except for this significant difference: we consider both stages to be high-dimensional while the second stage parameters in Belloni et al. (2012) are low-dimensional. To focus on presenting the main idea and inference steps, we ignore approximation errors in the model but one can add approximation errors in both (1) and (2) at the expense of more tedious derivations and notation (see the discussion of Theorem 6.3 in Bühlmann and van de Geer, 2011).

In matrix notation, we write

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

and

$$\boldsymbol{X} = \boldsymbol{D} + \boldsymbol{V} = \boldsymbol{Z}\mathbf{A} + \boldsymbol{V},$$

where the vectors $\boldsymbol{y}, \boldsymbol{u} \in \mathbb{R}^n$ consist of the responses y_i and the noise components u_i, respectively; the matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p_{\boldsymbol{x}}}$ has columns $\boldsymbol{x}^j = (x_{1j}, \ldots, x_{nj})^\top$ and rows $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip_{\boldsymbol{x}}})$; the matrix $\boldsymbol{D} = \mathrm{E}[\boldsymbol{X}|\boldsymbol{Z}] \in \mathbb{R}^{n\times p_{\boldsymbol{x}}}$ has columns $\boldsymbol{d}^j = (d_{1j}, \ldots, d_{nj})^\top$ and rows $\boldsymbol{d}_i = (d_{i1}, \ldots, d_{ip_{\boldsymbol{x}}})$; the matrix $\boldsymbol{Z} \in \mathbb{R}^{n\times p_{\boldsymbol{z}}}$ has columns $\boldsymbol{z}^k = (z_{1k}, \ldots, z_{nk})^\top$ and rows $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip_{\boldsymbol{z}}})$; and the matrix $\mathbf{A} \in \mathbb{R}^{p_{\boldsymbol{z}}\times p_{\boldsymbol{x}}}$ has columns given by $\boldsymbol{\alpha}^j$. We make the following assumption concerning the n-indexed sequence of regression parameters $\mathbf{A}$, $\boldsymbol{\beta}$.

Assumption 2.1 (*Regularity of* $\mathbf{A}$, $\boldsymbol{\beta}$)**.** The quantities $\|\mathbf{A}\|_{L_1}$ and $\|\boldsymbol{\beta}\|_1$ are bounded above by universal constants $m_{\mathbf{A}}, m_{\boldsymbol{\beta}} < \infty$, respectively.

We let $\widehat{\Sigma}_{\boldsymbol{z}} = \boldsymbol{Z}^\top\boldsymbol{Z}/n$ denote the empirical Gram matrix of the instrumental variables. As remarked earlier, the linear IV model has been studied extensively in the low-dimensional setting, where the number $p_{\boldsymbol{x}}$ of endogenous variables $\boldsymbol{x}^j$ is fixed. We are particularly concerned with the high-dimensional regime in which both $p_{\boldsymbol{x}}$ and the number $p_{\boldsymbol{z}}$ of instrumental variables $\boldsymbol{z}^k$ increase with n. Our results generalize to the low-dimensional case in which $p_{\boldsymbol{z}}$ and $p_{\boldsymbol{x}}$ are held fixed with respect to n, but we do not treat this case explicitly in the present essay. Regardless of whether the model is high-dimensional, we require that $p_{\boldsymbol{x}} \leq p_{\boldsymbol{z}}$ in order to maintain identifiability of $\mathrm{E}[\boldsymbol{x}_i|\boldsymbol{z}_i]$.

We study a sub-Gaussian regime for the noise components as well as the instrumental variables, which we treat as random throughout and for which we give marginal results. This regime encompasses the typical Gaussian model considered in the high-dimensional literature and allows for flexibility in modeling assumptions.

Please cite this article as: D. Gold, J. Lederer and J. Tao, Inference for high-dimensional instrumental variables regression. Journal of Econometrics (2020), https://doi.org/10.1016/j.jeconom.2019.09.009.

Assumption 2.2 (*Specification of $\boldsymbol{z}_i$*)**.** The instrumental variables $\boldsymbol{z}_i$ are i.i.d. and sub-Gaussian with sub-Gaussian norm $\tau_{\boldsymbol{z}} := \|\boldsymbol{z}_i\|_{\psi_2}$ and satisfy $\mathrm{E}[\boldsymbol{z}_i] = 0$ for each $i \in [n]$. Considered as components of an n-indexed sequence of models, the quantities $\tau_{\boldsymbol{z}}$ are bounded away from zero and infinity.

Remark 2.3 (*Specification of $\boldsymbol{z}_i$*)**.** Assumption Assumption 2.2 is not restrictive. We require that the first-stage regressors $\boldsymbol{z}_i$ have mean zero in order to simplify the following exposition and to apply concentration results under more specific distributional assumptions, such as in Lemma 4.10. This assumption can be relaxed at the expense of brevity and given a sufficient reformulation of the required concentration results. Similarly, the condition that $\tau_{\boldsymbol{z}} = O(1)$ can be relaxed at the expense of introducing more complex growth conditions in later results.

Assumption 2.4 (*Specification of $\boldsymbol{v}^j$ and $\boldsymbol{u}$*)**.** The noise vectors $\boldsymbol{v}^j$ and $\boldsymbol{u}$ are sub-Gaussian with sub-Gaussian norms $\tau_{\boldsymbol{v}^j} := \|\boldsymbol{v}^j\|_{\psi_2}$ and $\tau_{\boldsymbol{u}} := \|\boldsymbol{u}\|_{\psi_2}$. Considered as components of an n-indexed sequence of models, the quantities $\tau_{\boldsymbol{v}^j}$ and $\tau_{\boldsymbol{u}}$ are bounded strictly away from zero and infinity.

Note that Assumption 2.4 makes no stipulations concerning the joint covariance structure of the u_i and $\boldsymbol{v}_i$. The assumption therefore allows for nontrivial covariance between the two stages of noise, which can be used to model endogeneity of the $\boldsymbol{x}_i$. Furthermore, Assumption 2.4 allows for heteroscedasticity amongst the components of the $\boldsymbol{v}^j$. We require homoscedasticity of the second-stage noise elements u_i for Theorem 3.4 and Lemma 4.11; all other results hold in the presence of heteroscedasticity.

2.2. Generic two-stage estimators

We formulate our proposed method of inference for the components β_j of the second-stage regression parameter $\boldsymbol{\beta}$ in terms of generic estimators that reflect the structure of the model described above. We now introduce notation that will be used in Section 3.

For each $j \in [p_{\boldsymbol{x}}]$, let $\hat{\boldsymbol{\alpha}}^j \equiv \hat{\boldsymbol{\alpha}}^j(\boldsymbol{x}^j, \boldsymbol{Z})$ denote a generic *first-stage estimator* of the first-stage regression vector $\boldsymbol{\alpha}^j$ based on the data $\boldsymbol{x}^j$ and $\boldsymbol{Z}$. We write $\widehat{\mathbf{A}} := (\hat{\boldsymbol{\alpha}}^1, \ldots, \hat{\boldsymbol{\alpha}}^{p_{\boldsymbol{x}}})$ for the matrix of estimated regression vectors. From such an estimator $\widehat{\mathbf{A}}$ we may predict the conditional means $\boldsymbol{d}_i = \mathrm{E}[\boldsymbol{x}_i|\boldsymbol{z}_i]$ for $i \in [n]$ with $\hat{\boldsymbol{d}}_i := \boldsymbol{z}_i^\top \widehat{\mathbf{A}}$; we write $\widehat{\boldsymbol{D}}$ for the predicted conditional mean matrix whose rows are given by the $\hat{\boldsymbol{d}}_i$, and we write $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} := \widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}/n$ and $\boldsymbol{\Sigma}_{\boldsymbol{d}} := \mathrm{E}[\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}]$. Our choice of the notation $\widehat{\boldsymbol{D}}$ reflects the fact that this quantity predicts and, under suitable conditions, approaches in probability the conditional mean matrix $\boldsymbol{D}$. We write $\hat{\boldsymbol{\beta}} \equiv \hat{\boldsymbol{\beta}}(\boldsymbol{y}, \widehat{\boldsymbol{D}})$ for a generic *second-stage estimator* of the second-stage regression parameter $\boldsymbol{\beta}$ based on the response $\boldsymbol{y}$ and the predicted conditional means $\widehat{\boldsymbol{D}}$.

3. Main proposal

Our main contribution is to develop a method for statistical inference for the components β_j of the second-stage regression vector $\boldsymbol{\beta}$. In general, statistical inference for high-dimensional regression parameters is a difficult problem. Regularized estimators, such as the Lasso and ridge regression, are often used for the purpose of high-dimensional parameter estimation but generally do not have asymptotic distributions suitable for inference (Knight and Fu, 2000; Pötscher and Leeb, 2009). In studying the model of Section 2.1, we must also account for the dependence of the second-stage estimator on the first-stage estimators.

The basis for our procedure is to adapt the parametric one-step update to the two-stage estimation procedure described in Section 2.2. We first briefly review the use of the one-step estimator in parametric models and its application to high-dimensional inference for the ordinary linear model. Then we adapt the one-step update to the two-stage estimation procedure described in Section 2.2. Section 3.3 discusses high-level conditions under which the scaled updated estimator is asymptotically normal.

3.1. One-step with endogeneity

In this section, we develop an adaptation of the one-step update that, under suitable high-level conditions, yields asymptotic pivots for the second-stage components β_j of the two-stage model described in Section 2.1. We note that the present development is valid for any initial second-stage estimator $\hat{\boldsymbol{\beta}}$. To demonstrate that the high-level conditions are satisfied requires consideration of particular estimators.

In summary, the one-step update is a general method for constructing efficient estimators for parametric and semiparametric models. Recall that the Newton–Raphson method for finding the root in $\boldsymbol{b}$ to a *target system* of $p_{\boldsymbol{x}}$ equations

$$\boldsymbol{h}(y_i, \boldsymbol{x}_i; \boldsymbol{b}) \equiv (h_1(y_i, \boldsymbol{x}_i; \boldsymbol{b}), \ldots, h_{p_{\boldsymbol{x}}}(y_i, \boldsymbol{x}_i; \boldsymbol{b}))^\top = \mathbf{0}$$

is to update an approximation $\boldsymbol{b}^k$ by the rule

$$\boldsymbol{b}^{k+1} = \boldsymbol{b}^k - \left[\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{b}}\bigg|_{\boldsymbol{b}=\boldsymbol{b}^k} \right]^{-1} \boldsymbol{h}(y_i, \boldsymbol{x}_i; \boldsymbol{b}^k),$$

where $\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{b}}\big|_{\boldsymbol{b}=\boldsymbol{b}^k}$ is the Jacobian matrix of $\boldsymbol{h}$ with respect to $\boldsymbol{b}$ evaluated at $\boldsymbol{b}^k$. In the ordinary least squares regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}\,, \tag{4}$$

the score function $\boldsymbol{h}(y_i, \boldsymbol{x}_i; \boldsymbol{b}) = -\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^\top \boldsymbol{b})$ satisfies $\mathrm{E}[\boldsymbol{h}(y_i, \boldsymbol{x}_i; \boldsymbol{\beta})] = \boldsymbol{0}$ given the *orthogonality condition*

$$\mathrm{E}[\boldsymbol{x}_i u_i] = \boldsymbol{0}\,. \tag{5}$$

The *one-step update* $\tilde{\boldsymbol{\beta}}$ to an initial estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is given by

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n\,,$$

where $\widehat{\boldsymbol{\Theta}}$ denotes the inverse of

$$\frac{\partial(-\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})/n)}{\partial \boldsymbol{b}}\bigg|_{\boldsymbol{b}=\hat{\boldsymbol{\beta}}} = \boldsymbol{X}^\top\boldsymbol{X}/n = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\,.$$

The case when the model is high-dimensional is less well studied. When $p_{\boldsymbol{x}} > n$, the empirical covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ is not invertible, and we have instead

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \widehat{\boldsymbol{\Theta}}\boldsymbol{X}^\top\boldsymbol{u}/n + \underbrace{(\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} - \boldsymbol{I})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}_{\boldsymbol{f}/\sqrt{n}}\,,$$

where $\widehat{\boldsymbol{\Theta}}$ denotes an approximate inverse of the Jacobian matrix. The latter term $\boldsymbol{f}/\sqrt{n}$ in the above display is the "remainder" after incomplete inversion of $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$. Thus, in the high-dimensional one-stage linear model, the one-step update satisfies

$$\sqrt{n}(\tilde{\beta}_j - \beta_j) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{\boldsymbol{\theta}}_j^\top \boldsymbol{x}_i u_i + f_j\,, \tag{6}$$

where $\hat{\boldsymbol{\theta}}_j$ is the jth row of $\widehat{\boldsymbol{\Theta}}$. The structure of the main term on the right-hand side above suggests to use $\sqrt{n}(\tilde{\beta}_j - \beta_j)/\hat{\omega}_j$, where $\hat{\omega}_j$ is an appropriate estimate of $\omega_j = (\mathrm{E}[\langle\hat{\boldsymbol{\theta}}_j, \boldsymbol{x}_i\rangle^2 u_i^2])^{1/2}$, as an asymptotic pivot for β_j.

When the initial estimator $\hat{\boldsymbol{\beta}}$ is the Lasso, the updated estimator $\tilde{\boldsymbol{\beta}}$ is sometimes called the desparsified (van de Geer et al., 2014) or de-biased (Javanmard and Montanari, 2014) Lasso, though these authors obtain the form of $\tilde{\boldsymbol{\beta}}$ by means other than the one-step update. The general upshot of their results is that if $\|\boldsymbol{f}\|_\infty = o_{\mathrm{P}}(1)$, and if $\hat{\boldsymbol{\theta}}_j$ and $\boldsymbol{x}_i$ are independent of u_i, then the updated Lasso estimator yields asymptotically Gaussian pivots for the parameter components.

In contrast to the ordinary linear regression model, the challenge we face in the case of high-dimensional IV model is that the condition in (5) does not hold. Instead, the conditional moment restriction $\mathrm{E}[u_i|\boldsymbol{z}_i] = 0$ in (3) entails the orthogonality condition $\mathrm{E}[\boldsymbol{d}_i u_i] = \boldsymbol{0}$ for the conditional means $\boldsymbol{d}_i = \mathrm{E}[\boldsymbol{x}_i|\boldsymbol{z}_i]$. This suggests that, to develop a one-step update for a generic second-stage estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ of the present model, we ought to take the empirical analogue

$$\mathbb{E}_n[\tilde{\boldsymbol{h}}(y_i, \boldsymbol{x}_i, \hat{\boldsymbol{d}}_i; \boldsymbol{b})] := \mathbb{E}_n[-\hat{\boldsymbol{d}}_i(y_i - \boldsymbol{x}_i^\top\boldsymbol{b})] = -\widehat{\boldsymbol{D}}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})/n = \boldsymbol{0}\,,$$

of $\mathrm{E}[\boldsymbol{d}_i u_i] = \boldsymbol{0}$ as the target system for which the root is sought via a Newton–Raphson update. We have elected to base the target system on the moment condition $\mathrm{E}[\boldsymbol{d}_i u_i] = \boldsymbol{0}$ in accordance with optimal weighting regimes for generalized method of moments (GMM) estimators; see Amemiya (1974, 1977), Hansen (1982) and Newey (1990). Further, since the $\boldsymbol{d}_i$ are generally unavailable, we instead use the predicted conditional mean matrix $\widehat{\boldsymbol{D}}$ in the target system above. The one-step update $\tilde{\boldsymbol{\beta}}$ to a second-stage estimator $\hat{\boldsymbol{\beta}}$ is then given by

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\Theta}}\mathbb{E}_n[\tilde{\boldsymbol{h}}(y_i, \boldsymbol{x}_i, \hat{\boldsymbol{d}}_i; \boldsymbol{b})] = \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n\,, \tag{7}$$

where we continue to let $\widehat{\boldsymbol{\Theta}}$ denote an (approximate) inverse to the Jacobian matrix in $\boldsymbol{b}$ of the score $\tilde{\boldsymbol{h}}(y_i, \boldsymbol{x}_i, \hat{\boldsymbol{d}}_i; \boldsymbol{b})$.

If one were to follow strictly the prescription of the Newton–Raphson method for selection of $\widehat{\boldsymbol{\Theta}}$ for the updated second-stage estimator $\tilde{\boldsymbol{\beta}}$, one would select $\widehat{\boldsymbol{\Theta}} \approx [\widehat{\boldsymbol{D}}^\top\boldsymbol{X}/n]^{-1}$ to approximate the inverse of the Jacobian of $\tilde{\boldsymbol{h}}$ evaluated at $\hat{\boldsymbol{\beta}}$. However, the decomposition obtained in the following lemma suggests that $\widehat{\boldsymbol{\Theta}}$ ought to control, say, the sup-norm of $\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}$, and hence aim to invert $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} := \widehat{\boldsymbol{D}}^\top\widehat{\boldsymbol{D}}/n$ rather than $\widehat{\boldsymbol{D}}^\top\boldsymbol{X}/n$. We emphasize that the one-step formulation, insofar as it follows the Newton–Raphson method, is merely a vehicle for producing an updated estimator $\tilde{\boldsymbol{\beta}}$. In particular, Lemma 3.1 is valid regardless of what convergence properties an actual Newton–Raphson algorithm incorporating a specific choice of $\widehat{\boldsymbol{\Theta}}$ may exhibit. We may choose $\widehat{\boldsymbol{\Theta}}$ in whatever manner is most appropriate for achieving our goal, which is to obtain a tractable limiting distribution for $\sqrt{n}(\tilde{\beta}_j - \beta_j)/\omega_j$, where ω_j is an appropriate scale factor. That said, the two suggestions for how to choose $\widehat{\boldsymbol{\Theta}}$ may be reconciled somewhat by noting that both $\widehat{\boldsymbol{D}}^\top\boldsymbol{Z}/n$ and $\widehat{\boldsymbol{D}}^\top\widehat{\boldsymbol{D}}/n$ are equal to the empirical Gram matrix $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}$ modulo additional terms whose sup-norms can be controlled given a rate for $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}$ and appropriate concentration results for $\|\boldsymbol{Z}^\top\boldsymbol{v}^j/n\|_\infty$. In turn, one finds $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty = o_{\mathrm{P}}(1)$ under appropriate growth restrictions on $p_{\boldsymbol{x}}$; see Lemma C.5.

Please cite this article as: D. Gold, J. Lederer and J. Tao, Inference for high-dimensional instrumental variables regression. Journal of Econometrics (2020), https://doi.org/10.1016/j.jeconom.2019.09.009.

For our purposes, we consider the matrix $\widehat{\Theta}$ primarily as an estimator of the population quantity $\Theta := \mathrm{E}[\boldsymbol{d}_i \boldsymbol{d}_i^\top]^{-1}$. In particular, we require good behavior of $\widehat{\Theta}$ as such an estimator to derive the asymptotic distribution of $\sqrt{n}(\tilde{\beta}_j - \beta_j)$.

The following lemma characterizes a similar decomposition of the updated estimator $\tilde{\boldsymbol{\beta}}$ as in the one-stage model.

Lemma 3.1 (*Decomposition of One-step Second-stage Estimator*)**.** *Consider the two-stage linear model described in Section 2.1. Let $\widehat{\boldsymbol{D}}$ be a prediction of the conditional mean matrix $\boldsymbol{D}$ from an estimate $\widehat{\mathbf{A}}$ of the first-stage regression matrix $\mathbf{A}$. Let $\hat{\boldsymbol{\beta}}$ be a second-stage estimator based on the predictions $\widehat{\boldsymbol{D}}$. Let $\widehat{\Theta}$ denote an estimator of $\Theta = \mathrm{E}[\boldsymbol{d}_i \boldsymbol{d}_i^\top]^{-1}$. The one-step second-stage estimator*

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \widehat{\Theta}\widehat{\boldsymbol{D}}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n$$

satisfies $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \Theta \boldsymbol{D}^\top \boldsymbol{u}/\sqrt{n} + \sum_{\ell=1}^{4} \boldsymbol{f}_\ell$, where

$$\boldsymbol{f}_1 = (\widehat{\Theta} - \Theta)\boldsymbol{D}^\top \boldsymbol{u}/\sqrt{n}\,, \qquad \boldsymbol{f}_2 = \widehat{\Theta}(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top \boldsymbol{u}/\sqrt{n}\,,$$
$$\boldsymbol{f}_3 = \widehat{\Theta}\widehat{\boldsymbol{D}}^\top(\boldsymbol{X} - \widehat{\boldsymbol{D}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})/\sqrt{n}\,, \qquad \boldsymbol{f}_4 = \sqrt{n}(\widehat{\Theta}\widehat{\Sigma}_{\boldsymbol{d}} - \boldsymbol{I})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\,.$$

As in the case of the main term for the ordinary one-step update discussed in (6), this observation is similar to that of Neykov et al. (2015), who derive a similar asymptotic linearization but do not account for prediction of the conditional means $\boldsymbol{d}_i$. Indeed, due to the two stages of estimation, the update incurs four remainder terms as opposed to the single term in Javanmard and Montanari (2014) and van de Geer et al. (2014). In Section 3.3, we show that the quantity $\sqrt{n}(\tilde{\beta}_j - \beta_j)/\omega_j$, where $\omega_j^2 := \mathrm{E}[\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle^2 u_i^2]$, converges weakly to a $\mathcal{N}(0, 1)$ random variable under high level conditions on the remainder terms $f_{\ell j}$ and that the limit continues to hold if ω_j is replaced by an appropriate estimator. From this result one may construct asymptotically valid confidence intervals for the regression components β_j.

We have described a strategy for inference for the components $\tilde{\beta}_j$. To implement the strategy for a specific choice of first- and second-stage estimators, one must identify the conditions under which the remainder terms $\boldsymbol{f}_\ell$ vanish in probability. We demonstrate such an implementation in Section 4. The conditions in turn depend on the properties of the estimator $\widehat{\Theta}$. In the following section, we introduce an estimator suitable for our purposes.

3.2. Estimating Θ

The one-step second-stage estimator $\tilde{\boldsymbol{\beta}}$ depends on an estimator $\widehat{\Theta}$ of $\Theta = \Sigma_{\boldsymbol{d}}^{-1}$, the population precision matrix of the conditional means $\boldsymbol{d}_i$. In general, estimating the population precision matrix incurs two main difficulties in the high-dimensional setting. First, the empirical covariance matrix $\widehat{\Sigma}_{\boldsymbol{d}}$ is singular when $p_{\boldsymbol{x}} > n$ and cannot be inverted to produce an estimator of Θ. Second, even if an inverse were available, one cannot naïvely use the continuous mapping theorem to derive asymptotic guarantees if $p_{\boldsymbol{x}} \to \infty$, since the sequence of population covariance matrices $\Sigma_{\boldsymbol{d}} \equiv \Sigma_{\boldsymbol{d},n}$ does not itself have a limit if $p_{\boldsymbol{x}} \to \infty$. In addition to these general difficulties, we must further contend with the fact that the conditional mean matrix $\boldsymbol{D}$ is unknown. Hence any estimator of $\widehat{\Theta}$ will depend on the prediction $\widehat{\boldsymbol{D}}$, and guarantees for such an estimator must account for such dependence.

We use a slight modification of the CLIME estimator of Cai et al. (2011) to contend with the challenges described above. The rows $\hat{\boldsymbol{\theta}}_j$ of the estimator $\widehat{\Theta}$ are obtained as solutions to the CLIME program codified below.

Program 3.2 (*Program for $\hat{\boldsymbol{\theta}}_j$*)**.**

$$\underset{\boldsymbol{\theta} \in \mathbb{R}^{p_{\boldsymbol{x}}}}{\text{minimize:}} \quad Q(\boldsymbol{\theta}) := \|\boldsymbol{\theta}\|_1\,, \qquad \text{subject to:} \quad \|\widehat{\Sigma}_{\boldsymbol{d}}\boldsymbol{\theta} - \boldsymbol{e}_j\|_\infty \le \mu\,,$$

where $\boldsymbol{e}_j$ denotes the jth canonical basis vector in $p_{\boldsymbol{x}}$ dimensions and $\mu > 0$ is a controlled tolerance.

The present estimator $\widehat{\Theta}$ differs in only one respect from that of the CLIME estimator of Cai et al. (2011). The latter authors symmetrize the matrix Θ with rows obtained as solutions to the aforementioned optimization problem, whereas we use the raw solutions. We omit the symmetrization step for simplicity; the ℓ_∞ and ℓ_1 guarantee that Cai et al. (2011) obtain for the estimation error of the CLIME estimator continue to hold. We include the requisite guarantees for the unsymmetrized estimator in Section A.2 of the Supplementary Materials.

The present estimator $\widehat{\Theta}$ also differs in an important respect from that of Javanmard and Montanari (2014). The latter also obtain an inverse Gram matrix approximation as a solution to a convex program with identical constraints as in Problem 3.2 but with objective function $Q(\boldsymbol{\theta}) = \mathbb{E}_n[\langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle^2]$. To our knowledge, however, it is currently unknown whether the choice of Q in Javanmard and Montanari (2014) yields guarantees comparable to those of the CLIME estimator.

The ℓ_1 bound for $\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j$, which we require for control of the remainder terms $\boldsymbol{f}_\ell$, depends on the following restriction on the class of population precision matrices Θ.

