High-dimensional instrumental variables regression and confidence sets

Eric Gautier, Christiern Rose

HIGH-DIMENSIONAL INSTRUMENTAL VARIABLES REGRESSION AND CONFIDENCE SETS

ERIC GAUTIER AND CHRISTIERN ROSE

Abstract. This article considers inference in linear models with d_X regressors, some or many of which could be endogenous, and d_Z instrumental variables (IVs). d_Z can range from less than d_X to any order smaller than an exponential in the sample size. For moderate d_X, identification robust confidence sets are obtained by solving a hierarchy of semidefinite programs. For large d_X, we propose the STIV estimator. The analysis of its error uses sensitivity characteristics introduced in this paper. Robust confidence sets are derived by solving linear programs. Results on rates of convergence, variable selection, and confidence sets which "adapt" to the sparsity are given. Generalizations include models with endogenous IVs and systems of equations with approximation errors. We also analyse confidence bands for vectors of linear functionals and functions using bias correction. The application is to a demand system with approximation errors, cross-equation restrictions, and thousands of endogenous regressors.

1. Introduction

The high-dimensional paradigm concerns inference in models where the number of regressors d_X is large relative to the number of observations n. A challenging situation is when d_X is much larger than n ($d_X \gg n$) but there is an unknown small set of important regressors. This can happen for various reasons. Researchers increasingly have access to large datasets and theory is often silent on the correct choice of regressors. The number of observations can be limited because data is costly to obtain, because there simply exist few units (*e.g.*, countries or states), or because the researcher is interested in a stratified analysis. Even if the relevant regressors are known, it is atypical to know the functional form, and so functions of regressors can be included. It is then tempting to use many functions to incur a small approximation error. A model can also be low-dimensional with questionable instrument exogeneity, which one justifies with control variables, giving rise to a second model where the instrument is exogenous but there is an additional nonparametric function. In social effects models with unobserved networks, individual and peer outcomes can be determined simultaneously, peer identities are unobserved, and the number of peers can be small if link formation is costly. Multiple treatment models can also exhibit high-dimensionality, for example if there are group level heterogeneous effects and many groups.

Keywords: Instrumental variables, sparsity, endogeneity, confidence sets, variable selection, unknown variance, robustness to identification, bias correction.

We acknowledge financial support from ERC POEMH and ANR-17-EURE-0010. We warmly thank Alexandre Tsybakov whose reflections on this paper have been fundamental. We thank James Stock, Elie Tamer, Ulrich Müller, and the referees for comments that greatly improved this paper. We thank the participants of the seminars at Bocconi, Brown, Cambridge, CEMFI, CREST, Compiègne, Harvard-MIT, Institut Henri Poincaré, LSE, Madison, Mannheim, Oxford, Paris 6 and 7, Pisa, Princeton, Queen Mary, Toulouse, UC Louvain, Valparaíso, Wharton, Yale, of the 2011 SPA, Saint-Flour, ERCIM, and 2012 CIREQ and French Econometrics conference for comments on the main paper; and, since March 2014, of the seminars at CORE, Duke, LSE, TSE, Toulouse Paul Sabatier, UCSD, Zurich, EC2, CIREQ Montreal, Vanderbildt/Cemmap, Asian Meetings, and EcoSta conferences for comments on the results of Section 6 which was originally a separate paper. Previous versions of this paper are on arXiv and we refer to them as (v1) to (v5)for 1105.2454 and (v2b) for 1812.11330v2. (v5) corresponds to the previous revision sent on March 17, 2018; this one was sent on November 4, 2019.

The usual econometrics toolbox and fixed d_X large n asymptotic framework are not appropriate when there is high-dimensionality. Comparing models for all subsets of regressors is impossible when d_X is moderately large. The high-dimensional literature proposes methods that are computationally feasible. For high-dimensional regression, the Lasso ([31]) replaces the penalty on the number of nonzeros in the parameter by itsℓ_1-norm. The Dantzig Selector of [16] is a linear program.

To address endogeneity in a cross-section, researchers usually rely on instrumental variables (henceforth IVs). This paper concerns the high-dimensional linear structural model where some or many regressors are endogenous but the structural parameter β is *sparse*, meaning it has few nonzero entries, or *approximately sparse*, meaning it is well approximated by a sparse vector. It allows for a number of IVs $d_Z \gg n$ but of any order less than $\exp(n)$. It also allows for $d_Z < d_X$ when β is sparse.

If there are weak and/or many IVs, inference based on asymptotic approximations can fail even if d_X is small. Strong IVs are often scarce, particularly when there are many endogenous regressors. For these reasons, this paper pays particular attention to robustness to identification and finite sample validity. We make use of a ℓ_∞-norm statistic derived from the moment condition. This is close in spirit to identification robust test inversion in which the exogeneity is the null hypothesis and a confidence set is formed by all parameters which are not rejected. The Anderson-Rubin test is one such example. In practice, tests are conducted over a grid, which is only feasible for small d_X. This paper does not rely on grids but on various well-structured convex relaxations (linear, conic or semidefinite), hence on polynomial time optimization routines. Our approach does not rely on sparsity or approximate sparsity of a high-dimensional first-stage (as in two-stage least-squares, henceforth 2SLS). It does not estimate a first-stage reduced form to allow for uniformity in the first-stage coefficients.

Our main contributions are as follows. First, we provide confidence sets for a vector of functions of β. Some of our confidence sets are uniform over identifiable parameter vectors and the distribution of the data among classes which leave the dependence structure between the regressors and IVs unrestricted, implying robustness to identification. Second, we propose the *Self Tuned Instrumental Variables* (henceforth STIV) estimator, and establish error bounds and rates of convergence. These are based on stronger assumptions, including on features of the joint distribution of the regressors and IVs. The STIV estimator can also be a pilot (first-stage) estimator. We use it to perform variable selection and obtain confidence sets which adapt to the sparsity. We also use it to conduct joint inference on approximately linear functions of β based on a data-driven bias correction, itself obtained with a variant of STIV. Our extensions are to models where a few IVs can be endogenous, to systems of models with approximation errors (hence nonparametric IV) and cross-equation restrictions. The proposed methods jointly estimate standard errors of the structural error or of the moments, making the tuning parameters data-driven. This paper restricts attention to linear moments for computational reasons.

2. Preliminaries

Notations. The symbol $\triangleq$ means is defined as. Inequality between vectors is defined entrywise. $(e_k)_{k\in[d_X]}$ is the canonical basis of $\mathbb{R}^{d_X}$ and $(f_l)_{l\in[d_Z]}$ for $\mathbb{R}^{d_Z}$, $\mathcal{M}_{d,d'}$ the set of $d \times d'$ matrices, and 0 (resp. 1) sometimes denote a vector of 0 (resp. 1). For $1 \leq p \leq \infty$ and a matrix Δ, $|\Delta|_p$ is the ℓ_p-norm of the vectorized Δ. $\Delta_{f,k}$ is the element at row f and column k, $\Delta_{f,\cdot}$ (resp. $\Delta_{\cdot,k}$) is the f^{th} row (resp. column) of Δ. For $S \subseteq [d] \times [d']$, $|S|$ is its cardinality and S^c its complement. For $\Delta \in \mathcal{M}_{d,d'}$, let $S(\Delta) = \{(k,l) \in [d] \times [d'] : \Delta_{k,l} \neq 0\}$ be the support of Δ, and, for $S \subseteq [d] \times [d']$, we define $\Delta_S \triangleq (\Delta_{k,l}\mathbb{1}\{(k,l) \in S\})_{k\in[d],l\in[d']}$. For $a \in \mathbb{R}$, we set $a_+ \triangleq \max(0,a)$. We use the conventions $a/0 = \infty$ for $a > 0$.

The data comprises outcomes $\mathbf{Y} \in \mathbb{R}^n$, regressors $\mathbf{X} \in \mathcal{M}_{n,d_X}$, and IVs $\mathbf{Z} \in \mathcal{M}_{n,d_Z}$. $\mathbb{P}$ is the distribution of the data and $\mathbb{E}[\cdot]$ the expectation under $\mathbb{P}$. The set $S_I \subseteq [d_X]$ collects the indices of

the regressors included in the list of IVs and $S_Q \subseteq [d_X]$ of size d_Q collects the indices of the regressors for which the relevance is questionable. The vector $\varphi(\beta)$ of functions of β on which we make inference has dimension d_F. Some results are asymptotic in $n \to \infty$ in which case d_Z, d_X, d_Q, d_F, and s can increase with n. For $d \in \mathbb{N}$ and a random matrix $\mathbf{R} \in \mathcal{M}_{n,d}$, the sample and population means are $\mathbb{E}_n[R] \triangleq \sum_{i\in[n]} \mathbf{R}_{i,\cdot}/n$ and $\mathbb{E}[R] \triangleq \sum_{i\in[n]} \mathbb{E}\left[\mathbf{R}_{i,\cdot}\right]/n$ and, for $k \in [d]$ and $p \geq 0$, $\mathbb{E}_n\left[R_k^p\right]$ and $\mathbb{E}\left[R_k^p\right]$ are obtained by replacing $\mathbf{R}_{i,k}$ by $\mathbf{R}_{i,k}^p$, $\mathbf{D_R}$ is the diagonal matrix in $\mathcal{M}_{d,d}$ with entries $\mathbb{E}_n\left[R_k^2\right]^{-1/2}$ for $k \in [d]$ and we write D_R for the population counterpart. We define the function $b \in \mathbb{R}^{d_X} \to \mathbf{U}(b) \triangleq \mathbf{Y} - \mathbf{X}b \in \mathbb{R}^n$. For $b \in \mathbb{R}^{d_X}$, $\mathbb{P}(b)$ is the distribution of $(\mathbf{X}, \mathbf{Z}, \mathbf{U}(b))$ implied by $\mathbb{P}$. We write $\widehat{\Psi} \triangleq \mathbf{D_Z}\mathbf{Z}^\top\mathbf{X}\mathbf{D_X}/n$, and $\Psi \triangleq D_Z\mathbb{E}[ZX^\top]D_X$. For a mean zero random variable A, $\sigma_A \triangleq \mathbb{E}[A^2]^{1/2}$.

Baseline Model. The linear IV model is

$$\forall i \in [n],\ \mathbb{E}\left[\mathbf{Z}_{i,\cdot}^\top \mathbf{U}_i(\beta)\right] = 0 \tag{2.1}$$

$$\beta \in \mathcal{B},\ \mathbb{P}(\beta) \in \mathcal{P}, \tag{2.2}$$

where $\mathcal{B} \subseteq \mathbb{R}^{d_X}$ can account for restrictions on β and $\mathcal{P}$ is a nonparametric class. We write $\forall i \in [n]$ in (2.1) is because the data need not be identically distributed. The set $\mathcal{I}$ collects the vectors which satisfy (2.1)-(2.2). The researcher believes there exists a true structural parameter $\beta^* \in \mathcal{I}$. Because $\mathcal{I}$ might not be a singleton, our results (*e.g.*, confidence sets) are for all $\beta \in \mathcal{I}$, so they hold for $\beta = \beta^*$.

Sparsity Certificate. We call a *sparsity certificate* an a priori bound $s \in [d_Q]$ on the sparsity. To make this explicit, we maintain the original $\mathcal{B}$ and write $\mathcal{I}_s \triangleq \mathcal{I} \cap \left\{\beta \in \mathbb{R}^{d_X} : |S(\beta) \cap S_Q| \leq s\right\}$. This is the set of *s-sparse identifiable parameters*. Clearly, for $s \leq s' \leq d_Q$, $\mathcal{I}_s \subseteq \mathcal{I}_{s'} \subseteq \mathcal{I}_{d_Q} = \mathcal{I}$. Let us present, for i.d. data, two conditions under which a sparsity certificate yields identification. Assume $d_Z < d_X$ (*e.g.*, one is uncertain about some exclusion restrictions as in [23]), $S_Q = [d_X]$, $\mathcal{B} = \{\gamma \in \mathbb{R}^{d_X} : q = R\gamma\}$ for $q \in \mathbb{R}^{d_R}$ and $R^\top \in \mathcal{M}_{d_X,d_R}$ of rank d_R. When $d_Z + d_R + d_Q - s > d_X$, (2.1) yields $\binom{d_Z+d_R+d_Q-d_X}{s}$ overdetermined systems and identification is achieved if there exists a solution for only one system and it is unique. When $\mathcal{B} = \mathbb{R}^{d_X}$, a sufficient condition is an extension of a condition in [16]: $\mathcal{I}_s$ is a singleton if all sub-matrices formed from $2s$ columns of $\mathbb{E}[ZX^\top]$ have full column rank (see [22]).

The ℓ_∞-norm statistic. Our confidence sets and estimators use slacked versions of (2.1) based on the statistic $\widehat{t}(b) \triangleq \left|\mathbf{D}(b)\mathbf{Z}^\top\mathbf{U}(b)\right|_\infty/n$ for $b \in \mathbb{R}^{d_X}$. The diagonal matrix $\mathbf{D}(b)$ is introduced for scale invariance and to obtain finite sample results. It is either; (1) the diagonal matrix with positive diagonal elements $1/\widetilde{\sigma}_l(b)$ for $l \in [d_Z]$, where $\widetilde{\sigma}_l(b)^2 \triangleq \mathbb{E}_n\left[(Z_l U(b))^2\right]$, or (2) $\mathbf{D_Z}/\widehat{\sigma}(b)$ where $\widehat{\sigma}(b)^2 \triangleq \mathbb{E}_n[U(b)^2]$. We use, for $\beta \in \mathcal{I}$, the events

$$\mathcal{G}_0 \triangleq \left\{\widehat{t}(\beta) \leq r_0(n)\right\} \text{ (in case (1)) and } \mathcal{G} \triangleq \left\{\widehat{t}(\beta) \leq \widehat{r}\right\} \text{ (in case (2))}.$$

For a chosen confidence level α, $r_0(n)$ (resp. $\widehat{r}$) is adjusted so that $\mathbb{P}(\mathcal{G}_0) \geq 1 - \alpha - \alpha_B(n)$ (resp. $\mathbb{P}(\mathcal{G}) \geq 1 - \alpha - \alpha_B(n)$), where $\alpha_B(n)$ is the coverage error, uniformly over $\mathbb{P}$ and $\beta \in \mathcal{I}$. The expressions of $r_0(n)$, $\widehat{r}$, and $\alpha_B(n)$, for 5 large classes $\mathcal{P}$ are given in Section A.1.1. These restrict the joint distribution of $\mathbf{Z}$ and $\mathbf{U}(\beta)$, classes 1-4 allow for conditional heteroscedasticty, and all leave the dependence between $\mathbf{X}$ and $\mathbf{Z}$ unrestricted. For classes 1-3 we have $\alpha_B(n) = 0$. A reference behavior for $r_0(n)$ is $\sqrt{\ln(d_Z)/n}$, so the price to pay for many IVs is modest. The reference behaviour of $\widehat{r}$ is

slightly larger than for $r_0(n)$, *e.g.*, with an extra logarithmic factor.

Examples. In each of the following examples, the data is i.i.d., the regressors and IVs have mean zero and variance 1, $S_Q = [d_X]$, and $\mathcal{B} = \mathbb{R}^{d_X}$.
Example E1. All regressors are endogenous ($S_I = \varnothing$) and uncorrelated with one another. For $\beta \in \mathcal{I}$, regressors with index $k \in S(\beta)$ are correlated with $l_k \in [d_Z]$ IVs, with correlations $\rho_{j,k}$ for $j = 1, \dots, l_k$, and $\rho_k \triangleq \max_{j \in [l_k]} |\rho_{j,k}| > 0$. IVs correlated with regressor k are uncorrelated with all other regressors. This example permits $d_Z < d_X$. A particular case of Example E1 with a geometric decay rate of the entries of ρ is Example O2 in (v5), where the regressors are functions of $\widetilde{\mathbf{X}}_i$ and the IVs are functions of $\widetilde{\mathbf{Z}}_i$. If additionally $\mathbb{E}[ZZ^\top] = I$, then $\Psi^\top$ is the best linear predictor of the regressors given the instruments, and the support of its rows of index in $S(\beta)$ do not overlap.
Example E2. The last regressor is endogenous ($S_I^c = \{d_X\}$), $\Psi = (I \ \rho)$, and there is no excluded IV ($d_Z = d_X - 1$). In this example the support of the rows of $\Psi^\top$ overlap.

