Fast, Robust Inference for Linear Instrumental Variables Models using Self-Normalized Moments

Eric Gautier (Toulouse School of Economics),

Christiern Rose (University of Queensland)*

Abstract

We propose and implement an approach to inference in linear instrumental variables models which is simultaneously robust and computationally tractable. Inference is based on self-normalization of sample moment conditions, and allows for (but does not require) many (relative to the sample size), weak, potentially invalid (i.e., the exclusion restriction fails) or potentially endogenous instruments, as well as for many regressors and conditional heteroskedasticity. Our coverage results are uniform and can deliver a small sample guarantee. We develop a new computational approach based on semidefinite programming, which we show can equally be applied to rapidly invert existing tests (AR, LM, CLR, etc.).

Key words: High-dimensional inference, weak instruments, many instruments, invalid instruments, endogenous instruments.

1. Introduction

Instrumental variables are widely used in applied econometrics, yet computationally tractable robust inference remains challenging. It is well known that inference robust to weak instruments can be conducted by inverting robust tests. However, to our knowledge there do not exist inference methods which are simultaneously robust to weak instruments, many instruments/regressors (e.g., larger than the sample size), potentially invalid instruments (i.e., an exclusion restriction may be violated) and potentially endogenous instruments, nor which can offer coverage guarantees in small samples. Moreover, inverting robust tests can be computationally challenging (Andrews et al., 2019) because their non-rejection regions are not convex (Mikusheva, 2010) and a grid search is infeasible with just a handful of regressors (see Andrews (2016), supplementary material).

We address both the statistical and computational challenges. First, we provide an approach to inference based on self-normalized sample moment conditions, which we refer to as Self-Normalized Instrumental Variables (SNIV). Due to the minimal assumptions it requires, SNIV simultaneously allows for weak instruments and conditional heteroskedasticity, and can be applied equally to the standard low-dimensional setting (large sample size, few regressors, few instruments) and to the high-dimensional setting in which the number of instruments and/or regressors can be large, possibly much larger than the sample size. For example, in a model with a single endogenous regressor, SNIV could be used to construct a confidence interval which is simultaneously robust to many and to weak instruments.[1] We

---

[1]To further motivate our framework, there are several reasons to expect models with multiple endogenous regressors to become increasingly popular. Potential applications include demand systems with many goods and endogenous expenditure (Gautier and Rose, 2021) or prices (Belloni et al., 2022), as well as models of peer effects with unknown peer relationships (Rose, 2018). More broadly, due to increasing availability of rich datasets and the potential to allow for heterogeneous treatment effects by using interactions of the treatment with individual characteristics, applied research has recently considered models with multiple exogenous variables of interest (e.g., Farrell et al. (2020)). A natural extensionis to allow for endogenous treatment (Belloni et al., 2022).

extend SNIV to settings in which one or more instrument may be 'invalid' (i.e., the exclusion restriction fails, see Kolesár et al. (2015); Kang et al. (2016)) or endogenous without requiring a pre-test, and propose the use of an a-priori upper bound on the number of invalid/endogenous instruments for settings in which the set of identifiable parameters would otherwise be unbounded.

Second, we provide a computational implementation of SNIV which we show can also be applied to rapidly invert other robust tests. This is because test inversion can typically be cast as a semi-algebraic optimization problem, hence we can apply methods from the literature on semi-algebraic optimization. The basic idea to deal with computational intractability is to attempt to solve a hiearchy of semidefinite optimization problems. Solving each optimization problem delivers an outer bound on the confidence region. As we proceed up the hierarchy, the bounds become sharper at the expense of greater computational burden. This allows the researcher to effectively trade off sharpness with available computational resources. In practice, the bounds obtained towards the beginning of the hierarchy are often exact. A simple diagnostic informs the researcher if exact bounds have been attained. Similar computational approaches have been applied by Gautier et al. (2018), Lee (2020) and Auerbach (2022).

In contrast to a grid search, our approach can be applied to settings with multiple regressors. In contrast to heuristic/local optimization methods, our approach guarantees an outer bound on the confidence region, hence does not risk compromising the coverage guarantee. We illustrate how our computational approach can be used to rapidly invert existing robust tests by combining it with the results of Guggenberger et al. (2012), Guggenberger et al. (2019) and Guggenberger et al. (2021) to obtain weak instrument robust Anderson-Rubin (AR) confidence intervals which can be computed near instantaneously, even when a grid search is infeasible. We also show that our approach can be applied to rapidly invert other

3

robust tests such as the Lagrange-multiplier (LM) test (Kleibergen, 2002; Moreira, 2002) and the Conditional Likelihood Ratio (CLR) test (Moreira, 2003).

We conduct a Monte-Carlo experiment in which we demonstrate SNIV and AR confidence regions are both easily implemented in a setting in which a grid search is computationally intractable. SNIV has similar coverage to the AR test in designs with either strong or weak instruments. With many instruments, SNIV maintains coverage close to the nominal level but AR does not. We also show that SNIV can be applied to conduct informative inference with invalid instruments and endogenous instruments in challenging designs, to which existing approaches cannot be applied.

1.1. Related literature

Our work is related to the literature on many instruments (e.g., Bekker (1994); Angrist et al. (1999); Donald and Newey (2001); Anderson (2005); Chao and Swanson (2005); Stock and Yogo (2005); Hansen et al. (2008); Ackerberg and Devereux (2009); van Hasselt (2010); Chao et al. (2012); Hausman et al. (2012); Anatolyev (2013); Hansen and Kozbur (2014); Kolesár (2018); see Anatolyev (2019) for a recent review) and weak instruments (e.g., Anderson and Rubin (1949); Kleibergen (2002); Moreira (2002, 2003); Mikusheva (2010); Guggenberger et al. (2012); Andrews (2016); Guggenberger et al. (2019, 2021); see Andrews et al. (2019) for a recent review), but allows simultaneously for weak instruments and for the number of regressors and/or instruments to be large, possibly much larger than the sample size. This is because we conduct inference based on moderate deviations of self-normalized sample moments (e.g., Pinelis (1994); Bertail et al. (2008); Jing et al. (2003)) instead of using a Central Limit Theorem.

The most closely related papers are Gautier and Tsybakov (2011), Belloni et al. (2012), Gold et al. (2020), Gautier and Rose (2021) and Belloni et al. (2022), all

of which also consider the linear instrumental variables model in a potentially high-dimensional setting. Gautier and Tsybakov (2011) and Gautier and Rose (2021) suggest to combine a point estimator with lower bounds on its sensitivity characteristics to perform robust inference, but their confidence region is larger than ours and the authors do not provide a disciplined way to trade off computational complexity and sharpness when implementing their approach. Belloni et al. (2012) discuss inference based on self-normalization, but do not propose a practical computational solution, nor allow for invalid/endogenous instruments. Gold et al. (2020), Gautier and Rose (2021) and Belloni et al. (2022) propose confidence regions for a subset of parameters of interest (e.g., confidence intervals) but these rely on stronger assumptions than we use below. For example, these papers propose methods which are not robust to weak instruments and do not allow for potentially invalid nor endogenous instruments. We view our work as complementary to Gold et al. (2020), Gautier and Rose (2021) and Belloni et al. (2022), providing the applied researcher with a more robust alternative, but one which may sometimes yield wider bounds in practice.

Finally, our extensions of SNIV are related to the literature on invalid and endogenous instruments. Regarding invalid instruments (e.g., Kolesár et al. (2015); Kang et al. (2016)), we allow for a setting with multiple endogenous regressors, potentially weak instruments, and for the number of instruments and/or regressors to be larger than the sample size. Regarding endogenous instruments (e.g., Sargan (1958); Hansen (1982); Anatolyev and Gospodinov (2011); Lee and Okui (2012); Chao et al. (2014)), we do not perform specification tests, but instead perform inference directly, accounting for potential endogeneity of an unknown subset of instruments.

We proceed as follows. Section 2 sets out our model. Section 3 defines SNIV, establishes its coverage guarantee, and provides extensions to potentially invalid and

endogenous instruments. Section 4 presents our computational method, which we apply to invert existing robust tests in Section 5. Section 6 presents a Monte-Carlo experiment and Section 7 concludes. All proofs are gathered in the appendix.

1.2. Setup & notation

To simplify the exposition we consider an i.i.d. sample of size n. The i.i.d. setting is not critical for our results and can be relaxed by using an alternative choice of r_n below, for which we provide appropriate references. The population model comprises an outcome Y, regressors $X \in \mathbb{R}^{d_X}$, and instrumental variables $Z \in \mathbb{R}^{d_Z}$ of joint distribution $\mathbb{P}$. $\mathbb{E}$ is the expectation under $\mathbb{P}$ and $\mathbb{E}_n$ is its sample counterpart. Our results apply to a sequence of models indexed by n. For simplicity of exposition we do not make this explicit, but we occasionally note that certain objects can depend on n. To allow for high-dimensional data, the relative magnitudes of n, d_X and d_Z are unrestricted, and both d_X and d_Z can grow with n. For $b \in \mathbb{R}^{d_X}$, $U(b) \triangleq Y - X^\top b$, $\mathbb{P}(b)$ is the distribution of $(X, Z, U(b))$ implied by $\mathbb{P}$. For $S \subseteq [d] \triangleq \{1, 2, ..., d\}$, $|S|$ is its cardinality and S^c its complement. For $\Delta \in \mathbb{R}^d$, $S(\Delta) \triangleq \{k \in [d] : \Delta_k \neq 0\}$ and $|\Delta|_p$ is the ℓ^p-norm of Δ. For a polynomial p, $\deg(p)$ is its degree. We use $\mathbf{M} \succcurlyeq \mathbf{0}$ to say that the matrix $\mathbf{M}$ is positive semidefinite.