Definition 3.3 (*Uniformity Class*)**.** Following Cai et al. (2011), we define the *uniformity class* of population precision matrices $\Theta = \Sigma_{\boldsymbol{d}}^{-1}$ relative to the controlled tolerance $q \in [0, 1)$ and the generalized sparsity level $s_\Theta > 0$ by

$$\mathcal{U}(m_\Theta, q, s_\Theta) := \Big\{ \Theta = (\theta_{jk})_{j,k=1}^{p_{\boldsymbol{x}}} \succ \mathbf{0} : \|\Theta\|_{L_1} \le m_\Theta; \max_{j \in [p_{\boldsymbol{x}}]} \sum_{k \in [p_{\boldsymbol{x}}]} |\theta_{jk}|^q \le s_\Theta \Big\}\,.$$

In the sequel, we assume as part of high-level regularity conditions that $\boldsymbol{\Theta} \in \mathcal{U}(m_{\boldsymbol{\Theta}}, q, s_{\boldsymbol{\Theta}})$ and that the model parameters $m_{\boldsymbol{\Theta}}$ and $s_{\boldsymbol{\Theta}}$ are well-behaved as functions of n. These parameters appear in the rates for the remainder terms in our analysis of the two-stage Lasso of Section 4. For high-level results, we also assume that the probability that the rows $\boldsymbol{\theta}_j$ of the population precision matrix are feasible for Problem 3.2 approaches one. To express this requirement formally, we define the event

$$\mathcal{T}_{\boldsymbol{\Theta}}(\mu) := \left\{ \|\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty \leq \mu \right\} \tag{8}$$

where $\mu > 0$ is the tolerance of Problem 3.2, and require that $\mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu) \to 1$ as $n \to \infty$. We identify a theoretical choice of μ that satisfies the latter requirement in Lemma 4.10; we discuss a practical method for selecting μ in Section 5. Note that, given the event $\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$, the rows $\boldsymbol{\theta}_j$ of $\boldsymbol{\Theta}_j$ are each feasible for the respective Problem 3.2.

Since the quantity μ appears in the rates for the remainder terms $\boldsymbol{f}_\ell$, it must be chosen carefully so as to balance the growth of $\mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$ with the decay of the $\|\boldsymbol{f}_\ell\|_\infty$. The appropriate choice of μ depends on both the distribution of the $\boldsymbol{z}_i$ as well as the rate for $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}$ — see Lemma 4.10.

3.3. Asymptotic normality

We saw in Section 3.1 that the updated estimator $\tilde{\boldsymbol{\beta}}$ satisfies $\sqrt{n}(\tilde{\beta}_j - \beta_j) = \sqrt{n}\mathbb{E}_n[\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle u_i] + \sum_{\ell=1}^4 f_{\ell,j}$. If the remainder terms vanish in probability, then $\sqrt{n}(\tilde{\beta}_j - \beta_j)$ shares the same weak limit, if it exists, as $\sqrt{n}\mathbb{E}_n[\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle u_i]$. If the model were fixed in n, the Central Limit Theorem would entail that the latter quantity converges weakly to a $\mathcal{N}(0, \omega_j^2)$, where $\omega_j^2 = \mathrm{E}[\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle^2 u_i^2]$. In Theorem 3.4, we provide conditions under which $\sqrt{n}(\tilde{\beta}_j - \beta_j)/\omega_j$ converges weakly to a standard Normal random variable when the model is not fixed in n. We also show that the limit continues to hold if ω_j is replaced by an estimator $\hat{\omega}_j$ that satisfies $|\hat{\omega}_j - \omega_j| = o_{\mathrm{P}}(1)$. Note that Theorem 3.4 gives conditions under which the limit holds given homoscedastic Gaussian noise (Condition (5)) as well as conditions under which the limit holds given generic i.i.d. noise (Condition (6)). Condition (5) is unnecessarily restrictive in practice. Nonetheless, we include the result under Condition (5) because it serves as a benchmark and requires subsequently weaker assumptions concerning limiting rates. The latter result includes the case of Assumption 2.4, as well as any other i.i.d. second-stage noise regime.

Theorem 3.4 (*Weak Limits*)**.** *Suppose that*

(1) the quantity $\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty$ *vanishes in probability, where* $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} := \boldsymbol{D}^\top \boldsymbol{D}/n$*;*
(2) the remainder terms satisfy $\|\boldsymbol{f}_\ell\|_\infty = o_{\mathrm{P}}(1)$ *for each* $1 \leq \ell \leq 4$*;*
(3) $\boldsymbol{\Theta}_{jj} > \vartheta$ *for some universal constant* $\vartheta > 0$ *and each* $j \in [p_{\boldsymbol{x}}]$*;*
(4) $\max_{j \in [p_{\boldsymbol{x}}]} \|\boldsymbol{\theta}_j\|_1 \leq m_{\boldsymbol{\Theta}}$ *for some universal constant* $m_{\boldsymbol{\Theta}} < \infty$*.*

If either

(5) the noise elements u_i *satisfy* $u_i \mid \boldsymbol{z}_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma_u)$*, where* σ_u *is bounded away from zero and infinity uniformly in n, or*
(6) the $\boldsymbol{z}_i$ *and* u_i *are i.i.d. with* $\mathrm{E}[u_i^2 | \boldsymbol{z}_i] = \sigma_u^2$*, where* σ_u *is bounded away from zero and infinity uniformly in n, and there exist* $0 < \zeta < 1/2$ *and* $\nu > 0$ *such that*

(a) $\mathrm{P}\left\{ |\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle| > n^\zeta \right\} = o(1)$ *and*
(b) $\mathrm{E}[|u_i|^{2+\nu}] \lesssim \sigma_u^{2+\nu}$*,*

then

$$\sqrt{n}(\tilde{\beta}_j - \beta_j)/\omega_j \rightsquigarrow Z_j \sim \mathcal{N}(0, 1)\,.$$

Furthermore, the limit continues to hold if ω_j *is replaced by an estimator* $\hat{\omega}_j$ *that satisfies* $|\hat{\omega}_j - \omega_j| = o_{\mathrm{P}}(1)$*.*

The proof of Theorem 3.4 can be found in Section A.3 of the Supplementary Materials. We note that, unlike the setting of Bickel et al. (1998, Sections 2.5), we do not require $\sqrt{n}$-consistency of the initial estimator $\hat{\boldsymbol{\beta}}$ in order for the updates $\tilde{\beta}_j$ to be asymptotically Gaussian. Indeed, the rates of convergence required are dictated by the strategies used to bound the quantities $\|\boldsymbol{f}_\ell\|_\infty$. See Section 4.3 for an example of sufficient rates for the two-stage Lasso routine studied in Section 4.

The main application of Theorem 3.4 is the construction of asymptotically valid confidence intervals under a wide variety of noise regimes. Given $j \in [p_{\boldsymbol{x}}]$ and a confidence level τ, an asymptotic $100(1-\tau)\%$ confidence interval $\hat{\mathcal{I}}_{\tau,j}$ is given by

$$\hat{\mathcal{I}}_{\tau,j} := \left[\tilde{\beta}_j - z_\tau \hat{\omega}_j,\ \tilde{\beta}_j + z_\tau \hat{\omega}_j \right], \tag{9}$$

where $z_\tau = \Phi^{-1}(1-\tau/2)$ and $\hat{\omega}_j$ satisfies the conditions of Theorem 3.4. We present a simulation study of the finite-sample properties of this procedure for the updated two-stage Lasso estimator in Section 5.

Similarly as for the de-biasing approach in linear regression (van de Geer et al., 2014, Corollary 2.1), the results in Theorem 3.4 hold uniformly over a class of parameters as long as the assumptions also hold uniformly over that class.

Please cite this article as: D. Gold, J. Lederer and J. Tao, Inference for high-dimensional instrumental variables regression. Journal of Econometrics (2020), https://doi.org/10.1016/j.jeconom.2019.09.009.

This requires in particular that the remainder terms are $o_P(1)$ uniformly. In the Lasso case considered in Section 4, such uniformity is known for parameter sets with bounded sparsity — we refer again to van de Geer et al. (2014) for details.

Note of the scale factor ω_j that

$$\omega_j^2 = \mathrm{E}[\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle^2 u_i^2] = \sigma_u^2 \boldsymbol{\theta}_j^\top \mathrm{E}[\boldsymbol{d}_i \boldsymbol{d}_i^\top] \boldsymbol{\theta}_j = \sigma_u^2 \boldsymbol{\theta}_j^\top \boldsymbol{\Sigma}_{\boldsymbol{d}} \boldsymbol{\theta}_j = \sigma_u^2 \Theta_{jj} ,$$

which, if the second-stage model were fixed in n, would equal the asymptotic variance of the optimal linear IV estimator (Amemiya, 1974; Chamberlain, 1987; Newey, 1990). Indeed, Belloni et al. (2012) show that, in such a case, the linear IV estimator still attains the semi-parametric efficiency bound if the first-stage mean model is unknown but well-approximated with a high-dimensional linear model. While the display above suggests a similar optimality result for our estimator as well, such a property is difficult to state formally in the present framework. The aforementioned authors study target model parameters assumed to be identical in n or that converge to some fixed limit. In the high-dimensional setting of the present paper, the number of model parameters is allowed to grow infinitely and thus cannot converge to a fixed limit within the parameter space for any given model. (Note that this is not the same issue as contending with an infinite-dimensional nuisance parameter.) Thus it is impossible to compare asymptotic covariance matrices directly. Indeed, we do not prove any results concerning the asymptotic variance of our estimator but rather that of the derived t-statistic in Theorem 3.4.

One might be tempted to impose regularity conditions that would allow for such comparisons. For instance, one might require that, for any $p_0 \in \mathbb{N}$, the n-indexed sequence of parameter subsets $\beta_{1,n}, \ldots, \beta_{p_0,n}$ converge to a well-defined p_0-dimensional vector. Then one might conjecture along the lines of: "For each such p_0, the respective p_0-length subset of $\tilde{\boldsymbol{\beta}}$ has asymptotic variance identical to the optimal variance in a fixed p_0-dimensional linear IV model". However, the scope of such asymptotic statements is relative to a *fixed* p_0; the approach does not describe the asymptotic behavior of $\tilde{\boldsymbol{\beta}}$ under *joint* growth of n and $p_{\boldsymbol{x}}$ when the latter may depend on the former, which is the setting we study. Hence we do not pursue this approach here. Instead, we believe that a proper study of optimal inference in high-dimensional linear IV models should follow along the lines of Jankova and van de Geer (2018). We conjecture that the scale factor in the asymptotic pivot of Theorem 3.4 is optimal in the sense of Jankova and van de Geer (2018, Theorem 3) and that the updated second-stage estimator achieves a similar efficiency bound. A more thorough investigation of this matter is required for future work.

We conclude the present section with a brief discussion of the feasibility of select conditions of Theorem 3.4. Condition (3) can be derived as a consequence of the standard assumption that the minimal and maximal eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{d}}$ be bounded strictly away from zero and infinity; see Proposition C.4. The feasibility of Conditions (1) and (6) depends on the distribution of the conditional means $\boldsymbol{d}_i$ and hence of the instrumental variables $\boldsymbol{z}_i$. The following Lemma shows that both conditions are satisfied if the $\boldsymbol{z}_i$ are sub-Gaussian.

Lemma 3.5 (*Feasibility of Conditions (1) and (6)a*)**.** *Suppose that (i) the instrumental variables $\boldsymbol{z}_i$ satisfy Assumption 2.2, (ii) we have $\|\boldsymbol{\Sigma}_{\boldsymbol{z}}\|_\infty = O(1)$, (iii) we have $\|\boldsymbol{\Theta}\|_{L_1} \leq m_{\boldsymbol{\Theta}}$ for $m_{\boldsymbol{\Theta}} = O(1)$, (iv) $\sqrt{(\log p_{\boldsymbol{x}})/n} = o(1)$, and (v) $\exp(-n^{2\zeta}/\sqrt{\log p_{\boldsymbol{x}}}) = o(1)$ for some $0 < \zeta < 1/2$. Then Conditions (1) and (6)a of Theorem 3.4 are satisfied.*

The proof of Lemma 3.5 is found in Section A.3 of Appendix. The requirements that the quantities τ, $\|\boldsymbol{\Sigma}_{\boldsymbol{z}}\|_\infty$, and $m_{\boldsymbol{\Theta}}$ in the statement of Lemma 3.5 be of constant order can be relaxed at the cost of introducing more complex growth conditions. Condition (v) of Lemma 3.5 is satisfied under reasonable constraints on the growth of $p_{\boldsymbol{x}}$ — for instance, if $\sqrt{\log p_{\boldsymbol{x}}}/n^{2\zeta} \lesssim 1$.

4. Two-stage Lasso

Theorem 3.4 depends on high-level assumptions that ensure good behavior of the remainder terms $\boldsymbol{f}_\ell$ and standard error estimate $\hat{\omega}_j$. In this section, we demonstrate how such conditions may be satisfied in the high-dimensional setting. In particular, we introduce in Section 4.1 a two-stage Lasso estimation procedure, for which we provide theoretical bounds in Section 4.2. The rates for the second-stage estimation error are particularly involved due to the dependence on the predicted conditional means from the first-stage estimation. In Section 4.3, we identify conditions under which the remainder terms $\boldsymbol{f}_\ell$ vanish in probability under the two-stage Lasso procedure. We also propose estimators of the standard errors and provide conditions under which these estimators converge to the true standard errors.

4.1. Two-stage estimator

For $j \in [p_{\boldsymbol{x}}]$, we let $\hat{\boldsymbol{\alpha}}^j$ denote the *first-stage Lasso estimator*

$$\hat{\boldsymbol{\alpha}}^j \in \operatorname*{arg\,min}_{\boldsymbol{a} \in \mathbb{R}^{p_{\boldsymbol{z}}}} \left\{ \|\boldsymbol{x}^j - \boldsymbol{Z}\boldsymbol{a}\|_2^2/(2n) + r_j \|\boldsymbol{a}\|_1 \right\} . \tag{10}$$

We let $\hat{d}_{ij} := \boldsymbol{z}_i^\top \hat{\boldsymbol{\alpha}}^j$ denote the predicted conditional mean of x_{ij} given $\boldsymbol{z}_i$ based on the estimates $\hat{\boldsymbol{\alpha}}^j$ and write $\widehat{\boldsymbol{D}} = \boldsymbol{Z}\widehat{\mathbf{A}}$, where the matrix $\widehat{\boldsymbol{D}}$ has columns given by $\hat{\boldsymbol{d}}^j := (\hat{d}_{1j}, \ldots, \hat{d}_{nj})^\top$ and the matrix $\widehat{\mathbf{A}} \in \mathbb{R}^{p_{\boldsymbol{z}} \times p_{\boldsymbol{x}}}$ has columns given by the $\hat{\boldsymbol{\alpha}}^j$.

We define the *second-stage Lasso estimator* to be

$$\hat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{b} \in \mathbb{R}^{p_{\boldsymbol{x}}}} \left\{ \|\boldsymbol{y} - \widehat{\boldsymbol{D}}\boldsymbol{b}\|_2^2/(2n) + r_{\boldsymbol{\beta}} \|\boldsymbol{b}\|_1 \right\} . \tag{11}$$

Please cite this article as: D. Gold, J. Lederer and J. Tao, Inference for high-dimensional instrumental variables regression. Journal of Econometrics (2020), https://doi.org/10.1016/j.jeconom.2019.09.009.

In Section 4.2.2, we develop sparsity-based results that require the following quantities. We write $S_j = \operatorname{supp} \boldsymbol{\alpha}^j$ for the *active sets* of the first-stage regression parameters, and we write $s_{\boldsymbol{\alpha}^j} := |S_j|$ and $s_{\mathbf{A}} := \max_{j\in[p_{\boldsymbol{x}}]} s_{\boldsymbol{\alpha}^j}$; we write $S_{\boldsymbol{\beta}} := \operatorname{supp} \boldsymbol{\beta}$ for the active set of the second-stage regression parameter and $\boldsymbol{s}_{\boldsymbol{\beta}} := |S_{\boldsymbol{\beta}}|$. We note that ℓ_0 sparsity is not a limitation in principle and that more general regression vectors may be considered at the price of additional complexity (Bühlmann and van de Geer, 2011, Sections 6.2.3-4), (Belloni et al., 2012).

Remark 4.1 (*Model Identifiability*)**.** Let $\boldsymbol{\alpha}^j_{S_j} := \{\alpha^j_k : \alpha^j_k \neq 0\}$ denote the restriction of $\boldsymbol{\alpha}^j$ to S_j, and let $S = \cup_j S_j$. Write $\mathbf{A}_S$ for the $|S| \times p_{\boldsymbol{x}}$ matrix with columns $\boldsymbol{\alpha}^j_{S_j}$ and $\boldsymbol{Z}_S$ for the $n \times |S|$ matrix with columns (re-indexed as necessary) corresponding to the columns of $\mathbf{A}_S$. Note that $\boldsymbol{D} = \boldsymbol{Z}_S\mathbf{A}_S$ and hence that $\boldsymbol{\Sigma}_{\boldsymbol{d}} = \operatorname{cov}(\boldsymbol{D}) = \mathbf{A}_S^\top \operatorname{cov}(\boldsymbol{Z}_S)\mathbf{A}_S$. We require that $\boldsymbol{\Sigma}_{\boldsymbol{d}}$ be invertible and hence of full rank — for model identifiability, probabilistic guarantees we discuss in the following sections, and because we work directly with the inverse covariance matrix $\boldsymbol{\Theta}$. Thus we must have $p_{\boldsymbol{x}} = \operatorname{rank} \boldsymbol{\Sigma}_{\boldsymbol{d}} \leq \min(\operatorname{rank} \mathbf{A}_S, \operatorname{rank} \operatorname{cov}(\boldsymbol{Z}_S)) \leq |S|$. This is a relatively strong assumption, which we implicitly require throughout the sequel. However, we emphasize that this requirement is an artifact of the ℓ_0-sparsity-based methods by which we derive the bounds for the first- and second-stage estimation errors. As noted above, we chose these methods to simplify our exposition and that one can obtain morally similar but more complex bounds even when the $\boldsymbol{\alpha}^j$ are not sparse (Bühlmann and van de Geer, 2011, Sections 6.2.3-4). Thus, in general, this restriction is not impractical.

4.2. Estimation error bounds

In this section, we present estimation error bounds for the first- and second-stage estimators described in Section 4.1. Both such bounds depend on the same fundamental strategy for proving finite-sample guarantees for ℓ_1-regularized estimators. This strategy consists of two parts. The first part is the *oracle inequality*, which establishes a deterministic bound for the estimation and prediction performance of the Lasso on a particular set of interest. The second part is the control of the *empirical process term*, which defines the set of interest. We include such prerequisites in Section B.1 of Appendix.

Before we present the estimation error bounds for the first- and second-stage Lasso estimators, we first discuss the compatibility condition, which is required in the proof of the oracle inequality.

4.2.1. Compatibility condition

The oracle bounds rely on the good behavior of certain moduli of continuity of the empirical Gram matrices $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} = \boldsymbol{Z}^\top\boldsymbol{Z}/n$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \widehat{\boldsymbol{D}}^\top\widehat{\boldsymbol{D}}/n$. We codify this requirement in the following definition.

Definition 4.2 (*Compatibility Condition*)**.** For a given index set $S \subseteq [p]$, $p \in \mathbb{N}$, define the double-cone

$$\mathcal{C}(S) := \left\{\boldsymbol{\delta} \in \mathbb{R}^p \setminus \mathbf{0} : \|\boldsymbol{\delta}_{S^\complement}\|_1 \leq 3\|\boldsymbol{\delta}_S\|_1\right\}. \tag{12}$$

We say that the *compatibility condition* holds for the matrix $\boldsymbol{M} \in \mathbb{R}^{n\times p}$ relative to the index set S and the constant $\phi^2 > 0$ defined as

$$\phi^2 = \inf_{\boldsymbol{\delta}\in\mathcal{C}(S)} \frac{|S|\|\boldsymbol{M}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}_S\|_1^2} \tag{13}$$

if the latter is greater than zero. We call the quantity ϕ^2 the *compatibility constant*.

The compatibility condition is so named because it interfaces between the ℓ_1 norm of the estimation error and the ℓ_2 prediction error of the Lasso estimator. It is a standard tool in the ℓ_1-regularized estimation literature to ensure identifiability: it limits the correlations among the predictors such that the estimator can discriminate between the "relevant" parameters with index in S and the remaining parameters. For this purpose, the index set S is taken to be the active set of the target regression parameter (Bühlmann and van de Geer, 2011, Chapter 6).

The set $\mathcal{C}(S)$ increases with increasing active set S; hence, the larger S, the more restrictive the compatibility condition becomes. This means that in such theories, identifiability and sparsity are closely intertwined.

The constant 3 is arbitrary. Alternative choices require adjustment of other constants that appear in the bounds (Bickel et al., 2009). A related, slightly stronger condition known as the *restricted eigenvalue condition* is elsewhere used for the same end (Bickel et al., 2009; Bühlmann and van de Geer, 2011; van de Geer and Bühlmann, 2009). The only known task where such conditions can be avoided is prediction (Dalalyan et al., 2017; Hebiri and Lederer, 2013; Lederer et al., 2019; van de Geer and Lederer, 2013; Zhuang and Lederer, 2018).

The compatibility and restricted eigenvalue conditions are sometimes defined more generally in terms of the cardinality s of the index set S rather than a specific index set. For instance, (Bickel et al., 2009, Assumption RE(s, c_0)) require for their restricted eigenvalue condition that the quantity

$$\kappa(s, c_0) := \min_{S\subseteq[p]:|S|\leq s} \min_{\substack{\boldsymbol{\delta}\neq\mathbf{0}:\\ \|\boldsymbol{\delta}_{S^\complement}\|_1\leq c_0\|\boldsymbol{\delta}_S\|_1}} \|\boldsymbol{M}\boldsymbol{\delta}\|_2/(\sqrt{n}\|\boldsymbol{\delta}_S\|_2) \tag{14}$$

Please cite this article as: D. Gold, J. Lederer and J. Tao, Inference for high-dimensional instrumental variables regression. Journal of Econometrics (2020), https://doi.org/10.1016/j.jeconom.2019.09.009.

be bounded away from zero. The rationale for taking the minimum over all such index sets S is that the true support of $\boldsymbol{\beta}$ is unknown. See also the discussion of Rudelson and Zhou (2012). We note also that the compatibility and restricted eigenvalue conditions can be replaced by slightly weaker assumptions at the cost of more involved definitions (Dalalyan et al., 2017).

Practical use of the first- and second-stage Lasso estimators requires selection of the tuning parameter r_j and $r_{\boldsymbol{\beta}}$ for $j \in [p_{\boldsymbol{x}}]$. A number of proposals for theoretical choices of tuning parameters (Belloni and Chernozhukov, 2013; Belloni et al., 2011; Bickel et al., 2009; Bühlmann and van de Geer, 2011) for the ordinary (one-stage) linear model exist in the regularized regression literature. The bounds of Section 4.2.2 are based on oracle choices of the tuning parameters r_j, $r_{\boldsymbol{\beta}}$, which depend explicitly on inestimable quantities. It would be preferable to give results for *data-adaptive* tuning parameters such as Chételat et al. (2017) and Chichignoud et al. (2016) for which the respective Lasso problems can feasibly be implemented. For the present work, we are content to demonstrate that there exist sequences of oracle tuning parameters that tend to zero sufficiently fast to ensure that the remainder terms $\boldsymbol{f}_\ell$ are asymptotically negligible. In practice, cross-validated choices of Lasso tuning parameters and μ chosen according to the scheme described in Section 3.2 suffice in favorable parameter configurations. We provide evidence for this claim in Section 5.

4.2.2. Estimation error bounds

Simultaneous control of the first-stage estimation errors $\hat{\boldsymbol{\alpha}}^j - \boldsymbol{\alpha}^j$ is a straightforward consequence of the standard theoretical results for the Lasso. We include these bounds in Section B.2 of the Supplementary Materials. On the other hand, the bounds for the second-stage estimation error $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, which we study in the present section, are more involved due to the dependence of $\hat{\boldsymbol{\beta}}$ on the predicted conditional means $\hat{\boldsymbol{d}}_i$. Our strategy is to write $\boldsymbol{y} = \widehat{\boldsymbol{D}}\boldsymbol{\beta} + \tilde{\boldsymbol{u}}$, where $\tilde{\boldsymbol{u}} := \boldsymbol{u} + [(\boldsymbol{D} - \widehat{\boldsymbol{D}}) + \boldsymbol{V}]\boldsymbol{\beta}$ and apply concentration results to bound the probability of the event $\{4\|\widehat{\boldsymbol{D}}^\top \tilde{\boldsymbol{u}}\|_\infty \leq r_{\boldsymbol{\beta}}\}$, allowing us to adapt oracle inequality arguments for the Lasso to the present case.

We require the following assumption for the bounds of this section.

Assumption 4.3 (*Compatibility Conditions*)**.**

(1) There exists for each active set S_j of the first-stage model a constant $\phi_j^2 > 0$ such that $\boldsymbol{Z}$ satisfies the compatibility condition with respect to S_j and ϕ_j^2. We assume that the n-indexed sequences of such constants ϕ_j^2 are bounded strictly away from zero uniformly in n. We write $\phi_{\mathbf{A}}^2 := \max_{j\in[p_{\boldsymbol{x}}]} \phi_j^2$.