Roadmap. The paper is organized to progressively strengthen the assumptions. Here, we demonstrate our methods with the simplifications $S_Q = [d_X]$, $\mathbf{D_X} = D_X = I$, the data is i.i.d. and, for $i \in [n]$ and $\beta \in \mathcal{I}$, $\mathbf{U}_i(\beta)|\mathbf{Z}_{i,\cdot}^\top$ is normally distributed with mean 0 and known variance σ^2. Our results do not require any of these. The simplifications permit us to use $\mathbf{D}(b)$ of type (2) replacing $\widehat{\sigma}(b)$ with σ and $\widehat{r} = r = -\Phi^{-1}(\alpha/(2d_Z))/\sqrt{n}$, which is smaller than $\sqrt{\ln(2d_Z/\alpha)/n}$ when $2\sqrt{e}d_Z \geq \alpha$. This gives coverage error $\alpha_B(n) = 0$ for $\mathcal{G}$.

Suppose that the functional of interest is $\varphi(b) = b$. Given a sparsity certificate s, a natural starting point is to obtain bounds by minimizing and maximizing the entries of $b \in \mathbb{R}^{d_X}$ subject to $\widehat{t}(b) \leq r$ and $|S(b)| \leq s$. This is the principle underlying the *SNIV* confidence sets presented in Section 3. Supersets based on convex relaxations can be computed when d_X is up to around 100. For larger d_X the researcher can find $b \in \mathbb{R}^{d_X}$ which minimizes $|b|_1$ subject to $\widehat{t}(b) \leq r$. A solution to this problem, denoted $\widehat{\beta}$, is a type of *STIV* estimator, which we introduce in Section 4. It can be computed for very large d_X and, is typically sparse. To analyse its estimation error and construct confidence sets, we use *sensitivitity characteristics*, introduced in Section 4.1. To explain their role, we now take a $\beta \in \mathcal{I}$ ($\beta = \beta^*$ if $\mathcal{I}$ is a singleton). Since $\widehat{\beta}$ is a minimizer, on the event $\mathcal{G}$ we can use $|\widehat{\beta}|_1 \leq |\beta|_1$, $\widehat{t}(\widehat{\beta}) \leq r$ and $\widehat{t}(\beta) \leq r$. Letting $\widehat{\Delta} \triangleq \widehat{\beta} - \beta$, the first inequality implies $|\widehat{\Delta}_{S(\beta)^c}|_1 \leq |\widehat{\Delta}_{S(\beta)}|_1$. The last two imply $|\widehat{\Psi}\widehat{\Delta}|_\infty \leq 2\sigma r$. For $k \in [d_X]$ we introduce a sensitivity

$$\widehat{\kappa}^*_{e_k,S(\beta)} = \min_{\Delta \in \mathbb{R}^{d_X} : |\Delta_k| = 1, \quad |\Delta_{S(\beta)^c}|_1 \leq |\Delta_{S(\beta)}|_1} |\widehat{\Psi}\Delta|_\infty,$$

which by its definition gives $|\widehat{\Delta}_k| \leq |\widehat{\Psi}\widehat{\Delta}|_\infty / \widehat{\kappa}^*_{e_k,S(\beta)} \leq 2\sigma r(n)/\widehat{\kappa}^*_{e_k,S(\beta)}$. We do not know $S(\beta)$ but if we know $|S(\beta)| \leq s$, we replace $\widehat{\kappa}^*_{e_k,S(\beta)}$ with a lower bound $\widehat{\kappa}^*_{e_k}(s)$ obtained by solving linear programs. This yields the bounds $\widehat{\beta}_k \pm 2\sigma r(n)/\widehat{\kappa}^*_{e_k}(s)$, which gives us a second type of confidence set. An attractive feature of both confidence sets is that the coverage guarantee is uniform among classes $\mathcal{P}$ which leave the dependence between $\mathbf{X}$ and $\mathbf{Z}$ unrestricted. Moreover, $\mathcal{I}_s$ need not be a singleton. For example, one can have $d_Z < d_X$. This means that they are robust to any "weakness" of the IVs and lack of identification.

In Section 4.4 we obtain rates of convergence for $\widehat{\beta}$ by replacing sensitivities with their population analogues, which depend on Ψ. The rates are fast if there is a strong IV for every endogenous regressor

which is relevant. In Example E1, even if $d_Z < d_X$, we obtain

$$\left|\widehat{\beta} - \beta\right|_1 \lesssim r(n)|S(\beta)| \left(\min_{k \in S(\beta)} \rho_k\right)^{-1}, \quad \left|\widehat{\beta}_k - \beta_k\right| \lesssim r(n)\rho_k^{-1} \quad \forall k \in [d_X].$$

Assuming the nonzero entries of β are large relative to their rate of estimation, we obtain results on estimation of $S(\beta)$, and construct confidence sets based on $\widehat{\kappa}^*_{e_k}(\widehat{S})$, where the estimated support $\widehat{S}$ is an estimator of $S(\beta)$.

In Section 5, we present extensions to the moment models: (2.1b) $\mathbb{E}[\mathbf{Z}_{i,\cdot}^\top \mathbf{U}_i(\beta) - \theta] = 0$ where $\theta_l \neq 0$ accounts for the failure of (2.1) for IV $l \in [d_Z]$ (*i.e.*, the l^{th} IV is endogenous); (2.1c) $\mathbf{U}(\beta) = \mathbf{W}(\beta) + \mathbf{V}(\beta)$ and (2.1) holds $\mathbf{W}(\beta)$ rather than $\mathbf{U}(\beta)$ and $\mathbf{V}(\beta)$ is a small approximation error. For (2.1b) we propose the SNIV sets, the *C-STIV* estimator and confidence sets, and a two-stage method. C-STIV uses a scaling matrix in the spirit of $\mathbf{D}(b)$ of type (1), adapted to (2.1b). For (2.1c) we propose the *E-STIV* and allow for $\mathbf{U}(\beta)$ to be a vector of residuals from a system of equations. This is important for our empirical application which has both approximation errors and cross-equation restrictions.

STIV is typically "biased" towards zero. This allows to simultaneously perform estimation and model selection but is not necessarily well suited for inference. In Section 6, we present confidence bands for vectors of approximately linear functionals which are based on bias correction of a STIV, C-STIV, or E-STIV estimator. These are obtained by applying a C-STIV for systems in a second step to estimate left-inverses of each of the linear functionals by Ψ, and combining the two estimators. In Example E1, to construct a confidence interval for regressor $k \in [d_X]$, the assumption that a left-inverse of e_k by Ψ exists is $\rho_k > 0$ and $f_l^\top / \rho_{l,k}$ is a sparse left-inverse, where IV l has the strongest absolute correlation with regressor k.

In Section 7, we conduct a Monte-Carlo experiment to study the relative merits of the methods, and apply them to the EASI demand system ([25]) using a second-order in prices approximation, which results in many endogenous regressors.

Practical Guidance. We recommend starting with the STIV estimator. If the researcher suspects there might not be a strong enough instrument for every relevant endogenous regressor, she can compute the STIV confidence sets based on a sparsity certificate. If these are wide then she can invest in implementing the SNIV sets, provided that d_X is not too large. Otherwise, she can confidently use STIV as a pilot estimator and compute the STIV confidence sets based on an estimated support and the two-stage confidence bands. In the empirical application we have $d_Z = d_X$ and we believe every endogenous regressor in the model has a strong instrument. For this reason we apply the confidence bands involving bias correction and report together with the bands the plug-in estimator and the bias-corrected one. When using a sparsity certificate, the researcher may be unsure as to what is an appropriate value of s. She can then construct nested confidence sets over different values, which can be used to assess the information content of progressively stronger assumptions on the sparsity.

References. Econometrics for high-dimensional sparse models has become an active field. To name a few; [4] uses Lasso type methods to estimate the optimal IV and make inference on a low-dimensional structural equation, [18] consider a nonconvex approach to IV estimation, [12, 13] consider GMM and large dimensions but do not handle the high-dimensional regime, and [38] studies a 2SLS approach. Inference for subvectors in high-dimension is an active topic to which Section 6 relates (see [6, 21, 33, 37], but also [10, 20, 14] in the case of IVs using a 2SLS or GMM approach, and [7] using the C-STIV estimator as a pilot and orthogonalisation, [8] reviews some of the results based on the

nonpivotal STIV and others). Applications and extensions of this paper, in the context of networks can be found in [19, 28, 3].

3. Self-Normalized IV Confidence Sets

In this section we use $\mathbf{D}(b)$ of type (1) and $\mathcal{P}$ can be one of classes 1-4. We propose *SNIV* confidence sets, defined as $\widehat{C}_\varphi(s) \triangleq \{\varphi(\beta) : \beta \in \widehat{C}(s)\}$, where

$$\widehat{C}(s) \triangleq \left\{b \in \mathcal{B} : \ |S(b) \cap S_Q| \leq s, \ \widehat{t}(b) \leq r_0(n)\right\}. \tag{3.1}$$

Because, for all $\beta \in \mathcal{I}_s$, $\mathbb{P}\left(\varphi(\beta) \in \widehat{C}_\varphi(s)\right) = \mathbb{P}\left(\widehat{t}(\beta) \leq r_0(n)\right)$, $\widehat{C}(s)$ satisfies

$$\inf_{s\in[d_Q]} \inf_{\beta,\mathbb{P}:\ \beta\in\mathcal{I}_s} \mathbb{P}\left(\varphi(\beta) \in \widehat{C}_\varphi(s)\right) \geq 1 - \alpha - \alpha_B(n). \tag{3.2}$$

A natural way to summarize the confidence set $\widehat{C}_\varphi(s)$ is to compute a confidence band comprising d_F intervals such that $\varphi(b)$ lies in the band when $\varphi(b) \in \widehat{C}_\varphi(s)$. We use, for all $j \in [d_F]$, the lower and upper bounds

$$\underline{\widehat{C}}_{\varphi_j}(s) \triangleq \min_{b\in\widehat{C}(s)} \varphi_j(b), \quad \widehat{\overline{C}}_{\varphi_j}(s) \triangleq \max_{b\in\widehat{C}(s)} \varphi_j(b). \tag{3.3}$$

A computational difficulty is that $\widehat{t}(b) \leq r_0(n)$ gives rise to d_Z inequalities, each requiring that a polynomial in b be nonnegative (henceforth referred to as polynomial inequalities). This usually does not define a convex set. Even if φ is linear, finding a global solution to such quadratically constrained linear programs is NP-hard. Local methods can only be guaranteed to reach local solutions, hence we would only obtain certain subsets of the confidence set and have no coverage guarantee. [4] (Section 4.5) propose to use a grid in the case where there is no sparsity constraint. This is not a practical solution even in moderate dimensions.

We use hierarchies of convex relaxations for the optimization problems in (3.3). These allow for the sparsity constraint, and can be applied when $\varphi(b)$ is a rational function (*i.e.*, a ratio of polynomials) and $\mathcal{B}$ can be written via polynomial inequalities. We apply the Sparse BSOS hierarchy of [7] which avoids the large scale semidefinite programs required by the SOS method[1] by using additional linear constraints. Solving large scale SDPs is currently computationally infeasible. When $\mathcal{B}$ is compact, the Sparse BSOS hierarchy of optimization problems provides monotonic sequences of lower (resp. upper) bounds which converge to $\underline{\widehat{C}}_{\varphi_j}(s)$ (resp. $\widehat{\overline{C}}_{\varphi_j}(s)$). A simple diagonostic indicates whether a solution obtained via the hierarchy is one of the original problem. Further details are in Section C.1. We show in Section 7.1 that the relaxed SNIV sets can give useful information even in the presence of endogenous IVs and when there are more regressors than IVs. This offers a practical solution to problems where even testing if $\mathcal{I}_s$ is a singleton is NP-Hard. In practice we find that SNIV is computationally feasible for moderate d_X, of dimension up to 100.

Remark 3.1. *If $\mathcal{B} = \mathbb{R}^{d_X}$, the set $\widehat{C}(s)$ is defined by polynomials of degree 2 inequalities, so it can be empty, unbounded or disconnected depending on the (random) values of the polynomial coefficients. The same occurs for Anderson-Rubin confidence sets (see* [39]*). The fact that they can be unbounded is unavoidable for confidence sets which are robust to weak IVs (see* [17]*).*

[1] In (v5) we apply Lasserre's SOS hierarchies (see [24]) of semidefinite programs.

4. Self-Tuned IV Estimator and Confidence Sets

To permit large d_X we introduce parameters to replace the denominators in the definition of $\mathbf{D}(b)$: $\widetilde{\sigma}_l(b)$, for $l \in [d_Z]$, in case (1) (resp. $\widehat{\sigma}(b)$ in case (2)) in place of σ_l (resp. σ). This gives rise to the statistic $\widehat{t}(b, \sigma_1, \ldots, \sigma_{d_Z})$ (resp. $\widehat{t}(b, \sigma)$). To obtain a convex set in case (1) we replace $\widehat{C}(s)$ with

$$\left\{b \in \mathcal{B} : \ \widehat{t}(b, \sigma_1, \ldots, \sigma_{d_Z}) \leq r_0(n), \ \widetilde{\sigma}_l(b) \leq \sigma_l, \ \forall l \in [d_Z]\right\}, \tag{4.1}$$

and use a convex objective function which induces sparsity on the solution $\widehat{\beta}$ and prevents the vector $(\widetilde{\sigma}_l(\widehat{\beta}))_{l\in[d_Z]}$ from having large entries. This is the idea behind the C-STIV estimator presented in Section 5. For the sake of exposition we present the more simple estimator corresponding to case (2). It is easier to compute and handles much larger d_Z in practice.

Definition 4.1. *The set of IV-constraints is defined, for $\sigma, r > 0$, by*

$$\widehat{\mathcal{I}}(r, \sigma) \triangleq \left\{b \in \mathcal{B}, \ \left|\frac{1}{n}\mathbf{D_Z}\mathbf{Z}^\top \mathbf{U}(b)\right|_\infty \leq r\sigma, \ \widehat{\sigma}(b) \leq \sigma\right\}. \tag{4.2}$$

Definition 4.2. *For $c, r > 0$, a Self-Tuned Instrumental Variables (STIV) estimator is any solution $(\widehat{\beta}, \widehat{\sigma})$ of the minimization problem*

$$\min_{b\in\widehat{\mathcal{I}}(r,\sigma),\sigma\geq 0} \left(\left|\mathbf{D}_\mathbf{X}^{-1} b_{S_Q}\right|_1 + c\sigma\right). \tag{4.3}$$

The ℓ_1-norm is a convex relaxation of $|S(b) \cap S_Q|$. The researcher applying STIV should choose a class $\mathcal{P}$ and the corresponding value of $r = \widehat{r}$ from Section A.1.1. The estimator also depends on c, which trades off sparsity with the relaxation of the sample moment conditions. The term $c\sigma$ favors small σ, hence increasing c tightens the set $\widehat{\mathcal{I}}(r, \sigma)$. If $c = 0$, (4.3) minimizes the ℓ_1-norm of b_{S_Q}, yielding $\widehat{\beta}_{S_Q} = 0$. The matrix $\mathbf{D}_\mathbf{X}^{-1}$ guarantees invariance to scale of the regressors. In this formulation, if $\mathcal{B}$ comprises linear (in)equality restrictions, a STIV estimator is computed by solving a convex (second-order) conic program.