2. MODEL

The linear instrumental variables model is

$$\mathbb{E}\left[ZU(\beta)\right] = 0, \tag{1}$$

$$\beta \in \mathcal{B}, \ \mathbb{P}(\beta) \in \mathcal{P}, \tag{2}$$

where $\mathcal{B} \subseteq \mathbb{R}^{d_X}$ is the parameter space and $\mathcal{P}$ is a nonparametric class. The set $\mathcal{I}$ collects the vectors which satisfy (1)-(2). As will be made clear below, our results are for all $\beta \in \mathcal{I}$, hence for the true value β^*. We use the class $\mathcal{P}$ to permit the use of results on moderate deviations of self-normalized sums for inference. We consider four classes, including

Class 1. *There exists δ in $(0,1]$ and $\mu_{2+\delta} > 0$ such that*

$$\left| \left(\left(\mathbb{E}\left[|Z_l U(\beta)|^{2+\delta} \right] \right) \left(\mathbb{E}\left[Z_l^2 U(\beta)^2 \right] \right)^{-(2+\delta)/2} \right)_{l \in [d_Z]} \right|_\infty \le \mu_{2+\delta},$$

and $d_Z \le \alpha/(2\Phi(-n^{1/2-1/(2+\delta)}\mu_{2+\delta}^{-1/(2+\delta)}))$,
where $\alpha \in (0,1)$ is a confidence level and Φ the normal CDF;

Class 2. $\exists \mu_4 > 0 : \max_{l \in [d_Z]} \mathbb{E}[Z_l^4 U(\beta)^4](\mathbb{E}[Z_l^2 U(\beta)^2])^{-2} \le \mu_4$, $d_Z < \alpha \exp(n/\mu_4)/(2e+1)$ *and* $n - \mu_4 \log(d_Z(2e+1)/\alpha) \ge n/2$;

Class 3. *$Z_l U(\beta)$ is symmetric for all $l \in [d_Z]$ and $d_Z < 9\alpha/(4e^3\Phi(-\sqrt{n}))$.*

Classes 1-2 require mild bounds on ratios of moments, whereas Class 3 requires no bounds but uses symmetry. Further classes allowing for dependence and non i.d. data can be found in Chen et al. (2016) and references therein. In Section 3.1 we consider a fourth class based on Gaussian approximation rather than self-normalization.

Remark 1 *The model in (1)-(2) can be obtained by first partialling-out a low-dimensional vector of exogenous regressors. To simplify the exposition we do not make this explicit.*

3. Self Normalized Instrumental Variables

The $1 - \alpha$ SNIV confidence set is

$$\widehat{\mathcal{C}} \triangleq \left\{ \beta \in \mathcal{B} : |\mathbf{D}(\beta)\mathbb{E}_n[ZU(\beta)]|_\infty \leq r_n \right\}, \tag{3}$$

where $\mathbf{D}(\beta)$ is the $d_Z \times d_Z$ positive, diagonal matrix with l^{th} diagonal element $\mathbb{E}_n[Z_l^2 U(\beta)^2]^{-1/2}$ used to self-normalize the d_Z moments and r_n depends on the class. Under Class 1 we set $r_n = -\Phi^{-1}\left(\alpha/(2d_Z)\right)/\sqrt{n}$. Under Class 2 we set $r_n = 2\sqrt{\log(d_Z(2e + 1)/\alpha)/n}$. Under Class 3 we set $r_n = -\Phi^{-1}(9\alpha/(4d_Z e^3))/\sqrt{n}$.

Proposition 1 *Consider the model in* (1)-(2)*. If $\mathcal{P}$ is Class 1 we have*

$$\lim_{n\to\infty} \inf_{(\beta,\mathbb{P}):\beta\in\mathcal{I}} \mathbb{P}(\beta \in \widehat{\mathcal{C}}) \geq 1 - \alpha. \tag{4}$$

If $\mathcal{P}$ is either Class 2 or Class 3 we have

$$\inf_{(\beta,\mathbb{P}):\beta\in\mathcal{I}} \mathbb{P}(\beta \in \widehat{\mathcal{C}}) \geq 1 - \alpha. \tag{5}$$

Proposition 1 shows that the coverage of SNIV is at least the nominal level uniformly over the identifiable parameters and the distributions of the data they imply. Beyond the class used, no further assumptions are needed. Classes 1-3 allow for conditional heteroscedasticity, do not restrict the joint distribution of X and Z (hence are robust to weak instruments) and have very mild requirements on the relative magnitudes of n and d_Z (hence are robust to many regressors and/or instruments). For Class 1 the coverage guarantee is asymptotic in n in such a way that d_X and d_Z can grow with (and be much larger than) n. For Classes 2-3 the coverage guarantee is for any n.

The SNIV confidence set collects vectors for which the ℓ_∞ deviation from zero

of the self-normalized sample moment is at most r_n. The core components which deliver uniformity, finite sample validity and robustness to identification are the ℓ_∞-norm and self-normalization of the moments. The ℓ_∞-norm is crucial so as to allow for d_Z larger than n because it permits r_n to be of the order $\log(d_Z)/\sqrt{n}$. This means that the SNIV confidence set can be small even when the number of instruments is much larger than the sample size.[2] Note also that d_X can be arbitrarily large with respect to n.

Remark 2 *If $\mathcal{B}$ is defined by polynomial (in)equalities of degree at most 2, the SNIV confidence set is defined by polynomial inequalities of degree at most 2 (we show this in Proposition 3), so it can be empty, unbounded or disconnected depending on the (random) values of the polynomial coefficients. Possible unboundedness is unavoidable for confidence sets which are robust to identification (Dufour, 1997).*

3.1. Gaussian approximation

The SNIV confidence set may be conservative when the instruments are strongly correlated with one another because r_n is based on a union bound over d_Z self-normalized sample moments. Gautier and Rose (2021) propose an alternative to self-normalization based around the multiplier bootstrap of Chernozhukov et al. (2013), which we implement here. We modify the SNIV confidence set by replacing $\mathbf{D}(\beta)$ by $\mathbb{E}_n[U(\beta)]^{-1/2}\mathbf{D_Z}$ and r_n by the $1 - \alpha$ quantile of $|\mathbf{D_Z}\mathbb{E}_n[ZW]|_\infty$ (computed by simulation) in its definition, where $\mathbf{D_Z}$ is a $d_Z \times d_Z$ diagonal matrix with l^{th} diagonal element $\mathbb{E}_n[Z_l^2]^{-1/2}$ and W is a standard normal random variable which is independent of Z. The corresponding class is

Class 4. There exist constants C and c, and B_n such that, for all $(\beta, \mathbb{P})$: $\beta \in \mathcal{I}$,

---

[2]For Class 3, $r_n \leq 2\log\left(4d_Z e^3/(9\alpha)\right)/\sqrt{n}$, $\forall \alpha \in [0,1], d_Z \geq 1$(because $\Phi^{-1}(a) \geq 2\log(a)$ if $0 < a \leq \exp(-1/(4\pi))$).

$U(\beta) \perp Z$; $|Z|_\infty \leq B_n$ (a.s.); $\mathbb{E}[U(\beta)^4] \leq C$; and $B_n^4 \log(d_Z n)^7/n \leq Cn^{-c}$,

which delivers the following coverage guarantee.

Proposition 2 *Consider the model in* (1)-(2). *If* $\mathcal{P}$ *is Class 4, then, for* $\widehat{\mathcal{C}}$ *based on the multiplier bootstrap, we have*

$$\lim_{n \to \infty} \inf_{(\beta, \mathbb{P}):\beta \in \mathcal{I}} \mathbb{P}(\beta \in \widehat{\mathcal{C}}) \geq 1 - \alpha. \tag{6}$$

Further classes based on Gaussian approximation but allowing for non i.d. and dependent data can be found in Zhang and Wu (2017) and references therein.

3.2. Sparsity

In the high-dimensional setting with d_X larger than n, a natural and commonly used restriction is that β is *sparse*, meaning that it has many elements exactly equal to zero but the researcher does not know which ones. Sparsity implies that there exists an underlying parsimonious model which is unknown to the researcher. It can be used to motivate ℓ_1 penalized estimators such as the LASSO of Tibshirani (1996) for regression or the STIV of Gautier and Rose (2021) for instrumental variables. In the instrumental variables context, sparsity can be interpreted as imposing exclusion restrictions of unknown locations. As explained below, this is particularly useful in the underidentified case with $d_Z < d_X$, which can arise, for example, when there is uncertainty as to which candidate instruments can be excluded, implying that some instruments may be invalid (Kolesár et al., 2015; Kang et al., 2016).