(2) There exists a constant $\phi_{\boldsymbol{\beta}}$ such that $\widehat{\boldsymbol{D}}$ satisfies the compatibility condition with respect to $S_{\boldsymbol{\beta}}$ and $\phi_{\boldsymbol{\beta}}$. We assume that the n-indexed sequence of such constants is bounded strictly away from zero uniformly in n.

Assumption 4.3 imposes the compatibility condition on the random object $\widehat{\boldsymbol{D}}$. In Lemma 4.6, we provide sufficient conditions under which Assumption 4.3 holds with high probability.

Note that whether $\boldsymbol{Z}$ satisfies the compatibility condition with respect to one active set S_{j_1} does not bear directly on whether it satisfies the compatibility condition with respect to another active set S_{j_2} for $j_1, j_2 \in [p_{\boldsymbol{x}}]$. As such, it is non-trivial to assume that the compatibility condition as specified in Definition 4.2 holds for each active set S_j for $j \in [p_{\boldsymbol{x}}]$ when $p_{\boldsymbol{x}}$ tends to infinity. However, the condition that $\boldsymbol{Z}$ satisfies the compatibility condition with respect to each active set S_j is entailed by requiring that $\kappa(s, c_0)$ of (14) for $s = \max_{j\in[p_{\boldsymbol{x}}]} s_{\boldsymbol{\alpha}^j}$ and $c_0 = 3$ be bounded away from 0. Thus, the need to accommodate multiple active sets does not thereby significantly alter the treatment of the compatibility condition.

We will refer to the following choices of tuning parameters throughout the sequel.

Definition 4.4 (*Tuning Parameters*)**.** (1) For the first-stage Lasso estimator, set $r_j := c_j(\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty(\log p_{\boldsymbol{z}})/n)^{1/2}$, where $c_j > 0$ is a controlled quantity. We let $\boldsymbol{r} := (r_1, \ldots, r_{p_{\boldsymbol{x}}})$ denote the tuple of first-stage tuning parameters, and we write $r_{\mathbf{A}} := \max_{j\in[p_{\boldsymbol{x}}]} r_j = c_{\boldsymbol{V}}(\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty(\log p_{\boldsymbol{z}})/n)^{1/2}$, where $c_{\boldsymbol{V}} = \max_{j\in[p_{\boldsymbol{x}}]} c_j$. (2) For the second-stage Lasso estimator, set

$$r_{\boldsymbol{\beta}} = 16\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \big(4m_{\boldsymbol{\beta}}\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} + m_{\mathbf{A}}\big) + \big(4\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} + m_{\mathbf{A}}\big)\big(m_{\boldsymbol{\beta}}\lambda_{\boldsymbol{V}} + \lambda_{\boldsymbol{u}}\big),$$

where $m_{\mathbf{A}}$, $m_{\boldsymbol{\beta}}$ are as defined in Assumption 2.1 and

$$\lambda_{\boldsymbol{V}} := c_{\boldsymbol{V}}(\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty(\log p_{\boldsymbol{z}})/n)^{1/2}, \qquad \lambda_{\boldsymbol{u}} := c_{\boldsymbol{u}}(\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty(\log p_{\boldsymbol{z}})/n)^{1/2},$$

where $c_{\boldsymbol{u}} > 0$ is a controlled quantity.

Note that, unlike as in much of the related literature, the tuning parameters we identify above are random, in particular due to the term $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty$. Our theory handles the consequences of this allowance in Condition (5) of Assumption 4.7, for which we provide subsequent justification. As we note above, our practical choice of tuning parameters is guided by cross-validation.

We now present probabilistic bounds for the ℓ_1 estimation error for the second-stage Lasso estimator $\hat{\boldsymbol{\beta}}$.

Lemma 4.5 (*Bound for* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$)**.** *Suppose that Assumption 2.4 holds and that the compatibility conditions 4.3 are satisfied with probability at least* $t_n = o(1)$*. For each* $j \in [p_{\boldsymbol{x}}]$*, set* r_j *according to Definition 4.4; set* $r_{\boldsymbol{\beta}}$ *according to Definition 4.4.*

Then,

$$\mathrm{P}\Big\{\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 > 4\frac{\boldsymbol{s}_{\boldsymbol{\beta}}}{\phi^2_{\boldsymbol{\beta}}}\Big(4\frac{s_{\mathbf{A}}}{\phi^2_{\mathbf{A}}}c_{\boldsymbol{V}}\big(m_{\boldsymbol{\beta}}c_{\boldsymbol{V}}[16\frac{s_{\mathbf{A}}}{\phi^2_{\mathbf{A}}}\|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty+1]+c_{\boldsymbol{u}}\big)\|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty(\log p_{\boldsymbol{z}})/n + m_{\mathbf{A}}\big(16\frac{s_{\mathbf{A}}}{\phi^2_{\mathbf{A}}}\|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty c_{\boldsymbol{V}}+c_{\boldsymbol{u}}[m_{\boldsymbol{\beta}}+1]\big)(\|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty(\log p_{\boldsymbol{z}})/n)^{1/2}\Big)\Big\} \le ep_{\boldsymbol{z}}^{1-c^2_{\boldsymbol{u}}C_0/\tau^2_{\boldsymbol{u}}}+ep_{\boldsymbol{z}}^{2-C_0\min_{j\in[p_{\boldsymbol{x}}]}\{c_j^2/\|\boldsymbol{v}^j\|^2_{\psi_2}\}}+t_n\,,$$

where C_0 is as specified in Lemma C.1.

Lemma 4.5 entails that $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 = O_{\mathrm{P}}\big(\boldsymbol{s}_{\boldsymbol{\beta}}s^2_{\mathbf{A}}(\log p_{\boldsymbol{z}})/n+\boldsymbol{s}_{\boldsymbol{\beta}}s_{\mathbf{A}}\sqrt{(\log p_{\boldsymbol{z}})/n}\big)$. Thus, we see that the convergence rate of the second-stage Lasso estimator is slower than the typical rate of $\boldsymbol{s}_{\boldsymbol{\beta}}\sqrt{\log(p_{\boldsymbol{x}})/n}$ in the ordinary (sub-)Gaussian linear model. Zhu (2018) provides L_1 improved convergence rate of the second stage estimator with different set of assumptions. Whether the rate can be improved under the assumptions of the present paper is a direction for future work.

Since Lemma 4.5 requires that Assumption 4.3 holds, we need to demonstrate the latter's feasibility. The following Lemma provides such a guarantee. For other approaches to studying the empirical compatibility constants and restricted eigenvalues of random matrices, see Rudelson and Zhou (2012) and van de Geer and Muro (2014). Unlike the extant literature on the compatibility condition, however, we must account for the prediction error of $\widehat{\boldsymbol{D}}$.

Lemma 4.6 (*Second-stage Compatibility Constant*)**.** *Suppose that the $\boldsymbol{z}_i$ and $\boldsymbol{v}^j$ satisfy Assumptions 2.2 and 2.4, respectively. Set r_j according to Definition 4.4 for each $j\in[p_{\boldsymbol{x}}]$; set $r_{\mathbf{A}}=\max_{j\in[p_{\boldsymbol{x}}]}r_j$. Let $\sqrt{(\log p_{\boldsymbol{z}})/n}=o(1)$. Then, for n sufficiently large,*

$$\mathrm{P}\Big\{\phi^2(\widehat{\boldsymbol{D}},S_{\boldsymbol{\beta}})<\Lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{d}})-\big(a+384m_{\mathbf{A}}\boldsymbol{s}_{\boldsymbol{\beta}}\frac{s_{\mathbf{A}}}{\phi^2_{\mathbf{A}}}c_{\boldsymbol{V}}\|\widehat{\Sigma}_{\boldsymbol{z}}\|^{3/2}_\infty\big)\sqrt{(\log p_{\boldsymbol{z}})/n}\Big\} \le ep_{\boldsymbol{z}}^{2-C_0\min_{j\in[p_{\boldsymbol{x}}]}\{c_j^2/\|\boldsymbol{v}^j\|^2_{\psi_2}\}}+2p_{\boldsymbol{x}}^{2-a^2/(6e^2\kappa^2)}\,,$$

where $a>0$ is a controlled quantity, $\Lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{d}})$ denotes the minimal eigenvalue of $\boldsymbol{\Sigma}_{\boldsymbol{d}}$, and $\kappa=m^2_{\mathbf{A}}(2\tau^2_{\boldsymbol{z}}+\|\boldsymbol{\Sigma}_{\boldsymbol{z}}\|_\infty/\log 2)$.

Lemma 4.6 entails that, under mild growth conditions $\phi^2(\widehat{\boldsymbol{D}},S_{\boldsymbol{\beta}})$ is bounded below by a sequence of quantities approaching $\Lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{d}})$, and hence that $\widehat{\boldsymbol{D}}$ satisfies the compatibility condition with probability approaching one. Note that we must also choose the controlled quantity a above so that the exponent $2-a^2/(6e^2\kappa^2)$ is negative. The sub-Gaussian regime of the present paper entails that we may choose such an a of constant order. Throughout the present essay we use the symbol a in various bounds to denote a controlled quantity that plays this role as above.

4.3. Remainder terms and scale factors

The asymptotic results of Section 3.3 depend on the high-level assumption that the remainder terms $\boldsymbol{f}_\ell$ and satisfy $\|\boldsymbol{f}_\ell\|_\infty=o_{\mathrm{P}}(1)$. In this Section, we identify the specific conditions under which this assumption is satisfied for the two-stage Lasso. The primary goal of these conditions, which we present in Assumption 4.7, is to ensure the ℓ_1 consistency of the first- and second-stage estimators and of the estimator $\widehat{\boldsymbol{\Theta}}$ specified in Section 3.2. We implicitly refer to n-indexed sequences of all quantities mentioned below.

Assumption 4.7 (*Model Regularity for Inference of β_j*)**.**

(1) Assumption 4.3 holds;
(2) The growth condition $\max_{j\in[p_{\boldsymbol{x}}]}s_{\boldsymbol{\omega}^j}r_j=o(1)$ holds;
(3) The sequence of population quantities $\boldsymbol{\Theta}$ satisfies $\boldsymbol{\Theta}\in\mathcal{U}(m_{\boldsymbol{\Theta}},q,s_{\boldsymbol{\Theta}})$ for a universal constant $m_{\boldsymbol{\Theta}}$, $s_{\boldsymbol{\Theta}}>0$ and controlled $q\in[0,1)$;
(4) The condition $\mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu)^{\complement}=o(1)$ holds;
(5) The quantity $\|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty=\|\mathbb{E}_n[\boldsymbol{z}_i\boldsymbol{z}_i^\top]\|_\infty$ satisfies $\lim_{n\to\infty}\mathrm{P}\{\|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty>m_{\boldsymbol{Z}}\}=0$ for a universal constant $m_{\boldsymbol{Z}}$;

and

(6) The following growth conditions hold:
  (a) $\mu^{1-q}s_{\boldsymbol{\Theta}}\sqrt{\log p_{\boldsymbol{z}}}=o(1)$;
  (b) $s^3_{\mathbf{A}}\boldsymbol{s}_{\boldsymbol{\beta}}(\log p_{\boldsymbol{z}})^{3/2}/n+s^2_{\mathbf{A}}\boldsymbol{s}_{\boldsymbol{\beta}}(\log p_{\boldsymbol{z}})/\sqrt{n}=o(1)$
  (c) $\mu\boldsymbol{s}_{\boldsymbol{\beta}}\big(s^2_{\mathbf{A}}\log p_{\boldsymbol{z}}/\sqrt{n}+s_{\mathbf{A}}\sqrt{\log p_{\boldsymbol{z}}}\big)=o(1)$.

Condition (1) is a prerequisite for the bounds on the first- and second-stage estimation errors; we discuss the feasibility of these assumptions in Section 4.2.2. Condition (2) is required for asymptotic control of the remainder terms and is comparable to typical growth rates required for Lasso consistency. Condition (3) is required to control $\hat{\boldsymbol{\theta}}_j-\boldsymbol{\theta}_j$ under ℓ_∞ and ℓ_1 norms as discussed in Section 3.2. Condition (4) is a high-level requirement for asymptotic negligibility of the

remainder terms $\boldsymbol{f}_\ell$; it can be obtained as a consequence of specific model assumptions as in Lemma 4.10. Condition (5) is similarly a high-level condition required for asymptotic negligibility of the remainder terms: it ensures that the empirical quantity $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty = \|\boldsymbol{Z}^\top \boldsymbol{Z}/n\|_\infty$ behaves in probability as of constant order. It can be derived as a consequence of the standard requirement that the minimal and maximal eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ be bounded away from zero and infinity uniformly in n if $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} - \boldsymbol{\Sigma}_{\boldsymbol{z}}\|_\infty = o_{\mathrm{P}}(1)$; the latter condition can in turn be derived from distributional assumptions on the $\boldsymbol{z}_i$ using the tools of, say, Vershynin (2012). Condition (6) lists the model parameter growth conditions required for asymptotic negligibility of the remainder terms under the sub-Gaussian noise regime of Assumption 2.4. We can compare these conditions with the requirement $s \log p/\sqrt{n} = o(1)$ in Javanmard and Montanari (2014) and van de Geer et al. (2014) for negligibility of the single remainder term that occurs under the ordinary linear model. The conditions on the sparsity here are generally similar to those of the ordinary linear model, yet slightly more strict when comparing the powers at which the sparsity factors enter. It is not clear if our conditions can be relaxed further, or whether there are more fundamental reasons for the differences.

Note that while the quantity $m_{\boldsymbol{Z}}$ of Condition (5) appears in the bounds of Lemmas B.8–B.14 of the Appendix, which give the rates for the remainder terms, we do not include it in the growth conditions of Condition (6). This is because, under the presently studied regime, $m_{\boldsymbol{Z}}$ is assumed of constant order. One could consider more general scenarios where the maximum entry of $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}$ is not bounded in probability and include the quantity $m_{\boldsymbol{Z}}$ in the aforementioned growth conditions. Doing so would in turn affect the rate at which $s_{\mathbf{A}}$, $\boldsymbol{s}_{\boldsymbol{\beta}}$, and $p_{\boldsymbol{z}}$ may be allowed to grow with n while maintaining asymptotic negligibility of the remainder terms $\boldsymbol{f}_\ell$.

In addition to the model regularity conditions of Assumption 4.7, we require appropriate choices of the first- and second-stage Lasso tuning parameters and the estimator $\widehat{\boldsymbol{\Theta}}$. We codify such choices in the following Assumption and then conclude the asymptotic negligibility of the remainder terms.

Assumption 4.8 (*Specification of Estimators*)**.** Let $\widehat{\mathbf{A}}$ and $\hat{\boldsymbol{\beta}}$ be the first- and second-stage Lasso estimators, respectively. The tuning parameters under the sub-Gaussian noise regime of Assumption 2.4 are chosen according to

(i) Definition 4.4 for the first-stage tuning parameters $\boldsymbol{r} = (r_j)_{j=1}^{p_{\boldsymbol{x}}}$ and the quantity $r_{\mathbf{A}}$;
(ii) Definition 4.4 for the second-stage tuning parameter $r_{\boldsymbol{\beta}}$ and quantities $\lambda_{\boldsymbol{u}}$ and $\lambda_{\boldsymbol{V}}$;
(iii) let $\widehat{\boldsymbol{\Theta}}$ be an estimator of $\boldsymbol{\Theta}$ with rows $\hat{\boldsymbol{\theta}}_j$ given by solutions to Problem 3.2.

Lemma 4.9 (*Negligibility of Remainders for Theorem 3.4*)**.** *Suppose that Assumption 2.4 and Conditions (1)–(6) of Assumption 4.7 hold and that the estimators $\widehat{\mathbf{A}}$, $\hat{\boldsymbol{\beta}}$, and $\widehat{\boldsymbol{\Theta}}$ are chosen according to Assumption 4.8(i)–(iii). Then, $\|\boldsymbol{f}_\ell\|_\infty = o_{\mathrm{P}}(1)$ for $\ell \in [4]$.*

The primary use of Lemma 4.9 is to verify Condition (2) of Theorem 3.4. Indeed, the result justifies the use of the one-step update to the second-stage Lasso estimator to construct asymptotically valid confidence intervals for the components β_j according to (9).

We note that the quantity μ, which we recall is the tolerance parameter for Problem 3.2, must be given careful consideration. Conditions (6)a and (6)b of Assumption 4.7 require μ to be of small order $(s_{\boldsymbol{\Theta}}\sqrt{\log p_{\boldsymbol{z}}})^{\frac{1}{q-1}}$ and $(\boldsymbol{s}_{\boldsymbol{\beta}} s_{\mathbf{A}} \log p_{\boldsymbol{z}})^{-1}$, respectively. However, μ must not tend to zero so fast that the probability that $\boldsymbol{\Theta}$ is feasible for Problem 3.2, which we recall is formally denoted by $\mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$, becomes bounded away from zero. The following Lemma identifies a choice of μ that satisfies these competing objectives for sub-Gaussian $\boldsymbol{z}_i$ and first-stage noise elements.

Lemma 4.10 (*Probability of $\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$*)**.** *Suppose that (i) the $\boldsymbol{z}_i$ and $\boldsymbol{v}^j$ satisfy Assumptions 2.2 and 2.4, respectively; (ii) $\widehat{\mathbf{A}}$ consists of first-stage Lasso estimates of the $\boldsymbol{\alpha}^j$ with tuning parameters r_j chosen according to Definition 4.4. Set*

$$\mu = \frac{m_{\boldsymbol{\Theta}}}{\sqrt{n}}\left(a\sqrt{\log p_{\boldsymbol{x}}} + 12 m_{\mathbf{A}} c_{\boldsymbol{V}} \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty^{3/2} \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2}\sqrt{\log p_{\boldsymbol{z}}}\right),$$

where $c_{\boldsymbol{V}}$ is as in Definition 4.4. Then, for n sufficiently large,

$$\mathrm{P}\left\{\|\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty > \mu\right\} \le 2p_{\boldsymbol{x}}^{2-a^2/(6e^2\eta^2)} + e p_{\boldsymbol{z}}^{2 - C_0 \min_{j\in[p_{\boldsymbol{x}}]}\{c_j^2/\|\boldsymbol{v}^j\|_{\psi_2}^2\}},$$

where $\eta = m_{\mathbf{A}}^2(2\tau_{\boldsymbol{z}}^2 + \|\boldsymbol{\Sigma}_{\boldsymbol{z}}\|_\infty/\log 2)$ and C_0 is as defined in Lemma C.1.

If there exists $m_{\boldsymbol{Z}} = O(1)$ that satisfies $\mathrm{P}\{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty > m_{\boldsymbol{Z}}\} = o(1)$, we may substitute the former quantity into the specification of μ in Lemma 4.10 to obtain

$$\mathrm{P}\left\{\|\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty > \mu\right\} \le 2p_{\boldsymbol{x}}^{2-a^2/(6e^2\eta^2)} + 2p_{\boldsymbol{z}}^{1-c_{\mathrm{ep}}} + \mathrm{P}\{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty > m_{\boldsymbol{Z}}\},$$

which tends to zero under appropriate specification of the controlled quantities a and c. Note that $\mu \lesssim s_{\mathbf{A}}\sqrt{\log(p_{\boldsymbol{z}})/n}$; Condition (6)a of Assumption 4.7 then becomes

$$s_{\boldsymbol{\Theta}} s_{\mathbf{A}}^{1-q}(\log p_{\boldsymbol{z}})^{1-\frac{q}{2}}/n^{\frac{1-q}{2}} = o(1), \tag{15}$$

and Condition (6)c becomes

$$s_\beta s_{\mathbf{A}}^3 (\log p_z)^{3/2}/n + s_\beta s_{\mathbf{A}}^2 \log(p_z)/\sqrt{n} \;=\; o(1)\,, \tag{16}$$

which is identical to Condition (6)b.

Recall that Theorem 3.4 specifies the conditions under which $\sqrt{n}(\tilde{\beta}_j - \beta_j)/\omega_j$ converges weakly to a $\mathcal{N}(0,1)$ random variable and demonstrates that the limit continues to hold if ω_j is replaced with an estimator $\hat{\omega}_j$ that satisfies $|\hat{\omega}_j - \omega_j| = o_{\mathrm{P}}(1)$. In practice, ω_j is not available, hence it is crucial to demonstrate the existence of an estimator $\hat{\omega}_j$ that satisfies the foregoing condition. The following Lemma identifies such an estimator $\hat{\omega}_j$ and the conditions under which it is appropriate for use with Theorem 3.4.

Lemma 4.11 (*Existence of Appropriate $\hat{\omega}_j$*)**.** *Suppose that (i) (a) the second-stage noise elements u_i satisfy $\mathrm{E}[u_i^2 \mid \boldsymbol{z}_i] = \sigma_u^2$ and $\|\boldsymbol{u}\|_2^2/n - \sigma_u^2 = o_{\mathrm{P}}(1)$, (b) the $\boldsymbol{z}_i$ and $\boldsymbol{v}^j$ satisfy Assumptions 2.2 and 2.4, (c) we have that $s_\beta s_{\mathbf{A}}^2 \log(p_z)/n + s_\beta s_{\mathbf{A}} \sqrt{\log(p_z)/n} = o(1)$ and $\max_{j\in[p_x]} \max_{i\in[n]} \mathrm{E}[x_{ij}^2] = O(1)$. Let $\widehat{\mathbf{A}}$, $\hat{\boldsymbol{\beta}}$, and $\widehat{\boldsymbol{\Theta}}$ be as specified in Assumption 4.8. Define the estimator $\hat{\sigma}_u$ of the second-stage noise level σ_u by*

$$\hat{\sigma}_u^2 \;:=\; \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2/n\,. \tag{17}$$

Then $\hat{\sigma}_u - \sigma_u = o_{\mathrm{P}}(1)$. If, in addition to the conditions in (i), we have (ii) (a) Condition (4) of Assumption 4.7 holds, (b) the sequence of minimal and maximal eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{d}}$, denoted respectively by $\Lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{d}})$ and $\Lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{d}})$, are bounded away from zero and infinity uniformly in n; and (c) the tolerance μ satisfies $\mu = o(1)$, then $\hat{\omega}_j$ defined by

$$\hat{\omega}_j^2 \;:=\; \hat{\sigma}_u^2 \widehat{\Theta}_{jj} \tag{18}$$

satisfies $\hat{\omega}_j - \omega_j = o_{\mathrm{P}}(1)$.

5. Numerical experiments

In this section, we present a Monte Carlo simulation study of the finite-sample properties of the inferential procedure developed in Section 3 using the two-stage Lasso studied in Section 4. Our objective is to test this method under a variety of parameter configurations chosen to reflect settings of practical interest. In Section 5.1, we describe the general scheme according to which the data for each trial are generated and the metrics gathered for each configuration. In Section 5.2, we enumerate the specific parameter configurations studied and discuss the results.

5.1. General experimental design

Each trial contains a data-generation step and an estimation step. We specify the regression parameters $\boldsymbol{\beta}$ and $\mathbf{A}$ for the data-generation step as follows. For each configuration, we set the second-stage regression parameter $\boldsymbol{\beta}$ according to $\beta_j = 1$ for $j \in S_\beta$ and $\beta_j = 0$ otherwise, where $S_\beta \subset [p_x]$ is a random set of s_β generated by uniformly random draws from $[p_x]$ without replacement. Similarly, we set the first-stage regression parameters $\boldsymbol{\alpha}^j$ for $j \in [p_x]$ according to $\alpha_k^j = 1$ for $k \in S_j$ and $\alpha_k = 0$ otherwise, where $S_j \subset [p_z]$ is a random set of $s_{\mathbf{A}}$ generated by uniformly random draws from $[p_z]$ without replacement. We let $s_\beta, s_{\mathbf{A}}$ vary over configurations.

Having specified the regression parameters, we then draw n i.i.d. observations $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ according to

$$\begin{aligned} \boldsymbol{z}_i \;&\sim\; \mathcal{N}_{p_z}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{z}})\,, & x_{ij} \;&=\; \langle \boldsymbol{z}_i, \boldsymbol{\alpha}^j \rangle + v_{ij}\,, \\ (u_i, \boldsymbol{v}_i) \mid \boldsymbol{z}_i \;&\sim\; \mathcal{N}_{1+p_x}(\mathbf{0}, \boldsymbol{\Sigma}_{uv}) & y_i \;&=\; \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle + u_i\,, \end{aligned}$$

where $n, p_x, p_z, \boldsymbol{\beta}, \{\boldsymbol{\alpha}^j\}_{j\in[p_x]}$, and the structure of $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ vary amongst configurations. For all configurations, we set

$$\boldsymbol{\Sigma}_{uv} \;=\; \begin{pmatrix} \sigma_u & \boldsymbol{\sigma}_{uv}^\top \\ \boldsymbol{\sigma}_{uv} & \sigma_v \boldsymbol{I} \end{pmatrix}$$

where $\sigma_u, \sigma_v = \sqrt{.7}$ are held fixed and $\boldsymbol{\sigma}_{uv} = (\sigma_{uv^1}, \ldots, \sigma_{uv^{p_x}})$ is given as follows. For each configuration, we set one σ_{uv^j} chosen at random equal to .5, nine σ_{uv^j} chosen at random equal to .25, and the remaining σ_{uv^j} equal to .05. The present covariance structure for the noise reflects the constraint that $\boldsymbol{\Sigma}_{uv}$ be positive-definite; our choices of σ_{uv^j} are an attempt to balance this requirement with the goal of studying non-trivial correlations between the first- and second-stage noise elements.