If the data is i.i.d., $\widehat{\sigma}(\widehat{\beta})$, $\widehat{\sigma}$, and $\overline{\sigma} \triangleq (\widehat{\sigma} + \widehat{\sigma}(\widehat{\beta}))/2$ are estimators of the standard error of the structural error, which STIV does not require the researcher to know. Moreover, the researcher need not know an upper bound, nor have a preliminary estimator. STIV generalizes the Dantzig Selector to unknown variance.

If we have an upper bound σ on the standard error of the structural error, we can remove $\widehat{\sigma}(b) \leq \sigma$ from (4.2) and $+c\sigma$ from (4.3), as presented in Section 2. This is the nonpivotal STIV analyzed in Section 7.1 in (v1). The STIV estimator can be used as a pilot to estimate the standard error before applying the nonpivotal STIV. The proof of Theorem 4.3 provides rates of convergence (see (A.46) and (A.47)) for estimation of the standard error. However, relying on a pilot estimator does not provide robustness to identification, and so we focus the exposition on STIV. Consider now

$$\mathcal{O}(b) \triangleq \max\left(\widehat{\sigma}(b), \frac{1}{r}\left|\frac{1}{n}\mathbf{D_Z}\mathbf{Z}^\top \mathbf{U}(b)\right|_\infty\right).$$

The first component of $\mathcal{O}(b)$ is the square-root of the least-squares objective function. The second is derived from the exogeneity of the IVs and uses a ℓ_∞- norm statistic divided by $r(n)$ (the size of its fluctuations). Minimizing $\mathcal{O}(b)$ trades-off the two, which is desirable in the presence of weak IVs (see [1] for references comparing 2SLS and OLS). It is simple to check that the STIV estimator is a solution of a penalized version:

$$\widehat{\beta} \in \operatorname{argmin}_{b\in\mathcal{B}}\left(\frac{1}{c}\left|\mathbf{D}_\mathbf{X}^{-1} b_{S_Q}\right|_1 + \mathcal{O}(b)\right), \ \widehat{\sigma} = \mathcal{O}(\widehat{\beta}). \tag{4.4}$$

The STIV estimator can also be obtained as a solution of the convex program

$$\min_{(b,\sigma)\in\mathcal{B}\times(0,\infty)} \left(\frac{2}{c} \left| \mathbf{D}_{\mathbf{X}}^{-1} b_{S_Q} \right|_1 + \sigma + \frac{1}{\sigma} \mathcal{O}(b)^2 \right), \tag{4.5}$$

which can be obtained by iteratively minimizing over b and σ using

Algorithm 4.1. *Initialize at* $\left(\widehat{\beta}^{(0)}, \widehat{\sigma}^{(0)}\right)$. *At iteration* t, *solve*

$$\widehat{\beta}^{(t)} \in \operatorname{argmin}_{b\in\mathcal{B}} \left(\frac{2\widehat{\sigma}^{(t-1)}}{c} \left| \mathbf{D}_{\mathbf{X}}^{-1} b_{S_Q} \right|_1 + \mathcal{O}(b)^2 \right), \ \widehat{\sigma}^{(t)} = \mathcal{O}\left(\widehat{\beta}^{(t)}\right),$$

then replace t *by* $t+1$, *and iterate until convergence.*

Section C.3 gives details and presents a numerical acceleration of the minimization over b. If we remove the second term in the maximum in the definition of $\mathcal{O}(b)$ and take $S_Q = [d_X]$, (4.4) is the Square-root Lasso (see [5]), (4.5) is the Concomitant Lasso (see [27]), and Algorithm 4.1 is the Scaled Lasso (see [30]).

Remark 4.1. *The STIV estimator is not necessarily unique. When* $\mathcal{O}(b)^2$ *is a differentiable and strictly convex function of* $\mathbf{W}b$, [32] *shows that minimizers in problems such as Algorithm 4.1 are unique if the entries of* $\mathbf{W}$ *are drawn from a continuous distribution and discusses regularization paths. This can be useful when one does not know how to adjust the level of the penalty (*$2\widehat{\sigma}^{(t-1)}/c$ *in Algorithm 4.1). The assumptions are not satisfied for STIV. Non uniqueness also occurs for LIML which minimizes the Anderson-Rubin statistic. We emphasize however, that our analysis and inference methods are valid for all minimizers and that determination of the penalty level is also less of an issue because STIV is pivotal and our results hold for all* c.

4.1. **Sensitivity Characteristics.** If $\mathbf{Z} = \mathbf{X}$, the minimal eigenvalue of $\mathbf{X}^\top\mathbf{X}/n$ can be used to obtain error bounds. It is the minimum of $b^\top\mathbf{X}^\top\mathbf{X}b/(n|b|_2^2)$ over $b \in \mathbb{R}^{d_X}$. If the structural parameter is sparse and the estimator uses an ℓ_1 penalty, $\mathbb{R}^{d_X}$ can be replaced with a subset. This is typically expressed via the restricted isometry property of [16] or the restricted eigenvalue condition of [9]. These cannot be used for models with endogenous regressors because $\mathbf{Z}^\top\mathbf{X}/n$ can be rectangular. We introduce scalar sensitivity characteristics related to the action of $\widehat{\Psi}$ on a restricted set $\widehat{K}_S$ (a cone) for $S \subseteq [d_X]$. Let $\mathcal{L}$ be the set of continuous functions from $\mathbb{R}^{d_X}$ to $[0,\infty)$ which are homogeneous of degree 1. We define the *sensitivity*

$$\widehat{\kappa}_{l,S} \triangleq \min_{\Delta\in\widehat{K}_S:\ l(\Delta)=1} \left| \widehat{\Psi}\Delta \right|_\infty. \tag{4.6}$$

$|\widehat{\Psi}\Delta|_\infty$ arises due to the ℓ_∞-norm in the definition of $\widehat{\mathcal{I}}(r,\sigma)$ and $l \in \mathcal{L}$ should be interpreted as a loss function. Due to the ℓ_∞-norm in (4.6), one strong IV is enough to ensure a large sensitivity. Additional IVs can only increase $|\widehat{\Psi}\Delta|_\infty$, even if they are weak. The cost of additional IVs is mild since it appears in a logarithmic factor (see the typical behavior at the beginning of Section 4.4).

The cone comes from the form of the objective function in (4.3) which is such that, for every $\beta \in \mathcal{I}$, on the event $\mathcal{G}$, $\widehat{\Delta} \in \widehat{K}_{S(\beta)}$, where

$$\widehat{K}_S \triangleq \left\{ \Delta \in \mathbb{R}^{d_X} :\ \Delta_{S^c\cap S(\widehat{\beta})^c} = 0,\ \left|\Delta_{S^c\cap S_Q}\right|_1 \le \left|\Delta_{S\cap S_Q}\right|_1 + c\widehat{g}(\Delta) \right\} \tag{4.7}$$

and $\widehat{g}(\Delta) \triangleq \widehat{r}|\Delta_{S_I}|_1 + \left|\Delta_{S_I^c}\right|_1$. When $\mathcal{B} \subsetneq \mathbb{R}^{d_X}$, we replace $\Delta \in \mathbb{R}^{d_X}$ by $\mathbf{D}_{\mathbf{X}}\Delta \in \mathcal{B}_D \triangleq \{b_1 - b_2,\ \forall b_1, b_2 \in \mathcal{B}\}$. The error bounds for STIV in Section 4.2 are small when the sensitivities are sufficiently bounded away from zero, hence it is important that $\widehat{K}_S$ be as small as possible. The researcher's knowledge components $\mathcal{B}$, S_I and S_Q serve this purpose. The form of $\widehat{g}$ comes from the fact that, by convexity and

because the regressors of index in S_I are used as IVs, $\widehat{\sigma}(\beta) - \widehat{\sigma}(\widehat{\beta}) \leq \widehat{g}(\widehat{\Delta})$. If we ignore exogeneity or if all regressors are endogenous, we obtain $\widehat{\sigma}(\beta) - \widehat{\sigma}(\widehat{\beta}) \leq |\widehat{\Delta}|_1$. Because $\widehat{r}$ is small, accounting for S_I yields a smaller set. The nonpivotal STIV does not contain c and σ, so $c\widehat{g}(\Delta)$ is omitted from (4.7). If we omit $\Delta_{S^c \cap S(\widehat{\beta})^c} = 0$, take $S_Q = [d_X]$, $\mathcal{B} = \mathbb{R}^{d_X}$, and replace $\widehat{g}(\Delta)$ by $|\Delta|_1$ we obtain the simple cone $\left\{\Delta \in \mathbb{R}^{d_X} : (1-c)|\Delta_{S^c}|_1 \leq (1+c)|\Delta_S|_1\right\}$ used in (v1). This is $\mathbb{R}^{d_X}$ if $c \geq 1$. In our Monte-Carlo experiment, we find that STIV performs better for $c > 1$, and so we use (4.7) instead. Notice also that if there are no endogenous regressors ($S_I = [d_X]$), we have $\widehat{K}_S \subseteq \left\{\Delta \in \mathbb{R}^{d_X} : (1 - c\widehat{r})|\Delta_{S^c}|_1 \leq (1 + c\widehat{r})|\Delta_S|_1\right\}$. The restriction $\Delta_{S^c \cap S(\widehat{\beta})^c} = 0$ can be removed, for example to obtain rates of convergence. We use this restriction to construct confidence sets which ought naturally to depend on $\widehat{\beta}$. Doing so yields smaller confidence sets in practice.

If β is not sparse, $\widehat{K}_{S(\beta)}$ can be large (*e.g.*, $\mathbb{R}^{d_X}$ when $S(\beta) = [d_X]$). To handle such cases, the sensitivities are defined by replacing $\widehat{K}_S$ with

$$\widehat{K}_{\gamma,S} \triangleq \left\{\Delta \in \mathbb{R}^{d_X} : \left|\Delta_{S^c \cap S_Q}\right|_1 \leq 2\left(\left|\Delta_{S \cap S_Q}\right|_1 + c\widehat{g}(\Delta)\right) + \left|\Delta_{S_Q^c}\right|_1\right\}.$$

They are denoted by $\widehat{\gamma}$ instead of $\widehat{\kappa}$. Due to the additional terms on the right-hand side of the inequality, $\widehat{K}_{\gamma,S}$ is larger than $\widehat{K}_S$. However, in our analysis the sensitivities need not be computed at $S = S(\beta)$. The slackness allows $\widehat{\Delta} \in \widehat{K}_{\gamma,S}$, on the event $\mathcal{G}$ provided that $\left|\mathbf{D}_{\mathbf{X}}^{-1}\beta_{S^c \cap S_Q}\right|_1$ is sufficiently small. The form of the additional terms is related to the factor 3 which appears in Theorem 4.2.

List of Sensitivities. For $1 \leq q \leq \infty$, $S_0 \subseteq [d_X]$, and $l(\Delta) = |\Delta_{S_0}|_q$, we define the ℓ_q-S_0 *block sensitivity* as $\widehat{\kappa}_{q,S_0,S}$. By convention, we set $\widehat{\kappa}_{q,\varnothing,S} = \infty$ and, when $S_0 = [d_X]$, we use the shorthand notation $\widehat{\kappa}_{q,S}$ and call this the ℓ_q *sensitivity*. For sparse vectors we make use of the loss $\widehat{g}$ and for nonsparse vectors of $\widehat{h}(\Delta) \triangleq \min(|\Delta_{S_Q}|_1, \frac{1}{2}(3|\Delta_{S \cap S_Q}|_1 + c\widehat{r}|\Delta_{S_I}|_1 + c|\Delta_{S_I^c}|_1 + |\Delta_{S_Q^c}|_1))$ and refer to $\widehat{\kappa}_{\widehat{g},S}$ and $\widehat{\gamma}_{\widehat{h},S}$ as the corresponding sensitivities. By taking $\omega \in \mathbb{R}^{d_X}$ and $l(\Delta) = |\omega^\top \Delta|$ we obtain the sensitivity $\widehat{\kappa}^*_{\omega,S}$ for a linear combination. If $\omega = e_k$ this is the sensitivity $\widehat{\kappa}^*_{e_k,S}$ which we refer to as *coordinate-wise sensitivity*. Proposition A.2 shows how the sensitivities can be related to one another.

The sensitivities are not only core elements to analyse the performance of STIV. They also provide sharper results than existing quantities for the analysis of the Dantzig Selector and Lasso[2]. These differ from the quantities introduced by [36] in the definitions of the cone $\widehat{K}_S$, the matrix $\widehat{\Psi}$, and in that it does not involve a scaling by $|S|^{1/q}$ because, the sensitivities typically depend on S in a more complex manner than simply via $|S|$.

To get the intuition, consider the sensitivity $\widehat{\kappa}^*_{e_k,S}$ for $k \in [d_X]$. If $S_Q = \varnothing$, $\mathcal{B} = \mathbb{R}^{d_X}$, and $\widehat{\Psi}_{\cdot,-k}$ is obtained from $\sum_{i\in[n]} \mathbf{Z}_{i,\cdot}^\top \mathbf{X}_{i,\cdot}$ by removing the kth column, we have $\widehat{K}_S = \mathbb{R}^{d_X}$ and $\widehat{\kappa}^*_{e_k,S} = \min_{\Delta \in \mathbb{R}^{d_X - 1}} |\mathbf{D}_{\mathbf{Z}}(\sum_{i\in[n]} \mathbf{Z}_{i,\cdot}^\top \mathbf{X}_{i,k} - \widehat{\Psi}_{\cdot,-k}\Delta)|_\infty / n$. It is zero if $\sum_{i\in[n]} \mathbf{Z}_{i,\cdot}^\top \mathbf{X}_{i,k}/n$ is in the range of $\widehat{\Psi}_{\cdot,-k}$, which has probability zero if $d_Z \geq d_X$ and $\mathbf{Z}^\top \mathbf{X}$ has a continuous distribution. When the minimization is carried over a cone, certain combinations of the columns of $\widehat{\Psi}_{\cdot,-k}$ are ruled out. To be precise, if $l \in \mathcal{L}$ is point-separating, $\widehat{\kappa}_{l,S} > 0$ iff $\ker(\widehat{\Psi}) \setminus \{0\} \subseteq \widehat{K}_S^c$.

4.2. **Basic Error Bounds.**

[2]See Section A.4 in (v5).

Theorem 4.1. *For all $\beta, \mathbb{P}$ such that $\beta \in \mathcal{I}$ and all STIV estimators $(\widehat{\beta}, \widehat{\sigma})$, $l \in \mathcal{L}$, and $c > 0$, we have, on $\mathcal{G}$,*

$$l\left(\mathbf{D}_{\mathbf{X}}^{-1}\left(\widehat{\beta}-\beta\right)\right) \leq \frac{2\widehat{r}}{\widehat{\kappa}_{l,S(\beta)}} \min\left(\overline{\sigma}\left(1-\frac{\widehat{r}}{\widehat{\kappa}_{\widehat{g},S(\beta)}}\right)_+^{-1}, \widehat{\sigma}(\beta)\left(1-\frac{\widehat{r}}{c\widehat{\kappa}_{1,S(\beta)\cap S_Q,S(\beta)}}\right)_+^{-1}\right).$$

The first term in the minimum is used for confidence sets and the second for rates of convergence.