The SNIV confidence set can easily accommodate sparsity. We define $S_Q \subseteq [d_X]$ as the indices of the regressors of questionable relevance (i.e., whose entry of β^* may

10

be zero). We denote by $d_Q \triangleq |S_Q|$ and modify $\mathcal{I}$ and $\widehat{\mathcal{C}}$ to include the restriction
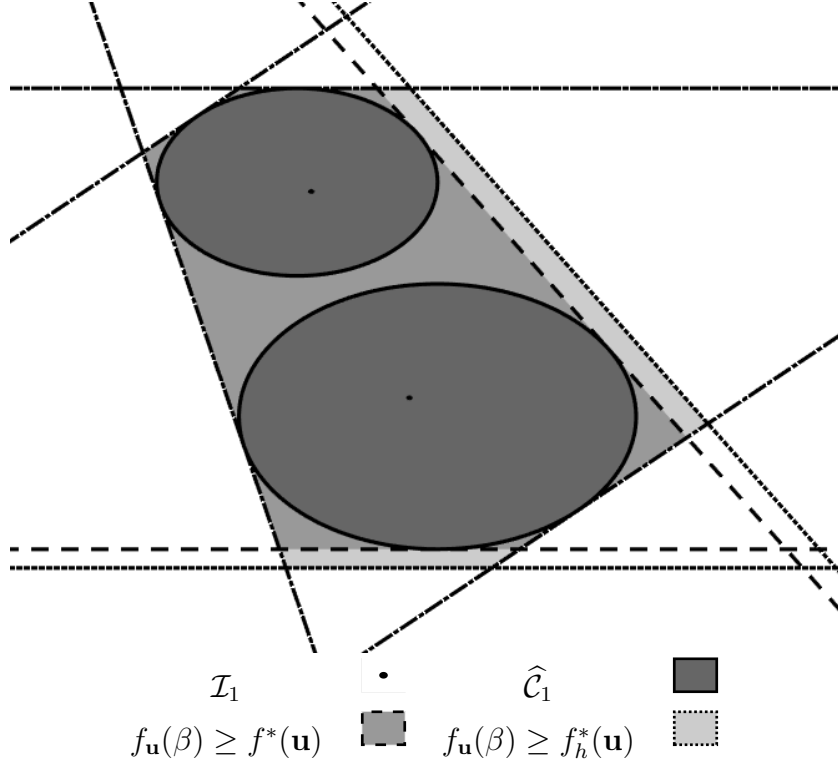
$$|S(\beta) \cap S_Q| \leq s, \tag{7}$$

where $s \in [d_Q]$ is an upper bound chosen by the researcher, and recalling that $S(\beta) \subseteq [d_X]$ is the support of β. Though we do not make it explicit, both s and S_Q can depend on n. Since the choice of s provides a guarantee on the sparsity, we refer to it as a *sparsity certificate*. When the sparsity certificate s is used, we use the notation $\mathcal{I}_s$ and $\widehat{\mathcal{C}}_s$ in place of $\mathcal{I}$ and $\widehat{\mathcal{C}}$. Clearly, $\mathcal{I}_{d_Q} = \mathcal{I}$ and $\widehat{\mathcal{C}}_{d_Q} = \widehat{\mathcal{C}}$.

Remark 3 *The restriction in (7) is weaker than imposing $d_Q - s$ exclusion restrictions on the parameters in S_Q.*

Remark 4 *In practice, the researcher may not know how to choose the sparsity certificate. In this case nested confidence sets can be computed by varying s over reasonable alternatives. This allows for an assessment of the information content of progressively stronger assumptions on the sparsity.*

Interestingly, $\mathcal{I}_s$ can be a singleton even when $\mathcal{I}$ is not. This means that sparsity can lead to point identification even in 'underidentified' models (i.e., when $d_Z < d_X$). In general, $\mathcal{I}_s$ is a singleton if there is a solution for only one of the $\binom{d_Q}{s}$ overdetermined systems based on (1)-(2) and it is unique. For example, $\mathcal{I}_s$ can be a singleton when $s + d_X - d_Q < d_Z < d_X$ and sparsity implies that some exogenous regressors have a zero coefficient (i.e., they are excluded, see Kang et al. (2016)). The basic idea is that excluded exogenous regressors can serve as instruments for included endogenous regressors, but we need not necessarily know *which* regressors are excluded. Finally, if $S_Q = [d_X]$, $\mathcal{I}_s$ is a singleton if all matrices formed from $2s$ columns of $\mathbb{E}[ZX^\top]$ have rank $2s$ (Candes and Tao, 2007). The corresponding order condition is $s \leq d_Z/2$, which does not depend on d_X.

Figure 1: *The Identified Set, SNIV Confidence Set and its Outer Approximations*

| $\mathcal{I}_1$ | $\cdot$ | $\widehat{\mathcal{C}}_1$ | ■ |
| $f_{\mathbf{u}}(\beta) \geq f^*(\mathbf{u})$ | ▨ | $f_{\mathbf{u}}(\beta) \geq f_h^*(\mathbf{u})$ | ▨ |

Notes: In this example, $d_Z = 1$, $d_X = 2$, $\mathcal{B} = \mathbb{R}^{d_X}$, $d_Q = [d_X]$ and $s = 1$.

However, SNIV does not require that $\mathcal{I}_s$ be a singleton. For example, $\mathcal{I}_s$ can comprise a finite union of singletons. Figure 1 depicts such an example with $d_Z = 1$, $d_X = 2$, $\mathcal{B} = \mathbb{R}^{d_X}$, $d_Q = [d_X]$ and $s = 1$, in which case $\mathcal{I}_s$ is the intersection of the line $\mathbb{E}[Zy] = \mathbb{E}[ZX^\top]\beta$ with the set $\{\beta \in \mathbb{R}^2 : \beta_1 = 0 \text{ or } \beta_2 = 0\}$. The SNIV confidence set allows for such partially identified cases due to the uniformity over s and $\mathcal{I}_s$ in the coverage guarantee, which is obtained by replacing $\inf_{(\beta,\mathbb{P}):\beta\in\mathcal{I}} \mathbb{P}(\beta \in \widehat{\mathcal{C}})$ by $\min_{s\in[d_Q]} \inf_{(\beta,\mathbb{P}):\beta\in\mathcal{I}_s} \mathbb{P}(\beta \in \widehat{\mathcal{C}}_s)$ in Propositions 1 and 2.

3.3. Endogenous instruments

Testing instrument exogeneity is a classical problem to which our framework can be applied. Introducing $\theta \in \mathbb{R}^{d_Z}$ to account for the possible failure of exogeneity,

we replace (1)-(2) by

$$\mathbb{E}[ZU(\beta) - \theta] = 0, \tag{8}$$

$$(\beta, \theta) \in \mathcal{B} \times \Theta, \ \mathbb{P}(\beta, \theta) \in \mathcal{P}, \tag{9}$$

where $\theta_l \neq 0$ means that Z_l is endogenous, $\mathbb{P}(b, t)$ is the distribution of $(X, Z, ZU(b) - t)$ implied by $\mathbb{P}$ and $\Theta \subseteq \mathbb{R}^{d_Z}$ encodes restrictions on θ. For example, Θ may be such that the sign of the correlation of a regressor and the structural error is known. An important restriction encoded by Θ is $\theta_{S_\perp} = 0$ for $S_\perp \subseteq [d_Z]$, which indexes the instruments known to be exogenous. The remaining instruments are potentially endogenous. We can use a sparsity certificate to place an upper bound on the number of endogenous instruments, given by

$$|S(\theta)| \leq \widetilde{s}, \tag{10}$$

for a given $\widetilde{s} \in [\widetilde{d}_Q]$, where $\widetilde{d}_Q \triangleq d_Z - |S_\perp|$. Thus, though the identities of the endogenous instruments may not be known, their number can be restricted. The counterpart of $\mathcal{I}_s$, denoted by $\mathcal{I}_{s,\widetilde{s}}$, collects the vectors which satisfy (8)-(9) and the sparsity restrictions in (7) and (10). Under Classes 1-3,[3] SNIV is

$$\widehat{\mathcal{C}}_{s,\widetilde{s}} \triangleq \left\{ (\beta, \theta) \in \mathcal{B} \times \Theta : \begin{array}{l} |\mathbf{D}(\beta, \theta)(\mathbb{E}_n[ZU(\beta)] - \theta)|_\infty \leq r_n, \\ |S(\beta) \cap S_Q| \leq s, |S(\theta)| \leq \widetilde{s} \end{array} \right\}, \tag{11}$$

where $\mathbf{D}(\beta, \theta)$ is the $d_Z \times d_Z$ positive, diagonal matrix with l^{th} diagonal element $\mathbb{E}_n[(Z_l U(\beta) - \theta_l)^2]^{-1/2}$. This set allows one to simutaneously perform inference on β^* and θ^*, without requiring, for example, a pilot estimator and subsequent test of instrument exogeneity. The coverage guarantee is obtained by replacing $\inf_{(\beta, \mathbb{P}): \beta \in \mathcal{I}} \mathbb{P}(\beta \in \widehat{\mathcal{C}})$ by $\min_{\widetilde{s} \in [\widetilde{d}_Q]} \min_{s \in [d_Q]} \inf_{(\beta, \theta, \mathbb{P}): (\beta, \theta) \in \mathcal{I}_{s,\widetilde{s}}} \mathbb{P}((\beta, \theta) \in \widehat{\mathcal{C}}_{s,\widetilde{s}})$ in

---

[3]Class 4 is not applicable with possibly endogenous instruments.

Proposition 1. Both $\widetilde{S}_Q$ and $\widetilde{s}$ can depend on n.

4. COMPUTATION

To implement SNIV the researcher needs some way to summarize the vectors which lie in the confidence set. Belloni et al. (2012) propose to use a grid for a confidence set with no sparsity constraints nor potentially endogenous instruments. This involves checking whether the inequalities in the definition of the SNIV confidence set are verified for every β on a grid over $\mathcal{B}$, and is a practical solution when d_X is small. However, a grid search quickly becomes infeasible for moderate d_X. In a low-dimensional setting, one can first partial-out a small number of exogenous regressors (see Remark 1) so that d_X is the number of endogenous regressors, which may be sufficiently small so as to use a grid. Otherwise we require an alternative.