We consider two forms for the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{z}}$. The first form is a Toeplitz (TZ) structure given by

$$\boldsymbol{\Sigma}_{\boldsymbol{z}}^{\mathrm{TZ}}\big|_{jk} \;=\; \rho^{|j-k|}\,, \qquad j,k \in [p_z]\,, \qquad \rho = 0.8\,.$$

The second is a circulant-symmetric (CS) structure given for $j \le k$ by

$$\boldsymbol{\Sigma}_{\boldsymbol{z}}^{\mathrm{CS}}\big|_{jk} \;=\; \begin{cases} 1 & k = j, \\ 0.1 & k \in \{j+1, \ldots, j+5\} \cup \{j+p_z-5, \ldots, j+p_z-1\}, \\ 0 & \text{otherwise.} \end{cases}$$

Please cite this article as: D. Gold, J. Lederer and J. Tao, Inference for high-dimensional instrumental variables regression. Journal of Econometrics (2020), https://doi.org/10.1016/j.jeconom.2019.09.009.

Table 1
Simulation results.

$(n,\ p_{\boldsymbol{x}},\ p_{\boldsymbol{z}})$	$(\boldsymbol{s}_{\boldsymbol{\beta}}, s_{\boldsymbol{\alpha}^j})$	$\boldsymbol{\Sigma}_{\boldsymbol{z}}^{\mathrm{CS}}$			$\boldsymbol{\Sigma}_{\boldsymbol{z}}^{\mathrm{TZ}}$		
		$\widehat{\mathrm{cvg}}$	$\widehat{\mathrm{len}}$	$\mathrm{MSE}(\hat{\boldsymbol{\beta}})$	$\widehat{\mathrm{cvg}}$	$\widehat{\mathrm{len}}$	$\mathrm{MSE}(\hat{\boldsymbol{\beta}})$
(100, 125, 150)	(3, 5)	0.942	0.225	0.004	0.895	0.201	0.005
	(5, 10)	0.941	0.211	0.004	0.672	0.212	0.014
	(10, 15)	0.930	0.190	0.003	0.545	0.219	0.030
(200, 250, 275)	(3, 5)	0.947	0.157	0.002	0.942	0.140	≤ 0.001
	(5, 10)	0.941	0.171	0.002	0.673	0.192	0.011
	(10, 15)	0.930	0.190	0.003	0.545	0.219	0.030
(300, 400, 500)	(3, 5)	0.947	0.094	≤ 0.001	0.952	0.092	≤ 0.001
	(5, 10)	0.955	0.085	≤ 0.001	0.945	0.082	≤ 0.001
	(10, 15)	0.961	0.067	≤ 0.001	0.927	0.064	≤ 0.001
(500, 600, 700)	(3, 5)	0.947	0.094	≤ 0.001	0.952	0.092	≤ 0.001
	(5, 10)	0.951	0.082	≤ 0.001	0.950	0.088	≤ 0.001
	(10, 15)	0.961	0.067	≤ 0.001	0.927	0.064	≤ 0.001

Within a configuration study, the random quantities $\boldsymbol{z}_i$, u_i, and $\boldsymbol{v}_i$ are re-drawn for each trial; the quantities $\boldsymbol{\beta}$, $\{\boldsymbol{\alpha}^j\}_{j=1}^{p_{\boldsymbol{x}}}$, $\boldsymbol{\Sigma}_{\boldsymbol{z}}$, $\boldsymbol{\sigma}_{uv}$, n, $p_{\boldsymbol{x}}$, $p_{\boldsymbol{z}}$ are held fixed.

For the estimation step of each trial, we compute the first- and second-stage Lasso as defined in Section 4.1 using the `glmnet` package (Friedman et al., 2010). Tuning parameters r for the Lasso estimators are selected by 10-fold cross-validation over a grid $\{r_\ell\}_{\ell=1}^{L}$, where $L = 100$, $r_L = .01 r_1$, and r_1 is the least quantity for which the respective Lasso estimator is identically 0. The tuning parameter $r_{\boldsymbol{\beta}}$ is chosen by similar cross-validation procedure.

The rows $\hat{\boldsymbol{\theta}}_j$ of $\widehat{\boldsymbol{\Theta}}$ are obtained as solutions to the respective Problem 3.2 with tuning parameter μ_j chosen as follows. For each $j \in [p_{\boldsymbol{x}}]$, we set $\mu_j := \kappa \times \inf_{\boldsymbol{\theta} \in \mathbb{R}^{p_{\boldsymbol{x}}}} \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} \boldsymbol{\theta} - \boldsymbol{e}_j\|_\infty$, where $\boldsymbol{e}_j$ denotes the jth canonical basis vector in $p_{\boldsymbol{x}}$ dimensions and $\kappa > 1$ is chosen at our discretion. Note that, under this choice of μ_j, the respective Problem 3.2 is guaranteed feasible. The factor κ is chosen to balance the performance of $\widehat{\boldsymbol{\Theta}}$ as a surrogate inverse for $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}$, for which a smaller κ is desirable, with the size of the objective function $\|\boldsymbol{\theta}\|_1$, for which a larger κ is desirable. The following results were obtained under $\kappa = 1.2$. To obtain the infimum, we cast $\mathrm{minimize}_{\boldsymbol{\theta} \in \mathbb{R}^{p_{\boldsymbol{x}}}} \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} \boldsymbol{\theta} - \boldsymbol{e}_j\|_\infty$ as a linear programming problem, which we solve using MOSEK optimization software (MOSEK AsP, 2017).

In a given trial t, $t = 1, \ldots, T$, we set $\tau = 0.05$ and compute the respective $100(1-\tau)\%$ confidence interval

$$\hat{\mathcal{I}}_{\tau,t,j} = \left[\tilde{\beta}_{t,j} - z_\tau \widehat{\mathrm{SE}}(\tilde{\beta}_{t,j}),\ \tilde{\beta}_j + z_\tau \widehat{\mathrm{SE}}(\tilde{\beta}_{t,j})\right],$$

for each component $\tilde{\beta}_{t,j}$ of $\tilde{\boldsymbol{\beta}}_t$, where $z_\tau = \Phi^{-1}(1-\tau/2)$ and

$$\widehat{\mathrm{SE}}(\tilde{\beta}_j)^2 = \mathbb{E}_n[(y_i - \hat{\boldsymbol{\beta}}\boldsymbol{X})^2 \langle \hat{\boldsymbol{\theta}}_j, \hat{\boldsymbol{d}}_i \rangle^2].$$

For each configuration of n, $p_{\boldsymbol{x}}$, $p_{\boldsymbol{z}}$, $\boldsymbol{s}_{\boldsymbol{\beta}}$, $s_{\mathbf{A}}$, $\boldsymbol{\Sigma}_{\boldsymbol{z}}$, we generate $T = 100$ trials and calculate the average coverage $\widehat{\mathrm{cvg}}$ for the 95% confidence intervals $\hat{\mathcal{I}}_{\tau,j}$ about components of $\tilde{\boldsymbol{\beta}}$ and the average interval length $\widehat{\mathrm{len}}$ given by

$$\widehat{\mathrm{cvg}} = \frac{1}{p_{\boldsymbol{x}}} \sum_{j=1}^{p_{\boldsymbol{x}}} \frac{1}{T} \sum_{t=1}^{T} 1\left\{\beta_{t,j} \in \hat{\mathcal{I}}_{\tau,t,j}\right\}, \qquad \widehat{\mathrm{len}} = \frac{1}{p_{\boldsymbol{x}}} \sum_{j=1}^{p_{\boldsymbol{x}}} \frac{1}{T} \sum_{t=1}^{T} \mathrm{len}(\hat{\mathcal{I}}_{\tau,t,j}).$$

For each configuration, we also provide the average mean squared error of the second-stage Lasso estimator given by $\widehat{\mathrm{MSE}} = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{p_{\boldsymbol{x}}} \sum_{j=1}^{p_{\boldsymbol{x}}} (\tilde{\beta}_{t,j} - \beta_{t,j})^2$. We present the results for the study described above in Table 1 in Section 5.2.

5.2. Specifications and results

We conduct simulations according the design described in Section 5.1 for all configurations belonging to

$$\underbrace{\begin{pmatrix}(100, 125, 150)\\(200, 250, 275)\\(300, 400, 500)\\(500, 600, 700)\end{pmatrix}}_{(n, p_{\boldsymbol{x}}, p_{\boldsymbol{z}})} \times \underbrace{\begin{pmatrix}(3, 5)\\(5, 10)\\(10, 15)\end{pmatrix}}_{(\boldsymbol{s}_{\boldsymbol{\beta}}, s_{\mathbf{A}})} \times \underbrace{\begin{pmatrix}\boldsymbol{\Sigma}_{\boldsymbol{z}}^{\mathrm{CS}}\\ \boldsymbol{\Sigma}_{\boldsymbol{z}}^{\mathrm{TZ}}\end{pmatrix}}_{\boldsymbol{\Sigma}_{\boldsymbol{z}}}.$$

The results, which are presented in Table 1, show that our estimator achieves close to nominal coverage under a variety of configurations. We also see that arguably the greatest determinant of coverage is the relative magnitude of $p_{\boldsymbol{x}}$ and $p_{\boldsymbol{z}}$ to the size of the active set $\boldsymbol{s}_{\boldsymbol{\beta}}$. As the latter grows, coverage diverges from the nominal level. This phenomenon is expected and has been found in ordinary linear regression models as well van de Geer et al. (2014) and Javanmard and Montanari (2014), since the bounds for the estimation error of the Lasso are proportional to the size of the active set. Nevertheless, the

Please cite this article as: D. Gold, J. Lederer and J. Tao, Inference for high-dimensional instrumental variables regression. Journal of Econometrics (2020), https://doi.org/10.1016/j.jeconom.2019.09.009.

performance improves significantly when we increase the sample size. Finally, we observe that the covariance structure of the instrumental variables $\boldsymbol{z}_i$ has a strong influence on coverage: the Toeplitz structure features greater correlation among the instrumental variables in general, and this is reflected in coverage that tends to be farther from the nominal level than in the case of the circulant-symmetric covariance structure. These results suggest that our proposed method of inference for the low-dimensional components of a high-dimensional regression vector is relevant to practical scenarios that may exhibit non-trivial degrees of correlation between the noise components and nontrivial correlation among the instrumental variables.

6. Conclusion

In this paper, we propose inference methods for the components of a high-dimensional instrumental variables regression parameter despite possible endogeneity of each regressor. We allow both the number of instruments and the number of regressors to be greater than the sample size. We construct asymptotically valid confidence intervals for the components of the second-stage regression coefficients. Though our estimator is not a nonlinear generalized method of moments (GMM) estimator (Hansen, 1982), we expect that our results can be extended to that more general setting.

Our Sections 2 and 3 comprise a general pipeline for estimation and inference, while the remainder is then an exemplification with a Lasso approach. Therefore, it would be interesting to use our pipeline with other regularized estimators as well (one could also use different estimators for the two different stages). Candidates include, for example, SCAD (Fan and Li, 2001) and MCP (Zhang, 2010). We finally refer to our software on github.com/LedererLab/HDIV.

Appendix A. Materials required for Section 3

A.1. Materials required for Section 3.1

Proof of Lemma 3.1. Note that

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\Theta}}\mathbb{E}_n[\tilde{\boldsymbol{h}}(\hat{\boldsymbol{\beta}})]\\ &= \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n\\ &= \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top(\boldsymbol{X}[\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}] + \boldsymbol{u})/n\\ &= \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top(\widehat{\boldsymbol{D}}[\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}] + [\boldsymbol{X} - \widehat{\boldsymbol{D}}][\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}] + \boldsymbol{u})/n\\ &= \boldsymbol{\beta} + \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top\boldsymbol{u}/n + \underbrace{\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top(\boldsymbol{X} - \widehat{\boldsymbol{D}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})/n}_{\boldsymbol{f}_3/\sqrt{n}} + \underbrace{(\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}_{\boldsymbol{f}_4/\sqrt{n}}\,.\end{aligned}$$

Now decompose the second term on the right-hand side above as follows

$$\begin{aligned}\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top\boldsymbol{u}/n &= \widehat{\boldsymbol{\Theta}}\boldsymbol{D}^\top\boldsymbol{u}/n + \widehat{\boldsymbol{\Theta}}(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top\boldsymbol{u}/n\\ &= \boldsymbol{\Theta}\boldsymbol{D}^\top\boldsymbol{u}/n + \underbrace{(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})\boldsymbol{D}^\top\boldsymbol{u}/n}_{\boldsymbol{f}_1/\sqrt{n}} + \underbrace{\widehat{\boldsymbol{\Theta}}(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top\boldsymbol{u}/n}_{\boldsymbol{f}_2/\sqrt{n}}\end{aligned}$$

to complete the proof. □

A.2. Materials required for Section 3.2

We present the properties of the estimators $\hat{\boldsymbol{\theta}}_j$ required for the results of Sections 3.3 and 4.3. Lemma A.1, which gives an ℓ_∞ bound for the estimation error $\hat{\boldsymbol{\theta}}_j - \hat{\boldsymbol{\theta}}$, is comparable to Cai et al. (2011, Theorem 4); the proofs are similar but depend on different conditions on the covariance estimator $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}$. The proof of Lemma A.2 follows part of that of Cai et al. (2011, Theorem 6).

Lemma A.1. *Suppose that: (i) the quantity $\|\boldsymbol{\Theta}\|_{L_1}$ is bounded above by a constant $m_{\boldsymbol{\Theta}} < \infty$; and (ii) $\widehat{\boldsymbol{\Theta}}$ is an estimate of $\boldsymbol{\Theta} = \boldsymbol{\Sigma}_{\boldsymbol{d}}^{-1} = \mathrm{cov}(\boldsymbol{d}_i)^{-1}$ with rows $\hat{\boldsymbol{\theta}}_j$ obtained as solutions to Problem 3.2. Then, on the set $\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$ as defined in (8),*

$$\|\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\|_\infty \le 2m_{\boldsymbol{\Theta}}\mu$$

for each $j \in [p_{\boldsymbol{x}}]$.

Proof of Lemma A.1. First, observe that the conditions of the present lemma entail that

$$\|\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty \le \mu\,, \qquad \|\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty \le \mu\,.$$

Please cite this article as: D. Gold, J. Lederer and J. Tao, Inference for high-dimensional instrumental variables regression. Journal of Econometrics (2020), https://doi.org/10.1016/j.jeconom.2019.09.009.

Now, on the set $\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$, each row $\boldsymbol{\theta}_j$ is feasible for the respective Specific Problem 3.2. It then follows from the optimality of $\hat{\boldsymbol{\theta}}_j$ that $\|\hat{\boldsymbol{\theta}}_j\|_1 \leq \|\boldsymbol{\theta}_j\|_1$ for each $j \in [p_{\boldsymbol{x}}]$ and hence that $\max_{j\in[p_{\boldsymbol{x}}]} \|\hat{\boldsymbol{\theta}}_j\|_1 \leq \|\boldsymbol{\Theta}\|_{L_1}$. Next, observe that

$$\begin{aligned}\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}} \;=\; \boldsymbol{\Theta}(\boldsymbol{I} - \boldsymbol{\Sigma}_{\boldsymbol{d}}\widehat{\boldsymbol{\Theta}}) \;&=\; \boldsymbol{\Theta}(\boldsymbol{I} + (\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma}_{\boldsymbol{d}})\widehat{\boldsymbol{\Theta}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}\widehat{\boldsymbol{\Theta}}) \\ &=\; \boldsymbol{\Theta}(\boldsymbol{I} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}\widehat{\boldsymbol{\Theta}}) - \boldsymbol{\Theta}(\boldsymbol{\Sigma}_{\boldsymbol{d}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}})\widehat{\boldsymbol{\Theta}} \\ &=\; \underbrace{\boldsymbol{\Theta}(\boldsymbol{I} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}\widehat{\boldsymbol{\Theta}})}_{\mathrm{I}_1} - \underbrace{(\boldsymbol{I} - \boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}})\widehat{\boldsymbol{\Theta}}}_{\mathrm{I}_2}\,.\end{aligned}$$

From Hölder,

$$\|\mathrm{I}_2\|_\infty \;\leq\; \|\boldsymbol{I} - \boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}\|_\infty \|\widehat{\boldsymbol{\Theta}}\|_{L_1} \;\leq\; m_{\boldsymbol{\Theta}}\mu\,.$$

The matrix ℓ_∞- and (L_1- norm if the argument is symmetric) are invariant under transposition of their arguments. Use this fact and Hölder to obtain

$$\begin{aligned}\|\mathrm{I}_1\|_\infty \;&=\; \|(\boldsymbol{I} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}\widehat{\boldsymbol{\Theta}})^\top \boldsymbol{\Theta}^\top\|_\infty \\ &\leq\; \|(\boldsymbol{I} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}\widehat{\boldsymbol{\Theta}})^\top\|_\infty \|\boldsymbol{\Theta}^\top\|_{L_1} \\ &\leq\; \|\boldsymbol{I} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}\widehat{\boldsymbol{\Theta}}\|_\infty \|\boldsymbol{\Theta}\|_{L_1} \;\leq\; m_{\boldsymbol{\Theta}}\mu\,,\end{aligned}$$

where the final line follows from the fact that both $\boldsymbol{I} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}\widehat{\boldsymbol{\Theta}}$ and $\boldsymbol{\Theta}$ are symmetric. Thus

$$\|\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}\|_\infty \;\leq\; \|\mathrm{I}_1\|_\infty + \|\mathrm{I}_2\|_\infty \;\leq\; 2m_{\boldsymbol{\Theta}}\mu\,,$$

as required. □

Lemma A.2. *Suppose in addition to the conditions of Lemma A.1 that $\boldsymbol{\Theta}$ belongs to the uniformity class $\mathcal{U}(m_{\boldsymbol{\Theta}}, q, s_{\boldsymbol{\Theta}})$. Then,*

$$\|\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\|_1 \;\leq\; 2c_q(2m_{\boldsymbol{\Theta}}\mu)^{1-q}s_{\boldsymbol{\Theta}}$$

for each $j \in [p_{\boldsymbol{x}}]$, where $c_q := 1 + 2^{1-q} + 3^{1-q}$.

Proof of Lemma A.2. See the proof of line (14) of Cai et al. (2011, Theorem 6). □

A.3. Materials required for Section 3.3

Proof of Theorem 3.4. The proof of the first claim consists of two steps. First, we show that the quantity

$$Z_{j,n} \;:=\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\langle\boldsymbol{\theta}_j, \boldsymbol{d}_i\rangle \mathrm{u}_i/\omega_j\,.$$

and $\sqrt{n}(\tilde{\beta}_j - \beta_j)/\omega_j$ share the same weak limit. Second, we show that $Z_{j,n} \rightsquigarrow \mathcal{N}(0,1)$. To establish the first step, we claim that

$$\lim_{n\to\infty} \mathrm{P}\{\sqrt{n}(\tilde{\beta}_j - \beta_j)/\omega_j \leq t\} \;\leq\; \lim_{n\to\infty} \mathrm{P}\{Z_{j,n} \leq t\}\,, \tag{A.1}$$

for all $t \in \mathbb{R}$. An analogous lower bound follows by a matching argument, which shows that $\sqrt{n}(\tilde{\beta}_j - \beta_j)/\omega_j$ and $Z_{j,n}$ share the same weak limit. To show the claim above, let $t \in \mathbb{R}$ be given, fix a controlled $\epsilon > 0$, and note that, by the decomposition of Lemma 3.1, we have

$$\begin{aligned}\mathrm{P}\big\{\sqrt{n}(\tilde{\beta}_j - \beta_j)/\omega_j \leq t\big\} \;&\leq\; \mathrm{P}\left\{Z_{j,n} + \sum_{\ell=1}^{4} f_{\ell,j}/\omega_j \leq t + 4\epsilon\right\} \\ &\leq\; \mathrm{P}\big\{Z_{j,n} \leq t + \epsilon\big\} + \sum_{\ell=1}^{4}\mathrm{P}\big\{f_{\ell,j}/\omega_j > \epsilon\big\}\,.\end{aligned}$$

By specification of σ_u and $\boldsymbol{\Theta}_{jj}$ in Conditions (5) and (6) and Condition (3) of the present theorem, it follows that ω_j is bounded strictly away from 0 uniformly in n. The assumptions of the present theorem then entail that $\mathrm{P}\{f_{\ell,j}/\omega_j > \epsilon\} = o(1)$ for all $\epsilon > 0$ and each ℓ. Letting ϵ tend to zero shows the claim of (A.1). It follows from the analogous lower bound that

$$\lim_{n\to\infty} \mathrm{P}\{\sqrt{n}(\tilde{\beta}_j - \beta_j)/\omega_j \leq t\} \;=\; \lim_{n\to\infty} \mathrm{P}\{Z_{j,n} \leq t\}$$

for all $t \in \mathbb{R}$, thus completing the first step.

Please cite this article as: D. Gold, J. Lederer and J. Tao, Inference for high-dimensional instrumental variables regression. Journal of Econometrics (2020), https://doi.org/10.1016/j.jeconom.2019.09.009.

Next, we show that, under each of Conditions (5) and (6) in the statement of the present theorem, $Z_{j,n} \rightsquigarrow Z_j \sim \mathcal{N}(0,1)$. To this end, we define the quantity

$$w_j^2 := \boldsymbol{\theta}_j^\top \overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} \boldsymbol{\theta}_j = \frac{1}{n} \sum_{i=1}^n \langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle^2 ,$$

where we recall that $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \boldsymbol{D}^\top \boldsymbol{D}/n$. We claim first that

$$\tilde{Z}_{j,n} := \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle u_i}{w_j \sigma_u} \rightsquigarrow Z_j \sim \mathcal{N}(0,1)$$

under each of Conditions (5) and (6) and second that $\sigma_u w_j/\omega_j \to_{\mathrm{P}} 1$. Note that $Z_{j,n} = \frac{w_j \sigma_u}{\omega_j} \tilde{Z}_{j,n}$, hence the desired limit follows from an application Slutsky's Lemma.

To show the first claim under Condition (5), note that, by specification of w_j, we have under Assumption 2.4 that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle u_i}{w_j \sigma_u} \mid \boldsymbol{Z} \sim \mathcal{N}(0,1) .$$

Thus

$$\lim_{n\to\infty} \mathrm{P}\{\tilde{Z}_{j,n} \le t\} = \lim_{n\to\infty} \mathrm{E}\big[\mathrm{P}\{\tilde{Z}_{j,n} \le t \mid \boldsymbol{Z}\}\big] = \lim_{n\to\infty} \mathrm{E}[\Phi(t) \mid \boldsymbol{Z}] = \Phi(t)$$

for all $t \in \mathbb{R}$, where Φ denotes the c.d.f. of a standard Normal random variable. This shows the desired weak limit under Condition (5).