Theorem 4.2. *For all $\beta, \mathbb{P}$ such that $\beta \in \mathcal{I}$ and all STIV estimators $(\widehat{\beta}, \widehat{\sigma})$, $q \in [1, \infty]$, $S_0, S \subseteq [d_X]$, and $c > 0$, we have, on $\mathcal{G}$,*

$$\left|\mathbf{D}_{\mathbf{X}}^{-1}\left(\widehat{\beta}-\beta\right)_{S_0}\right|_q \leq 2\max\Bigg(\frac{\widehat{r}}{\widehat{\gamma}_{q,S_0,S}} \min\left(\overline{\sigma}\left(1-\frac{\widehat{r}}{\widehat{\gamma}_{\widehat{g},S}}\right)_+^{-1}, \widehat{\sigma}(\beta)\left(1-\frac{\widehat{r}}{c\widehat{\gamma}_{\widehat{h},S}}\right)_+^{-1}\right), 3\left|\mathbf{D}_{\mathbf{X}}^{-1}\beta_{S^c\cap S_Q}\right|_1\Bigg).$$

Theorem 4.2 can give sharper results than Theorem 4.1 when $\mathcal{I}$ contains nonsparse vectors. It is the basis of the sparsity oracle inequality in Theorem 4.3 (iii), to which point we defer the comparison of Theorems 4.1 and 4.2.

To obtain a confidence set one needs to use the sensitivities for $S = S(\beta)$ for $\beta \in \mathcal{I}$, and should circumvent their dependence on the unknown $S(\beta)$ and provide a computationally feasible method. This is the focus of Section 4.3. For rates of convergence, one should relate the upper bounds in the above theorems to deterministic ones. This is the focus of Section 4.4.1.

4.3. **Computable Lower Bounds on the Sensitivities and Robust Confidence Sets.** We propose two means to bound the sensitivities from below; one based on an estimated set $\widehat{S}$, and another using a sparsity certificate. We postpone the discussion on how to obtain $\widehat{S}$ and the properties of the resulting confidence set to Section 4.4, in which we analyse variable selection. Even if $S(\beta)$ were known, computing the sensitivities is non-trivial since $\widehat{K}_{S(\beta)}$ is not a convex set.

If $\widehat{S} \supseteq S(\beta)$, by (i) in Proposition A.2, we can replace $S(\beta)$ with $\widehat{S}$. To handle non-convexity, we introduce a new decision variable $\mu \in \mathbb{R}^{d_X}$ to play the role of $|\Delta|$ and augment the constraints to include $-\mu \leq \Delta \leq \mu$, $\mu_{\widehat{S}^c\cap S(\widehat{\beta})^c} = 0$, and $\mu \leq |\Delta|_\infty$. For some $j \in [d_X]$ we have $|\Delta|_\infty = \eta\Delta_j$, where η is the sign of Δ_j. If we knew j and η we could replace the non-convex constraint $\mu \leq |\Delta|_\infty$ with $\mu \leq \eta\Delta_j$. Since $j \in [d_X]$ and $\eta = \pm 1$, for $k \in [d_X]$ we have,

$$\widehat{\kappa}^*_{e_k,S(\beta)} \geq \kappa^*_{e_k}(\widehat{S}) \triangleq \min_{j\in[d_X],\eta=\pm 1} \min_{\substack{-\mu\leq\Delta\leq\mu\leq\eta\Delta_j,\quad \mu_{\widehat{S}^c\cap S(\widehat{\beta})^c}=0 \\ \mu_{\widehat{S}^c\cap S_Q}^\top 1\leq(\mu_{\widehat{S}\cap S_Q}+c\widehat{r}\mu_{S_I}+c\mu_{S_I^c})^\top 1,\quad \mu_k=1}} |\widehat{\Psi}\Delta|_\infty$$

which is computed by solving $2d_X$ LPs. If instead of $\widehat{S}$ we have a sparsity certificate s, we use $|\Delta_{S(\beta)\cap S_Q}|_1 \leq s|\Delta|_\infty$, which also results in $2d_X$ LPs. Proposition 4.1 applies these ideas to other sensitivities, including those for the non-sparse case. Further details and other solutions are provided in Section C.2.

Proposition 4.1. *For all $S \subseteq \widehat{S} \subseteq [d_X]$, $|S \cap S_Q| \leq s$, and $c > 0$,*

$$\widehat{\kappa}_{\infty,S}, \geq \max\left(\widehat{\kappa}_\infty\left(\widehat{S}\right), \widehat{\kappa}_\infty(s)\right), \quad \widehat{\kappa}^*_{\omega,S} \geq \max\left(\widehat{\kappa}^*_\omega\left(\widehat{S}\right), \widehat{\kappa}^*_\omega(s)\right),$$
$$\widehat{\kappa}_{1,S} \geq \max\left(\widehat{\kappa}_1\left(\widehat{S}\right), \widehat{\kappa}_1(s)\right), \quad \widehat{\kappa}_{\widehat{g},S} \geq \max\left(\widehat{\kappa}_{\widehat{g}}\left(\widehat{S}\right), \widehat{\kappa}_{\widehat{g}}(s)\right),$$

$$\left(1-\frac{r}{\widehat{\kappa}_{\widehat{g},S}}\right)_+^{-1} \leq \min\left(\widehat{\theta}_\kappa\left(\widehat{S}\right), \widehat{\theta}_\kappa(s)\right),$$

where the quantities in the bounds are in Table 9. The same holds for the sensitivities $\widehat{\gamma}$ using, instead of $\widehat{B}$ and $\widehat{\theta}_\kappa$, the sets $\widehat{B}_\gamma$ and constant $\widehat{\theta}_\gamma$.

Based on these lower bounds, $\widehat{C}_{\varphi,\kappa}(s) \triangleq \{\varphi(b) : b \in \widehat{C}_\kappa(s)\}$, where

$$\widehat{C}_\kappa(s) \triangleq \left\{ b \in \mathcal{B} : \forall l \in \mathcal{L}, \forall c > 0, \quad l\left(\mathbf{D}_{\mathbf{X}}^{-1}\left(\widehat{\beta} - b\right)\right) \leq \frac{2\widehat{r}\widehat{\sigma}\widehat{\theta}_\kappa(s)}{\widehat{\kappa}_l(s)} \right\}, \tag{4.8}$$

defines a confidence set. It inherits the same robustness and uniformity properties as the *SNIV* confidence set (*i.e.* (3.2) applies). In practice we can summarize $\widehat{C}_{\varphi,\kappa}(s)$ by using a finite number of functions in $\mathcal{L}$ in (4.8). We focus on constructing a confidence band comprising d_F intervals in which $\varphi(b)$ lies when $\varphi(b) \in \widehat{C}_{\varphi,\kappa}(s)$. For example, if $\varphi(b) = b$ we can take d_X functions, $l(\Delta) = |e_k^\top \Delta|$ for $k \in [d_X]$, yielding d_X lower and upper bounds,

$$\underline{\widehat{C}}_{\varphi_k,\kappa}(s) = \max_{c>0}\left(\widehat{\beta}_k - \frac{2\widehat{r}\widehat{\sigma}\widehat{\theta}_\kappa(s)}{\widehat{\kappa}^*_{e_k}(s)\sqrt{\mathbb{E}_n[X_k^2]}}\right), \ \overline{\widehat{C}}_{\varphi_k,\kappa}(s) = \min_{c>0}\left(\widehat{\beta}_k + \frac{2\widehat{r}\widehat{\sigma}\widehat{\theta}_\kappa(s)}{\widehat{\kappa}^*_{e_k}(s)\sqrt{\mathbb{E}_n[X_k^2]}}\right). \tag{4.9}$$

Alternatively, we can take one function, $l(\Delta) = |\Delta|_\infty$, and replace $\widehat{\kappa}^*_{e_k}(s)$ with $\widehat{\kappa}_\infty(s)$ in (4.9). The latter yields wider bounds but is less computationally demanding. In the same way, if $\varphi(\beta) = \Omega\beta$ for some specified Ω, one can use either the sensitivities $\widehat{\kappa}^*_{\omega_j}(s)$ for $j \in [d_F]$ and $\omega_j = \mathbf{D}_{\mathbf{X}}\Omega^\top_{j,\cdot}$ or the sensitivity for loss $l(\Delta) = |\Omega\mathbf{D}_{\mathbf{X}}\Delta|_\infty$. Though we do not make it explicit, the bound in (4.8) depends on c. Increasing c decreases $\overline{\sigma}$ (by increasing the penalty on σ in the STIV objective function) but increases $\widehat{\theta}_\kappa(s)/\widehat{\kappa}_l(s)$ (by enlarging $\widehat{K}_S$). The optimal value for c does not admit a tractable form. Due to uniformity in c, in (4.9) we can maximize (resp. minimize) the lower (resp. upper) bound over a grid on c. In Section 7.1 we propose a rule of thumb to determine a single value of c. Even if c is determined from the data, the set has coverage at least $1-\alpha-\alpha_B(n)$ due to (4.8). Since it relies on conic and linear programs, the approach is computationally feasible even for large d_X. Unlike SNIV, the inequalities at the basis of the sets are linear, but the sets can be infinite due to $\widehat{\theta}_\kappa(s)$.

4.4. **Rates, Model Selection, and Refined Confidence Sets.** To obtain rates of convergence we use deterministic bounds to replace the random error bounds in Theorem 4.1 and Theorem 4.2. Our analysis relies on replacing $\widehat{\Psi}$ with Ψ and the sensitivities (and their lower bounds) with their population analogues. To obtain deterministic bounds, we further restrict $\mathcal{P}$ using Assumption A.2. The additional restrictions still do not restrict the dependence between $\mathbf{Z}$ and $\mathbf{X}$. These are simply mild assumptions on second moments which give finite sample bounds on the error made by replacing empirical averages by population ones. Assumption A.2 also allows $\widehat{r}$ and $\widehat{\sigma}(\beta)$ to be replaced by nonrandom $r(n)$and $\sigma_{U(\beta)}$. Our results are stated on $\mathcal{G} \cap \mathcal{G}_\Psi$ which has probability at least $1-\alpha-\alpha_D(n)$, where $\alpha_D(n) \to 0$. Asymptotic statements allow c to depend on n. Our statements involve a sequence $(\tau(n))_{n\in\mathbb{N}} \in (0,1)^{\mathbb{N}}$ converging to zero with n such that $\ln(\max(d_Z, d_X, d_F))/(n\tau(n)^2) \to 0$. This behaviour is adequate if $\mathbf{Z}_{i,\cdot}$ and $\mathbf{X}_{i,\cdot}$ are uniformly bounded. For other distributions $\tau(n)$ can have to decay more slowly to zero. The results hold for all sequence $(\tau(n))_{n\in\mathbb{N}}$. Details are in Section A.1.2. The sequence $r(n)$ is of the order of $\sqrt{\ln(d_Z)/n}$, $\sqrt{\ln(C(n)nd_Z)\ln(d_Z)/n}$, or $\sqrt{\ln(C(n)d_Z^2)\ln(d_Z)/n}$, where $C(n)$ is inversely proportional to a coverage error sequence, depending on the class $\mathcal{P}$.

4.4.1. *Population Sensitivities.* The population analogues of the sensitivities are key objects to establish the rate. The analogues of $\widehat{\kappa}, \widehat{\gamma}, \widehat{\theta}$ are denoted by κ, γ, θ. They are obtained by replacing $\widehat{\Psi}$, $\widehat{K}_S$ and

$\widehat{K}_{\gamma,S}$ with Ψ, K_S and $K_{\gamma,S}$, where

$$K_S \triangleq \left\{\Delta \in \mathbb{R}^{d_X} : 1_n \left|\Delta_{S^c \cap S_Q}\right|_1 \leq \left|\Delta_{S \cap S_Q}\right|_1 + cg(\Delta)\right\},$$
$$K_{\gamma,S} \triangleq \left\{\Delta \in \mathbb{R}^{d_X} : 1_n \left|\Delta_{S^c \cap S_Q}\right|_1 \leq 2\left(\left|\Delta_{S \cap S_Q}\right|_1 + cg(\Delta)\right) + \left|\Delta_{S_Q^c}\right|_1\right\},$$

g is defined as $\widehat{g}$ using $r(n)$ instead of $\widehat{r}$, $1_n \triangleq \sqrt{(1-\tau(n))/(1+\tau(n))}$, and we do not write the set $\mathcal{B}_D$ for simplicity. As with g, h is the counterpart of $\widehat{h}$ which replaces $\widehat{r}$ with $r(n)$. The results below hold for all c but if $c \geq 1_n/r(n)$ we have $K_S = \mathbb{R}^{d_X}$.

For $S \subseteq [d_X]$, we use $\overline{S} \triangleq (S \cap S_Q) \cup S_Q^c \cup S_I^c$ if $c \geq 1_n$, $\overline{S} = (S \cap S_Q) \cup S_Q^c$ if $c < 1_n$, $c_\kappa(S) \triangleq \min(c_{>,\kappa}(S), c_{<,\kappa}(S))$,

$$c_{>,\kappa}(S) \triangleq ((1+1_n)|S \cap S_Q| + 1_n|S_Q^c| + c(1-r(n))|S_I^c|)(1_n - cr(n))_+^{-1},$$
$$c_{<,\kappa}(S) \triangleq ((1+1_n)|S \cap S_Q| + 1_n|S_Q^c|)(1_n - c)_+^{-1},$$

and $c_\gamma(S)$ for the sensitivities γ replacing $(1+1_n)$ by $(2+1_n)$ and c by $2c$. If $c < 1_n$ and $S_Q = [d_X]$, we have $\overline{S} = S$, $c_\kappa(S) = u_\kappa|S|$, and, if $c < 1_n/2$, $c_\gamma(S) = u_\gamma|S|$, where $u_\kappa \triangleq (1+1_n)/(1_n - c)$ and $u_\gamma \triangleq (2+1_n)/(1_n - 2c)$. We refer to this condition together with $\mathcal{B} = \mathbb{R}^{d_X}$ as (IC).

Proposition 4.2. *We have, for all $S \subseteq [d_X]$, $S_0 \subseteq [d_X]$, and $q \in [1, \infty]$,*

(i) $\kappa_{q,S_0,S} \geq \kappa_{q,S}$,
(ii) $\kappa_{\infty,S_0,S} = \min_{k \in S_0} \kappa^*_{e_k,S}$,
(iii) $c_\kappa(S)^{-1/q}\kappa_{\infty,S} \leq \kappa_{q,S} \leq \kappa_{\infty,S}$,
(iv) $|S_0|^{-1/q}\kappa_{\infty,S_0,S} \leq \kappa_{q,S_0,S} \leq \kappa_{\infty,S_0,S}$,
(v) $\kappa_{1,S} \geq \kappa_{\infty,\overline{S},S}/c_\kappa(S)$,
(vi) $\kappa_{g,S} \geq \max(\kappa_{1,S}, \left(r(n)/\kappa_{1,S_I,S} + 1/\kappa_{1,S_I^c,S}\right)^{-1})$,
(vii) For all $S_0 \supseteq \overline{S}$,

$$\kappa_{\infty,S_0,S} \geq \min_{k \in S_0} \max_{\lambda \in \mathbb{R}^{d_Z} : |\lambda|_1 \leq 1} \left(|\lambda^\top \Psi_{\cdot,k}| - (c_\kappa(S) - 1) \max_{k' \neq k} |\lambda^\top \Psi_{\cdot,k'}|\right),$$

(viii) For all $k \in [d_X]$,

$$\kappa^*_{e_k,S} \geq \max_{\lambda \in \mathbb{R}^{d_Z} : |\lambda|_1 \leq 1} ((|\lambda^\top \Psi_{\cdot,k}| + \max_{k' \neq k} |\lambda^\top \Psi_{\cdot,k'}|) \left(\frac{c_\kappa(S) \max_{k' \neq k} |\lambda^\top \Psi_{\cdot,k'}|}{\kappa_{\infty,\overline{S},S}} + 1\right)^{-1}.$$

The above statements hold if we replace the sensitivities based on K_S by those based on $K_{\gamma,S}$, $c_\kappa(s)$ by $c_\gamma(s)$. We also have $\gamma_{h,S} \geq \gamma_{1,S}$.