We propose a method based on solving convex optimization problems. For a given direction $\mathbf{u} \in \mathbb{R}^{d_X+d_Z}$ normalized to satisfy $|\mathbf{u}|_2 = 1$ and the function $f_{\mathbf{u}}(\beta, \theta) \triangleq \mathbf{u}^\top (\beta^\top, \theta^\top)^\top$ we seek to compute

$$f^*(\mathbf{u}) \triangleq \inf_{(\beta,\theta)\in\widehat{\mathcal{C}}_{s,\widetilde{s}}} f_{\mathbf{u}}(\beta, \theta), \tag{12}$$

which is the support function of $\widehat{\mathcal{C}}_{s,\widetilde{s}}$. By solving (12) for all directions $\mathbf{u} \in \{\mathbf{u} \in \mathbb{R}^{d_X+d_Z} : |\mathbf{u}|_2 = 1\}$, we obtain the convex envelope of $\widehat{\mathcal{C}}_{s,\widetilde{s}}$ defined by the inequalities $f_{\mathbf{u}}(\beta, \theta) \geq f^*(\mathbf{u})$ for all $\mathbf{u} \in \{\mathbf{u} \in \mathbb{R}^{d_X+d_Z} : |\mathbf{u}|_2 = 1\}$.[4]

If $\widehat{\mathcal{C}}_{s,\widetilde{s}}$ is convex, solving (12) is straightforward. In general $\widehat{\mathcal{C}}_{s,\widetilde{s}}$ is not convex because none of the inequalities in its definition define a convex set. This is unavoidable because $\mathcal{I}_{s,\widetilde{s}}$ need not be convex. We now show that $\widehat{\mathcal{C}}_{s,\widetilde{s}}$ is a semi-algebraic set (i.e., a set defined by polynomial inequalities) and apply methods in semi-algebraic optimization to problem (12).

---

[4]In practice, we consider only a finite number of directions.

Proposition 3 *If $\mathcal{B}$ and Θ are semi-algebraic, the SNIV confidence set is semi-algebraic, taking the form*

$$\widehat{\mathcal{C}}_{s,\widetilde{s}} = \left\{ (\beta, \theta) : \exists \gamma \in [0,1]^{d_Q + \widetilde{d}_Q} : \widehat{\mathbf{g}}(\beta, \theta, \gamma) \geq 0 \right\} \tag{13}$$

where $\widehat{\mathbf{g}}$ is a $d_g \times 1$ vector of polynomials, the form of which is given in the proof. If $\mathcal{B}$ and Θ are defined by polynomial inequalities of degree at most e then $\widehat{\mathbf{g}}$ has degree $\max(2, e)$.

The requirement that $\mathcal{B}$ and Θ are semi-algebraic is mild. For example, $\mathcal{B} = \mathbb{R}^{d_x}$ is semi-algebraic. The additional parameter γ is required to model the sparsity constraints in (7) and (10). Proposition 3 implies that

$$f^*(\mathbf{u}) = \inf_{(\beta, \theta, \gamma) : \widehat{\mathbf{g}}(\beta, \theta, \gamma) \geq 0} f_{\mathbf{u}}(\beta, \theta), \tag{14}$$

which can be computed by solving a polynomial optimization problem. Due to non-convexity, exact computation of $f^*(\mathbf{u})$ is NP-hard. Instead, we focus on solving convex relaxations of (14). Convex relaxation is routinely used to construct computationally tractable estimators. For example, LASSO uses an ℓ_1 penalty as a convex relaxation of a sparsity constraint such as (7). We solve a sequence of convex relaxations, delivering a *hierarchy* of convex optimization problems. Following the seminal paper of Lasserre (2001), such hierarchies have attracted much attention in the optimization literature in recent years. We first provide a general summary of the approach, then explain the specific hierarchy we propose.

The most important feature of a hierarchy is that it is *disciplined*, meaning that it delivers a monotone sequence of lower bounds converging to $f^*(\mathbf{u})$. If $f_h^*(\mathbf{u})$ is the optimal value obtained by solving the hth convex optimization problem in the hierarchy, we have $f_h^*(\mathbf{u}) \leq f_{h+1}^*(\mathbf{u}) \leq f^*(\mathbf{u})$ for all $h \in \mathbb{N}$ and $f_h^*(\mathbf{u}) \to f^*(\mathbf{u})$ as $h \to \infty$. As h increases, though convex, the optimization problems become more

computationally intensive. It is also generically the case that there exists finite h^* such that $f_{h^*}^*(\mathbf{u}) = f^*(\mathbf{u})$, and that the researcher can identify when such h^* has been encountered.

Monotonicity of the sequence of lower bounds on $f^*(\mathbf{u})$ is crucial. This is because it allows us to construct bounds on the convex envelope of the SNIV confidence set defined by the linear inequalities $f_{\mathbf{u}}(\beta, \theta) \geq f_h^*(\mathbf{u})$ for all $\mathbf{u} \in \mathcal{U}$ and some $h \in \mathbb{N}$, where $\mathcal{U}$ is a finite collection of directions. Since we construct a superset, the coverage guarantee cannot fall below $1 - \alpha$. The larger is h, the closer the superset becomes to the SNIV confidence set. Hence, by varying h and $\mathcal{U}$, we can trade off the computational burden with the quality of the approximation without compromising the coverage guarantee. Such a trade-off cannot be achieved by local or heuristic optimization methods nor by adjusting the spacing of a grid, both of which may compromise the coverage guarantee. Figure 1 illustrates the SNIV confidence set and its outer approximations for a partially identified model with $d_Z = 1$, $d_X = 2$, $\mathcal{B} = \mathbb{R}^{d_X}$, $d_Q = [d_X]$, $s = 1$ and all instruments known to be exogenous.

4.1. Low-dimensional objects of interest

The method we propose can also be used to compute bounds on a polynomial function of interest $p(\beta^*, \theta^*)$. For example, to obtain a confidence interval for β_1^*, we can solve (12) for $f_{\mathbf{u}_1}$ and $f_{\mathbf{u}_2}^\top$, where $\mathbf{u}_1 = (1, 0, ..., 0)^\top$ and $\mathbf{u}_2 = -\mathbf{u}_1$ (i.e., use the projection method). More generally one can consider a vector $\mathbf{p}$ of functions of interest. If the dimension of $\mathbf{p}$ is small relative to d_X, it is well known that the projection method can be conservative. A leading case with $d_X = \dim(\mathbf{p}) = 1$ is a confidence interval in a model with one endogenous regressor. In this case, the projection method is not conservative and SNIV is simultaneously robust to weak instruments and to d_Z much larger than n.

4.2. A hierarchy of semidefinite optimization problems

Since it is straightforward to implement, we present our application of the seminal hierarchy first proposed by Lasserre (2001). This hierarchy is sufficiently computationally tractable to deal with problems of size likely to be encountered in empirical work. Recent advances allowing for even larger problems are provided by Lasserre et al. (2017) and Weisser et al. (2018).

To simplify the exposition, we denote the decision variable in problem (14) by $\delta \triangleq (\beta^\top, \theta^\top, \gamma^\top)^\top$ of size $d_\delta \triangleq d_X + d_Z + d_Q + \widetilde{d}_Q$. The hierarchy uses the decision variable μ, each entry of which represents a monomial of δ. For example, if $d_\delta = 2$ then $\mu = (1, \delta_1, \delta_2, \delta_1^2, \delta_1\delta_2, \delta_2^2, ...)^\top$, so the polynomial $p(\delta) = \delta_2 + 2\delta_1^2$ is equivalently expressed as $\mu_3 + 2\mu_4$. This allows us to define the *Riesz linear functional* of p as $L_\mu(p) = \mu_3 + 2\mu_4$.[5] Now let the vector $\mathbf{m}_e(\delta)$ comprise all monomials of δ of degree no larger than e. For example, $\mathbf{m}_1(\delta) = (1, \delta_1, \delta_2)^\top$. Then we can define the *moment matrix* $\mathbf{M}_e(\mu) \triangleq L_\mu(\mathbf{m}_e(\delta)\mathbf{m}_e(\delta)^\top)$. For example, if $d_\delta = 2$ and $e = 1$, we have

$$\mathbf{m}_1(\delta)\mathbf{m}_1(\delta)^\top = \begin{bmatrix} 1 & \delta_1 & \delta_2 \\ \delta_1 & \delta_1^2 & \delta_1\delta_2 \\ \delta_2 & \delta_1\delta_2 & \delta_2^2 \end{bmatrix} \Rightarrow \mathbf{M}_1(\mu) = \begin{bmatrix} \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_4 & \mu_5 \\ \mu_3 & \mu_5 & \mu_6 \end{bmatrix}. \qquad (15)$$

Given another polynomial $q(\delta)$, we can similarly define the *localizing matrix* $\mathbf{M}_e(q\mu) \triangleq L_\mu(q(\delta)\mathbf{m}_e(\delta)\mathbf{m}_e(\delta)^\top)$.

---

[5] We present the case in which he order of the entries of μ is a graded lexicographic order. Other orderings are possible. None of our results depend on the ordering used.