We use the Lindeberg–Feller Central Limit Theorem to show the limit under Condition (6). To begin, write

$$\tilde{Z}_{j,n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i , \qquad \xi_i := \langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle u_i/(w_j \sigma_u) .$$

Note that

$$\mathrm{E}[\xi_i] = \mathrm{E}\big[\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle/(w_j \sigma_u) \mathrm{E}[u_i | \boldsymbol{z}_i]\big] = 0$$

and that

$$\sigma_n^2 := \sum_{i=1}^n \mathrm{E}[\xi_i^2] = \mathrm{E}\left[\frac{1}{w_j^2} \sum_{i=1}^n \langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle^2 \mathrm{E}[(u_i/\sigma_u)^2 | \boldsymbol{z}_i]\right] = n .$$

To demonstrate Lindeberg's condition, let $\delta > 0$ be arbitrary and write

$$\begin{aligned} \sigma_n^{-2} \sum_{i=1}^n \mathrm{E}[\xi_i^2 \mathbb{1}\{|\xi_i| > \delta \sigma_n\}] &= \frac{1}{n} \sum_{i=1}^n \mathrm{E}\big[\mathrm{E}[\xi_i^2 \mathbb{1}\{|\xi_i| > \delta \sqrt{n}\} \mid \boldsymbol{z}_i]\big] \\ &= \mathrm{E}\left[\frac{1}{n} \frac{1}{w_j^2} \sum_{i=1}^n \langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle^2 \mathrm{E}[(u_i/\sigma_u)^2 \mathbb{1}\{|\xi_i| > \delta \sqrt{n}\} \mid \boldsymbol{z}_i]\right] , \end{aligned}$$

where we have substituted the definition of ξ_i and extracted the factor $\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle^2/w_j^2$ from the conditional expectation to obtain the second line. Now substitute the definition of w_j below and note

$$\frac{1}{w_j^2} \sum_{i=1}^n \langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle^2 = \frac{n}{\sum_{i=1}^n \langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle^2} \times \sum_{i=1}^n \langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle^2 = n .$$

Substitute the above result into the second line of two displays prior and use the law of iterated expectation to obtain

$$\sigma_n^{-2} \sum_{i=1}^n \mathrm{E}[\xi_i^2 \mathbb{1}\{|\xi_i| > \delta \sigma_n\}] = \mathrm{E}\big[(u_1/\sigma_u)^2 \mathbb{1}\{|\xi_1| > \delta \sqrt{n}\}\big] ,$$

where we cite condition (6) that the $\boldsymbol{z}_i$ and u_i are i.i.d. with $\mathrm{E}[u_i^2 | \boldsymbol{z}_i] = \sigma_u^2$ to reduce to the case of u_1. For brevity, we write $\tilde{u}_1 := u_1/\sigma_u$. Introduce the set $\mathcal{T} := \{|\langle \boldsymbol{\theta}_j, \boldsymbol{d}_1 \rangle| \le n^\zeta\}$ and note, since $|\xi_1| \le w_j^{-1} |\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle| |\tilde{u}_1|$, that

$$\big\{|\xi_1| > \delta \sqrt{n}\big\} \cap \mathcal{T} \subseteq \big\{|\tilde{u}_1| > \delta w_j n^{1/2-\zeta}\big\} ,$$

and hence that $\mathbb{1}\{|\xi_1| > \delta \sqrt{n}\} \cap \mathcal{T} \le \mathbb{1}\{|\tilde{u}_1| > \delta w_j n^{1/2-\zeta}\}$. Combine this inequality and the law of total probability with the result of two displays previous to obtain

$$\sigma_n^{-2} \sum_{i=1}^n \mathrm{E}[\xi_i^2 \mathbb{1}\{|\xi_i| > \delta \sigma_n\}] = \mathrm{E}\big[\tilde{u}_1^2 (\mathbb{1}\{|\xi_1| > \delta \sqrt{n}\} \cap \mathcal{T} + \mathbb{1}\{|\xi_1| > \delta \sqrt{n}\} \cap \mathcal{T}^\complement)\big]$$

$$= \underbrace{\mathrm{E}[\tilde{u}_1^2 \mathbb{1}\{|\tilde{u}_1| > \delta w_j n^{1/2-\zeta}\}]}_{\mathrm{I}_1} + \underbrace{\mathrm{E}[\tilde{u}_1^2 \mathbb{1}_{\mathcal{T}^c}]}_{\mathrm{I}_2},$$

where the substitution of indicators in the final line above is permitted since $\tilde{u}_1^2 \geq 0$. To show Lindeberg's condition, it sufficies to show that I_1 and I_2 are each $o(1)$. To treat I_1, consider the event $\{w_j \leq \vartheta^{1/2}/\sqrt{2}\}$ and write

$$\begin{aligned}\mathbb{1}\{|\tilde{u}_1| > \delta w_j n^{1/2-\zeta}\} &= \mathbb{1}\{|\tilde{u}_1| > \delta w_j n^{1/2-\zeta}\} \cap \{w_j \leq \vartheta^{1/2}/\sqrt{2}\} \\ &\quad + \mathbb{1}\{|\tilde{u}_1| > \delta w_j n^{1/2-\zeta}\} \cap \{w_j > \vartheta^{1/2}/\sqrt{2}\} \\ &\leq \mathbb{1}\{w_j \leq \vartheta^{1/2}/\sqrt{2}\} + \mathbb{1}\{|\tilde{u}_1| > \delta \vartheta^{1/2} n^{1/2-\zeta}\}\end{aligned}$$

so that

$$\mathrm{I}_1 \leq \underbrace{\mathrm{E}[\tilde{u}_1^2 \mathbb{1}\{w_j \leq \vartheta^{1/2}/\sqrt{2}\}]}_{\mathrm{I}_{1a}} + \underbrace{\mathrm{E}[\tilde{u}_1^2 \mathbb{1}\{|\tilde{u}_1| > \delta \vartheta^{1/2} n^{1/2-\zeta}\}]}_{\mathrm{I}_{1b}}.$$

Observe that

$$\begin{aligned}\mathrm{I}_{1a} &= \mathrm{E}\big[\mathbb{1}\{w_j \leq \vartheta^{1/2}/\sqrt{2}\}\mathrm{E}[\tilde{u}_1^2 \mid \boldsymbol{z}_i]\big] \\ &\lesssim \mathrm{P}\{w_j \leq \vartheta^{1/2}/\sqrt{2}\} \\ &= \mathrm{P}\{w_j^2 - \boldsymbol{\Theta}_{jj} \leq \vartheta/2 - \boldsymbol{\Theta}_{jj}\} \\ &\leq \mathrm{P}\{w_j^2 - \boldsymbol{\Theta}_{jj} \leq -\vartheta/2\} \\ &\leq \mathrm{P}\{|w_j^2 - \boldsymbol{\Theta}_{jj}| \geq \vartheta/2\},\end{aligned}$$

where the inference to the second line of the display above follows from specification of $\mathrm{E}[\tilde{u}_1^2 \mid \boldsymbol{z}_i]$ in condition (6) of the present Theorem. Now note that

$$\begin{aligned}|w_j^2 - \boldsymbol{\Theta}_{jj}| &= |\boldsymbol{\theta}_j^\top \overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} \boldsymbol{\theta}_j - \boldsymbol{\Theta}_{jj}| \\ &= |\boldsymbol{\theta}_j^\top (\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma}_{\boldsymbol{d}}) \boldsymbol{\theta}_j| \leq \|\boldsymbol{\theta}_j\|_1^2 \|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty.\end{aligned}$$

Thus

$$\begin{aligned}\mathrm{I}_{1a} &\leq \mathrm{P}\{|w_j^2 - \boldsymbol{\Theta}_{jj}| \geq \vartheta/2\} \\ &\leq \mathrm{P}\big\{\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty \geq \vartheta/(2\|\boldsymbol{\theta}_j\|_1^2)\big\} = o(1)\end{aligned} \tag{A.2}$$

by Conditions (1) and (4) of the present theorem.

We now show that

$$\mathrm{I}_{1b} = \mathrm{E}[\tilde{u}_1^2 \mathbb{1}\{|\tilde{u}_1| > \delta \vartheta^{1/2} n^{1/2-\zeta}\}] = o(1).$$

Consider the event $\{|\tilde{u}_1| > \delta \vartheta^{1/2} n^{1/2-\zeta}\}$, which is the index of the indicator function above. Exponentiate both sides of the inequality in that set by ν to obtain that

$$\big\{|\tilde{u}_1| > \delta \vartheta^{1/2} n^{1/2-\zeta}\big\} = \big\{|\tilde{u}_1|^\nu > (\delta \vartheta^{1/2})^\nu n^{\nu(1/2-\zeta)}\big\}.$$

Conclude of the respective indicator functions that

$$\mathbb{1}\{|\tilde{u}_1| > \delta \vartheta^{1/2} n^{1/2-\zeta}\} = \mathbb{1}\{|\tilde{u}_1|^\nu > (\delta \vartheta^{1/2})^\nu n^{\nu(1/2-\zeta)}\}.$$

Note by direct manipulation that, given the event that indexes the right-hand indicator above, it holds that

$$(\delta \vartheta^{1/2})^{-\nu} n^{-\nu(1/2-\zeta)} |\tilde{u}_1|^\nu > 1.$$

Thus, on the support of the right-hand side two displays prior, it holds that

$$\mathbb{1}\{(\delta \vartheta^{1/2})^{-\nu} n^{-\nu(1/2-\zeta)} |\tilde{u}_1|^\nu > 1\} \leq \big(\delta \vartheta^{1/2}\big)^{-\nu} n^{-\nu(1/2-\zeta)} |\tilde{u}_1|^\nu.$$

Hence, by the monotonicity and linearity of expectations,

$$\begin{aligned}\mathrm{I}_{1b} &= \mathrm{E}[\tilde{u}_1^2 \mathbb{1}\{|\tilde{u}_1| > \delta \vartheta^{1/2} n^{1/2-\zeta}\}] \\ &\leq \mathrm{E}\big[\tilde{u}_1^2 \big(\delta \vartheta^{1/2}\big)^{-\nu} n^{-\nu(1/2-\zeta)} |\tilde{u}_1|^\nu\big] = (\delta \vartheta^{1/2})^{-\nu} n^{-\nu(1/2-\zeta)} \mathrm{E}[|\tilde{u}_1|^{2+\nu}].\end{aligned}$$

Condition (6) of the present theorem stipulates that $\mathrm{E}[|\tilde{u}_1|^{2+\nu}]$ is of constant order, hence,

$$\lim_{n\to\infty} \mathrm{I}_{1b} \lesssim \lim_{n\to\infty} (\delta \vartheta^{1/2})^{-\nu} n^{-\nu(1/2-\zeta)} = 0$$

by the specification that $\nu > 0$ and $\zeta < 1/2$ in Condition (6) of the present theorem.

Please cite this article as: D. Gold, J. Lederer and J. Tao, Inference for high-dimensional instrumental variables regression. Journal of Econometrics (2020), https://doi.org/10.1016/j.jeconom.2019.09.009.

To show as much for I_2, observe that

$$\lim_{n\to\infty} \mathrm{I}_2 \;=\; \lim_{n\to\infty} \mathrm{E}\big[\mathbb{1}_{\{|\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i\rangle| > n^\zeta\}} \mathrm{E}[\tilde{u}_1^2 \mid \boldsymbol{z}_i]\big] \\ \lesssim \lim_{n\to\infty} \mathrm{P}\{|\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i\rangle| > n^\zeta\} \;=\; 0$$

by Condition (a) of the present theorem. This concludes the demonstration of Lindeberg's condition. It follows that $\tilde{Z}_{j,n} \rightsquigarrow \mathcal{N}(0,1)$. To show as much for $Z_{j,n}$ and hence for $\sqrt{n}(\tilde{\beta}_j - \beta_j)/\omega_j$, it suffices to show that $w_j\sigma_u/\omega_j \to_{\mathrm{P}} 1$. Note that $\omega_j = \sigma_u \boldsymbol{\Theta}_{jj}$ and hence that $w_j\sigma_u/\omega_j = w_j/\boldsymbol{\Theta}_{jj}$. Since $\boldsymbol{\Theta}_{jj}$ is bounded strictly away from zero uniformly in n, we have $|w_j/\boldsymbol{\Theta}_{jj} - 1| = |w_j - \boldsymbol{\Theta}_{jj}|/\boldsymbol{\Theta}_{jj}$ and hence that it suffices to show that $|w_j - \boldsymbol{\Theta}_{jj}| = o_{\mathrm{P}}(1)$. But, as we established above, we have for arbitrary $\epsilon > 0$

$$\mathrm{P}\{|w_j - \boldsymbol{\Theta}_{jj}| > \epsilon\} \;\le\; \mathrm{P}\{\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty > \epsilon/m_{\boldsymbol{\Theta}}^2\} \;=\; o(1)$$

by Condition (1) of the present theorem. Thus $w_j\sigma_u/\omega_j \to_{\mathrm{P}} 1$ hence $Z_{j,n} \rightsquigarrow \mathcal{N}(0,1)$ under each Condition (5) and (6) of the present theorem.

It remains to show that the limit holds when ω_j is replaced by an estimator $\hat{\omega}_j$ that satisfies $|\hat{\omega}_j - \omega_j| = o_{\mathrm{P}}(1)$. Suppose that $\hat{\omega}_j$ is such an estimator. We claim that $\hat{\omega}_j/\omega_j - 1 = o_{\mathrm{P}}(1)$. To see as much, note first that, from the hypotheses of the present theorem, $\omega_j = \sigma_u \boldsymbol{\Theta}_{jj}$ is bounded strictly away from zero uniformly in n. It then follows that $|\hat{\omega}_j/\omega_j - 1| = |\hat{\omega}_j - \omega_j|/\omega_j = o_{\mathrm{P}}(1)$ by the specification of $\hat{\omega}_j$. By the continuous mapping theorem, it holds that $\omega_j/\hat{\omega}_j \to_{\mathrm{P}} 1$ and hence by Slutsky's Lemma that

$$\sqrt{n}(\tilde{\beta}_j - \beta_j)/\hat{\omega}_j = (\omega_j/\hat{\omega}_j) Z_{j,n} \rightsquigarrow \mathcal{N}(0,1)\,,$$

as claimed. □

Proof of Lemma 3.5. We first show that, under the conditions of the present Lemma,

$$\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty = o_{\mathrm{P}}(1)\,,$$

thereby demonstrating Condition (1). Note that Lemma C.5 gives

$$\mathrm{P}\big\{\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty > a\sqrt{(\log(p_{\boldsymbol{x}} \vee n))/n}\big\} \;\le\; 2(p_{\boldsymbol{x}} \vee n)^{2 - a^2/(6e^2\kappa^2)}$$

for n sufficiently large, where a is a controlled constant and $\kappa = m_{\mathbf{A}}^2(2\tau_{\boldsymbol{z}}^2 + \|\boldsymbol{\Sigma}_{\boldsymbol{z}}\|_\infty/\log 2)$. Condition (1) then follows given that $\sqrt{(\log p_{\boldsymbol{x}})/n} = o(1)$ and by choosing $a = O(1)$ large enough so that $a^2/(6e^2\kappa^2) > 2$.

We now show that, under the conditions of the present Lemma, $\mathrm{P}\{|\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i\rangle| > n^\zeta\} = o(1)$ for $0 < \zeta < 1/2$ and hence that Condition (a) is satisfied. First, write

$$\mathrm{P}\big\{|\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i\rangle| > n^\zeta\big\} \;\le\; \mathrm{P}\big\{\|\boldsymbol{\theta}_j\|_1 \|\boldsymbol{d}_i\|_\infty > n^\zeta\big\} \;\le\; \mathrm{P}\big\{m_{\boldsymbol{\Theta}}\|\boldsymbol{d}_i\|_\infty > n^\zeta\big\}\,.$$

Next, (van der Vaart and Wellner, 1996, Chapters 2.1.3 and 2.2, Pages 90–91 and 95–97) and the proof of Lemma C.5 to infer that

$$\|\max_{j\in[p_{\boldsymbol{x}}]} |d_{ij}|\|_{\psi_2} \;\le\; C_2\sqrt{\log p_{\boldsymbol{x}} + 1} \max_{j\in[p_{\boldsymbol{x}}]} \|d_{ij}\|_{\psi_2} \;\le\; C_2\tau_{\boldsymbol{z}} m_{\mathbf{A}}\sqrt{\log p_{\boldsymbol{x}} + 1}\,,$$

where C_2 is an absolute constant. The exponential Markov bound then yields

$$\mathrm{P}\big\{\|\boldsymbol{d}_i\|_\infty > t\big\} \;\le\; \exp\Big(1 - \frac{C_3 t^2}{C_2\tau_{\boldsymbol{z}} m_{\mathbf{A}}\sqrt{\log p_{\boldsymbol{x}} + 1}}\Big)$$

for $t \ge 0$, where C_3 is an absolute constant. Combine the above result with that of three displays previous, choose $t = n^\zeta/m_{\boldsymbol{\Theta}}$ and cite the growth conditions of the present lemma to conclude that

$$\mathrm{P}\big\{|\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i\rangle| > n^\zeta\big\} \;\lesssim\; \exp\Big(-\frac{n^{2\zeta}}{\sqrt{\log p_{\boldsymbol{x}}}}\Big) \overset{n\to\infty}{\to} 0\,,$$

as required for Condition (a). □

Appendix B. Materials required for Section 4

B.1. Materials required for Section 4.2

Our guarantees for estimating $\mathbf{A}$ and $\boldsymbol{\beta}$ consist of two parts. The first is the *oracle inequality*, which bounds the ℓ_1 estimation error of a generic Lasso estimator conditional on the occurrence of a special set $\mathcal{T}$. The oracle inequality is a fixture of the ℓ_1 regularized estimation literature; see for instance (Bühlmann and van de Geer, 2011, Chapter 6). We present it for the sake of completeness. The oracle inequality itself is specific to neither the first- nor the second- stage

estimators of the present work. Indeed, we require the result to derive bounds for both estimators. As such, we present the theorem in terms of a generic model that shares notation with neither the first- nor second- stage models described in Section 2.1 except for the number of observations n.

Theorem B.1 (*Oracle Inequality*)**.** *Consider the generic linear model*

$$\boldsymbol{g} \;=\; \boldsymbol{W}\boldsymbol{\gamma} + \boldsymbol{h}\,,$$

where $\boldsymbol{g} \in \mathbb{R}^n$ is a vector of univariate responses, $\boldsymbol{W} \in \mathbb{R}^{n\times p}$ is a design matrix with rows $\boldsymbol{w}_i$, $\boldsymbol{\gamma} \in \mathbb{R}^n$ is a noise vector with arbitrary distribution. Let $\hat{\boldsymbol{\gamma}}$ denote the Lasso estimator given by

$$\hat{\boldsymbol{\gamma}} \;\in\; \operatorname*{argmin}_{\boldsymbol{a}\in\mathbb{R}^p}\left\{\|\boldsymbol{g} - \boldsymbol{W}\boldsymbol{a}\|_2^2/(2n) + r\|\boldsymbol{a}\|_1\right\}.$$

Let $S_{\boldsymbol{\gamma}} := \operatorname{supp}\boldsymbol{\gamma}$, and let $s_{\boldsymbol{\gamma}} := |S_{\boldsymbol{\gamma}}|$. Suppose that $\boldsymbol{W}$ satisfies the compatibility Condition with respect to the index set $S_{\boldsymbol{\gamma}}$ and compatibility constant $\phi_{\boldsymbol{\gamma}} > 0$. Then, on the set

$$\mathcal{T}(r) \;:=\; \left\{4\|\boldsymbol{W}^\top\boldsymbol{h}/n\|_\infty \le r\right\}, \tag{B.1}$$

the bound

$$\|\boldsymbol{W}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\|_2^2/n + r\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 \;\le\; 4s_{\boldsymbol{\gamma}}r^2/\phi_{\boldsymbol{\gamma}}^2$$

holds.

Proof of Theorem B.1. The proof is algebra. See Theorem 6.1 of Bühlmann and van de Geer (2011). □

Theorem B.1 provides a deterministic guarantee for the ℓ_1 estimation error of a generic Lasso estimator $\hat{\boldsymbol{\gamma}}$ on the set $\mathcal{T}(r) = \left\{4\|\boldsymbol{W}^\top\boldsymbol{h}/n\|_\infty \le r\right\}$. Consequently, it holds that

$$\mathrm{P}\left\{\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 > 4s_{\boldsymbol{\gamma}}r/\phi_{\boldsymbol{\gamma}}^2\right\} \;\le\; \mathrm{P}\,\mathcal{T}(r)^\complement\,.$$

The quantity $\|\boldsymbol{W}^\top\boldsymbol{h}/n\|_\infty$ is sometimes called the *empirical process term*; for instance, (Bühlmann and van de Geer, 2011, Chapter 6). Thus, upper bounds for $\mathrm{P}\,\mathcal{T}(r)^\complement$ yield probabilistic guarantees for the ℓ_1 estimation error. We provide such bounds in Section C of the main Appendix.

B.2. Materials required for Section 4.2.2

Recall that the model for $\boldsymbol{x}^j$ is given by

$$\boldsymbol{x}^j \;=\; \boldsymbol{Z}\boldsymbol{\alpha}^j + \boldsymbol{v}^j\,,$$

where $\boldsymbol{v}^j$ has nontrivial covariance with the noise $\boldsymbol{u}$. It suffices for our purposes to take a naïve approach to bounding the quantity $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} \;=\; \max_{j\in[p_{\boldsymbol{x}}]}\|\hat{\boldsymbol{\alpha}}^j - \boldsymbol{\alpha}^j\|_1$. That is, we simultaneously bound the estimation error of each individual task. One could use a more complex approach such as Liu et al. (2015) to treat different patterns of joint sparsity amongst the first-stage regression vectors. We make the following assumption.

The following lemma provides finite-sample guarantees for $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}$ under the choice of tuning parameters in Definition 4.4.

The following generic bound for $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}$ can be combined with concentration results for specific distributions of the first-stage noise elements. We present it separately for the sake of modularity with respect to such assumptions.

Lemma B.2 (*Generic Bound for $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}$*)**.** *Suppose that Assumption 4.3 holds. For each $j \in [p_{\boldsymbol{x}}]$, let the sets $\mathcal{T}_{\boldsymbol{v}^j}(\lambda)$ be as defined in Lemma C.3. It then holds on the set $\bigcap_{j\in[p_{\boldsymbol{x}}]}\mathcal{T}_{\boldsymbol{v}^j}(r_j)$ that*

$$\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} \;\le\; 4s_{\mathbf{A}}r_{\mathbf{A}}/\phi_{\mathbf{A}}^2 \;=\; 4\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2}r_{\mathbf{A}}\,,$$

where r_j is the tuning parameter for the respective first-stage Lasso problem, $r_{\mathbf{A}}$ and $s_{\mathbf{A}}$ are as defined in Section 2 and $\phi_{\mathbf{A}}$ is as defined in Assumption 4.3.

Proof of Lemma B.2. For each $j \in [p_{\boldsymbol{x}}]$, Bühlmann and van de Geer (2011, Theorem 6.1, Lemma 6.2) entails that

$$\begin{aligned}\mathcal{T}_{\boldsymbol{v}^j}(r_j) \;=\; \left\{4\|\boldsymbol{Z}^\top\boldsymbol{v}^j/n\|_\infty \le r_j\right\} \;&\subseteq\; \left\{\|\hat{\boldsymbol{\alpha}}^j - \boldsymbol{\alpha}^j\|_1 \le 4s_{\boldsymbol{\alpha}^j}r_j/\phi_j^2\right\}\\ &\subseteq\; \left\{\|\hat{\boldsymbol{\alpha}}^j - \boldsymbol{\alpha}^j\|_1 \le 4s_{\mathbf{A}}r_{\mathbf{A}}/\phi_{\mathbf{A}}^2\right\},\end{aligned}$$

where the latter containment follows by specification of $s_{\mathbf{A}}$, $r_{\mathbf{A}}$, and $\phi_{\mathbf{A}}$. Take intersections over $j \in [p_{\boldsymbol{x}}]$ on both sides of the above display to conclude. □

Lemma B.3 (*Bound for* $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}$). *Suppose that Assumptions 2.4 and 4.3 hold. Set* r_j *according to Definition 4.4. Then,*

$$\mathrm{P}\big\{\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} > 4\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} c_{\boldsymbol{V}} (\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}\|_\infty (\log p_{\boldsymbol{Z}})/n)^{1/2}\big\} \;\le\; e p_{\boldsymbol{Z}}^{2 - C_0 \min_{j\in[p_{\boldsymbol{X}}]}\{c_j^2/\|\boldsymbol{v}^j\|_{\psi_2}^2\}},$$

where $c_{\boldsymbol{V}}$ *is as specified in Definition 4.4 and* C_0 *is as defined in Lemma C.1.*

Proof of Lemma B.3. Lemma B.2 entails that

$$\mathrm{P}\{\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} > 4 s_{\mathbf{A}} r_{\mathbf{A}}/\phi_{\mathbf{A}}^2\} \;\le\; \mathrm{P}\Big\{\bigcap_{j\in[p_{\boldsymbol{X}}]} \mathcal{T}_{\boldsymbol{v}^j}(r_j)\Big\}^{\complement}.$$

Apply the estimate of Lemma C.3 for the right-hand side to conclude. □

Note that, in order to obtain a rate of convergence, we must choose each c_j so that the quantity $\min_{j\in[p_{\boldsymbol{X}}]} c_j^2 C_0/\tau_{\boldsymbol{v}^j}^2 - 2$ is bounded strictly away from zero. Such a task may not be feasible in practice. Our empirical results, which we present in Section 5, suggest that cross-validated choices of the tuning parameters for the first- and second-stage Lasso estimators suffice for good behavior of the resultant updated estimator $\tilde{\boldsymbol{\beta}}$. For the sake of our theory, we assume that such appropriate choices of c_j have been made. Given such an assumption, Lemma B.3 entails that $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} = O_{\mathrm{P}}(s_{\mathbf{A}}\sqrt{\log(p_{\boldsymbol{Z}})/n})$, essentially identical to the Lasso rate for single task regression problems.

The following bound is required for Lemma 4.5

Lemma B.4 (*Control of* $\|\widehat{\boldsymbol{D}}^\top \tilde{\boldsymbol{u}}/n\|_\infty$). *Let* $\tilde{\boldsymbol{u}} = \boldsymbol{u} + [(\boldsymbol{D} - \widehat{\boldsymbol{D}}) + \boldsymbol{V}]\boldsymbol{\beta}$. *Then,*

$$\begin{aligned}\|\widehat{\boldsymbol{D}}^\top \tilde{\boldsymbol{u}}/n\|_\infty \;\le\; & \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}\|_\infty (\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}\|\boldsymbol{\beta}\|_1 + \|\mathbf{A}\|_{L_1}) \\ & + (\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} + \|\mathbf{A}\|_{L_1})(\|\boldsymbol{Z}^\top \boldsymbol{V}/n\|_\infty \|\boldsymbol{\beta}\|_1 + \|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty).\end{aligned}$$

Proof of Lemma B.4. Write $\widehat{\boldsymbol{D}}^\top = (\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top + \boldsymbol{D}^\top$ to find that

$$\begin{aligned}\|\widehat{\boldsymbol{D}}^\top \tilde{\boldsymbol{u}}/n\|_\infty \;=\; & \big\|\big[(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top + \boldsymbol{D}^\top\big]\big[(\boldsymbol{D} - \widehat{\boldsymbol{D}})\boldsymbol{\beta} + (\boldsymbol{V}\boldsymbol{\beta} + \boldsymbol{u})\big]/n\big\|_\infty \\ \le\; & \|(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top (\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{\beta}/n\|_\infty + \|(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top (\boldsymbol{V}\boldsymbol{\beta} + \boldsymbol{u})/n\|_\infty \\ & + \|\boldsymbol{D}^\top (\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{\beta}/n\|_\infty + \|\boldsymbol{D}^\top (\boldsymbol{V}\boldsymbol{\beta} + \boldsymbol{u})/n\|_\infty \\ :=\; & \mathrm{I}_1 + \mathrm{I}_2 + \mathrm{I}_3 + \mathrm{I}_4 \end{aligned} \tag{B.2}$$

We treat each quantity in the right-hand side above in turn.