Maximizing only on the vectors $(f_l)_{l \in [d_Z]}$, (vii) yields

$$\kappa_{\infty,S_0,S} \geq \min_{k \in S_0} \max_{l \in [d_Z]} \left(|\Psi_{l,k}| - (c_\kappa(S) - 1) \max_{k' \neq k} |\Psi_{l,k'}|\right). \tag{4.10}$$

The advantage of (vii) is that it allows for combinations of the IVs, which can provide a tighter bound. These combinations do not need to be constructed in practice. The term $\max_{k' \neq k} |\lambda^\top \Psi_{\cdot,k'}|$ could be set to zero by taking λ in the orthogonal complement of the vector space spanned by the columns $\Psi_{\cdot,k'}$ for $k' \neq k$. Doing so, one could then maximize $|\lambda^\top \Psi_{\cdot,k}|$ for vectors λ in that space such that $|\lambda|_1 \leq 1$ (which is not trivial if $d_X - 1 < d_Z$) to optimally combine the IVs. The lower bound in (vii) is sharper and achieves a trade-off, which is desirable when the orthogonal complement is too small to deliver a strong IV. Results (vii), (iii), and (v) yield a lower bound on $\kappa_{q,S}$ for $q \in [1, \infty]$. The lower bound (viii) for coordinate-wise sensitivities is useful to analyse variable selection. Result (ii) with $S_0 = \overline{S} \cup \{k\}$

also provides bounds for the coordinate-wise sensitivities $\kappa^*_{e_k,S}$ which could be sharper than (viii). In case (IC), we have

$$\kappa_{1,S} \geq (u_\kappa|S|)^{-1}\,\kappa_{\infty,S,S}, \quad \kappa_{q,S} \geq (u_\kappa|S|)^{-q}\,\kappa_{\infty,S}. \tag{4.11}$$

It is usual, in the absence of endogeneity, to assume that certain quantities such as eigenvalues of all sub-blocks of Ψ of size at most s (sparse eigenvalues) or restricted eigenvalues are bounded away from zero (see, *e.g.*, [9]) in which case $\kappa_{1,S} \geq \kappa/s$ for some $\kappa > 0$ which could depend on n (see Section A.4 in (v5)). In contrast, the lower bounds on $\kappa_{1,S}$ and $\kappa_{\infty,S,S}$ only depend on the regressors with index in S. This can give much sharper lower bounds than when we pay the price of a minimum over all sets of size s. Proposition A.3 provides an alternative lower bound in case (IC).

We now illustrate these bounds in case (IC). The examples can be simple because, due to (vii), the identity of the IVs or their linear combination is irrelevant. In Example E1, by (4.10) and (viii), we have $\kappa_{\infty,S,S} \geq \min_{k\in S}\rho_k$ and $\kappa^*_{e_k,S} \geq \rho_k$. Due to the form of $\kappa_{\infty,S,S}$, its magnitude and those of the sensitivities we deduce from it (*e.g.*, $\kappa_{1,S}$) depend on S beyond its cardinality. For each relevant regressor, what matters is the strength of the strongest IV. The sensitivity $\kappa_{\infty,S,S}$ can be small if some regressors of index in S do not have sufficiently strong IVs, but not if this happens for regressors outside S.

In Example E2, taking the entries of the vector λ to be constant on a set $\widetilde{S} \subseteq [d_Z]$ and zero elsewhere for $k = d_X$ and $\lambda = f_k$ for $k \in S \cap S_I$, (vii) yields $\kappa_{\infty,S,S} \geq \min(\max_{\widetilde{S}\subseteq[d_Z]}(|\rho_{\widetilde{S}}|_1 - (u_\kappa|S| - 1))/|\widetilde{S}|, 1 - (u_\kappa|S| - 1)|\rho_{S\cap S_I}|_\infty)$. From the first term in the minimum it is important that $u_\kappa|S| - 1$ be small relative to $|\rho|_1$. This occurs if there are sufficiently many exogenous regressors with index outside S which are sufficiently correlated with the endogenous regressor. Due to the second term in the minimum, the regressors of index in $S \cap S_I$ should have a sufficiently small correlation with the endogenous regressor.

4.4.2. *Rates of Convergence and Confidence Sets with Estimated Support.* Theorem 4.3 establishes the rate of convergence for $\widehat{\beta}$. Rates of convergence for $\widehat{\sigma}$ and $\widehat{\sigma}(\widehat{\beta})$ are given in (A.46) and (A.47). For $S \subseteq [d_X]$, we use

$$\Theta_\kappa(S) \triangleq (1+\tau(n))\left(1 - \frac{\tau(n)}{\kappa_{1,S}} - \frac{r(n)(1+\tau(n))}{c\kappa_{1,S\cap S_Q,S}}\right)^{-1}_+$$

and $\Theta_\gamma(S)$ which is obtained by replacing $\kappa_{1,S}$ and $\kappa_{1,S\cap S_Q,S}$ by $\gamma_{1,S}$ and $\gamma_{h,S}$. We also sometimes use, for vectors $\omega \in \mathbb{R}^{d_X}$, the restricted set of identified vectors to comprise vectors with sufficiently large nonzero entries

$$\mathcal{I}(\omega) \triangleq \mathcal{I} \bigcap \left\{\beta : \forall k \in S(\beta),\ 1_n\sqrt{\mathbb{E}[X_k^2]}\,|\beta_k| > \omega_k\right\}.$$

Theorem 4.3. *Let $c > 0$. Under Assumption A.2, for all $\beta, \mathbb{P}$ such that $\beta \in \mathcal{I}$ and all STIV estimators $(\widehat{\beta}, \widehat{\sigma})$, we have, on $\mathcal{G} \cap \mathcal{G}_\Psi$,*

(i) For all $l \in \mathcal{L}$,

$$1_n l\left(D_X^{-1}\left(\widehat{\beta} - \beta\right)\right) \leq \frac{2r(n)\sigma_{U(\beta)}}{\kappa_{l,S(\beta)}}\Theta_\kappa(S(\beta));$$

(ii) If $\beta \in \mathcal{I}(\omega)$ where, for all $k \in S(\beta)$, $\omega_k = 2r(n)\sigma_{U(\beta)}\Theta_\kappa(S(\beta))/\kappa^*_{e_k,S(\beta)}$, then $S(\beta) \subseteq S(\widehat{\beta})$;

(iii) For all $q \in [1,\infty]$ and $S_0 \subseteq [d_X]$,

$$1_n\left|D_X^{-1}\left(\widehat{\beta} - \beta\right)_{S_0}\right|_q \leq 2\min_{S\subseteq[d_X]}\max\left(\frac{2r(n)\sigma_{U(\beta)}}{\gamma_{q,S_0,S}}\Theta_\gamma(S), 3\left|D_X^{-1}\beta_{S^c\cap S_Q}\right|_1\right).$$

For a given $\widehat{\beta}$, one can take the infimum over $\beta \in \mathcal{I}$ on both sides of the inequality in (i) or (iii). The left-hand side can then be viewed as the distance to a set, and the right-hand side defines the elements of $\mathcal{I}$ to which $\widehat{\beta}$ is closest. The discussion below uses such β. If $\mathcal{I}$ is a singleton, $\beta = \beta^*$. Due to Proposition 4.2 (i), $\Theta_\kappa(S(\beta))$ is close to 1 when $\max(r(n), \tau(n))/\kappa_{1,S(\beta)} \to 0$. For Example E1, using (4.11), this occurs if

$$\max(r(n), \tau(n))|S(\beta)| \left(\min_{k \in S(\beta)} \rho_k \right)^{-1} \to 0. \tag{4.12}$$

If $\Theta_\kappa(S(\beta))$ is close to 1, then (i) implies $|D_X^{-1}(\widehat{\beta} - \beta)|_1 \lesssim r(n)/\kappa_{1,S(\beta)}$. We might ask ourselves if this converges to zero as n goes to infinity. If d_X increases with n, this is a bound on the norm of an increasingly large vector. For Example E1, we obtain $|D_X^{-1}(\widehat{\beta} - \beta)|_1 \lesssim r(n)|S(\beta)|(\min_{k \in S(\beta)} \rho_k)^{-1}$ and "consistency" in ℓ_1-norm simply relies on (4.12), hence on $\min_{k \in S(\beta)} \rho_k \gg \max(r(n), \tau(n))|S(\beta)|$. This means that there should exist a strong enough IV for every regressor with index in $S(\beta)$. The rate is not affected if there is not a strong IV for a regressor which is irrelevant. When ρ is constant, (4.12) requires $\max(r(n), \tau(n))|S(\beta)| \to 0$ which imposes an upper bound on the sparsity of the structural vector. When the entries of ρ are in decreasing order and the largest index in $S(\beta)$ is k_*, then consistency requires $\max(r(n), \tau(n))|S(\beta)|\rho_{k_*}^{-1} \to 0$. The magnitude of $\rho_{k_*}^{-1}$ depends on the rate of decay of the components of ρ but also on the identity of the relevant regressors. It is independent of d_X. Even if d_X is very large, the rate of convergence of STIV in ℓ_1-norm can be very small.

Remark 4.2. *If in Example E1 we also have $\mathbb{E}[ZZ^\top] = I$ and, for all $k \in S(\beta)$, $\rho_{j,k} = 1/\sqrt{l_k}$ when it is nonzero (so the population R^2 does not change with l_k), one should have $\max(r(n), \tau(n))\sqrt{|l_{S(\beta)}|_\infty}|S(\beta)| \to 0$. If we take the first typical behavior for $r(n)$, we cannot show that STIV is consistent when $|l_{S(\beta)}|_\infty \ln(d_Z d_X)/n$ converges to a nonzero constant. When $d_X = 1$ this becomes $d_Z \ln(d_Z)/n$ converges to a nonzero constant. This is a case where 2SLS is consistent (it is not when we remove $\ln(d_Z)$).*

If we consider different losses for Example E1, we have a different behavior. For example, $|\widehat{\beta}_k - \beta_k| \lesssim r(n)\rho_k^{-1}$ and the sparsity does not play a role. This means that we can have heterogenous rates of convergence for different components. Moreover, $\max_{k \in S_0} |\widehat{\beta}_k - \beta_k| \lesssim r(n) \max_{k \in S_0} \rho_k^{-1}$. This echoes [26] for linear regression, except for the additional $\max_{k \in S_0} \rho_k^{-1}$ factor. Importantly STIV can be consistent when $d_Z < d_X$ including for Example E2.

The condition $\beta \in \mathcal{I}(\omega)$ in (ii) is a beta-min condition. It makes sense in applications such as social interactions when the network is sparse. There, the coefficients also have a clear interpretation and are meaningful parameters of interest. The beta-min assumption is not aimed to be used when the regressors are functions of baseline regressors introduced to approximate a structural function. The value of ω_k corresponds to the upper bound on $1_n\sqrt{\mathbb{E}[X_k^2]}|\widehat{\beta}_k - \beta_k|$. In Example E1, the magnitude of ω_k is proportional to ρ_k^{-1}. Result (ii) means that, for all $\beta \in \mathcal{I}(\omega)$, the STIV estimator recovers a superset of the regressors. Based on (ii), the set $\widehat{C}_{\varphi,\kappa} \triangleq \{\varphi(b) : b \in \widehat{C}_\kappa\}$, where

$$\widehat{C}_\kappa \triangleq \left\{ b : \forall l \in \mathcal{L},\ l\left(\mathbf{D}_{\mathbf{X}}^{-1}\left(\widehat{\beta} - b\right)\right) \leq \frac{2\widehat{r}\widehat{\sigma}\widehat{\theta}_\kappa(S(\widehat{\beta}))}{\widehat{\kappa}_l(S(\widehat{\beta}))} \right\}, \tag{4.13}$$

is such that

$$\inf_{\beta, \mathbb{P}:\ \beta \in \mathcal{I}(\omega)} \mathbb{P}\left(\varphi(\beta) \in \widehat{C}_{\varphi,\kappa}\right) \geq 1 - \alpha - \alpha_D(n). \tag{4.14}$$

The confidence set $\widehat{C}_{\varphi,\kappa}$ is not robust to identification because ω depends on Ψ, hence on the joint distribution of $\mathbf{Z}$ and $\mathbf{X}$, through the population sensitivities.

If $\beta \in \mathcal{I}$ is sparse, the upper bound in (iii), which holds for all $S \subseteq [d_X]$, also applies to $S = S(\beta)$ in which case the second term in the maximum is zero. We are then left with a bound which is similar to the right-hand side of (i). The second term is an alternative bound on the loss when, on $\mathcal{G}$, $\beta \notin K_{\gamma,S}$. When $q = 1$ and $S_0 = S_Q = [d_X]$, it is 6 times the error we would make by having $\widehat{\beta} = \beta_S$ (estimating perfectly the components in S). This is the sense in which the second term is an approximation error. The STIV estimator performs a data-driven trade-off for nonsparse vectors. This trade-off is desirable when a subvector $\beta_{S^c \cap S_Q}$ cannot be distinguished from zero, in which case it is better to estimate it as zero and make a smaller error on the important entries. One gets from (iii) that, for an optimal set $S_* \subseteq [d_X]$ (not necessarily $S(\beta)$),

$$1_n \left| D_X^{-1} \left(\widehat{\beta} - \beta \right) \right|_1 \leq \frac{2r(n)\sigma_{U(\beta)}}{\gamma_{1,S_*}} \Theta_\gamma(S_*). \tag{4.15}$$

This allows us to define formally approximately sparse vectors as vectors which are sufficiently well approximated by a sparse vector so that the right-hand side of (4.15) is small. In the case of Example E1 with a decreasing sequence ρ_k (this is without loss of generality because (iii) concerns nonlinear approximation) and an ordered sequence of subsets of the form $S = [K]$ for $K \in [d_X]$, we obtain

$$\left| D_X^{-1} \left(\widehat{\beta} - \beta \right) \right|_1 \lesssim \min_{K \in [d_X]} \max \left(\frac{sr(n)}{\rho_K}, \sum_{k>K} \left| \left(D_X^{-1} \beta \right)_k \right| \right).$$

To avoid the right-hand sides of (i) and (iii) being infinite, or the set of vectors in $\mathcal{I}(\omega)$ being empty for some $\mathbb{P} \in \mathcal{P}$ (due to the sensitivities being too small), one can restrict further the class $\mathcal{P}$. The results are no longer robust to identification nor uniform in the sense that Ψ can no longer be arbitrary.