At level h of the hierachy we solve the semidefinite program

$$f_h^*(\mathbf{u}) \triangleq \inf_\mu L_\mu(f_\mathbf{u}) \quad \text{subject to} \quad \mathbf{M}_h(\mu) \succcurlyeq \mathbf{0},$$

$$\mathbf{M}_{h-e_j}(\widehat{\mathbf{g}}_j \mu) \succcurlyeq \mathbf{0}, \quad j \in [d_g],$$

$$\mu_1 = 1, \tag{16}$$

where e_j is the smallest integer which is at least as large as $\deg(\widehat{\mathbf{g}}_j)/2$ for $j \in [d_g]$. This program has a linear objective function and $d_g + 1$ semidefinite constraints. The semidefinite constraint on $\mathbf{M}_h(\mu)$ arises because $\mathbf{m}_h(\delta)\mathbf{m}_h(\delta)^\top$ has rank 1. In principle we would like to impose that $\mathbf{M}_h(\mu)$ has rank 1. However, the set of rank 1 matrices is not convex. To obtain a convex problem, we use instead the set of positive semidefinite matrices. The intuition is the same for the other d_g semidefinite constraints because the polynomials $\widehat{\mathbf{g}}$ are restricted to be non-negative.

Corollary 1 *If $\mathcal{B}$ and Θ are compact then $f_h^*(\mathbf{u}) \leq f_{h+1}^*(\mathbf{u}) \leq f^*(\mathbf{u})$ for all $h \in \mathbb{N}$ and $f_h^*(\mathbf{u}) \to f^*(\mathbf{u})$ as $h \to \infty$.*

Corollary 1 follows from Proposition 3 due to Theorem 4.2 of Lasserre (2001). The only assumption beyond the class $\mathcal{P}$ is a technical assumption requiring that the parameter space be compact. Compactness is useful because it allows us to find B sufficiently large such that the redundant polynomial constraint

$$B - |\beta|_2^2 - |\theta|_2^2 \geq 0 \tag{17}$$

holds. In practice, we augment the constraints $\widehat{\mathbf{g}}(\beta, \theta, \gamma) \geq 0$ to include (17) prior to applying the semidefinite hierarchy.

Though compactness is a common technical assumption, in practice we may often not have compact $\mathcal{B}$ and Θ. For example, we may have $\mathcal{B} = \mathbb{R}^{d_X}$. In this case we suggest increasing B until (17) ceases to bind at the solution. If the SNIV

confidence set is unbounded in direction $\mathbf{u}$, (17) will always bind. In practice this is of little consequence since there is little distinction between $f^*(\mathbf{u})$ being $-\infty$ or an arbitrarily small finite constant. Thus, when the parameter space is not compact, our approach characterizes the intersection of the SNIV confidence set with an arbitrarily large ℓ_2 ball.

The intuition for the result that $f_h^*(\mathbf{u}) \leq f^*(\mathbf{u})$ for all $h \in \mathbb{N}$ comes from convex relaxation. By replacing rank 1 constraints for the moment and localizing matrices by positive semidefinite constraints, we minimize over a larger set, hence it must be that we obtain a lower bound on the optimal value. The intuition for $f_h^*(\mathbf{u}) \leq f_{h+1}(\mathbf{u})$ for all $h \in \mathbb{N}$ is that increasing h reduces the size of the set over which we minimize, hence must always deliver a larger optimal value. The computational trade-off is also clear from the form of problem (16) because the dimension of the moment matrix $\mathbf{M}_h(\mu)$ is $\binom{d_\delta + h}{h}$, which is increasing in h. Similarly, the dimensions of the localizing matrices are combinatorically increasing in h. Thus, increasing h delivers a tigher bound but at increased computational cost.

We implement the hierarchy using the following algorithm proposed by Lasserre (2015).

Algorithm 1 *Inititialize $h = 1$ and the largest level of the hierachy $\overline{h} \in \mathbb{N}$. Then,*

1. *Solve the semidefinite optimization problem in (16) to obtain optimal value $f_h^*(\mathbf{u})$ and optimal solution μ^* (if it exists).*

2. *If there is no optimal solution μ^* and $h < \overline{h}$ then increase h by one and go to step 1. If there is no optimal solution μ^* and $h = \overline{h}$ then terminate the algorithm with bound $f_{h-1}^*(\mathbf{u})$.*

3. *If rank $\mathbf{M}_h(\mu^*) = $ rank $\mathbf{M}_{h-e}(\mu^*)$ (where $e \triangleq \max_{j \in [d_g]} e_j$) then we know $f_h^*(\mathbf{u}) = f^*(\mathbf{u})$. Terminate the algorithm with the exact bound $f_h^*(\mathbf{u})$.*

4. If rank $\mathbf{M}_h(\mu^*) \neq$ rank $\mathbf{M}_{h-e}(\mu^*)$ and $h < \overline{h}$ then increase h by one and go to step 1. Else, if $h = \overline{h}$, terminate the algorithm with the lower bound $f_h^*(\mathbf{u})$.

The basic idea is to begin with the most computationally tractable semidefinite program with $h = 1$ and continue to increase h until either we know that $f_h^*(\mathbf{u}) = f^*(\mathbf{u})$ (step 3) or we hit the largest computationally feasible level of the hierarchy ($\overline{h}$). In practice, $\overline{h}$ is determined by the size of the problem and the available computational resources. In our Monte-Carlo experiment we use $\overline{h} = 2$ on a standard desktop machine. Step 3 provides a stopping criterion which can be used to establish finite convergence of the hierarchy. For brevity, we do not provide technical conditions under which finite convergence is possible, which can be found in Lasserre (2015) (see Theorem 6.5) and involve standard Karusch-Kuhn-Tucker conditions for an optimal solution to be a local minimizer of a nonlinear program. In fact, these conditions imply that finite convergence is achieved generically (Lasserre, 2015) (see Theorem 7.6), though there is no guarantee that it is achieved for small values of h. In our Monte-Carlo experiment we achieve finite convergence with high frequency in some designs but with low frequency in others.

5. Inverting other robust tests

In the standard linear instrumental variables setting with large n and fixed $d_Z \geq d_X$, robust inference can be conducted by inverting robust tests (Andrews et al., 2019). In this section we show that the computational approach of Section 4 can be applied to do so. There are myriad such tests, including but not limited to, the Anderson Rubin (AR) test (Anderson and Rubin, 1949), Lagrange-multiplier (LM) test (Kleibergen, 2002; Moreira, 2002) and the Conditional Likelihood Ratio (CLR) test (Moreira, 2003). All of these tests have a non-rejection region (i.e., a confidence

set) of the form

$$\widetilde{\mathcal{C}} \triangleq \{\beta \in \mathcal{B} : \widehat{p}(\beta) \leq \widehat{q}_\alpha(\beta)\}, \tag{18}$$

where $\widehat{p}$ and $\widehat{q}_\alpha$ are polynomials and the coefficients of $\widehat{q}_\alpha$ depend on the confidence level α. As with the SNIV confidence set, $\widetilde{\mathcal{C}}$ is semi-algebraic whenever $\mathcal{B}$ is. The polynomial inequality in the definition of $\widetilde{\mathcal{C}}$ can be degree 2 (AR test, CLR test with $d_X = 1$) or larger (LM test, CLR test with $d_X > 1$). For example, the AR test under homoskedasticity uses $\widehat{p}(\beta) = \widehat{p}_{AR}(\beta)$ and $\widehat{q}_\alpha(\beta) = \widehat{q}_{AR,\alpha}(\beta)$, where

$$\widehat{p}_{AR}(\beta) \triangleq \mathbf{U}(\beta)^\top \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{U}(\beta), \quad \widehat{q}_{AR,\alpha}(\beta) \triangleq C_\alpha(d_Z)\widehat{q}_{AR}(\beta), \tag{19}$$

$$\widehat{q}_{AR}(\beta) \triangleq (1, -\beta^\top)(\mathbf{y}, \mathbf{X})^\top(\mathbf{I}_n - \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top)(\mathbf{y}, \mathbf{X})(1, -\beta^\top)^\top/(n - d_Z), \tag{20}$$

$\mathbf{y}$ is the $n \times 1$ vector of outcomes, $\mathbf{X}$ is the $n \times d_X$ matrix of regressors, $\mathbf{U}(\beta) \triangleq \mathbf{y} - \mathbf{X}\beta$, $\mathbf{Z}$ is the $n \times d_Z$ matrix of instruments and $C_\alpha(d)$ is the $1 - \alpha$ quantile of the χ_d^2 distribution. None of the above tests yield a convex confidence set (Mikusheva, 2010), making a grid search computationally demanding (see Andrews (2016), supplementary material). Alternatives to a grid search (e.g., Mikusheva (2010) for the CLR test) can also be computationally intensive for moderate d_X. Hierarchies of semidefinite optimization problems provide a practical alternative.