For I_1, write

$$\mathrm{I}_1 \;=\; \|(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top (\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{\beta}/n\|_\infty \;\le\; \|(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top (\widehat{\boldsymbol{D}} - \boldsymbol{D})/n\|_\infty \|\boldsymbol{\beta}\|_1 .$$

Recall that $\widehat{\boldsymbol{D}} - \boldsymbol{D} = \boldsymbol{Z}(\widehat{\mathbf{A}} - \mathbf{A})$ and write

$$\begin{aligned}\|(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top (\widehat{\boldsymbol{D}} - \boldsymbol{D})/n\|_\infty \;&=\; \|(\widehat{\mathbf{A}} - \mathbf{A})^\top \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}} (\widehat{\mathbf{A}} - \mathbf{A})\|_\infty \\ &\le\; \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}^2 \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}\|_\infty \;=\; \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}^2 \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}\|_\infty ,\end{aligned}$$

where the second line follows from repeated application of Hölder's inequality. Combine the two previous displays to conclude that

$$\mathrm{I}_1 \;\le\; \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}^2 \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Z}}\|_\infty \|\boldsymbol{\beta}\|_1 . \tag{B.3}$$

For I_2, write

$$\begin{aligned}\mathrm{I}_2 \;&=\; \|(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top (\boldsymbol{V}\boldsymbol{\beta} + \boldsymbol{u})/n\|_\infty \\ &=\; \|(\widehat{\mathbf{A}} - \mathbf{A})^\top \boldsymbol{Z}^\top (\boldsymbol{V}\boldsymbol{\beta} + \boldsymbol{u})/n\|_\infty \\ &\le\; \underbrace{\|(\widehat{\mathbf{A}} - \mathbf{A})^\top \boldsymbol{Z}^\top \boldsymbol{V}\boldsymbol{\beta}/n\|_\infty}_{\mathrm{I}_{2,\mathrm{a}}} + \underbrace{\|(\widehat{\mathbf{A}} - \mathbf{A})^\top \boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty}_{\mathrm{I}_{2,\mathrm{b}}} .\end{aligned}$$

Applications of Hölder's inequality yield

$$\mathrm{I}_{2,\mathrm{a}} \;\le\; \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} \|\boldsymbol{Z}^\top \boldsymbol{V}/n\|_\infty \|\boldsymbol{\beta}\|_1$$

and

$$\mathrm{I}_{2,\mathrm{b}} \;\le\; \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} \|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty \;=\; \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} \|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty .$$

Combine the previous three displays to conclude that

$$\mathrm{I}_2 \;\le\; \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} (\|\boldsymbol{Z}^\top \boldsymbol{V}/n\|_\infty \|\boldsymbol{\beta}\|_1 + \|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty), \tag{B.4}$$

Please cite this article as: D. Gold, J. Lederer and J. Tao, Inference for high-dimensional instrumental variables regression. Journal of Econometrics (2020), https://doi.org/10.1016/j.jeconom.2019.09.009.

For the quantity I_3, write

$$\mathrm{I}_3 \;=\; \|\boldsymbol{D}^\top(\widehat{\boldsymbol{D}}-\boldsymbol{D})\boldsymbol{\beta}/n\|_\infty \;\le\; \|\boldsymbol{D}^\top(\widehat{\boldsymbol{D}}-\boldsymbol{D})/n\|_\infty\|\boldsymbol{\beta}\|_1$$

and observe that

$$\boldsymbol{D}^\top(\widehat{\boldsymbol{D}}-\boldsymbol{D})/n \;=\; \mathbf{A}^\top\boldsymbol{Z}^\top\boldsymbol{Z}(\widehat{\mathbf{A}}-\mathbf{A})/n \;=\; \mathbf{A}^\top\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}(\widehat{\mathbf{A}}-\mathbf{A})\,,$$

which yields

$$\begin{aligned}\|\boldsymbol{D}^\top(\widehat{\boldsymbol{D}}-\boldsymbol{D})/n\|_\infty \;&=\; \|\mathbf{A}^\top\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}(\widehat{\mathbf{A}}-\mathbf{A})\|_\infty\\ &\le\; \|\mathbf{A}\|_{L_1}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\|(\widehat{\mathbf{A}}-\mathbf{A})\|_{L_1}\end{aligned}$$

after repeated application of Hölder's inequality. Conclude from the previous three displays that

$$\mathrm{I}_3 \;\le\; \|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\|\mathbf{A}\|_{L_1}\,. \tag{B.5}$$

For the quantity I_4, write

$$\begin{aligned}\mathrm{I}_4 \;&=\; \|\boldsymbol{D}^\top(\boldsymbol{V}\boldsymbol{\beta}+\boldsymbol{u})/n\|_\infty\\ &=\; \|\mathbf{A}^\top\boldsymbol{Z}^\top(\boldsymbol{V}\boldsymbol{\beta}+\boldsymbol{u})/n\|_\infty\\ &\le\; \|\mathbf{A}\|_{L_1}\big(\|\boldsymbol{Z}^\top\boldsymbol{V}\boldsymbol{\beta}/n\|_\infty+\|\boldsymbol{Z}^\top\boldsymbol{u}/n\|_\infty\big)\\ &\le\; \big(\|\boldsymbol{Z}^\top\boldsymbol{V}/n\|_\infty\|\boldsymbol{\beta}\|_1+\|\boldsymbol{Z}^\top\boldsymbol{u}/n\|_\infty\big)\|\mathbf{A}\|_{L_1}\,.\end{aligned} \tag{B.6}$$

The original claim follows from line (B.2) and lines (B.3)–(B.6). □

The following generic bound for $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1$ can be combined with concentration results for specific distributions of the first- and second-stage noise elements. We present it separately for the sake of modularity with respect to such assumptions.

Lemma B.5 (*Generic Bound for $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1$*). *Suppose that Assumption 4.3 holds. Let $\lambda_{\boldsymbol{V}}, \lambda_{\boldsymbol{u}} > 0$ be arbitrary. Set $r_{\boldsymbol{\beta}}$ according to Definition 4.4. Then, on the set $\mathcal{T}_{\boldsymbol{V}}(\boldsymbol{r})\cap\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}})\cap\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})$, where $\mathcal{T}_{\boldsymbol{V}}$ and $\mathcal{T}_{\boldsymbol{u}}$ are defined as in Lemmas C.2 and C.3, respectively, and $\boldsymbol{r}$ is the tuple of first-stage tuning parameters, we have*

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \;\le\; 4\frac{s_{\boldsymbol{\beta}}}{\phi^2_{\boldsymbol{\beta}}}r_{\boldsymbol{\beta}}\,.$$

Proof of Lemma B.5. By Theorem B.1, we have

$$\mathcal{T}_{\tilde{\boldsymbol{u}}} = \big\{4\|\widehat{\boldsymbol{D}}^\top\tilde{\boldsymbol{u}}/n\|_\infty \le r\big\} \;\subseteq\; \big\{\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \le 4s_{\boldsymbol{\beta}}r/\phi^2\big\}\,.$$

It therefore suffices to show that $\mathcal{T}_{\boldsymbol{V}}(\boldsymbol{r})\cap\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})\subseteq\mathcal{T}_{\tilde{\boldsymbol{u}}}(r_{\boldsymbol{\beta}})$ for the present choice of $r_{\boldsymbol{\beta}}$. Lemma B.4 gives the bound

$$\begin{aligned}\|\widehat{\boldsymbol{D}}^\top\tilde{\boldsymbol{u}}/n\|_\infty \;\le\; &\underbrace{\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty(m_{\boldsymbol{\beta}}\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}+m_{\mathbf{A}})}_{\mathrm{I}_1}\\ &+\underbrace{(\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}+m_{\mathbf{A}})(\|\boldsymbol{Z}^\top\boldsymbol{V}/n\|_\infty\|\boldsymbol{\beta}\|_1+\|\boldsymbol{Z}^\top\boldsymbol{u}/n\|_\infty)}_{\mathrm{I}_2}\,.\end{aligned}$$

Cite Lemma B.2 to conclude that, on the set $\mathcal{T}_{\boldsymbol{V}}(\boldsymbol{r})$,

$$\mathrm{I}_1 \;\le\; 4\frac{s_{\mathbf{A}}}{\phi^2_{\mathbf{A}}}r_{\mathbf{A}}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty(4m_{\boldsymbol{\beta}}\frac{s_{\mathbf{A}}}{\phi^2_{\mathbf{A}}}r_{\mathbf{A}}+m_{\mathbf{A}})\,.$$

Note that, on the set $\mathcal{T}_{\boldsymbol{V}}(\boldsymbol{r})\cap\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}})\cap\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})$,

$$\mathrm{I}_2 \;\le\; \frac{1}{4}\big(4\frac{s_{\mathbf{A}}}{\phi^2_{\mathbf{A}}}r_{\mathbf{A}}+m_{\mathbf{A}}\big)\big(m_{\boldsymbol{\beta}}\lambda_{\boldsymbol{V}}+\lambda_{\boldsymbol{u}}\big)$$

by specification. Multiply the two previous displays by 4 and combine with the third-previous display to conclude that $\mathcal{T}_{\boldsymbol{V}}(\boldsymbol{r})\cap\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}})\cap\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})\subseteq\mathcal{T}_{\tilde{\boldsymbol{u}}}(r_{\boldsymbol{\beta}})$ for the present choice of $r_{\boldsymbol{\beta}}$, as required. □

Proof of Lemma 4.5. Note first that

$$\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}}) \;=\; \bigcap_{j=1}^{p_{\boldsymbol{x}}}\big\{4\|\boldsymbol{Z}^\top\boldsymbol{v}^j/n\|_\infty > r_j\big\} \subseteq \mathcal{T}_{\boldsymbol{V}}(\boldsymbol{r})\,.$$

Let $\mathcal{T}_{\rfloor\rfloor}$ be the set on which Assumption 4.3 holds. Lemma B.5 then entails that

$$\big\{\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 > 4\frac{\boldsymbol{s}_{\boldsymbol{\beta}}}{\phi^2_{\boldsymbol{\beta}}} r_{\boldsymbol{\beta}}\big\} \cap \mathcal{T}_{\rfloor\rfloor} \subseteq (\mathcal{T}_{\boldsymbol{V}}(\boldsymbol{r}) \cap \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}}))^{\complement} = \mathcal{T}_{\boldsymbol{V}}(\boldsymbol{r})^{\complement} \cup \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})^{\complement}\,.$$

Thus,

$$\mathrm{P}\big\{\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 > 4\frac{\boldsymbol{s}_{\boldsymbol{\beta}}}{\phi^2_{\boldsymbol{\beta}}} r_{\boldsymbol{\beta}}\big\} \le \mathrm{P}\,\mathcal{T}_{\boldsymbol{V}}(\boldsymbol{r})^{\complement} + \mathrm{P}\,\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})^{\complement} + t_n\,.$$

Now substitute the present choices of tuning parameters and cite the estimates of Lemmas C.2 and C.3. □

Lemma B.6 (*Second-Stage Compatibility Constant*)**.** *Let $S \subseteq [p]$ be an arbitrary index set with $s = |S|$. For a given matrix $\boldsymbol{M} \in \mathbb{R}^{n\times p}$, define the quantity*

$$\phi^2(\boldsymbol{M}, S) = \inf_{\delta\in\mathcal{C}(S)} \frac{s\|\boldsymbol{M}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}_S\|_1^2}\,.$$

Let $\epsilon_1, \epsilon_2 > 0$ be arbitrary. Then,

$$\begin{aligned}\mathrm{P}\big\{\phi^2(\widehat{\boldsymbol{D}}, S) < \Lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{d}}) - \epsilon_2 - \epsilon_1\big\} &\le \mathrm{P}\big\{16s(2m_{\mathbf{A}}\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1} + \|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}^2)\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty > \epsilon_1\big\} \\ &\quad + \mathrm{P}\big\{16s\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty > \epsilon_2\big\}\,,\end{aligned}$$

where $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \boldsymbol{D}^\top\boldsymbol{D}/n$.

Proof of Lemma B.6. Let S, s be as in the statement of Lemma B.6, and let $\boldsymbol{\delta} \in \mathbb{R}^{p_{\boldsymbol{x}}} \setminus \{\mathbf{0}\}$ satisfying $\|\boldsymbol{\delta}_{S^\complement}\|_1 \le 3\|\boldsymbol{\delta}_S\|_1$ be arbitrary. Write $\widehat{\boldsymbol{D}} = \boldsymbol{D} + (\widehat{\boldsymbol{D}} - \boldsymbol{D})$, so that

$$\begin{aligned}\|\widehat{\boldsymbol{D}}\boldsymbol{\delta}\|_2^2 &= \|[\boldsymbol{D} + (\widehat{\boldsymbol{D}} - \boldsymbol{D})]\boldsymbol{\delta}\|_2^2 \\ &= \big\langle[\boldsymbol{D} + (\widehat{\boldsymbol{D}} - \boldsymbol{D})]\boldsymbol{\delta}, [\boldsymbol{D} + (\widehat{\boldsymbol{D}} - \boldsymbol{D})]\boldsymbol{\delta}\big\rangle \\ &= \|\boldsymbol{D}\boldsymbol{\delta}\|_2^2 + 2\langle\boldsymbol{D}^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{\delta},\ \boldsymbol{\delta}\rangle + \langle(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{\delta}, \boldsymbol{\delta}\rangle\,.\end{aligned}$$

Thus,

$$\begin{aligned}\|\widehat{\boldsymbol{D}}\boldsymbol{\delta}\|_2^2/n \ge \|\boldsymbol{D}\boldsymbol{\delta}\|_2^2/n &- \underbrace{2|\langle\boldsymbol{D}^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{\delta}/n, \boldsymbol{\delta}\rangle|}_{\mathrm{I}_1} \\ &- \underbrace{|\langle(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{\delta}/n, \boldsymbol{\delta}\rangle|}_{\mathrm{I}_2}\,.\end{aligned} \tag{B.7}$$

We now obtain bounds for the quantities $\mathrm{I}_1, \mathrm{I}_2$. From repeated applications of Hölder's inequality, write

$$\begin{aligned}\mathrm{I}_1 \lesssim |\langle\boldsymbol{D}^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{\delta}/n, \boldsymbol{\delta}\rangle| &\le \|\boldsymbol{D}^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})/n\|_\infty\|\boldsymbol{\delta}\|_1^2 \\ &= \|\mathbf{A}^\top\boldsymbol{Z}^\top\boldsymbol{Z}(\widehat{\mathbf{A}} - \mathbf{A})/n\|_\infty\|\boldsymbol{\delta}\|_1^2 \\ &\le \|\mathbf{A}\|_{L_1}\|\boldsymbol{Z}^\top\boldsymbol{Z}/n\|_\infty\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}\|\boldsymbol{\delta}\|_1^2 \\ &\le m_{\mathbf{A}}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}\|\boldsymbol{\delta}\|_1^2\end{aligned}$$

and

$$\begin{aligned}\mathrm{I}_2 &= |\langle(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{\delta}/n, \boldsymbol{\delta}\rangle| \\ &\le \|(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})/n\|_\infty\|\boldsymbol{\delta}\|_1^2 \\ &= \|(\widehat{\mathbf{A}} - \mathbf{A})^\top\boldsymbol{Z}^\top\boldsymbol{Z}(\widehat{\mathbf{A}} - \mathbf{A})/n\|_\infty\|\boldsymbol{\delta}\|_1^2 \\ &\le \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}\|\boldsymbol{Z}^\top\boldsymbol{Z}/n\|_\infty\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}\|\boldsymbol{\delta}\|_1^2 \\ &\le \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}^2\|\boldsymbol{\delta}\|_1^2\,.\end{aligned}$$

Combine the previous two displays with (B.7) to find that

$$\|\widehat{\boldsymbol{D}}\boldsymbol{\delta}\|_2^2/n \ge \|\boldsymbol{D}\boldsymbol{\delta}\|_2^2/n - (2m_{\mathbf{A}}\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} + \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}^2)\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\|\boldsymbol{\delta}\|_1^2\,.$$

By assumption, we have $\|\boldsymbol{\delta}\|_1 \le 4\|\boldsymbol{\delta}_S\|_1$. Substitute this expression in the right-hand side above and multiply through by $s/\|\boldsymbol{\delta}_S\|_1^2$ to obtain

$$\frac{s\|\widehat{\boldsymbol{D}}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}_S\|_1^2} \ge \frac{s\|\boldsymbol{D}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}_S\|_1^2} - 16s\big(2m_{\mathbf{A}}\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} + \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}^2\big)\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\,.$$

Thus, on the set $\left\{16s\big(2m_{\mathbf{A}}\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}+\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}^2\big)\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \le \epsilon_1\right\}$, we have

$$\begin{aligned}\frac{s\|\widehat{\boldsymbol{D}}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}_S\|_1^2} &\ge \frac{s\|\boldsymbol{D}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}_S\|_1^2}-\epsilon_1\\ &= \left(\frac{s\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_{\boldsymbol{d}}\boldsymbol{\delta}}{\|\boldsymbol{\delta}_S\|_1^2}-\frac{s\boldsymbol{\delta}^\top(\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}}-\boldsymbol{\Sigma}_{\boldsymbol{d}})\boldsymbol{\delta}}{\|\boldsymbol{\delta}_S\|_1^2}\right)-\epsilon_1\\ &\ge \frac{s\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_{\boldsymbol{d}}\boldsymbol{\delta}}{\|\boldsymbol{\delta}_S\|_1^2}-\frac{s\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}}-\boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty\|\boldsymbol{\delta}\|_1^2}{\|\boldsymbol{\delta}_S\|_1^2}-\epsilon_1\\ &\ge \frac{s\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_{\boldsymbol{d}}\boldsymbol{\delta}}{\|\boldsymbol{\delta}_S\|_1^2}-16s\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}}-\boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty-\epsilon_1\,.\end{aligned}$$

From Cauchy-Schwartz we have $\|\boldsymbol{\delta}_S\|_1\le\sqrt{s}\|\boldsymbol{\delta}_S\|_2$ and hence that $\|\boldsymbol{\delta}_S\|_1^2\le s\|\boldsymbol{\delta}\|_2^2$. Substitute this bound into the first term on the right-hand side above to obtain

$$\begin{aligned}\frac{s\|\widehat{\boldsymbol{D}}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}_S\|_1^2} &\ge \frac{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_{\boldsymbol{d}}\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_2^2}-16s\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}}-\boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty-\epsilon_1\\ &\ge \Lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{d}})-16s\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}}-\boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty-\epsilon_1\,,\end{aligned}$$

where $\Lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{d}})$ denotes the minimal eigenvalue of $\boldsymbol{\Sigma}_{\boldsymbol{d}}$. The right-hand side above does not depend on $\boldsymbol{\delta}$, so we may take the infimum of the left-hand side above over $\boldsymbol{\delta}\in\mathcal{C}(S)$ to write, for any $\epsilon_1,\epsilon_2>0$ as in the statement of the present Lemma,

$$\begin{aligned}&\mathrm{P}\left\{\phi^2(\widehat{\boldsymbol{D}},S)<\Lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{d}})-\epsilon_2-\epsilon_1\right\}\\ &\qquad\le\ \mathrm{P}\big\{16s(2m_{\mathbf{A}}\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}+\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}^2)\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty>\epsilon_1\big\}\\ &\qquad\quad+\mathrm{P}\big\{16s\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}}-\boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty>\epsilon_2\big\}\,,\end{aligned}$$

as claimed. □

Lemma B.6 may be combined with results for the maximum first-stage estimation error $\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}$ and the maximum entry-wise difference $\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}}-\boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty$ to obtain specific bounds for $\phi^2(\widehat{\boldsymbol{D}},S_{\boldsymbol{\beta}})$ under different error and design matrix regimes, such as in Lemma 4.6.

Proof of Lemma 4.6. Set ϵ_1 in the statement of Lemma B.6 as

$$\epsilon_1\ =\ 128s_{\boldsymbol{\beta}}(m_{\mathbf{A}}\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2}r_{\mathbf{A}}+2\frac{s_{\mathbf{A}}^2}{\phi_{\mathbf{A}}^4}r_{\mathbf{A}}^2)\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\,,$$

so that

$$\mathrm{P}\big\{16s_{\boldsymbol{\beta}}(2m_{\mathbf{A}}\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}+\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}^2)\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty>\epsilon_1\big\}\ \le\ \mathrm{P}\big\{\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}>4\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2}r_{\mathbf{A}}\big\}\,.$$

The present choices of tuning parameters, along with the estimates of Lemma B.3 and Lemma B.6, entail that

$$\begin{aligned}&\mathrm{P}\left\{\phi^2(\widehat{\boldsymbol{D}},S_{\boldsymbol{\beta}})<\Lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{d}})-\epsilon_2-128s_{\boldsymbol{\beta}}(m_{\mathbf{A}}\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2}r_{\mathbf{A}}+2\frac{s_{\mathbf{A}}^2}{\phi_{\mathbf{A}}^4}r_{\mathbf{A}}^2)\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\right\}\\ &\quad\le\ ep_{\boldsymbol{z}}^{\,2-C_0\min_{j\in[p_{\boldsymbol{x}}]}\{c_j^2/\|\boldsymbol{v}^j\|_{\psi_2}^2\}}+\mathrm{P}\big\{16s\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}}-\boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty>\epsilon_2\big\}\,.\end{aligned}$$

Cite the growth assumptions of the present lemma to observe that, for n sufficiently large, we have $m_{\mathbf{A}}\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2}r_{\mathbf{A}}+2\big(\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2}\big)^2r_{\mathbf{A}}^2\le 3m_{\mathbf{A}}\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2}r_{\mathbf{A}}$, from which it follows that, for such n,

$$\begin{aligned}&\mathrm{P}\big\{\phi^2(\widehat{\boldsymbol{D}},S_{\boldsymbol{\beta}})<\Lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{d}})\ -\ \epsilon_2-384m_{\mathbf{A}}s_{\boldsymbol{\beta}}\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2}c_{\boldsymbol{V}}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty^{3/2}\sqrt{(\log p_{\boldsymbol{z}})/n}\big\}\\ &\qquad\qquad\le\ ep_{\boldsymbol{z}}^{\,2-C_0\min_{j\in[p_{\boldsymbol{x}}]}\{c_j^2/\|\boldsymbol{v}^j\|_{\psi_2}^2\}}+\mathrm{P}\big\{16s_{\boldsymbol{\beta}}\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}}-\boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty>\epsilon_2\big\}\,,\end{aligned}$$

Cite a slight modification of Lemma C.5 to conclude that

$$\begin{aligned}&\mathrm{P}\big\{\phi^2(\widehat{\boldsymbol{D}},S_{\boldsymbol{\beta}})<\Lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{d}})\ -\ \big(a+384m_{\mathbf{A}}s_{\boldsymbol{\beta}}\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2}c_{\boldsymbol{V}}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty^{3/2}\big)\sqrt{(\log p_{\boldsymbol{z}})/n}\big\}\\ &\qquad\qquad\le\ ep_{\boldsymbol{z}}^{\,2-C_0\min_{j\in[p_{\boldsymbol{x}}]}\{c_j^2/\|\boldsymbol{v}^j\|_{\psi_2}^2\}}+2p_{\boldsymbol{x}}^{2-a^2/(6e^2\kappa^2)}\,,\end{aligned}$$

where $a>0$ is a controlled quantity, as claimed. □

B.3. Materials required for Section 4.3

Lemmas B.7, B.9, B.11, and B.13 provide finite-sample bounds for the quantities $\|\boldsymbol{f}_\ell\|_\infty$ for $\ell \in [4]$ that are generic over various noise regimes. We present them separately for the sake of modularity with respect to such assumptions. Lemmas B.8, B.10, B.12, and B.14 in turn provide specific rates for the $\|\boldsymbol{f}_\ell\|_\infty$ under the sub-Gaussian noise regime of Assumption 2.4.

Lemma B.7 (*Control of $\boldsymbol{f}_1$*). *Suppose that Assumption 4.3 and Conditions (2), (3) of Assumption 4.7 hold and that $\widehat{\boldsymbol{\Theta}}$ is chosen according to Assumption 4.8. Then, on the set $\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}}) \cap \mathcal{T}_{\boldsymbol{\Theta}}(\mu)$, where $\lambda_{\boldsymbol{u}} > 0$ is arbitrary, the remainder term*

$$\boldsymbol{f}_1 \;=\; (\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})^\top \boldsymbol{D}^\top \boldsymbol{u}/\sqrt{n}$$

satisfies

$$\|\boldsymbol{f}_1\|_\infty \;\leq\; 2^{-q}\sqrt{n} m_{\mathbf{A}} c_q (m_{\boldsymbol{\Theta}}\mu)^{1-q} s_{\boldsymbol{\Theta}} \lambda_{\boldsymbol{u}}\,,$$

where c_q is as in Lemma A.2.