4.4.3. *Selection of Variables and Confidence Sets with Estimated Support.* Theorem 4.3 (ii) provides a superset of regressors. Under a stronger beta-min condition exact selection can be performed. For this purpose, we use a thresholded STIV estimator $\widehat{\beta}^\omega$ which uses a sparsity certificate. It is defined by

$$\widehat{\beta}_k^\omega \triangleq \widehat{\beta}_k 1\!\!1 \left\{ \sqrt{\mathbb{E}_n[X_k^2]} \left| \widehat{\beta}_k \right| > \widehat{\omega}_k(s) \right\}, \ \widehat{\omega}_k(s) \triangleq \frac{2\widehat{r}\widehat{\sigma}\widehat{\theta}_\kappa(s)}{\widehat{\kappa}^*_{e_k}(s)}. \tag{4.16}$$

for $k \in [d_X]$. The constants $\kappa_l(s)$ and $\kappa^*_\omega(s)$ are population analogues of the lower bounds in Proposition 4.1. Due to the duality between testing and confidence sets, for the same value of c, the thresholded STIV estimator is equivalently obtained by setting to zero the entries of $\widehat{\beta}$ for which 0 lies in the bounds in (4.9). The following theorem shows that, based on thresholding a STIV estimator, we achieve sign consistency and hence, $S(\widehat{\beta}^\omega) = S(\beta)$ for all $\beta \in \mathcal{I}_s \cap \mathcal{I}(2\omega(s))$. It uses $\text{sign}(b) \triangleq (\text{sign}(b_k))_{k \in [d_X]}$, where $\text{sign}(t) \triangleq 1\!\!1\{t > 0\} - 1\!\!1\{t < 0\}$. We make use of the following population counterparts to establish the result

$$\omega_k(s) \triangleq \frac{2r(n)\sqrt{1+\tau(n)}\overline{\theta}_\kappa(s)}{\overline{\kappa}^*_{e_k}(s)} \sup_{\beta \in \mathcal{I}_s} \left(\sigma_{U(\beta)} \left(2 \left(1 - \frac{r(n)(1+\tau(n))}{c\kappa_{1,S(\beta)\cap S_Q,S(\beta)}} \left(1 - \frac{\tau(n)}{\kappa_{1,S(\beta)}} \right)_+^{-1} \right)_+^{-1} - 1 \right) \right),$$

for all $k \in [d_X]$, where $\overline{\theta}_\kappa(s)$ and $\overline{\kappa}^*_{e_k}(s)$ are defined in Proposition A.1.

Theorem 4.4. *Let $s \in [d_Q]$. Under Assumption A.2, for all $\beta, \mathbb{P}$ such that $\beta \in \mathcal{I}_s \cap \mathcal{I}(2\omega(s))$ and all STIV estimators $(\widehat{\beta}, \widehat{\sigma})$, we have, on $\mathcal{G} \cap \mathcal{G}_\Psi$, $\text{sign}(\widehat{\beta}^\omega) = \text{sign}(\beta)$.*

The confidence sets corresponding to Theorem 4.4 are obtained by replacing $S(\widehat{\beta})$ by $S(\widehat{\beta}^\omega)$ in (4.13), and satisfy (4.14) with $\mathcal{I}_s \cap \mathcal{I}(2\omega(s))$ in place of $\mathcal{I}(\omega)$ in the infimum. The sparsity certificate s can be large, and possibly equal to d_X. The width of the sets matches the error bound in Theorem

4.1 as regards the dependence on $S(\beta)$. For this reason the confidence set adapts to the sparsity. To achieve this we need to remove a small set from $\mathcal{B}$ (the vectors which are too close to $|S(\beta)|$-sparse vectors). This is related to impossibility results for adaptive confidence sets in nonparametric and high-dimensional statistics.

5. Other Moment Models

5.1. **Endogenous IVs.** Testing IV exogeneity in the presence of overidentification is a classical problem studied since [29][3]. We replace (2.1) with (2.1b), (2.2) with $(\beta,\theta)\in\mathcal{B}$, $\mathbb{P}(\beta,\theta)\in\mathcal{P}$, where $\mathbb{P}(\beta,\theta)$ is the distribution of $\left(\mathbf{X}_{i,\cdot},\mathbf{Z}_{i,\cdot},\mathbf{Z}_{i,\cdot}^\top\mathbf{U}_i(\beta)-\theta\right)_{i\in[n]}$ implied by $\mathbb{P}$, and modify $\mathcal{I}\subseteq\mathbb{R}^{d_X+d_Z}$ accordingly. The class $\mathcal{P}$ can be any of classes 1-4 in Section A.1.1, replacing $\mathbf{Z}_{i,\cdot}^\top\mathbf{U}_i(\beta)$ with $\mathbf{Z}_{i,\cdot}^\top\mathbf{U}_i(\beta)-\theta$. The set $\mathcal{B}\subseteq\mathbb{R}^{d_X+d_Z}$ accounts for the restrictions on (β,θ). For example, the researcher could know the sign of the correlation between an endogenous regressor and the structural error. An important restriction is $\theta_{S_\perp}=0$ for $S_\perp\subseteq[d_Z]$ which corresponds to the indices of the IVs known to be exogenous. We denote by $d_{EX}\triangleq|S_\perp|$ and $\mathcal{I}_{s,\widetilde{s}}=\mathcal{I}\cap\{(\beta,\theta)\in\mathcal{B}:|S(\beta)\cap S_Q|\leq s,|S(\theta)|\leq\widetilde{s}\}$, where $\widetilde{s}\in[d_Z-d_{EX}]$ is a sparsity certificate for the possibly endogenous IVs. We now present confidence sets for $\varphi(\beta,\theta)$. They do not require $d_{EX}\geq d_X$.

First, the SNIV confidence set in (3.1) is easy to modify, yielding

$$\widehat{C}(s,\widetilde{s})\triangleq\left\{(b,t)\in\mathcal{B}:|S(b)\cap S_Q|\leq s,|S(t)|\leq\widetilde{s},\ \max_{l\in[d_Z]}\left|\frac{1}{n}\frac{\mathbf{Z}_{\cdot,l}^\top\mathbf{U}(b)-nt_l}{\widetilde{\sigma}_l(b,t)}\right|\leq r_0(n)\right\},$$

where $\widetilde{\sigma}_l(b,t)^2\triangleq\mathbb{E}_n[(Z_lU(b)-t_l)^2]$. Further details are in Section C.1. A second approach relies on the C-STIV estimator (see Section B.1.1).

Definition 5.1. *For $c\in(0,1)$, the C-STIV estimator $(\widehat{\beta},\widehat{\theta},\widehat{\sigma})$ is any solution of*

$$\min_{(b,t)\in\widehat{\mathcal{I}}_C(r_0(n),\sigma),\sigma\geq0}\left|\mathbf{D}_\mathbf{X}^{-1}b_{S_Q}\right|_1+\left|\mathbf{D}_\mathbf{Z}t_{S_\perp^c}\right|_1+c\sigma,$$

where

$$\widehat{\mathcal{I}}_C(r_0(n),\sigma)\triangleq\left\{(b,t)\in\mathcal{B}:\ \left|\mathbf{D}_\mathbf{Z}\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{U}(b)-t\right)\right|_\infty\leq r_0(n)\sigma,\ \widehat{F}(b,t)\leq\sigma\right\},$$

$$\forall(b,t)\in\mathbb{R}^{d_X+d_Z},\ \widehat{F}(b,t)\triangleq\max_{l\in[d_Z]}\widehat{\sigma}_l(b,t),\ \text{and}\ \widehat{\sigma}_l(b,t)^2\triangleq(\mathbf{D}_\mathbf{Z})_{l,l}^2\,\mathbb{E}_n\left[(Z_lU(b)-t_l)^2\right].$$

When $d_{EX}=d_Z$ the C-STIV is an alternative to the STIV estimator[4]. Its analysis is similar to that of STIV. The main computational difference is that there are as many second-order cones as d_Z. This makes it much harder to solve. However, we use a smoothing approach to approximate it which makes the computations much faster (see Section C.3). This development is essential for the confidence bands in Section 6 to handle easily thousands of regressors and instruments (as in the application in Section 7.2) which would else be limited to a few hundreds. The C-STIV estimator is also used to estimate certain left inverses of Ψ in Section 6.

C-STIV obtains $(\widehat{\beta},\widehat{\theta})$ simultaneously. Alternatively, one can apply STIV or C-STIV to obtain a pilot estimator $\widehat{\beta}$ using IVs of index in $S_\perp$, and obtain $\widehat{\theta}$ using the NV-STIV estimator (see Section B.1.2). NV-STIV is a modification of C-STIV, in which $\widehat{\mathcal{I}}_C(r_0(n),\sigma)$ is altered to include $b=\widehat{\beta}$, $r_0(n)\sigma$ is replaced by $r_0(n)\sigma+\widehat{b}$ and $\widehat{F}(b,t)\leq\sigma$ by $\widehat{F}(b,t)\leq\sigma+\widehat{b}^\sigma$. The terms $\widehat{b},\widehat{b}^\sigma$ are constructed using

[3]This question is addressed in (v1), see (v5) for references.

[4]The idea appeared in (6.5) in (v1) and C-STIV was called the STIV when the paper was first revised (see (v2b)).

lower bounds on sensitivities and such that for β in $\mathcal{I}_s$

$$\left|\left(\widehat{\Psi}\mathbf{D}_{\mathbf{X}}^{-1}\left(\widehat{\beta}-\beta\right)\right)_{S_\perp^c}\right|_\infty \leq \widehat{b}, \quad \widehat{\rho}_{ZX,S_\perp^c}\left|\mathbf{D}_{\mathbf{X}}^{-1}\left(\widehat{\beta}-\beta\right)\right|_1 \leq \widehat{b}^\sigma, \tag{5.1}$$

with high probability, where

$$\widehat{\rho}_{ZX,S_\perp^c} \triangleq \max_{l\in S_\perp^c,\ k\in[d_X]} (\mathbf{D}_{\mathbf{Z}})_{l,l}\,(\mathbf{D}_{\mathbf{X}})_{k,k}\sqrt{\mathbb{E}_n\left[Z_l^2X_k^2\right]},$$

which is simply denoted by $\widehat{\rho}_{ZX}$ when we replace $l \in S_\perp^c$ by $l \in [d_Z]$. One then uses sensitivities for the NV-STIV to construct confidence sets for θ.

5.2. **Approximation Errors and Systems of Equations.** We replace (2.1) to (2.1c) and assume that $\sigma_{V(\beta)} \leq v(d_X)$, for $v(d_X)$ decaying to zero with d_X. The assumptions previously made on $(\mathbf{X},\mathbf{Z},\mathbf{U}(\beta))$ are made on $(\mathbf{X},\mathbf{Z},\mathbf{W}(\beta))$ and $\mathcal{I}$ is modified accordingly and incorporates $\sigma_{V(\beta)} \leq v(d_X)$. The class $\mathcal{P}$ can be any of classes 1-5. The model with approximation errors allows for the structural equation

$$\forall i \in [n],\ \mathbf{Y}_i = f(\widetilde{\mathbf{X}}_i) + \mathbf{W}_i,\ \mathbb{E}[\mathbf{W}_i|\widetilde{\mathbf{Z}}_i] = 0, \tag{5.2}$$

where $f \in \mathcal{S}$, and for functions $(\varphi_k)_{k\in\mathbb{N}}$ and a decreasing sequence $(v(d_X))_{d_X\in\mathbb{N}}$,

$$\forall d_X \in \mathbb{N},\ \sup_{g\in\mathcal{S}}\inf_{b\in\mathbb{R}^{d_X}} \mathbb{E}\left[\left(g\left(\widetilde{\mathbf{X}}_i\right) - \sum_{k=1}^{d_X}\varphi_k\left(\widetilde{\mathbf{X}}_i\right)b_k\right)^2\right] \leq v(d_X)^2. \tag{5.3}$$

The rate of decay of $(v(d_X))_{d_X\in\mathbb{N}}$ is usually taken slow so that $\mathcal{S}$ can be large. This amounts to assuming minimum smoothness. The function f can lie in a class of much smoother functions contained in $\mathcal{S}$. The model with approximation error is obtained by taking $f \in \mathcal{S}$, $\mathbf{X}_{i,\cdot} = (\varphi_1(\widetilde{\mathbf{X}}_i),\ldots,\varphi_{d_X}(\widetilde{\mathbf{X}}_i))$, $\mathbf{V}_i(\beta) = f(\widetilde{\mathbf{X}}_i) - \sum_{k=1}^{d_X}\varphi_k(\widetilde{\mathbf{X}}_i)\beta_k$, and $\sigma_{V(\beta)} \leq v_{d_X}$. The term $\mathbf{V}_i(\beta)$ is the error made by approximating the function in the high-dimensional space, and $v(d_X) = o(n^{-1/2})$ for d_X sufficiently large. For well chosen classes $\mathcal{S}$ and functions $(\varphi_k)_{k\in\mathbb{N}}$, the vector $\beta \in \mathcal{I}$ is approximately sparse. We use IVs which comprise functions of $\widetilde{\mathbf{Z}}_i$. This approach can be generalized to system where $\beta \in \mathcal{M}_{d_X,d_G}$, $S_Q, S_I \subseteq [d_X]\times[d_G]$, $\mathbf{U}(\beta),\mathbf{V}(\beta),\mathbf{W}(\beta) \in \mathcal{M}_{n,d_G}$, and $\sigma_{V(\beta)} \leq v(d_X)$ for $v(d_X) \in \mathbb{R}^{d_G}$. We now define the estimator.

Definition 5.2. *For $c, \widehat{v} > 0$, the* E-STIV *estimator $\left(\widehat{\beta},\widehat{\sigma}\right)$ is any solution of*

$$\min_{b\in\widehat{\mathcal{I}}_E(\widehat{r},\sigma),\sigma\geq 0}\left(\left|\mathbf{D}_{\mathbf{X}}^{-1}b_{S_Q}\right|_1 + c|\sigma|_1\right), \tag{5.4}$$

where

$$\widehat{\mathcal{I}}_E(\widehat{r},t) \triangleq \left\{b\in\mathcal{B},\ \forall g\in[d_G],\ \left|\frac{1}{n}\mathbf{D}_{\mathbf{Z}}\mathbf{Z}^\top(\mathbf{Y}_{\cdot,g} - \mathbf{X}b_{\cdot,g})\right|_\infty \leq \widehat{r}\sigma_g + (\widehat{r}+1)\widehat{v}_g,\ \widehat{\sigma}_g(b) \leq t_g\right\},$$
$$\forall b \in \mathcal{M}_{d_X,d_G},\ \forall g\in[d_G],\ \widehat{\sigma}_g(b)^2 \triangleq \mathbb{E}_n[U_g(b)^2].$$

For the nonparametric model (5.2) with a single equation, one can take $\widehat{v} = \sqrt{1+\tau(n)}v(d_X)$. The E-STIV estimator can also be used for other approximate models such as bracketed data (then $\widehat{v}$ can be determined from the data, see (v5)). We analyse the E-STIV estimator in Section B.2. It allows for cross-equation restrictions, and the number of equations d_G can depend on n.

6. Confidence Bands using Bias Correction

The confidence sets presented thus far can be conservative if d_F is small relative to d_X. In this section, for $\Omega \in \mathcal{M}_{d_F,d_X}$, we propose an estimator and confidence band for $\varphi(b) = \Omega b$. Specifically,

we construct a confidence band comprising d_F intervals and directly control the probability that $\Omega\beta$ lies inside. Taking $\Omega = e_k^\top$ for $k \in [d_X]$ yields a confidence interval as a special case. If the structural equation is (5.2), we can use this to construct a confidence band around f at d_F grid points (*i.e.*, around $f(t_1), f(t_2), \ldots, f(t_{d_F})$ where $(t_j)_{j=1}^{d_F}$ is a grid on the support of $\widetilde{\mathbf{X}}$) by taking $\Omega = (\varphi_k(t_j))_{j\in[d_F],k\in[d_X]}$ and we control instead the probability that $\Omega\beta + \overline{V}(\beta)$, where $\overline{V}(\beta)$ is the approximation error at those grid points, lies in the band. The number of grid points is permitted to grow with n.