Sometimes the object of interest may be a function of β^* of dimension smaller than d_X (see Section 4.1). For example, one may be interested in a sub-vector of β^*. To obtain coverage probability $1 - \alpha$ for a confidence set for a sub-vector (e.g., a confidence interval), we need to adjust $\widehat{q}_\alpha$. Guggenberger et al. (2012), Guggenberger et al. (2019) and Guggenberger et al. (2021) provide appropriate adjustments for the AR test. Decomposing $\beta = (\beta_1^\top, \beta_2^\top)^\top$, the results of Guggenberger et al. (2012) imply that an asymptotic $1 - \alpha$ confidence set for β_1^* under

homoskedasticity is

$$\widetilde{\mathcal{C}}_{AR} \triangleq \left\{ \beta_1 : \inf_{\beta_2} \frac{\widehat{p}_{AR}(\beta)}{\widehat{q}_{AR}(\beta)} \leq C_\alpha(d_Z - d_X + d_{X_1}) \right\}, \tag{21}$$

where d_{X_1} is the dimension of β_1.[6] Thus, for a given direction $\mathbf{u}_1 \in \mathbb{R}^{d_{X_1}}$ normalized to have $|\mathbf{u}_1|_2 = 1$ we seek to compute $\inf_{\beta_1 \in \widetilde{\mathcal{C}}_{AR}} \mathbf{u}_1^\top \beta_1$, equivalently expressed as

$$\inf_{\beta : \widehat{p}_{AR}(\beta) \leq C_\alpha(d_Z - d_X + d_{X_1})\widehat{q}_{AR}(\beta)} \mathbf{u}^\top \beta, \tag{22}$$

where $\mathbf{u} = (\mathbf{u}_1^\top, \mathbf{0}^\top)^\top$ is $d_X \times 1$. This is a polynomial optimization problem, hence we can apply convex hierarchies to find a monotonic sequence of lower bounds. For the special case of a confidence interval we have $d_{X_1} = 1$ hence only need consider $\mathbf{u}_1 = \pm 1$. Guggenberger et al. (2019) replace $C_\alpha(d_Z - d_X + d_{X_1})$ with an alternative which delivers a less conservative confidence set in a finite sample and Guggenberger et al. (2021) extend the approach to allow for conditional heterokskedasticity. Thus, we can apply hierarchies of semidefinite optimization problems to (22) in order to obtain AR confidence intervals which can be rapidly computed.

6. Monte-Carlo

To illustrate our approach, we consider a setting with $d_X = 10$ endogenous regressors. We choose this design because d_X is large enough to render a grid search infeasible yet small enough to permit many replications of our experiment on a standard desktop machine within a reasonable timeframe.

We consider an i.i.d. sample of size $n = 2000$ satisfying (1)-(2). The instruments are related to the regressors according to $\mathbb{E}[ZV(\Pi)] = 0$ where $V(\Pi) \triangleq X - \Pi Z$ and Π is $d_X \times d_Z$. We set $\beta^* = (1, -1, 0, ..., 0)^\top$ and vary Π^* by design, as

---

[6]This decomposition is without loss of generality because the regressors can be reordered.

explained below. The instruments follow $\mathcal{N}(0, I_{d_Z})$ and the error terms verify $(U(\beta^*), V(\Pi^*)^\top)^\top \sim \mathcal{N}(0, \Omega)$, where $\Omega_{11} = 1$ (homoskedasticity) or $\Omega_{11} = Z_1^2$ (conditional heteroskedasticity), $\Omega_{1j} = (-1)^j (1 - \pi^*)/5$, $\Omega_{jj} = 1 - \pi^*$ for $j > 1$ and all other entries are equal to zero. The parameter $\pi^* \in [0, 1]$ determines the fraction of the variance of each regressor which is due to the instruments. In all designs, the variances of each regressor and the structural error are equal to 1.

To compute the SNIV confidence set we choose r_n using Class 1 and Class 3 with $\alpha = 0.05$. To implement the hierarchy, we use $\overline{h} = 2$ and $B = 1000$. We compute the coverage probability for the SNIV confidence set, and, for designs in which it is feasible, the AR confidence set and confidence intervals. For the SNIV and AR confidence sets, we also report the coverage probability for their outer approximations obtained by solving hierarchies of semidefinite optimization problems, defined by $f_{\mathbf{u}}(\beta) \geq f_h^*(\mathbf{u})$ for all $\mathbf{u} \in \mathcal{U}$, where, for $\mathcal{U}$ we use a grid of 1600 points over the surface of an ℓ_2 ball of radius 1.[7]

Our results are collected in Table 1. We focus the discussion of SNIV on Class 1, which, identically to AR, provides an asymptotic coverage guarantee. Class 3 provides a finite guarantee, hence a larger confidence set in all designs. Nevertheless, we find that whenever Class 1 provides an informative confidence set, so does Class 3.

Classical design. We set $d_Z = d_X = 10$, $\pi^* = 0.3$ and $\Pi^* = \sqrt{\pi^*} I_{d_Z}$ and do not impose any sparsity constraint. The SNIV and AR confidence sets have similar coverage, both of which are marginally below the nominal level. The SNIV confidence set is marginally narrower than the AR confidence set. The coverage of the AR confidence intervals are almost exactly equal to the nominal level, and their width is narrower than either of the confidence sets, as expected.

---

[7]In practice we parallelize over $\mathbf{u} \in \mathcal{U}$.

Almost all optimization problems solved yielded an exact global optimum (i.e., $f_h^*(\mathbf{u}) = f^*(\mathbf{u})$). The time taken to solve an optimization problem is a little over one second for SNIV and the AR confidence set/interval. In the design with conditional heteroskedasticity the SNIV confidence set is marginally wider than under homoskedasticity and attains the nominal coverage.

Many instruments. We take the classical design and add 1989 redundant instruments, all drawn from the standard normal distribution. This yields $d_Z = 1999$ instruments and $n = 2000$ observations. The SNIV confidence sets have coverage almost identical to the nominal level but are wider than the classical design. However, they remain sufficiently narrow as to be informative on the sign of the nonzero entries of β^*. In contrast, the AR confidence sets and intervals do not have the correct coverage and are too wide so as to be informative. We also consider an identical design but with $d_Z = 2100$. The SNIV confidence sets are similar to the case of $d_Z = 1999$, whereas AR confidence sets and intervals are not defined. Almost all optimization problems solved yielded a global optimum. The SNIV optimization problems are solved more slowly than the classical design, taking around 4 seconds on average. In the designs with conditional heteroskedasticity the SNIV confidence set has nominal coverage but is marginally narrower than under homoskedasticity, likely because a greater fraction of the optimization problems yielded an exact global optimum.

Weak instruments. We take the classical design and set $\pi^* = 0.03$. The AR confidence sets are narrower than SNIV but have coverage further from the nominal level. The AR confidence intervals have coverage slightly larger than the nominal level. Almost all optimization problems solved yielded a global optimum. In the design with conditional heteroskedasticity the SNIV confidence set is

marginally wider than under homoskedasticity with coverage close to the nominal level. Computation timings are similar to the classical design.

Invalid instruments. We take the classical design with one endogenous regressor and $d_Z = 9$ instruments, all of which are included as regressors (i.e., $X_k = Z_{k-1}$ for $k = 2, 3, ..., d_X$). Hence there are $d_X = 10$ regressors, but only the first is endogenous. We set $\Pi^* = (0, 0, ..., 0, \sqrt{\pi^*}/2, -\sqrt{\pi^*}/2)^\top$ so that only the final two instruments are correlated with the endogenous regressor. We suppose that $S_Q = [d_X]$ (i.e., the relevance of all regressors is questionable) and consider the sparsity certificates $s \in \{2, 3\}$ (recalling that β^* has two nonzero entries). This design is such that $\mathcal{I}_2$ is a singleton, $\mathcal{I}_3$ is not a singleton but is bounded, and $\mathcal{I}_s$ is unbounded for $s > 3$. When $s = 3$, though $\beta_1^* = 1$ we have $\min_{\beta \in \mathcal{I}_3} \beta_1 = 0$ and $\max_{\beta \in \mathcal{I}_3} \beta_1 = 1$. The AR confidence sets and intervals cannot be computed.

The SNIV confidence set has coverage slightly larger than the nominal level. For $s = 2$, SNIV is sufficiently narrow so as to be informative on the sign of the nonzero entries of β^*. For $s = 3$, the width of SNIV for β_1 is around 1.25 on average, which is not sufficiently narrow so as to be informative on the sign of β_1^*. This is expected because β_1^* is not point identified (the identified set has width 1), as explained in the previous paragraph. In the design with conditional heteroskedasticity the SNIV confidence set is marginally wider than under homoskedasticity. Each optimization problem is solved in around 17-27 seconds depending on the sparsity certificate used. This is likely due to the non-convex nature of the sparsity constraint. Nevertheless, the problems are sufficiently tractable so as to allow informative inference on a standard machine.

Endogenous instruments. We take the classical design and but add the instruments $Z_{11} = X_1$ and $Z_{12} = X_2$ (i.e., include the first two regressors as additional

instruments). This results in $d_Z = 12$ instruments, two of which are endogenous, with $\theta_{11} = \Omega_{1,2}$ and $\theta_{12} = \Omega_{1,3}$. We suppose that the researcher questions the exogeneity of the two endogenous instruments and the final three exogenous instruments, hence $S_\perp = [7]$. This implies that are seven instruments known to be exogenous, whereas $d_X = 10$, hence the model using only the instruments known to be exogenous is underidentified. Thus, classical tests of overidentifying restrictions are infeasible.

Since the identified set is otherwise unbounded, we restrict the number of endogenous instruments using the sparsity certificate $\widetilde{s}$ on θ^*. We do not make any sparsity restriction on β^*. Using $\widetilde{s} = 2$ corresponds to the case where we assume that there are ten exogenous instruments, but we do not know all of their identities. This design is such that $\mathcal{I}_{d_X,2}$ is a singleton. In contrast, $\mathcal{I}_{d_X,\widetilde{s}}$ is unbounded for $\widetilde{s} > 2$. We compute the SNIV confidence set for $\widetilde{s} = 2$.

The SNIV confidence set has coverage almost exactly equal to the nominal level and is sufficiently narrow so as to be informative on the sign of the nonzero entries of β^*. Moreover, the SNIV confidence set allows the null hypotheses of $\theta_{S_\perp^c} = 0$ to be (correctly) rejected with probability 0.84. In the design with conditional heteroskedasticity the SNIV confidence set is marginally wider than under homoskedasticity. Each optimization problem is solved more slowly than under the classical design, taking around 36 seconds on average. Nevertheless, the problems are sufficiently tractable so as to allow informative inference on a standard machine.