Proof of Lemma B.7. Lemma A.2 entails that, on the set $\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$,

$$\max_{j\in[p_{\boldsymbol{x}}]} \|\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\|_1 \;\leq\; 2c_q(2m_{\boldsymbol{\Theta}}\mu)^{1-q} s_{\boldsymbol{\Theta}}\,.$$

We therefore find that

$$\begin{aligned}
\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})^\top \boldsymbol{D}^\top \boldsymbol{u}/\sqrt{n}\|_\infty \;&\leq\; \sqrt{n}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{L_1} \|\boldsymbol{D}^\top \boldsymbol{u}/n\|_\infty \\
&\leq\; \sqrt{n}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{L_1} \|\mathbf{A}\|_{L_1} \|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty \\
&\leq\; \sqrt{n} m_{\mathbf{A}} \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{L_1} \|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty \\
&\leq\; 2\sqrt{n} m_{\mathbf{A}} c_q (2m_{\boldsymbol{\Theta}}\mu)^{1-q} s_{\boldsymbol{\Theta}} \|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty\,.
\end{aligned}$$

On the set $\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})$ we have $\|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty \leq \lambda_{\boldsymbol{u}}/4$. From this bound and the previous display we conclude that, on the set $\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}}) \cap \mathcal{T}_{\boldsymbol{\Theta}}(\mu)$,

$$\|(\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}})^\top \boldsymbol{D}^\top \boldsymbol{u}/\sqrt{n}\|_\infty \;\leq\; 2^{-q}\sqrt{n} m_{\mathbf{A}} c_q (m_{\boldsymbol{\Theta}}\mu)^{1-q} s_{\boldsymbol{\Theta}} \lambda_{\boldsymbol{u}}\,,$$

as claimed. □

Lemma B.8 (*Control of $\boldsymbol{f}_1$, sub-Gaussian Noise*). *Suppose that (i) Assumption 4.3 and Conditions (2), (3) of Assumption 4.7 hold and (ii) Assumption 2.4 holds. Choose $\widehat{\boldsymbol{\Theta}}$ according to Assumption 4.8. Then,*

$$\mathrm{P}\left\{\|\boldsymbol{f}_1\|_\infty > 2^{-q} m_{\mathbf{A}} c_q c_{\boldsymbol{u}} (m_{\boldsymbol{\Theta}}\mu)^{1-q} s_{\boldsymbol{\Theta}} \sqrt{m_{\boldsymbol{Z}} \log p_{\boldsymbol{z}}}\right\} \;\leq\; e p_{\boldsymbol{z}}^{1 - c_{\boldsymbol{u}}^2 C_0/\tau_{\boldsymbol{u}}^2} + \mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu)^{\complement}\,,$$

where $c_{\boldsymbol{u}}$ is as in Definition 4.4 and C_0 is as defined in Lemma C.1. If Conditions (1), (4), (5), and (6)a of Assumption 4.7 also hold, then $\|\boldsymbol{f}_1\|_\infty = o_{\mathrm{P}}(1)$.

Proof of Lemma B.8. Lemma B.7 entails that

$$\mathrm{P}\left\{\|\boldsymbol{f}_1\|_\infty > 2^{-q}\sqrt{n} m_{\mathbf{A}} c_q (m_{\boldsymbol{\Theta}}\mu)^{1-q} s_{\boldsymbol{\Theta}} \lambda_{\boldsymbol{u}}\right\} \;\leq\; \mathrm{P}\,\mathcal{T}_{\boldsymbol{u}}^{\complement} + \mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu)^{\complement}\,.$$

Substitute $\lambda_{\boldsymbol{u}}$ chosen according to Definition 4.4 into the display above and cite the estimate of Lemma C.2 to deduce the original claim. □

Lemma B.9 (*Control of $\boldsymbol{f}_2$*). *Suppose that Assumption 4.3 and Condition (3) of Assumption 4.7 hold. Choose $\widehat{\boldsymbol{\Theta}}$ according to Assumption 4.8, set $r_{\mathbf{A}}$ according to Definition 4.4, and let $\lambda_{\boldsymbol{u}} > 0$ be arbitrary. Then, on the set $\mathcal{T}_{\boldsymbol{V}}(\boldsymbol{r}) \cap \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}}) \cap \mathcal{T}_{\boldsymbol{\Theta}}(\mu)$, the remainder term*

$$\boldsymbol{f}_2 \;=\; \widehat{\boldsymbol{\Theta}}^\top (\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top \boldsymbol{u}/\sqrt{n}$$

satisfies

$$\|\boldsymbol{f}_2\|_\infty \;\leq\; \sqrt{n} m_{\boldsymbol{\Theta}} \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} \lambda_{\boldsymbol{u}}$$

for n sufficiently large.

Proof of Lemma B.9. Observe that

$$\widehat{\boldsymbol{\Theta}}^\top (\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top \boldsymbol{u}/\sqrt{n} \;=\; \sqrt{n}\widehat{\boldsymbol{\Theta}}^\top (\widehat{\mathbf{A}} - \mathbf{A})^\top (\boldsymbol{Z}^\top \boldsymbol{u}/n)\,.$$

On the set $\mathcal{T}_\Theta(\mu)$, each row $\boldsymbol{\theta}$ is feasible for Problem 3.2. Then, $\|\hat{\boldsymbol{\theta}}_j\|_1 \le \|\boldsymbol{\theta}_j\|_1$ for each $j \in [p_{\boldsymbol{x}}]$ by specification. Lemmas B.2 and C.2 then entail that, on the set $\mathcal{T}_{\mathbf{A}}(\boldsymbol{r}) \cap \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}}) \cap \mathcal{T}_\Theta(\mu)$,

$$\begin{aligned}\|\widehat{\Theta}^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top \boldsymbol{u}/\sqrt{n}\|_\infty &\le \max_{j,k\in[p_{\boldsymbol{x}}]} \sqrt{n}\|\hat{\boldsymbol{\theta}}_j\|_1 \|\hat{\boldsymbol{\alpha}}^k - \boldsymbol{\alpha}^k\|_1 \|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty \\ &\le \sqrt{n} m_\Theta \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} \lambda_{\boldsymbol{u}} ,\end{aligned}$$

as claimed. □

Lemma B.10 (*Control of $\boldsymbol{f}_2$, sub-Gaussian Noise*)**.** *Suppose that (i) Assumption 4.3 and Condition (3) of Assumption 4.7 hold and (ii) Assumption 2.4 holds. Choose $\widehat{\Theta}$, $\boldsymbol{r} = (r_1, \dots, r_{p_{\boldsymbol{x}}})$, and $r_{\mathbf{A}}$ according to Assumption 4.8. Then,*

$$\begin{aligned}&\mathrm{P}\big\{\|\boldsymbol{f}_2\|_\infty > m_\Theta c_{\boldsymbol{V}} c_{\boldsymbol{u}} m_{\boldsymbol{Z}} \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} \log p_{\boldsymbol{z}}/\sqrt{n}\big\} \\ &\quad\le e p_{\boldsymbol{z}}^{2 - c_{\boldsymbol{V}}^2 C_0/\tau_{\boldsymbol{V}}^2} + e p_{\boldsymbol{z}}^{1 - c_{\boldsymbol{u}}^2 C_0/\tau_{\boldsymbol{u}}^2} + \mathrm{P}\big\{\|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty > m_{\boldsymbol{Z}}\big\} + \mathrm{P}\,\mathcal{T}_\Theta(\mu)^{\complement} ,\end{aligned}$$

where $c_{\boldsymbol{V}}$ is as in Definition 4.4, $c_{\boldsymbol{u}}$ is as in Definition 4.4, and C_0 is as defined in Lemma C.1. Consequently, if Conditions (1), (4), (5), and (6)b of Assumption 4.7 also hold, then $\|\boldsymbol{f}_2\|_\infty = o_{\mathrm{P}}(1)$.

Proof of Lemma B.10. Lemma B.9 entails that

$$\mathrm{P}\left\{\|\boldsymbol{f}_2\|_\infty > \frac{1}{4}\sqrt{n} m_\Theta \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} \lambda_{\boldsymbol{u}}\right\} \le \mathrm{P}\,\mathcal{T}_{\mathbf{A}}(\boldsymbol{r})^{\complement} + \mathrm{P}\,\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})^{\complement} + \mathrm{P}\,\mathcal{T}_\Theta(\mu)^{\complement} .$$

Substitute the present choices of $r_{\mathbf{A}}$ and $\lambda_{\boldsymbol{u}}$ into the previous display and cite the estimates of Lemmas C.2 and C.3 to deduce the original claim. □

Lemma B.11 (*Control of $\boldsymbol{f}_3$*)**.** *Suppose that Assumptions 4.3 and 4.3 and Conditions (2) and (3) of Assumption 4.7 hold. Choose $\widehat{\Theta}$ according to Assumption 4.8; let $\boldsymbol{r} = (r_1, \dots, r_{p_{\boldsymbol{x}}}) > \mathbf{0}$, $\lambda_{\boldsymbol{u}} > 0$, and $\lambda_{\boldsymbol{V}} > 0$ be arbitrary. Set*

$$r_\beta = 16 \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} \|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty \big(4 m_\beta \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} + m_{\mathbf{A}}\big) + \big(4 \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} + m_{\mathbf{A}}\big)\big(m_\beta \lambda_{\boldsymbol{V}} + \lambda_{\boldsymbol{u}}\big) .$$

Then, on the set

$$\mathcal{T}_{\boldsymbol{V}}(\boldsymbol{r}) \cap \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}}) \cap \mathcal{T}_\Theta(\mu) ,$$

the remainder term

$$\boldsymbol{f}_3 = \widehat{\Theta}\widehat{\boldsymbol{D}}^\top (\boldsymbol{X} - \widehat{\boldsymbol{D}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})/\sqrt{n}$$

satisfies

$$\|\boldsymbol{f}_3\|_\infty \le 8 m_\Theta m_{\mathbf{A}} \sqrt{n} (4 \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} + \lambda_{\boldsymbol{u}}) \frac{\boldsymbol{s}_\beta}{\phi_\beta^2} r_\beta .$$

Proof of Lemma B.11. We first observe that

$$\|\widehat{\Theta}\widehat{\boldsymbol{D}}^\top (\boldsymbol{X} - \widehat{\boldsymbol{D}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})/\sqrt{n}\|_\infty \le \sqrt{n}\|\widehat{\Theta}\|_{L_1} \|\widehat{\boldsymbol{D}}^\top (\boldsymbol{X} - \widehat{\boldsymbol{D}})/n\|_\infty \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 .$$

Now,

$$\begin{aligned}\widehat{\boldsymbol{D}}^\top (\boldsymbol{X} - \widehat{\boldsymbol{D}})/n &= \widehat{\boldsymbol{D}}^\top (\boldsymbol{D} + \boldsymbol{u} - \widehat{\boldsymbol{D}})/n \\ &= \widehat{\boldsymbol{D}}^\top (\boldsymbol{D} - \widehat{\boldsymbol{D}})/n + \widehat{\boldsymbol{D}}^\top \boldsymbol{u}/n \\ &= \widehat{\mathbf{A}}^\top (\boldsymbol{Z}^\top \boldsymbol{Z}/n)(\mathbf{A} - \widehat{\mathbf{A}}) + \widehat{\boldsymbol{D}}^\top \boldsymbol{u}/n .\end{aligned}$$

For the first term on the right-hand side above, write

$$\begin{aligned}\|\widehat{\mathbf{A}}^\top (\boldsymbol{Z}^\top \boldsymbol{Z}/n)(\mathbf{A} - \widehat{\mathbf{A}})\|_\infty &\le \|\mathbf{A}^\top (\boldsymbol{Z}^\top \boldsymbol{Z}/n)(\mathbf{A} - \widehat{\mathbf{A}})\|_\infty \\ &\quad + \|(\widehat{\mathbf{A}} - \mathbf{A})^\top (\boldsymbol{Z}^\top \boldsymbol{Z}/n)(\mathbf{A} - \widehat{\mathbf{A}})\|_\infty \\ &\le \|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty [m_{\mathbf{A}} \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} + \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}^2] .\end{aligned}$$

On the set $\mathcal{T}_{\boldsymbol{V}}(\boldsymbol{r})$, the right-hand side above is less than or equal to $2 m_{\mathbf{A}} \cdot \|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}$ for n sufficiently large by Lemma B.2 and the hypotheses of the present lemma. For the second term on the right-hand side of two displays previous, write

$$\|\widehat{\boldsymbol{D}}^\top \boldsymbol{u}/n\|_\infty = \|\widehat{\mathbf{A}}^\top \boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty$$

$$= \|(\mathbf{A} + [\widehat{\mathbf{A}} - \mathbf{A}])^\top (\boldsymbol{Z}^\top \boldsymbol{u}/n)\|_\infty$$
$$\leq 2m_{\mathbf{A}} \|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty ,$$

where the final line holds on the set $\mathcal{T}_{\boldsymbol{V}}(\boldsymbol{r})$ for such n by Lemma B.2. Thus, on the set $\mathcal{T}_{\boldsymbol{V}}(\boldsymbol{r}) \cap \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})$, we have

$$\|\widehat{\boldsymbol{D}}^\top (\boldsymbol{X} - \widehat{\boldsymbol{D}})/n\|_\infty \leq 2m_{\mathbf{A}} (\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} + \|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty)$$
$$\leq 2m_{\mathbf{A}} (4\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} + \lambda_{\boldsymbol{u}}) ,$$

where the latter substitutions are justified by Lemma B.2 and the definition of $\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})$.

On the set $\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$, each row $\boldsymbol{\theta}$ is feasible for Problem 3.2. Then, $\|\hat{\boldsymbol{\theta}}_j\|_1 \leq \|\boldsymbol{\theta}_j\|_1$ for each $j \in [p_{\boldsymbol{x}}]$ by specification.

Finally, Lemma B.5 entails that, for the present choice of $r_{\boldsymbol{\beta}}$,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \leq 4 \frac{\boldsymbol{s}_{\boldsymbol{\beta}}}{\phi_{\boldsymbol{\beta}}^2} r_{\boldsymbol{\beta}}$$

on the set $\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}}) \cap \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})$.

Combining the foregoing results, we see that, on the set

$$\mathcal{T}_{\mathbf{A}}(\boldsymbol{r}) \cap \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}}) \cap \mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}}) \cap \mathcal{T}_{\boldsymbol{\Theta}}(\mu) ,$$

it holds that

$$\|\widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{D}}^\top (\boldsymbol{X} - \widehat{\boldsymbol{D}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})/\sqrt{n}\|_\infty \leq 8m_{\boldsymbol{\Theta}} m_{\mathbf{A}} \sqrt{n} (4\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} + \lambda_{\boldsymbol{u}}) \frac{\boldsymbol{s}_{\boldsymbol{\beta}}}{\phi_{\boldsymbol{\beta}}^2} r_{\boldsymbol{\beta}} .$$

Now note that, under the present choices of $\boldsymbol{r}$ and $\lambda_{\boldsymbol{V}}$, the set $\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}})$ is contained in $\mathcal{T}_{\boldsymbol{V}}(\boldsymbol{r})$. □

Lemma B.12 (*Control of $\boldsymbol{f}_3$, sub-Gaussian Noise*)**.** *Suppose that Conditions (1), (2), and (3) of Assumption 4.7 hold and (ii) Assumption 2.4 holds. Choose $\widehat{\boldsymbol{\Theta}}$, $\widehat{\mathbf{A}}$ and $\hat{\boldsymbol{\beta}}$ according to Assumption 4.8. Then,*

$$\mathrm{P}\{\|\boldsymbol{f}_3\|_\infty > 8m_{\boldsymbol{\Theta}} m_{\mathbf{A}} \sqrt{n} (4m_{\boldsymbol{Z}} \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} + \lambda_{\boldsymbol{u}}) \frac{\boldsymbol{s}_{\boldsymbol{\beta}}}{\phi_{\boldsymbol{\beta}}^2} r_{\boldsymbol{\beta}}\}$$
$$\leq e p_{\boldsymbol{z}}^{2 - C_0 \min_{j \in [p_{\boldsymbol{x}}]} \{c_j^2 / \|\boldsymbol{v}^j\|_{\psi_2}^2\}} + e p_{\boldsymbol{z}}^{1 - c_{\boldsymbol{u}}^2 C_0 / \tau_{\boldsymbol{u}}^2} + \mathrm{P}\, \mathcal{T}_{\boldsymbol{\Theta}}(\mu)^{\complement} ,$$

where $c_{\boldsymbol{V}}$ is as in Definition 4.4, $c_{\boldsymbol{u}}$ is as in Definition 4.4, and C_0 is as defined in Lemma C.1. Consequently, if Conditions (4), (5), and (b) of Assumption 4.7 also hold, then $\|\boldsymbol{f}_3\|_1 = o_{\mathrm{P}}(1)$.

Proof of Lemma B.12. Lemma B.11 entails that

$$\mathrm{P}\{\|\boldsymbol{f}_3\|_\infty > 8m_{\boldsymbol{\Theta}} m_{\mathbf{A}} \sqrt{n} (4m_{\boldsymbol{Z}} \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} + \lambda_{\boldsymbol{u}}) \frac{\boldsymbol{s}_{\boldsymbol{\beta}}}{\phi_{\boldsymbol{\beta}}^2} r_{\boldsymbol{\beta}}\}$$
$$\leq \mathrm{P}\, \mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}})^{\complement} + \mathrm{P}\, \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})^{\complement} + \mathrm{P}\, \mathcal{T}_{\boldsymbol{\Theta}}(\mu)^{\complement} .$$

Substitute the present choices of tuning parameters into the display above and cite the estimates of Lemmas C.2 and C.3, to deduce the first claim. Expand the present choices of tuning parameters to find

$$8m_{\boldsymbol{\Theta}} m_{\mathbf{A}} \sqrt{n} (4m_{\boldsymbol{Z}} \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} + \lambda_{\boldsymbol{u}}) \frac{\boldsymbol{s}_{\boldsymbol{\beta}}}{\phi_{\boldsymbol{\beta}}^2} r_{\boldsymbol{\beta}}$$
$$\lesssim s_{\mathbf{A}}^3 s_{\boldsymbol{\beta}} (\log p_{\boldsymbol{z}})^{3/2}/n + s_{\mathbf{A}}^2 s_{\boldsymbol{\beta}} (\log p_{\boldsymbol{z}})/\sqrt{n}$$
$$+ s_{\mathbf{A}}^2 s_{\boldsymbol{\beta}} (\log p_{\boldsymbol{z}})^{3/2}/n + s_{\mathbf{A}} s_{\boldsymbol{\beta}} \log p_{\boldsymbol{z}}/\sqrt{n} ,$$

from which the latter claim follows. □

Lemma B.13 (*Control of $\boldsymbol{f}_4$*)**.** *Suppose that Assumption 4.3 and Condition (3) of Assumption 4.7 hold. Choose $\widehat{\boldsymbol{\Theta}}$ according to Assumption 4.8; let $\lambda_{\boldsymbol{V}}, \lambda_{\boldsymbol{u}} > 0$ be arbitrary; set*

$$r_{\boldsymbol{\beta}} = 16 \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \big(4m_{\boldsymbol{\beta}} \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} + m_{\mathbf{A}}\big) + \big(4 \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} + m_{\mathbf{A}}\big)\big(m_{\boldsymbol{\beta}} \lambda_{\boldsymbol{V}} + \lambda_{\boldsymbol{u}}\big) .$$

Then, on the set $\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}}) \cap \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}}) \cap \mathcal{T}_{\boldsymbol{\Theta}}(\mu)$, the remainder term

$$\boldsymbol{f}_4 = \sqrt{n} (\widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$$

satisfies

$$\|\boldsymbol{f}_4\|_\infty \leq 4\sqrt{n} \mu \frac{\boldsymbol{s}_{\boldsymbol{\beta}}}{\phi_{\boldsymbol{\beta}}^2} r_{\boldsymbol{\beta}} .$$

Please cite this article as: D. Gold, J. Lederer and J. Tao, Inference for high-dimensional instrumental variables regression. Journal of Econometrics (2020), https://doi.org/10.1016/j.jeconom.2019.09.009.

Proof of Lemma B.13. Note first that

$$\begin{aligned}\|\sqrt{n}(\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_\infty &\leq \sqrt{n}\|\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1 \\ &\leq \sqrt{n}\mu\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1 ,\end{aligned}$$

where the latter inequality follows from the specification of $\widehat{\boldsymbol{\Theta}}$ and the fact that Problem 3.2 is feasible given a. By Lemma B.5, on the set $\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}}) \cap \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}}) \cap \mathcal{T}_{\boldsymbol{\Theta}}(\mu)$,

$$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1 \leq 4\frac{\boldsymbol{s}_{\boldsymbol{\beta}}}{\phi^2_{\hat{\boldsymbol{\beta}}}} r_{\boldsymbol{\beta}} .$$

Combine the two previous displays to deduce the original claim. □

Lemma B.14 (*Control of $\boldsymbol{f}_4$, Gaussian Noise*)**.** *Suppose that (i) Assumption 4.3 and Condition (3) of Assumption 4.7 hold and (ii) Assumption 2.4 holds. Choose $\widehat{\boldsymbol{\Theta}}$, $\widehat{\mathbf{A}}$ and $\hat{\boldsymbol{\beta}}$ according to Assumption 4.8. Then,*

$$\mathrm{P}\Big\{\|\boldsymbol{f}_4\|_\infty > 4\sqrt{n}\mu\frac{\boldsymbol{s}_{\boldsymbol{\beta}}}{\phi^2_{\hat{\boldsymbol{\beta}}}} r_{\boldsymbol{\beta}}\Big\} \leq ep_{\boldsymbol{z}}^{2-c^2_{\boldsymbol{V}}C_0/\tau^2_{\boldsymbol{V}}} + ep_{\boldsymbol{z}}^{1-c^2_{\boldsymbol{u}}C_0/\tau^2_{\boldsymbol{u}}} + \mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu)^{\complement} ,$$

where $c_{\boldsymbol{V}}$ is as in Definition 4.4, $c_{\boldsymbol{u}}$ is as in Definition 4.4, and C_0 is as defined in Lemma C.1. Consequently, if Conditions (4), (5), and (c) of Assumption 4.7 also hold, then $\|\boldsymbol{f}_4\|_1 = o_{\mathrm{P}}(1)$.

Proof of Lemma B.14. Lemma B.13 entails that

$$\mathrm{P}\Big\{\|\boldsymbol{f}_4\|_\infty > 4\sqrt{n}\mu\frac{\boldsymbol{s}_{\boldsymbol{\beta}}}{\phi^2_{\hat{\boldsymbol{\beta}}}} r_{\boldsymbol{\beta}}\Big\} \leq \mathrm{P}\,\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}})^{\complement} + \mathrm{P}\,\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})^{\complement} + \mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu) .$$

Substitute the present choices of tuning parameters into the display above and cite the estimates of Lemmas C.2 and C.3 to deduce the first claim. Expand the present choices of tuning parameters to find

$$\sqrt{n}\mu\frac{\boldsymbol{s}_{\boldsymbol{\beta}}}{\phi^2_{\hat{\boldsymbol{\beta}}}} r_{\boldsymbol{\beta}} \lesssim \mu s_{\boldsymbol{\beta}}\big(s^2_{\mathbf{A}}\log p_{\boldsymbol{z}}/\sqrt{n} + s_{\mathbf{A}}\sqrt{\log p_{\boldsymbol{z}}}\big) ,$$

from which the second claim follows. □

Proof of Lemma 4.9. The result follows from Lemmas B.8, B.10, B.12, and B.14. □

The following two Lemmas are required for Lemma 4.10.

Lemma B.15. *For $\epsilon < m_{\mathbf{A}}$, it holds that*

$$\begin{aligned}&\mathrm{P}\Big\{\|\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty > t + 3m_{\boldsymbol{\Theta}}m_{\mathbf{A}}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\epsilon\Big\} \\ &\quad\leq \mathrm{P}\Big\{\|\boldsymbol{\Theta}\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty > t\Big\} + \mathrm{P}\Big\{\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} \leq \epsilon\Big\} ,\end{aligned}$$

where $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \mathbb{E}_n[\boldsymbol{d}_i\boldsymbol{d}_i^\top]$.