A first strategy to estimate $\Omega\beta$ is to use the plug-in estimator $\Omega\widehat{\beta}$, where $\widehat{\beta}$ is an estimator from the STIV family. A second strategy is to rely on the assumption

$$\exists \Lambda \in \mathcal{M}_{d_F,d_Z} : \ \forall i \in [n], \ \Lambda\mathbb{E}\left[\mathbf{Z}_{i,\cdot}^\top \mathbf{X}_{i,\cdot}\right] = \Omega. \tag{6.1}$$

The statement $\forall i \in [n]$ is to allow for independent but non i.i.d. data. To understand this condition, let us consider the case of i.i.d. data. We have $\Lambda\mathbb{E}\left[ZX^\top\right]\beta = \Lambda\mathbb{E}[ZY]$, so $\Omega\beta = \Lambda\mathbb{E}[ZY]$. This yields a second plug-in strategy to estimate $\Omega\beta$ by using an estimator of Λ. Condition (6.1), however, can be problematic. It states that there exists a solution to a system of $d_F d_X$ equations in $d_F d_Z$ unknowns. If $d_Z < d_X$, the system is overdetermined, and so a solution may only exist for particular choices of Ω. If there exists a unique solution, it might not be sparse. If $d_Z > d_X$ there are more unknowns than equations. If a solution exists, the set of Λ which satisfy (6.1) is an affine space, and so Λ is not point identified but there can exist for example a sparsest solution. In case of $\mathbb{E}\left[ZX^\top\right]\beta = \mathbb{E}[ZY]$ and (6.1), one has to deal with the multiplication of the parameter to estimate with $\mathbb{E}\left[ZX^\top\right]$, either to the left or to the right. The dimensions d_Z and d_X play opposite roles: having $d_Z \geq d_X$ is desirable to estimate β, while having $d_X \geq d_Z$ is desirable to estimate Λ. To deal with both underidentified situations sparsity is a useful device. We now denote, for $L \in \mathcal{M}_{d_F,d_Z}$, by $\mathbf{T}_i(L) \triangleq \Omega - L\mathbf{Z}_{i,\cdot}^\top \mathbf{X}_{i,\cdot}$.

Using either of these plug-in strategies poses problems in high-dimensions because STIV types estimators perform shrinkage and are "biased" towards zero. They can also converge slowly or even fail to converge as we have seen in the case of many equally weak IVs. The approach that we take is to combine the two estimators and form the bias corrected estimator

$$\widehat{\Omega\beta} \triangleq \Omega\widehat{\beta} + \frac{1}{n}\widehat{\Lambda}\mathbf{Z}^\top \mathbf{U}(\widehat{\beta}). \tag{6.2}$$

It amounts to estimating d_F linear combinations of the IVs and interacting them with the estimated residuals. It is at the basis of the confidence bands in this section. They can yield less conservative inference than the previous confidence sets when d_F is small.

Remark 6.1. *The bias-corrected estimator is close in spirit to* [21]*. Another motivation for it is that the mapping O defined as $O(b, L) \triangleq \Omega b + L(\mathbb{E}[Z^\top Y] - \mathbb{E}[Z^\top X]b)$ has partial derivatives with respect to b and L which are zero at, respectively, identified β and Λ (due to $\mathbb{E}\left[ZX^\top\right]\beta = \mathbb{E}[ZY]$ and* (6.1)*) and $O(\beta, \Lambda) = \Omega\beta$. This is a type of double-robustness (see,* e.g.*,* [15]*). In the approach of this paper, $\widehat{\beta}$ is not obtained by estimating regressions (hence no machine learning) but directly in a robust way, allowing all regressors to be endogenous. Another difference is that our analysis is based on* (B.21) *which does not include a term which is the product of two estimation errors. For this reason, we can have much weaker requirements on rates of convergence.*

Remark 6.2. *The assumptions on $(\widehat{\beta}, \widehat{\sigma})$ and $(\widehat{\Lambda}, \widehat{\nu})$ will not be symmetric. If the researcher thinks they are more likely to hold if we interchange their role, she can use $\widetilde{\Omega\beta} \triangleq (\widehat{\Lambda}\mathbf{Z}^\top\mathbf{Y} + \sum_{i\in[n]} \mathbf{T}(\widehat{\Lambda})\widehat{\beta})/n$. The mapping O' defined as $O'(b, L) \triangleq L\mathbb{E}[Z^\top Y] + (\Omega - L\mathbb{E}[Z^\top X])b$ satisfies the same double-robustness property as O.*

The previous methods can be used to estimate Λ because (6.1) is simply another system of moment conditions. Condition (6.1) can hold approximately like in Section B.2. In this section we use

the C-STIV estimator for a system of d_F equations. For $\lambda \in (0,1)$, we obtain $\widehat{\Lambda}$ by solving

$$\min_{L \in \widehat{\mathcal{I}}_C(r_0'(n),\nu),\nu>0} \left|L\mathbf{D}_{\mathbf{Z}}^{-1}\right|_1 + \frac{\lambda\nu}{\widehat{\rho}_{ZX}}, \tag{6.3}$$

where

$$\widehat{\mathcal{I}}_C(r_0'(n),\nu) \triangleq \left\{L \in \mathcal{M}_{d_F,d_Z} : \left|\frac{1}{n}\sum_{i\in[n]} \mathbf{T}_i(L)\mathbf{D}_{\mathbf{X}}\right|_\infty \leq r_0'(n)\nu,\ \widehat{F}(L) \leq \nu\right\},$$

$$\forall L \in \mathcal{M}_{d_F,d_Z},\ \widehat{F}(L) \triangleq \max_{(f,k)\in[d_F]\times[d_X]} \widehat{\sigma}_{f,k}(L),\ \text{and } \widehat{\sigma}_{f,k}(L)^2 \triangleq (\mathbf{D}_{\mathbf{X}})_{k,k}^2\, \mathbb{E}_n\left[(T(\Lambda))_{f,k}^2\right].$$

If $\Omega = I$ and $\mathbf{Z} = \mathbf{X}$, $\widehat{\Lambda}$ is an approximate inverse of the matrix $\mathbf{X}^\top\mathbf{X}/n$, which improves on the CLIME estimator of [11] by additionally estimating standard errors. The rates of convergence are given in Theorem B.3.

The sequences $c, \widehat{r}$ and α for the preliminary estimator can vary with n, the last converging to zero. $\mathcal{P}'$ is a nonparametric class for $\mathbb{P}(\beta,\Lambda)$ the distribution of $(\mathbf{X},\mathbf{Z},\mathbf{U}(\beta),(\mathbf{T}_i(\Lambda))_{i\in[n]},\mathbf{Z}\Lambda^\top D_{\Lambda Z})$. The one on which our analysis is based defined in Assumption B.1. To choose $r_0'(n)$, we use a suitable modification of the classes presented in Section A.1.1. For ease of exposition we focus on class 4, in which case $r_0'(n)$ is defined identically to $r_0(n)$, replacing α with a sequence converging to zero and d_Z with $d_F d_X$. We define the set of identifiable parameters

$$\mathcal{I}_\Omega \triangleq \left\{\beta \in \mathcal{I}, \Lambda : \forall i \in [n], \Lambda\mathbb{E}\left[\mathbf{Z}_{i,\cdot}^\top \mathbf{X}_{i,\cdot}\right] = \Omega, \mathbb{P}(\beta,\Lambda) \in \mathcal{P}'.\right\},$$

Since it is needed for the empirical application, we present our results below for the case in which $\widehat{\beta}$ is a pilot estimator of the parameters of a system of d_G equations as the one studied in generalization (2.1d) or the C-STIV estimator for system used here to obtain $\widehat{\Lambda}$. $\widehat{\beta}$ is a $d_X \times d_G$ matrix and d_G can depend on n. For $\alpha \in (0,1)$, we define

$$\widehat{C}_{\Omega,g} \triangleq \left[\underline{\widehat{C}}_{\Omega,g}, \overline{\widehat{C}}_{\Omega,g}\right],\ \underline{\widehat{C}}_{\Omega,g} \triangleq \widehat{\Omega\beta}_{\cdot,g} - \widehat{q},\ ,\overline{\widehat{C}}_{\Omega,g} \triangleq \widehat{\Omega\beta}_{\cdot,g} + \widehat{q}, \tag{6.4}$$

$$\widehat{q} \triangleq \frac{\left(q_{G_\Omega|\mathbf{U}(\widehat{\beta})\mathbf{Z}\widehat{\Lambda}^\top}(1-\alpha) + 2\zeta(n)\right)}{\sqrt{n}} \mathbf{D}^{-1}_{\mathbf{U}(\widehat{\beta})\mathbf{Z}\widehat{\Lambda}^\top} 1 + \frac{\overline{v}(n)}{\sqrt{n}}, \tag{6.5}$$

where $q_{G_\Omega|\mathbf{U}(\widehat{\beta})\mathbf{Z}\widehat{\Lambda}^\top}(1-\alpha)$ is the $1-\alpha$ quantile of the distribution of $G_\Omega = |\mathbf{D}_{\mathbf{U}(\widehat{\beta})\mathbf{Z}\widehat{\Lambda}^\top}\widehat{\Lambda}\mathbf{Z}^\top\mathbf{U}(\widehat{\beta})\mathbf{E}|_\infty/\sqrt{n}$ given $\mathbf{U}(\widehat{\beta})\mathbf{Z}\widehat{\Lambda}^\top$ (computed by simulation), $\mathbf{E}$ is a Gaussian vector in $\mathbb{R}^n$ with mean zero and variance I and independent from $\mathbf{U}(\widehat{\beta})\mathbf{Z}\widehat{\Lambda}^\top$, and $(\zeta(n))_{n\in\mathbb{N}}$, $(\overline{v}(n))_{n\in\mathbb{N}}$, and $(\alpha_\Omega(n))_{n\in\mathbb{N}}$ are sequences appearing in Assumption B.2. To have a unified framework, the result below is for the vector of approximately linear functionals $\Omega\beta_{\cdot,g} + \overline{V}_g(\beta)$.

Theorem 6.1. *Let assumptions B.1 and B.2 hold. Then, for all $\mathbb{P},(\beta,\Lambda)$ such that $(\beta,\Lambda) \in \mathcal{I}_\Omega$ and $n \in \mathbb{N}$, every collection $(\widehat{C}_{\Omega,g})_{g\in[d_G]}$ satisfies*

$$\mathbb{P}\left(\Omega\beta_{\cdot,g} + \overline{V}_g(\beta) \in \widehat{C}_{\Omega,g}\right) \geq 1 - \alpha - \alpha_\Omega(n) \quad \forall g \in [d_G].$$

The sequence $(\overline{v}(n))_{n\in\mathbb{N}}$ in (6.5) allows to obtain a result which is uniform in n but the researcher can use $\overline{v}(n) = 0$. The confidence bands have asymptotically coverage at least $1-\alpha$ if $\alpha_\Omega(n) \to 0$ and $\overline{v}(n) \to 0$. The first condition is very mild and requires that $(\zeta(n))_{n\in\mathbb{N}}$ does not converge too fast to 0 (see (i) in Assumption B.2), $\alpha_\beta(n) + \alpha_\Omega(n) \to 0$, and d_F is of smaller order than an exponential in n. Due to item (iii) in Assumption B.2, the second condition holds if

$$\sqrt{n}(1+\overline{v}_D(n))\left(\left|D_{\Lambda Z W(\beta)}\right|_\infty r_0'(n) v_{F(\Lambda)}(n) v_{\beta,1}(n) + \left|D_{\Lambda Z W(\beta)}\right|_\infty |\overline{V}(\beta)|_\infty\right. \tag{6.6}$$

$$+\left(v_{\Lambda,3}(n)+\left|D_{\Lambda ZW(\beta)}D_{\Lambda Z}^{-1}\right|_\infty\right)|v(d_X)|_\infty\Big)\to 0,$$

$$\overline{v}_D(n)\triangleq\min\Big(\Big(\left|D_{\Lambda ZW(\beta)}\Lambda D_Z^{-1}\right|_{\infty,\infty}+\left|D_{\Lambda ZW(\beta)}\right|_\infty v_{\Lambda,1}(n)\Big)\left(\rho_{ZX}(n)v_{\beta,1}(n)+\rho_Z v(d_X)\right)+v_{\Lambda,2}(n),$$
$$\left|D_{\Lambda ZU(\beta)}\right|_\infty v_{\Lambda,1}(n)\rho_{ZX}\left(v_{\beta,1}(n)+\left|D_X^{-1}\beta\right|_1\right)+\max_{f\in[d_F]}\sqrt{\mathbb{E}\left[\Big(\left(D_{\Lambda ZW(\beta)}\Lambda\right)_{f,\cdot}ZV(\beta)\Big)^2\right]}+v_{\beta,2}(n)\Big),$$

where the additional sequences are introduced in Section B.3. The sequences of the form $v_{\Lambda,1}(n)$, $v_{\Lambda,2}(n)$, and $v_{\Lambda,3}(n)$ are rates of convergence for the estimation of Λ for three different losses, $v_{F(\Lambda)}(n)$ is a bound for $\widehat{\nu}$ (hence is not converging to zero), and $v_{\beta,1}(n)$ and $v_{\beta,2}(n)$ are rates of convergence for the estimation of β. In the absence of approximation error and for linear functionals, the condition simplifies because $v(d_X)=0$ and $\overline{V}(\beta)=0$. Omitting the second term in the minimum in the definition of $\overline{v}_D(n)$, a sufficient condition is that

(1) $(|D_{\Lambda ZW(\beta)}\Lambda D_Z^{-1}|_{\infty,\infty}+|D_{\Lambda ZW(\beta)}|_\infty v_{\Lambda,1}(n))\rho_{ZX}(n)v_{\beta,1}(n)+v_{\Lambda,2}(n)=O(1)$,
(2) $|D_{\Lambda ZW(\beta)}|_\infty r_0'(n)v_{F(\Lambda)}(n)v_{\beta,1}(n)=o(1/\sqrt{n})$.

Item 1 can hold can hold even if $v_{\Lambda,1}(n)$ diverges but $v_{\Lambda,2}(n)$ remains bounded. So it is not required that $\widehat{\Lambda}$ even converges, for any of the two losses for which $v_{\Lambda,1}(n)$ and $v_{\Lambda,2}(n)$ account, to an element Λ such that $(\beta,\Lambda)\in\mathcal{I}_\Omega$. Assuming $\left|D_{\Lambda ZW(\beta)}\right|_\infty v_{F(\Lambda)}(n)=O(1)$, which again does not require the consistency of $\widehat{\Lambda}$, item 2 holds if $r_0'(n)v_{\beta,1}(n)=o(1/\sqrt{n})$. So the estimation error on β should be $o(\ln(d_Fd_X)^{-1/2})$. This is very mild and much weaker than what is usually needed with doubly robust methods (the product of the rates for estimation of Λ and β is $o(1/\sqrt{n})$). Another advantage of the bias correction analysed in this section is that the researcher does not need to find the form of robust moments and can rely on a simple data-driven method.

In nonidentified cases it is possible that two elements of $\mathcal{I}$ give rise to the same $\Omega\beta$ and what then matters is that $\widehat{\beta}$ converges to an approximately sparse element of $\mathcal{I}$ giving rise to the same functional. When Ψ is such that we cannot prove consistency of $\widehat{\beta}$ then the coverage of the confidence bands might not be correct. This is because these bands are not robust to identification. Also, the confidence bands can be valid even if condition 1 does not hold and the quantity diverges, if $(v_{\beta,1}(n))_{n\in\mathbb{N}}$ converges faster to 0. Condition (6.6) can be restrictive due to terms like $|D_{\Lambda ZW(\beta)}\Lambda D_Z^{-1}|_{\infty,\infty}$ in $\overline{v}_D(n)$. The second term in the minimum in the definition of $\overline{v}_D(n)$ can then be useful because it involves instead $|D_X^{-1}\beta|_1$. In Section B.3 we show that if we assume that the structural errors are conditional homoscedastic and consider alternative confidence bands, those conditions can be relaxed significantly. When the researcher uses the debiaising technique of Remark 6.2, the role of the estimation error of β and of Λ are exchanged. Section 8.2 in (v5) proposes an alternative solution to handle non identified matrices Λ. Section 8.3 in (v5) combines the confidence bands with an upper bound on the bias obtained from the identification robust confidence sets in case we suspect the "bias" of the debiased estimator might not be negligible.