7. Conclusion

We use self-normalization of sample moments to conduct robust, computationally tractable inference in linear instrumental variables models. We also show that our computational approach is not unique to self-normalzation, and can be applied to

perform fast inversion of other tests. In our view there are two avenues for future work.

First, though SNIV requires minimal assumptions and has desirable statistical and computational properties, when d_X is large it can be conservative when the object of interest is low dimensional (e.g., a single treatment effect). Though this is not an issue in the leading case in which d_X is small (e.g., one endogenous regressor, possibly after partialing our a small number of exogenous regressors) and d_Z may be large (with possibly weak instruments), future work may seek to adapt our approach to perform robust inference directly on a *sub-vector* of parameters of interest.

Second, we believe that our computational approach is applicable beyond the instrumental variables context. An obvious setting to which our results may be applied is that of inference in partially identified models, which is often based on solving programming problems such as (12). A simple example in which the optimization problem is semi-algebraic (hence our approach is applicable) is the 2×2 entry game considered by Kaido et al. (2019).

REFERENCES

ACKERBERG, D. A. AND P. J. DEVEREUX, "Improved JIVE estimators for overidentified linear models with and without heteroskedasticity," *The Review of Economics and Statistics* 91 (2009), 351–362.

ANATOLYEV, S., "Instrumental variables estimation and inference in the presence of many exogenous regressors," *The Econometrics Journal* 16 (2013), 27–72.

———, "Many instruments and/or regressors: A friendly guide," *Journal of Economic Surveys* 33 (2019), 689–726.

ANATOLYEV, S. AND N. GOSPODINOV, "Specification testing in models with many instruments," *Econometric Theory* 27 (2011), 427–441.

ANDERSON, T., "Origins of the limited information maximum likelihood and two-stage least squares estimators," *Journal of econometrics* 127 (2005), 1–16.

ANDERSON, T. W. AND H. RUBIN, "Estimation of the parameters of a single equation in a complete system of stochastic equations," *The Annals of Mathematical Statistics* 20 (1949), 46–63.

ANDREWS, I., "Conditional linear combination tests for weakly identified models," *Econometrica* 84 (2016), 2155–2182.

ANDREWS, I., J. H. STOCK AND L. SUN, "Weak instruments in instrumental variables regression: Theory and practice," *Annual Review of Economics* 11 (2019), 727–753.

ANGRIST, J. D., G. W. IMBENS AND A. B. KRUEGER, "Jackknife instrumental variables estimation," *Journal of Applied Econometrics* 14 (1999), 57–67.

AUERBACH, E., "Testing for Differences in Stochastic Network Structure," *Econometrica* 90 (2022), 1205–1223.

BEKKER, P. A., "Alternative approximations to the distributions of instrumental variable estimators," *Econometrica: Journal of the Econometric Society* (1994), 657–681.

BELLONI, A., D. CHEN, V. CHERNOZHUKOV AND C. HANSEN, "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica* 80 (2012), 2369–2429.

BELLONI, A., C. HANSEN AND W. NEWEY, "High-dimensional linear models with many endogenous variables," *Journal of Econometrics* (2022).

BERTAIL, P., E. GAUTHERAT AND H. HARARI-KERMADEC, "Exponential bounds for multivariate self-normalized sums," *Electronic Communications in Probability* 13 (2008), 628–640.

CANDES, E. AND T. TAO, "The Dantzig selector: Statistical estimation when p is much larger than n," *The annals of Statistics* 35 (2007), 2313–2351.

CHAO, J. C., J. A. HAUSMAN, W. K. NEWEY, N. R. SWANSON AND T. WOUTERSEN, "Testing overidentifying restrictions with many instruments and heteroskedasticity," *Journal of econometrics* 178 (2014), 15–21.

CHAO, J. C. AND N. R. SWANSON, "Consistent estimation with a large number of weak instruments," *Econometrica* 73 (2005), 1673–1692.

CHAO, J. C., N. R. SWANSON, J. A. HAUSMAN, W. K. NEWEY AND T. WOUTERSEN, "Asymptotic distribution of JIVE in a heteroskedastic IV regression with many instruments," *Econometric Theory* 28 (2012), 42–86.

CHEN, X., Q.-M. SHAO, W. B. WU AND L. XU, "Self-normalized Cramér-type moderate deviations under dependence," *The Annals of Statistics* 44 (2016), 1593–1617.

CHERNOZHUKOV, V., D. CHETVERIKOV AND K. KATO, "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors," *The Annals of Statistics* 41 (2013), 2786–2819.

DONALD, S. G. AND W. K. NEWEY, "Choosing the number of instruments," *Econometrica* 69 (2001), 1161–1191.

DUFOUR, J.-M., "Some impossibility theorems in econometrics with applications to structural and dynamic models," *Econometrica* (1997), 1365–1387.

FARRELL, M. H., T. LIANG AND S. MISRA, "Deep learning for individual heterogeneity: an automatic inference framework," *arXiv preprint arXiv:2010.14694* (2020).

FENG, M., J. E. MITCHELL, J.-S. PANG, X. SHEN AND A. WÄCHTER, "Complementarity formulations of l0-norm optimization problems," *Industrial Engineering and Management Sciences. Technical Report. Northwestern University, Evanston, IL, USA* 5 (2013).

GAUTIER, E. AND C. ROSE, "High-dimensional instrumental variables regression and confidence sets," *arXiv preprint arXiv:1105.2454* (2021).

GAUTIER, E., C. ROSE AND A. TSYBAKOV, "High-dimensional instrumental variables regression and confidence sets," *https://arxiv.org/abs/1105.2454v5* (2018).

GAUTIER, E. AND A. TSYBAKOV, "High-dimensional instrumental variables regression and confidence sets," *arXiv preprint arXiv:1105.2454* (2011).

GOLD, D., J. LEDERER AND J. TAO, "Inference for high-dimensional instrumental variables regression," *Journal of Econometrics* 217 (2020), 79–111.

GUGGENBERGER, P., F. KLEIBERGEN AND S. MAVROEIDIS, "A more powerful subvector Anderson Rubin test in linear instrumental variables regression," *Quantitative Economics* 10 (2019), 487–526.

———, "A Powerful Subvector Anderson Rubin Test in Linear Instrumental Variables Regression with Conditional Heteroskedasticity," *arXiv preprint arXiv:2103.11371* (2021).

GUGGENBERGER, P., F. KLEIBERGEN, S. MAVROEIDIS AND L. CHEN, "On

the asymptotic sizes of subset Anderson–Rubin and Lagrange multiplier tests in linear instrumental variables regression," *Econometrica* 80 (2012), 2649–2666.

HANSEN, C., J. HAUSMAN AND W. NEWEY, "Estimation with many instrumental variables," *Journal of Business & Economic Statistics* 26 (2008), 398–422.

HANSEN, C. AND D. KOZBUR, "Instrumental variables estimation with many weak instruments using regularized JIVE," *Journal of Econometrics* 182 (2014), 290–308.

HANSEN, L. P., "Large sample properties of generalized method of moments estimators," *Econometrica: Journal of the econometric society* (1982), 1029–1054.

HAUSMAN, J. A., W. K. NEWEY, T. WOUTERSEN, J. C. CHAO AND N. R. SWANSON, "Instrumental variable estimation with heteroskedasticity and many instruments," *Quantitative Economics* 3 (2012), 211–255.

JING, B.-Y., Q.-M. SHAO AND Q. WANG, "Self-normalized Cramér-type large deviations for independent random variables," *The Annals of probability* 31 (2003), 2167–2215.

KAIDO, H., F. MOLINARI AND J. STOYE, "Constraint qualifications in partial identification," *Econometric Theory* (2019), 1–24.

KANG, H., A. ZHANG, T. T. CAI AND D. S. SMALL, "Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization," *Journal of the American statistical Association* 111 (2016), 132–144.

KLEIBERGEN, F., "Pivotal statistics for testing structural parameters in instrumental variables regression," *Econometrica* 70 (2002), 1781–1803.

KOLESÁR, M., "Minimum distance approach to inference with many instruments," *Journal of Econometrics* 204 (2018), 86–100.

KOLESÁR, M., R. CHETTY, J. FRIEDMAN, E. GLAESER AND G. W. IMBENS, "Identification and inference with many invalid instruments," *Journal of Business & Economic Statistics* 33 (2015), 474–484.

LASSERRE, J. B., "Global optimization with polynomials and the problem of moments," *SIAM Journal on optimization* 11 (2001), 796–817.

———, *An introduction to polynomial and semi-algebraic optimization*, volume 52 (Cambridge University Press, 2015).

LASSERRE, J. B., K.-C. TOH AND S. YANG, "A bounded degree SOS hierarchy for polynomial optimization," *EURO Journal on Computational Optimization* 5 (2017), 87–117.

LEE, W., "Identification and estimation of dynamic random coefficient models," *Working paper* (2020).

LEE, Y. AND R. OKUI, "Hahn–Hausman test as a specification test," *Journal of Econometrics* 167 (2012), 133–139.

MIKUSHEVA, A., "Robust confidence sets in the presence of weak instruments," *Journal of Econometrics* 157 (2010), 236–247.

MOREIRA, M. J., *Tests with correct size in the simultaneous equations model*, Ph.D. thesis, University of California, Berkeley (2002).

———, "A conditional likelihood ratio test for structural models," *Econometrica* 71 (2003), 1027–1048.

PINELIS, I., "Extremal probabilistic problems and Hotelling's T2 test under a symmetry condition," *The Annals of Statistics* (1994), 357–368.