Proof of Lemma B.15. Note that

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \widehat{\boldsymbol{D}}^\top\widehat{\boldsymbol{D}}/n &= \boldsymbol{D}^\top\boldsymbol{D}/n + (\widehat{\mathbf{A}} - \mathbf{A})^\top\boldsymbol{Z}^\top\boldsymbol{Z}\mathbf{A}/n + \widehat{\mathbf{A}}^\top\boldsymbol{Z}^\top\boldsymbol{Z}(\widehat{\mathbf{A}} - \mathbf{A})/n \\ &= \overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} + (\widehat{\mathbf{A}} - \mathbf{A})^\top\boldsymbol{Z}^\top\boldsymbol{Z}\mathbf{A}/n + \mathbf{A}^\top\boldsymbol{Z}^\top\boldsymbol{Z}(\widehat{\mathbf{A}} - \mathbf{A})/n \\ &\quad + (\widehat{\mathbf{A}} - \mathbf{A})^\top\boldsymbol{Z}^\top\boldsymbol{Z}(\widehat{\mathbf{A}} - \mathbf{A})/n ,\end{aligned}$$

so that

$$\begin{aligned}\|\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty &\leq \|\boldsymbol{\Theta}(\boldsymbol{D}^\top\boldsymbol{D}/n) - \boldsymbol{I}\|_\infty + \|\boldsymbol{\Theta}(\widehat{\mathbf{A}} - \mathbf{A})^\top\boldsymbol{Z}^\top\boldsymbol{Z}\mathbf{A}/n\|_\infty \\ &\quad + \|\boldsymbol{\Theta}\mathbf{A}^\top\boldsymbol{Z}^\top\boldsymbol{Z}(\widehat{\mathbf{A}} - \mathbf{A})/n\|_\infty \\ &\quad + \|\boldsymbol{\Theta}(\widehat{\mathbf{A}} - \mathbf{A})^\top\boldsymbol{Z}^\top\boldsymbol{Z}(\widehat{\mathbf{A}} - \mathbf{A})/n\|_\infty \\ &:= \|\boldsymbol{\Theta}(\boldsymbol{D}^\top\boldsymbol{D}/n) - \boldsymbol{I}\|_\infty + \mathrm{I}_1 + \mathrm{I}_2 + \mathrm{I}_3 .\end{aligned}$$

Note that

$$\begin{aligned}\mathrm{I}_1 = \|\boldsymbol{\Theta}(\widehat{\mathbf{A}} - \mathbf{A})^\top\boldsymbol{Z}^\top\boldsymbol{Z}\mathbf{A}/n\|_\infty &\leq \|\boldsymbol{\Theta}\|_{L_1}\|(\widehat{\mathbf{A}} - \mathbf{A})^\top\boldsymbol{Z}^\top\boldsymbol{Z}\mathbf{A}/n\|_\infty \\ &\leq m_{\boldsymbol{\Theta}}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}\|\mathbf{A}\|_{L_1} \\ &= m_{\boldsymbol{\Theta}}m_{\mathbf{A}}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} .\end{aligned}$$

The same bound holds for I_2 by symmetry of the ℓ_∞ norm under transposition of its argument.

For the term I_3, similar reasoning yields

$$\begin{aligned}\mathrm{I}_3 &= \|\Theta(\widehat{\mathbf{A}}-\mathbf{A})^\top \boldsymbol{Z}^\top \boldsymbol{Z}(\widehat{\mathbf{A}}-\mathbf{A})/n\|_\infty \\ &\le m_\Theta \|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty \|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}^2 = m_\Theta \|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty \|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}^2 .\end{aligned}$$

If $\epsilon < m_{\mathbf{A}}$, then, on the set $\{\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1} \le \epsilon\}$, it holds that $\mathrm{I}_3 \le m_\Theta m_{\mathbf{A}} \cdot \|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty \|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}$. Conclude that

$$\mathrm{P}\big\{\mathrm{I}_1+\mathrm{I}_2+\mathrm{I}_3 > 3m_\Theta m_{\mathbf{A}} \|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty \epsilon\big\} \le \mathrm{P}\{\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1} > \epsilon\}$$

and hence that

$$\begin{aligned}&\mathrm{P}\big\{\|\Theta\widehat{\Sigma}_{\boldsymbol{d}}-\boldsymbol{I}\|_\infty > t + 3m_\Theta m_{\mathbf{A}} \|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty \epsilon\big\} \\ &\quad \le \mathrm{P}\big\{\|\Theta\overline{\Sigma}_{\boldsymbol{d}}-\boldsymbol{I}\|_\infty > t\big\} + \mathrm{P}\{\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1} > \epsilon\},\end{aligned}$$

as claimed. □

Lemma B.16. *Suppose that the $\boldsymbol{z}_i$ satisfy Assumption 2.2. Set*

$$\mu = m_\Theta a\sqrt{\log(p_{\boldsymbol{x}})/n} + 3m_\Theta m_{\mathbf{A}} \|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty \epsilon$$

where $a > 0$ and $\epsilon > 0$ are controlled quantities. Then

$$\mathrm{P}\mathcal{T}_\Theta(\mu) \le 2p_{\boldsymbol{x}}^{2-a^2/(6e^2\kappa^2)} + \mathrm{P}\{\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1} > \epsilon\},$$

where $\kappa = m_{\mathbf{A}}^2(2\tau_{\boldsymbol{z}}^2 + \|\Sigma_{\boldsymbol{z}}\|_\infty/\log 2)$.

Proof of Lemma B.16. Lemma B.15 entails that

$$\begin{aligned}&\mathrm{P}\big\{\|\Theta\widehat{\Sigma}_{\boldsymbol{d}}-\boldsymbol{I}\|_\infty > t + 3m_\Theta m_{\mathbf{A}} \|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty \epsilon\big\} \\ &\quad \le \underbrace{\mathrm{P}\big\{\|\Theta\overline{\Sigma}_{\boldsymbol{d}}-\boldsymbol{I}\|_\infty > t\big\}}_{\mathrm{I}_1(t)} + \mathrm{P}\{\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1} > \epsilon\}.\end{aligned}$$

For $t > 0$. Now write

$$\Theta\overline{\Sigma}_{\boldsymbol{d}} - \boldsymbol{I} = \Theta\Sigma_{\boldsymbol{d}} + \Theta(\overline{\Sigma}_{\boldsymbol{d}} - \Sigma_{\boldsymbol{d}}) - \boldsymbol{I} = \Theta(\overline{\Sigma}_{\boldsymbol{d}} - \Sigma_{\boldsymbol{d}})$$

to infer that

$$\|\Theta\overline{\Sigma}_{\boldsymbol{d}}-\boldsymbol{I}\|_\infty = \|\Theta(\overline{\Sigma}_{\boldsymbol{d}} - \Sigma_{\boldsymbol{d}})\|_\infty \le \|\Theta\|_{L_1}\|\overline{\Sigma}_{\boldsymbol{d}} - \Sigma_{\boldsymbol{d}}\|_\infty \le m_\Theta\|\overline{\Sigma}_{\boldsymbol{d}} - \Sigma_{\boldsymbol{d}}\|_\infty$$

and hence that

$$\mathrm{I}_1(t) \le \mathrm{P}\big\{m_\Theta\|\overline{\Sigma}_{\boldsymbol{d}} - \Sigma_{\boldsymbol{d}}\|_\infty > t\big\}$$

Choose $t = m_\Theta a\sqrt{\log(p_{\boldsymbol{x}})/n}$ for a controlled quantity $a > 0$ and cite a slight modification of Lemma C.5 to find

$$\begin{aligned}\mathrm{I}_1\big(m_\Theta a\sqrt{\log(p_{\boldsymbol{x}})/n}\big) &\le \mathrm{P}\big\{\|\overline{\Sigma}_{\boldsymbol{d}} - \Sigma_{\boldsymbol{d}}\|_\infty > a\sqrt{\log(p_{\boldsymbol{x}})/n}\big\} \\ &\le 2p_{\boldsymbol{x}}^{2-a^2/(6e^2\kappa^2)},\end{aligned}$$

where $\kappa = m_{\mathbf{A}}^2(2\tau_{\boldsymbol{z}}^2 + \|\Sigma_{\boldsymbol{z}}\|_\infty/\log 2)$. Substitute the above bound into the first display of the present proof to conclude that

$$\begin{aligned}&\mathrm{P}\big\{\|\Theta\widehat{\Sigma}_{\boldsymbol{d}}-\boldsymbol{I}\|_\infty > m_\Theta a\sqrt{\log(p_{\boldsymbol{x}})/n} + 3m_\Theta m_{\mathbf{A}} \|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty \epsilon\big\} \\ &\quad \le 2p_{\boldsymbol{x}}^{2-a^2/(6e^2\kappa^2)} + \mathrm{P}\{\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1} > \epsilon\},\end{aligned}$$

as claimed. □

Proof of Lemma 4.10. Set $\epsilon = 4\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} c_{\boldsymbol{V}}(\|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty(\log p_{\boldsymbol{z}})/n)^{1/2}$, plug this choice into the result of Lemma B.16 and cite the estimate of Lemma B.3 to find

$$\begin{aligned}&\mathrm{P}\big\{\|\Theta\widehat{\Sigma}_{\boldsymbol{d}}-\boldsymbol{I}\|_\infty > \frac{m_\Theta}{\sqrt{n}}\big(a\sqrt{\log p_{\boldsymbol{x}}} + 12m_{\mathbf{A}}\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} c_{\boldsymbol{V}}\|\widehat{\Sigma}_{\boldsymbol{z}}\|_\infty^{3/2}\sqrt{\log p_{\boldsymbol{z}}}\big)\big\} \\ &\quad \le 2p_{\boldsymbol{x}}^{2-a^2/(6e^2\kappa^2)} + ep_{\boldsymbol{z}}^{2-C_0\min_{j\in[p_{\boldsymbol{x}}]}\{c_j^2/\|\boldsymbol{v}^j\|_{\psi_2}^2\}}\end{aligned}$$

for n sufficiently large, as claimed. □

Please cite this article as: D. Gold, J. Lederer and J. Tao, Inference for high-dimensional instrumental variables regression. Journal of Econometrics (2020), https://doi.org/10.1016/j.jeconom.2019.09.009.

Proof of Lemma 4.11. (i) We first show that $\hat{\sigma}_u - \sigma_u = o_{\mathrm{P}}(1)$. To begin, write

$$\begin{aligned}\hat{\sigma}_u^2 \;&=\; \|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|_2^2/n \;=\; \|\boldsymbol{u} + \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2/n \\ &=\; \|\boldsymbol{u}\|_2^2/n + 2\underbrace{\langle \boldsymbol{u}, \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\rangle/n}_{\mathrm{I}_1} + \underbrace{\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2/n}_{\mathrm{I}_2}\,.\end{aligned}$$

It follows that

$$\begin{aligned}(\|\boldsymbol{u}\|_2^2/n - \sigma_u^2) - 2|\mathrm{I}_1| + \mathrm{I}_2 \;&\le\; \hat{\sigma}_u^2 - \sigma_u^2 \\ &\le\; |\hat{\sigma}_u^2 - \sigma_u^2| \;\le\; \big|\|\boldsymbol{u}\|_2^2/n - \sigma_u^2\big| + 2|\mathrm{I}_1| + \mathrm{I}_2\,.\end{aligned}$$

We claim that I_1 and I_2 are each $o_{\mathrm{P}}(1)$. It then follows that $\hat{\sigma}_u^2 - \sigma_u^2 = o_{\mathrm{P}}(1)$, since $\|\boldsymbol{u}\|_2^2/n - \sigma_u^2 = o_{\mathrm{P}}(1)$ by assumption. To show the claim, note first that

$$\mathrm{I}_1^2 \;\le\; \big(\|\boldsymbol{u}\|_2^2/n\big)\big(\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2^2/n\big) \;=\; \big(\|\boldsymbol{u}\|_2^2/n\big)\mathrm{I}_2\,.$$

From the assumption that $\|\boldsymbol{u}\|_2^2/n - \sigma_u^2 = o_{\mathrm{P}}(1)$ we infer that $\|\boldsymbol{u}\|_2^2/n = O_{\mathrm{P}}(1)$. Thus, it suffices to show that $\mathrm{I}_2 = o_{\mathrm{P}}(1)$. To this end, note that

$$\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2 \;\le\; \|\boldsymbol{X}\|_{L_2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$$

hence $\mathrm{I}_2 \;\le\; \|\boldsymbol{X}\|_{L_2}^2/n\big(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1^2\big)$. Now cite Lemma C.6 to obtain that

$$\|\boldsymbol{X}\|_{L_2}^2/n \;=\; O_{\mathrm{P}}\Big(\max_{j\in[p_{\boldsymbol{X}}]}\mathrm{E}[x_{ij}^2] + a\sqrt{(\log p_{\boldsymbol{X}})/n}\Big)$$

for a suitably chosen a of constant order. It then follows from the growth conditions of the present lemma and the estimate of Lemma 4.5 for $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$ that $\mathrm{I}_2^2 = O_{\mathrm{P}}(1)o_{\mathrm{P}}(1) = o_{\mathrm{P}}(1)$, as required. Thus, $\hat{\sigma}_u^2 - \sigma_u^2 = o_{\mathrm{P}}(1)$. To show as much for $\hat{\sigma}_u - \sigma_u$, it suffices to show that $\mathrm{P}\{|\hat{\sigma}_u - \sigma_u| > \epsilon\} \to 0$ for each $\epsilon > 0$. We will show that $\mathrm{P}\{\hat{\sigma}_u - \sigma_u > \epsilon\} \to 0$; the matching limit follows from an analogous argument. Fix $\epsilon > 0$ and note that

$$\begin{aligned}\mathrm{P}\{\hat{\sigma}_u - \sigma_u > \epsilon\} \;&=\; \mathrm{P}\{\hat{\sigma}_u > \sigma_u + \epsilon\} \\ &=\; \mathrm{P}\{\hat{\sigma}_u^2 > \sigma_u^2 + 2\sigma_u\epsilon + \epsilon^2\} \\ &=\; \mathrm{P}\{\hat{\sigma}_u^2 - \sigma_u^2 > 2\sigma_u\epsilon + \epsilon^2\} \;\to\; 0\end{aligned}$$

as $n \to \infty$ since, by Assumption 2.4, σ_u is bounded strictly away from zero uniformly in n. The previous display and the matching limit for $\mathrm{P}\{\sigma_u - \hat{\sigma}_u > \epsilon\}$ entail that $\hat{\sigma}_u - \sigma_u = o_{\mathrm{P}}(1)$, as claimed.

(ii) We now show that $\hat{\omega}_j^2 = \hat{\sigma}_u^2\widehat{\boldsymbol{\Theta}}_{jj}$ satisfies $\hat{\omega}_j - \omega_j = o_{\mathrm{P}}(1)$. We first show as much for $\hat{\omega}_j^2 - \omega_j^2$; the original claim then follows from reasoning analogous to that above. To this end, note that, since the noise components u_i are homoscedastic, we have

$$\omega_j^2 \;=\; \mathrm{E}[\langle\boldsymbol{\theta}_j, \boldsymbol{d}_i\rangle^2 u_i^2] \;=\; \mathrm{E}\big[\langle\boldsymbol{\theta}_j, \boldsymbol{d}_i\rangle^2\mathrm{E}\big[[u_i^2 \mid \boldsymbol{z}_i]\big]\big] \;=\; \sigma_u^2\boldsymbol{\Theta}_{jj}\,.$$

Now write

$$\begin{aligned}\hat{\omega}_j^2 - \omega_j^2 \;&=\; \hat{\sigma}_u^2\widehat{\boldsymbol{\Theta}}_{jj} - \sigma_u^2\boldsymbol{\Theta}_{jj} \\ &=\; \sigma_u^2(\widehat{\boldsymbol{\Theta}}_{jj} - \boldsymbol{\Theta}_{jj}) + (\hat{\sigma}_u^2 - \sigma_u^2)(\widehat{\boldsymbol{\Theta}}_{jj} - \boldsymbol{\Theta}_{jj}) + (\sigma_u^2 - \hat{\sigma}_u^2)\boldsymbol{\Theta}_{jj}\,.\end{aligned}$$

Next, recall that, on the set $\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$, it holds due to Lemma A.1 that $\|\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\|_\infty \le 2m_{\boldsymbol{\Theta}}\mu$. Letting $\epsilon > 0$ be arbitrary, it then follows from the previous display that

$$\mathrm{P}\big\{|\hat{\omega}_j^2 - \omega_j^2| > 2m_{\boldsymbol{\Theta}}\mu(\sigma_u^2 + \epsilon) + \epsilon\boldsymbol{\Theta}_{jj}\big\} \;\le\; \mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu)^{\complement} + \mathrm{P}\{|\hat{\sigma}_u^2 - \sigma_u^2| > \epsilon\}\,.$$

Now note that, due to the first claim of the present lemma and condition (4) of Assumption 4.7, we have

$$\begin{aligned}&\lim_{n\to\infty}\mathrm{P}\big\{|\hat{\omega}_j - \omega_j| > 2m_{\boldsymbol{\Theta}}\mu(\sigma_u + \epsilon) + \epsilon\boldsymbol{\Theta}_{jj}\big\} \\ &\quad\le\; \lim_{n\to\infty}\mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu)^{\complement} + \mathrm{P}\{|\hat{\sigma}_u - \sigma_u| > \epsilon\} \;=\; 0\,.\end{aligned}$$

Finally, cite Proposition C.4 and the present assumption that $\mu = o(1)$ to find that

$$\lim_{n\to\infty} m_{\boldsymbol{\Theta}}\mu(\sigma_u + \epsilon) + \epsilon\boldsymbol{\Theta}_{jj} \;=\; 0\,.$$

Conclude that $\omega_j^2 - \hat{\omega}_j^2 = o_{\mathrm{P}}(1)$ and hence that $\hat{\omega}_j - \omega_j = o_{\mathrm{P}}(1)$, as claimed. □

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2019.09.009.

Please cite this article as: D. Gold, J. Lederer and J. Tao, Inference for high-dimensional instrumental variables regression. Journal of Econometrics (2020), https://doi.org/10.1016/j.jeconom.2019.09.009.

References

Amemiya, T., 1974. The nonlinear two-stage least-squares estimator. Journal of Econometrics 2, 105–110.
Amemiya, T., 1977. The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model. Econometrica 45, 955–968.
Angrist, J., Pischke, J.-S., 2009. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.
Belloni, A., Chen, D., Chernozhukov, V., Hansen, C., 2012. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80, 2369–2429.
Belloni, A., Chernozhukov, V., 2013. Least squares after model selection in high-dimensional sparse models. Bernoulli 19, 521–547.
Belloni, A., Chernozhukov, V., Hansen, C., 2011. Inference for high-dimensional sparse econometric models. In: Advances in Economics and Econometrics: Tenth World Congress Volume 3, Econometrics.
Belloni, A., Chernozhukov, V., Hansen, C., Newey, W., 2018. Simultaneous confidence intervals for high-dimensional linear models with many endogenous variables. arXiv preprint arXiv:1712.08102.
Bickel, P., Ritov, Y., Tsybakov, A., 2009. Simultaneous analysis of Lasso and Dantzig selector. Annals of Statistics 37, 1705–1732.
Bühlmann, P., van de Geer, S., 2011. Statistics for High-Dimensional Data. Springer.
Cai, T., Liu, W., Luo, X., 2011. A constrained ℓ_1 minimization approach to aparse precision matrix estimation. Journal of the American Statistical Association 106, 594–607.
Caner, M., Fan, Q., 2015. Hybrid generalized empirical likelihood estimators: Instrument selection with adaptive lasso. Journal of Econometrics 187, 256–274.
Caner, M., Kock, A.B., 2018a. Asymptotically honest confidence regions for high dimensional parameters by the desparsified conservative Lasso. Journal of Econometrics 203, 143–168.
Caner, M., Kock, A.B., 2018b. High dimensional linear GMM. arXiv preprint arXiv:1811.08779.
Chamberlain, G., 1987. Asymptotic efficiency in estimation with conditional moment restrictions. Journal of Econometrics 34, 305–334.
Cheng, X., Liao, Z., 2015. Select the valid and relevant moments: An information-based lasso for gmm with many moments. Journal of Econometrics 186, 443–464.
Chételat, D., Lederer, J., Salmon, J., 2017. Optimal two-step prediction in regression. Electronic Journal of Statistics 11, 2519–2546.
Chichignoud, M., Lederer, J., Wainwright, M., 2016. A practical scheme and fast algorithm to tune the Lasso with optimality guarantees. Journal of Machine Learning Research 17, 1–17.
Dalalyan, A., Hebiri, M., Lederer, J., 2017. On the prediction performance of the Lasso. Bernoulli 23, 552–581.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.
Fan, J., Liao, Y., 2014. Endogeneity in high dimensions. Annals of Statistics 42, 872–917.
Fan, Q., Zhong, W., 2018. Nonparametric additive instrumental variable estimator: a group shrinkage estimation perspective. Journal of Business and Economic Statistics 36 (3), 388–399.
Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33, 1–22.
Gautier, E., Tsybakov, A., 2014. High-dimensional instrumental variables regression and confidence sets. ArXiv:1105.2454v4.
Gautier, E., Tsybakov, A., 2018. High-dimensional instrumental variables regression and confidence sets. ArXiv e-prints.
van de Geer, S., Bühlmann, P., 2009. On the conditions used to prove oracle results for the Lasso. Electronic Journal of Statistics 3, 1360–1392.
van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. Annals of Statistics 42, 1166–1202.
van de Geer, S., Lederer, J., 2013. The Lasso, correlated design, and improved oracle inequalities. In: From Probability to Statistics and Back: High-Dimensional Models and Processes – A Festschrift in Honor of Jon A. Wellner. In: Collections, vol. 9, Institute of Mathematical Statistics, Beachwood, Ohio, USA, pp. 303–316.
van de Geer, S., Muro, A., 2014. On higher order isotropy conditions and lower bounds for sparse quadratic forms. Electronic Journal of Statistics 8, 3031–3061.
Giraud, C., 2014. Introduction to high-dimensional statistics. In: Monographs on Statistics and Applied Probability (Series), vol. 139, CRC Press, Taylor & Francis Group.
Hansen, L., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50, 1029–1054.
Hansen, C., Kozbur, D., 2014. Instrumental variables estimation with many weak instruments using regularized jive. Journal of Econometrics 182, 290–308.
Hastie, T., Tibshirani, R., Wainwright, M., 2015. Statistical learning with sparsity: The Lasso and generalizations. In: Monographs on Statistics and Applied Probability (Series), vol. 143, Boca Raton: CRC Press, Taylor & Francis Group.
Hebiri, M., Lederer, J., 2013. How correlations influence lasso prediction. IEEE Trans. Inform. Theory 59, 1846–1854.
Imbens, G., Donald, S., Newey, W., 2003. Empirical likelihood estimation and consistent tests with conditional moment restrictions. Journal of Econometrics 117, 55–93.
Jankova, J., van de Geer, S., 2018. Semiparametric efficiency bounds for high-dimensional models. Annals of Statistics 46 (5), 2336–2359.
Javanmard, A., Montanari, A., 2014. Confidence intervals and hypothesis testing for high-dimensional regression. Journal of Machine Learning Research 15, 2869–2909.
Knight, K., Fu, W., 2000. Asymptotics for lasso-type estimators. Annals of Statistics 28, 1356–1378.
Lederer, J., Yu, L., Gaynanova, I., 2019. Oracle inequalities for high-dimensional prediction. Bernoulli 25, 1225–1255.
Lin, W., Feng, R., Li, H., 2015. Regularization methods for high-dimensional instrumental variables regression with an application to genetical genomics. Journal of the American Statistical Association 110, 270–288.
Liu, H., Wang, L., Zhao, T., 2015. Calibrated multivariate regression with application to neural semantic basis discovery. J. Mach. Learn. Res. 16, 1579–1606.
MOSEK AsP, 2017. MOSEK Rmosek Package 8.1.0.34.
Newey, W., 1990. Efficient instrumental variables estimation of nonlinear models. Econometrica 58, 809–837.
Neykov, M., Ning, Y., Liu, J., Liu, H., 2015. A unified theory of confidence regions and testing for high dimensional estimating equations. arXiv:1510.08986.
Ning, Y., Liu, H., 2017. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. Annals of Statistics 45, 158–195.
Pötscher, B., Leeb, H., 2009. On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. Journal of Multivariate Analysis 100, 2065–2082.
Rudelson, M., Zhou, S., 2012. Reconstruction from anisotropic random measurements. In: JMLR Workshop Conf. Proc., Vol. 23, pp. 10.1–10.28.
van der Vaart, A., Wellner, J., 1996. Weak Convergence and Empirical Processes: With Applications to Statistics. Springer-Verlag New York.

Please cite this article as: D. Gold, J. Lederer and J. Tao, Inference for high-dimensional instrumental variables regression. Journal of Econometrics (2020), https://doi.org/10.1016/j.jeconom.2019.09.009.

Vershynin, R., 2012. Introduction to the non-asymptotic analysis of random matrices. In: Eldar, Y.C., Kutyniok, G. (Eds.), Compressed Sensing: Theory and Applications, Vol. 5. Cambridge University Press, pp. 210–268.
Zhang, C.-H., 2010. Nearly unbiased variable selection under minimax concave penalty. Annals of Statistics 38, 894–942.
Zhang, C.-H., Zhang, S., 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76, 217–242.
Zhu, Y., 2018. Sparse linear models and l1-regularized 2sls with high-dimensional endogenous regressors and instruments. Journal of Econometrics 202 (2), 196–213.
Zhuang, R., Lederer, J., 2018. Maximum regularized likelihood estimators: A general prediction theory and applications. Statistics 7, e186, sta4.186.

Please cite this article as: D. Gold, J. Lederer and J. Tao, Inference for high-dimensional instrumental variables regression. Journal of Econometrics (2020), https://doi.org/10.1016/j.jeconom.2019.09.009.