Rose, C., "Identification of Spillover Effects using Panel Data," *Working paper* (2018).

Sargan, J. D., "The estimation of economic relationships using instrumental variables," *Econometrica: Journal of the Econometric Society* (1958), 393–415.

Stock, J. and M. Yogo, *Asymptotic distributions of instrumental variables statistics with many instruments*, volume 6 (Chapter, 2005).

Tibshirani, R., "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1996), 267–288.

van Hasselt, M., "Many instruments asymptotic approximations under nonnormal error distributions," *Econometric Theory* 26 (2010), 633–645.

Weisser, T., J. B. Lasserre and K.-C. Toh, "Sparse-BSOS: a bounded degree SOS hierarchy for large scale polynomial optimization with sparsity," *Mathematical Programming Computation* 10 (2018), 1–32.

Zhang, D. and W. B. Wu, "Gaussian approximation for high dimensional time series," *The Annals of Statistics* 45 (2017), 1895–1919.

Appendix

Proof of Proposition 1. The result follows by applying a union bound to the bounds in Jing et al. (2003) (Class 1), Bertail et al. (2008) (Class 2) or Pinelis (1994) (Class 3), which yield the corresponding values of r_n in the main text. For Class 1, the coverage is asymptotic because $C_1 \mu_{2+\delta} \left(1 + \sqrt{n} r_n\right)^{2+\delta} n^{-\delta/2} \to 0$ where C_1 is an unknown universal constant. For Class 2, the results of Bertail et al. (2008) yield the bound $\sqrt{2/(n/\log(d_Z(2e+1)/\alpha) - \mu_4)}$ and we use $n - \mu_4 \log(d_Z(2e+1)/\alpha) \geq n/2$ to obtain r_n which does not depend on the

unknown μ_4. $\qquad\square$

Proof of Proposition 2. The result follows from Corollary 2.1 in Chernozhukov et al. (2013) and the fact that $\mathbb{E}_n[U(\beta)^2]$ is consistent for $\mathbb{E}[U(\beta)^2]$ under the conditions of Class 4. $\qquad\square$

Proof of Proposition 3. Under Classes 1-3, the first inequality in the definition of the SNIV confidence set can be rewritten as

$$\mathbb{E}_n[(Z_l U(\beta) - \theta_l)^2]^{-1/2}|\mathbb{E}_n[Z_l U(\beta)] - \theta_l| \le r_n \quad \forall l \in [d_Z]. \tag{23}$$

Squaring both sides and rearranging yields the equivalent degree 2 polynomial inequalities

$$r_n^2 \mathbb{E}_n[(Z_l U(\beta) - \theta_l)^2] - (\mathbb{E}_n[Z_l U(\beta)] - \theta_l)^2 \ge 0 \quad \forall l \in [d_Z]. \tag{24}$$

Under Class 4 (which is not applicable with potentially endogenous instruments), we obtain instead the degree 2 polynomial inequalities

$$\widehat{r}^2 \mathbb{E}_n[Z_l^2]\mathbb{E}_n[U(\beta)^2] - \mathbb{E}_n[Z_l U(\beta)]^2 \ge 0 \quad \forall l \in [d_Z]. \tag{25}$$

Without loss of generality, suppose that we order the indices of the regressors such that $S_Q = [d_Q]$. The second inequality in the definition of the SNIV confidence set is equivalently expressed using the polynomial (in)equalities

$$
\begin{aligned}
\exists \zeta \in [0,1]^{d_Q}: \quad & \zeta_k^a(1 - \zeta_k)^b = 0 && \forall k \in S_Q, (a,b) \in \mathbb{N}^2, \\
& (1 - \zeta_k)^a \beta_k = 0 && \forall k \in S_Q, a \in \mathbb{N}, \\
& s - \sum_{k \in S_Q} \zeta_k \ge 0, && \tag{26}
\end{aligned}
$$

where, due to the constraint $\zeta_k(1 - \zeta_k) = 0$, ζ comprises d_Q indicators for the nonzero entries of β_{S_Q} (see Feng et al. (2013)). The third inequality in the definition of the SNIV confidence set is obtained identically introducing $\eta \in [0,1]^{\widetilde{d}_Q}$. Since equalities can be defined using two inequalities of opposing directions we can stack all of the polynomial inequalities as $\widehat{\mathbf{g}}(\beta, \theta, \gamma) \geq 0$ where $\gamma \triangleq (\zeta^\top, \eta^\top)^\top$. $\square$

Table 1: *Monte Carlo*

	β_1	β_2	β_3	β_4	β_5	Cover	Exact	Time (s)	$\theta \neq 0$
Classical ($d_Z = 10, \pi^* = 0.3$)									
AR	0.361	0.361	0.360	0.361	0.360	0.942 (0.994)	1.000	1.08	
SNIV (Class 1)	0.343	0.343	0.341	0.342	0.343	0.944 (0.992)	1.000	1.34	
SNIV (Class 1, het.)	0.349	0.349	0.347	0.348	0.349	0.946 (0.996)	1.000	12.10	
SNIV (Class 3)	0.419	0.420	0.418	0.419	0.419	0.988 (1.000)	1.000	15.57	
AR (CI)	0.163	0.163	0.163	0.163	0.163	0.949	1.000	1.07	
Many instruments ($d_Z = 1999, \pi^* = 0.3$)									
AR	19.373	19.427	19.374	19.391	19.399	0.324 (1.000)	0.995	0.73	
SNIV (Class 1)	0.634	0.635	0.632	0.633	0.633	0.956 (1.000)	0.855	4.17	
SNIV (Class 1, het.)	0.608	0.610	0.607	0.608	0.606	0.954 (1.000)	0.874	8.29	
SNIV (Class 3)	0.734	0.736	0.733	0.733	0.733	0.988 (1.000)	0.692	9.48	
AR (CI)	19.373	19.427	19.374	19.391	19.399	1.000	0.993	0.73	
Many instruments ($d_Z = 2100, \pi^* = 0.3$)									
SNIV (Class 1)	0.635	0.632	0.635	0.632	0.634	0.954 (1.000)	0.863	4.41	
SNIV (Class 1, het.)	0.609	0.608	0.611	0.608	0.609	0.952 (1.000)	0.878	8.50	
SNIV (Class 3)	0.735	0.732	0.736	0.732	0.735	0.988 (1.000)	0.681	9.73	
Weak instruments ($d_Z = 10, \pi^* = 0.03$)									
AR	6.4959	6.314	6.208	6.253	6.189	0.942 (1.000)	0.991	1.22	
SNIV (Class 1)	12.776	12.833	12.782	12.892	12.875	0.944 (1.000)	0.999	1.59	
SNIV (Class 1, het.)	12.771	12.823	12.775	12.886	12.869	0.946 (1.000)	1.000	12.48	
SNIV (Class 3)	13.819	13.886	13.852	13.950	13.935	0.988 (1.000)	1.000	15.95	
AR (CI)	0.7021	0.695	0.680	0.688	0.688	0.963	1.000	1.23	
Invalid instruments ($d_Z = 9, \pi^* = 0.3$)									
SNIV (Class 1, $s = 2$)	0.276	0.136	0.000	0.000	0.000	0.968 (0.968)	0.000	17.09	
SNIV (Class 1, $s = 3$)	1.251	0.163	0.096	0.095	0.096	0.968 (0.968)	0.000	16.51	
SNIV (Class 1, het., $s = 2$)	0.280	0.220	0.000	0.000	0.000	0.972 (0.972)	0.000	25.26	
SNIV (Class 1, het., $s = 3$)	1.253	0.245	0.100	0.099	0.099	0.972 (0.972)	0.000	25.05	
SNIV (Class 3, $s = 2$)	0.335	0.157	0.000	0.000	0.000	0.988 (0.988)	0.000	27.21	
SNIV (Class 3, $s = 3$)	1.279	0.188	0.111	0.110	0.111	0.972 (0.988)	0.000	26.64	
Endogenous instruments ($d_Z = 12, \pi^* = 0.3$)									
SNIV (Class 1, $\widetilde{s} = 2$)	0.946	0.958	1.238	1.236	1.225	0.948 (0.976)	0.000	36.01	0.840 (0.304)
SNIV (Class 1, het., $\widetilde{s} = 2$)	0.956	0.959	1.237	1.253	1.236	0.956 (0.992)	0.000	36.42	0.824 (0.308)
SNIV (Class 3, $\widetilde{s} = 2$)	1.206	1.211	1.543	1.553	1.537	0.984 (1.000)	0.000	36.83	0.760 (0.296)

Notes: $d_X = 10, n = 2000, \beta^* = (1, -1, 0, ..., 0)^\top$. We report the mean width of the confidence region for $\beta_1, ..., \beta_5$. 'het' is the design with conditional heteroskedasticity. 'Cover' is for $(\beta^\top, \theta^\top)^\top$ for SNIV and β for AR. In parentheses, we include the coverage probability of the outer approximation of the confidence set $(f_{\mathbf{u}}(\beta) \geq f_h^*(\mathbf{u})$ for all $\mathbf{u} \in \mathcal{U})$. For $\mathcal{U}$ we use a grid of 1600 points over the surface of ℓ_2 ball of radius 1. 'Exact' is the fraction of optimization problems solved exactly $(f_h^* = f^*)$. 'Time (s)' is the mean time taken to solve an optimization problem in seconds. '$\theta \neq 0$' is the fraction of datasets in which an endogenous instrument was detected. In parentheses, both were detected. All programs use $\bar{h} = 2, B = 100$. 500 replications.