7. Inference Put Into Practice

7.1. **Simulation Study.** We consider the model (2.1)-(2.2) with i.i.d. data, $\mathcal{B}=\mathbb{R}^{d_X}$ and $S_Q=[d_X]$. We set $\beta^*=(1,-2,-0.5,0.25,0,\dots,0)^\top$ and let $\mathbf{Z}_{i,\cdot}^\top$ be a Gaussian vector in $\mathbb{R}^{d_Z}$ with mean zero and variance I. We take the exogenous regressors to be the first $|S_I|$ IVs. For an endogenous regressor of index k we set $\mathbf{X}_{i,k}=\mathbf{Z}_{i,\cdot}\Pi_{\cdot,k}+\mathbf{V}_{i,k}$ where $\Pi\in\mathcal{M}_{d_Z,|S_I^c|}$ is a matrix of first-stage parameters and $\mathbf{V}_{i,\cdot}\in\mathbb{R}^{|S_I^c|}$. We let $(\mathbf{U}_i,\mathbf{V}_{i,\cdot})^\top$ be a Gaussian vector in $\mathbb{R}^{1+|S_I^c|}$ with mean zero and variance Σ, with

entries

$$\Sigma_{i,j} = 1 \quad \text{if } i = j = 1, \quad \Sigma_{i,j} = 1 - \pi \quad \text{if } i = j > 1, \quad \Sigma_{i,j} = 0.05 \quad \text{otherwise.}$$

We set Π such that $|\Pi_{\cdot,k}|_2^2 = \pi$ for $k \in [|S_I^c|]$, and $\pi \in \{0.5, 0.8\}$ controls the collective IV strength. This implies that each regressor has unit variance. Since the IVs are uncorrelated with one another, a matrix version of the concentration parameter $\Pi^\top \mathbf{Z}^\top \mathbf{Z} \Pi/(1-\pi)$ has diagonal elements close to $n\pi/(1-\pi)$ and the degree of endogeneity is $0.05/\sqrt{1-\pi}$ (see [1]). In low dimensions this would be considered a strong IV situation. However, we consider cases in which the first-stage is not approximately sparse so even a first-stage Lasso estimator would not be consistent (so it is not possible to even estimate the concentration parameter and apply a method akin to 2SLS), most of the IVs have very weak correlation with the endogenous regressors, and identification fails in the absence of sparsity. We take $\Pi_{d_Z,1} = \Pi_{d_Z-1,2} = \cdots = \Pi_{d_Z-|S_I^c|+1,|S_I^c|} = \sqrt{3\pi/4}$. For the remaining entries we set

$$\Pi_{i,j} = \begin{cases} -\sqrt{(\pi/4)/(d_Z-1)} & \text{i is odd and j is odd, or i is even and } j \geq d_Z/2 \\ +\sqrt{(\pi/4)/(d_Z-1)} & \text{otherwise} \end{cases}$$

This means that there is one stronger IV and $d_Z - 1$ weaker IVs for each endogenous regressor. Though each weaker IV accounts for a small fraction of the variance in the endogenous regressors, collectively the weaker IVs account for fraction $\pi/4$ of the variance. Moreover, when $d_Z \leq d_X$ all of the IVs are strongly correlated with at least one regressor.

We construct confidence sets for $\varphi(b) = b$ and the corresponding confidence bands using $\Omega = I$. To construct confidence sets based on SNIV and lower bounds on the sensitivities we use $r_0(n)$ from class 4 with $\alpha = 0.05$ and invoke assumption A.1 to set $\widehat{r} = 1.01 r_0(n)$. We consider sparsity certificates $4, 5, 6, 7, 10$. For the choices of d_X, d_Z below, the design is such that $\mathcal{I}_s$ is a singleton for each sparsity certificate (*i.e.,* there is identification under sparsity). We construct the bounds in (4.9), replacing $c > 0$ with a grid, the construction of which is discussed below. For computational reasons we limit the grid to at most two points. Increasing the number of grid points (and/or loss functions) could lead to less conservative inference. We follow the same approach to construct the confidence set in (4.13) based on an estimated support, taking c equal to the first grid point and $S(\widehat{\beta})$ to be the indices of the elements of $\mathbf{D_X}^{-1}\widehat{\beta}$ with absolute value larger than 1e-4. For the confidence bands we use STIV for the pilot and class 5 to set $\widehat{r}$ (as mandated by Assumption B.1). Since the IVs are uncorrelated with one another, the values of $\widehat{r}$ corresponding to classes 4 and 5 are nearly identical. We use the STIV estimator with $c = 0.99/\widehat{r}$ for the pilot and set $\lambda = 0.99$ to estimate $\widehat{\Lambda}$. Computational details and software are in Section C.4.

Rule of Thumb for c. We start by applying STIV with $c = \widehat{r}^{-1}$, which corresponds to the least shrinkage of $\widehat{\beta}$. As c decreases the estimate remains almost unchanged, until a point after which $\overline{\sigma}$ increases. If the researcher wishes to use a single value of c, we recommend this value. As c decreases further, the estimator remains unchanged until a point after which there is a second increase in $\overline{\sigma}$. If we use a second point to construct a grid we use this value, and so on. Choosing the grid in this way means that we use the smallest possible c (yielding the largest possible sensitivities) for a given value of $\overline{\sigma}$.

Estimation. We consider estimation in a challenging setting with $n < d_Z < d_X$. We set $n = 750$, $d_X = 1750$ and $d_Z = 1500$. We take $S_I^c = \{1, 5, 1503, ..., 1750\}$, such that there are 250 endogenous regressors. Table 1 reports the results for different choices of c and π over 1000 replications. For sufficiently large c, the estimator performs well in terms of selecting nonzero entries of the parameter, and does not select entries of the parameter with values of zero. Due to the shrinkage, there is a

TABLE 1. STIV estimator

$d_Z = 1500, d_X = 1750, n = 750, \pi = 0.8$												
	$c\widehat{r} = 0.95$			$c\widehat{r} = 0.75$			$c\widehat{r} = 0.5$			$c\widehat{r} = 0.25$		
	p2.5	p50	p97.5	p2.5	p50	p97.5	p2.5	p50	p97.5	p2.5	p50	p97.5
$\beta_1^*(=0.5)$	0.8	0.88	0.95	0.74	0.82	0.91	0.67	0.78	0.88	0.27	0.55	0.78
$\beta_2^*(=-2)$	-1.9	-1.83	-1.75	-1.9	-1.83	-1.75	-1.9	-1.82	-1.74	-1.89	-1.81	-1.73
$\beta_3^*(=-0.5)$	-0.41	-0.33	-0.26	-0.41	-0.33	-0.26	-0.41	-0.33	-0.26	-0.41	-0.32	-0.25
$\beta_4^*(=0.25)$	0.01	0.08	0.16	0.01	0.08	0.16	0	0.08	0.16	0	0.08	0.16
$\beta_5^*(=0)$	0	0	0	0	0	0	0	0	0	0	0	0
$\beta_6^*(=0)$	0	0	0	0	0	0	0	0	0	0	0	0
$\sigma^*(=1)$	1	1.05	1.1	1.01	1.06	1.12	1.02	1.07	1.13	1.05	1.12	1.2
$\left\|(\widehat{\beta}-\beta^*)_{S(\beta^*)}\right\|_\infty$	0.15	0.2	0.25	0.16	0.21	0.27	0.17	0.23	0.33	0.23	0.45	0.73
$\left\|(\widehat{\beta}-\beta^*)_{S(\beta^*)^c}\right\|_\infty$	0	0	0.03	0	0	0	0	0	0.03	0	0.17	0.39
$S(\widehat{\beta}) \supseteq S(\beta^*)$	0.98			0.98			0.98			0.96		
$S(\widehat{\beta}) = S(\beta^*)$	0.62			0.95			0.91			0.06		
$d_Z = 1500, d_X = 1750, n = 750, \pi = 0.5$												
$\beta_1^*(=0.5)$	0	0.79	0.96	0	0.78	0.96	0	0.79	0.98	0	0.8	0.98
$\beta_2^*(=-2)$	-1.93	-1.83	-1.5	-1.91	-1.83	-1.48	-1.9	-1.83	-1.47	-1.93	-1.83	-1.49
$\beta_3^*(=-0.5)$	-0.39	-0.32	0	-0.4	-0.34	0	-0.4	-0.34	0	-0.4	-0.33	0
$\beta_4^*(=0.25)$	0	0.08	0.18	0	0.07	0.18	0	0.08	0.16	0	0.07	0.15
$\beta_5^*(=0)$	0	0	0	0	0	0	0	0	0	0	0	0
$\beta_6^*(=0)$	0	0	0	0	0	0	0	0	0	0	0	0
$\sigma^*(=1)$	1	1.09	3.3	0.99	1.09	3.42	0.99	1.08	3.4	0.98	1.08	3.38
$\left\|(\widehat{\beta}-\beta^*)_{S(\beta^*)}\right\|_\infty$	0.15	0.24	1	0.15	0.23	1	0.15	0.23	1	0.15	0.24	1
$\left\|(\widehat{\beta}-\beta^*)_{S(\beta^*)^c}\right\|_\infty$	0	0.02	0.42	0	0.02	0.43	0	0.02	0.45	0	0.02	0.43
$S(\widehat{\beta}) \supseteq S(\beta^*)$	0.82			0.82			0.80			0.78		
$S(\widehat{\beta}) = S(\beta^*)$	0.49			0.45			0.47			0.46		

1000 replications. $r_0(n) = 0.16$.

bias towards zero, which is decreasing in c. The shrinkage also leads to a slight upwards bias in the estimator of the variance.

Confidence Sets. We set $n = 2000$, $d_X = 50$, $d_Z \in \{2050, 49\}$, $S_I^c = \{1, 5\}$ and construct confidence sets for β^*. This design is challenging since there are two endogenous regressors and either $d_Z < d_X$ or $n < d_Z$. We limit d_X so as to permit application of all of our methods to the same design over 1000 replications. Below we modify the design to allow for $d_X > n$. Table 2 illustrates the SNIV confidence sets for a single dataset. The bounds are nested and contain the true value for all values of s. If $\pi = 0.8$ the bounds are sufficiently narrow so as to be informative on the sign of first three entries of β^*, which are the largest in magnitude. If $d_Z = 49$ there is no strong IV for regressor 5, yielding wide bounds. Table 3 reports results for the other confidence sets over 1000 replications. The confidence sets based on a sparsity certificate are are nested. If $d_Z = 2050$, they can be sufficiently narrow so as to be informative on the sign of the first three entries of β^*. Though robust to identification, the sets can also be conservative, with infinite volume if $d_Z = 49$ and coverage close to 1 if $d_Z = 2050$. The STIV estimator performs well in selecting the nonzero entries of the parameter, which results in less conservative confidence sets based on an estimated support. These are narrower than those based on the sparsity certificate $s = |S(\beta^*)| = 4$, even those based on SNIV. This is because they use information on both the cardinality and identities of the relevant regressors. Confidence sets based on estimated support have coverage below the nominal level when $d_Z = 2050$ and $\pi = 0.5$. This is because STIV applied using the rule of thumb value for c sometimes fails to distinguish $\beta_4^* = 0.25$ from zero. In the

TABLE 2. 0.95 *SNIV* confidence sets for β^*

	$d_Z = 49, d_X = 50, n = 2000, \pi = 0.8$								$d_Z = 2050, d_X = 50, n = 2000, \pi = 0.8$							
	Lower bound				Upper bound				Lower bound				Upper bound			
	SC 10	SC 6	SC 5	SC 4	SC 4	SC 5	SC 6	SC 10	SC 10	SC 6	SC 5	SC 4	SC 4	SC 5	SC 6	SC 10
$\beta_1^*(=0.5)$	0.40	0.40	0.59	0.60	1.33	1.48	1.50	1.55	-0.17	0.01	0.07	0.07	1.53	2.00	2.05	2.15
$\beta_2^*(=-2)$	-2.37	-2.34	-2.24	-2.23	-1.76	-1.66	-1.64	-1.61	-2.59	-2.57	-2.46	-2.42	-1.49	-1.33	-1.33	-1.27
$\beta_3^*(=-0.5)$	-0.87	-0.83	-0.80	-0.79	-0.18	-0.16	-0.14	-0.09	-0.93	-0.83	-0.83	-0.89	-0.05	-0.05	0.01	0.09
$\beta_4^*(=0.25)$	-0.28	-0.21	-0.19	-0.07	0.51	0.66	0.70	0.75	-0.42	-0.36	-0.36	-0.13	0.49	0.76	0.78	0.86
$\beta_5^*(=0)$	-3.59	-3.59	-3.54	-3.42	3.51	3.51	3.57	3.71	-0.53	-0.47	-0.46	-0.42	0.42	0.42	0.48	0.68
$\beta_6^*(=0)$	-0.34	-0.29	-0.27	-0.28	0.29	0.30	0.32	0.37	-0.54	-0.34	-0.34	-0.34	0.35	0.36	0.44	0.54
	$d_Z = 49, d_X = 50, n = 2000, \pi = 0.5$								$d_Z = 2050, d_X = 50, n = 2000, \pi = 0.5$							
$\beta_1^*(=0.5)$	-0.30	-0.15	-0.12	-0.09	1.94	1.98	2.09	2.21	-1.10	-0.92	-0.82	-0.64	2.64	2.70	2.91	3.05
$\beta_2^*(=-2)$	-2.53	-2.49	-2.48	-2.42	-1.62	-1.54	-1.42	-1.40	-2.92	-2.88	-2.84	-2.72	-0.99	-0.91	-0.80	-0.50
$\beta_3^*(=-0.5)$	-1.12	-1.05	-1.05	-0.98	0.01	0.06	0.08	0.18	-1.47	-1.36	-1.28	-1.24	0.25	0.34	0.38	0.55
$\beta_4^*(=0.25)$	-0.49	-0.42	-0.33	-0.29	0.74	0.75	0.82	0.93	-0.69	-0.61	-0.54	-0.52	0.87	0.91	0.98	1.08
$\beta_5^*(=0)$	-6.22	-6.00	-6.00	-5.73	5.99	6.64	6.46	6.76	-1.50	-1.36	-1.26	-1.17	1.14	1.22	1.42	1.59
$\beta_6^*(=0)$	-0.76	-0.63	-0.57	-0.54	0.47	0.47	0.53	0.69	-0.93	-0.83	-0.75	-0.74	0.61	0.69	0.74	0.90

1 replication. For $d_Z = 49$ (resp. 2050) $r_0(n) = 0.074$ (resp. 0.094). 'SC s' use sparsity certificate s.

other designs, confidence sets using an estimated support are typically sufficiently narrow so as to be informative on the signs of the first three entries of β^*. The bias corrected STIV reduces the shrinkage of STIV and centers its distribution on the true value. For $d_Z = 2050$, there exists a sparse Λ verifying (6.1), with $|S_I| + |S_I^c| d_Z = 4148$ nonzero entries out of a possible $d_X d_Z = 102500$. The coverage of the confidence bands is slightly below 0.95, but the bands are narrower than those based on sensitivities. Coverage below 0.95 stems from the shrinkage applied when estimating Λ. For $d_Z = 49$, there does not exist Λ verifying (6.1). This leads the coverage to be below 0.95, significantly so when $\pi = 0.5$.

Confidence Sets with $d_X > n$. We set $n = 4000$, $d_X = 4100$, $S_I^c = \{1, 5\}$ and $d_Z = 4100$. Confidence bands are computationally infeasible since $\widehat{\Lambda}$ is the solution of a second order cone program with $d_X d_Z = 4100^2$ cones. Table 4 reports the results. If $\pi = 0.8$, confidence sets with a sufficiently small sparsity certificate can be informative on the signs. For sparsity certificate $s = 7$, the confidence set has infinite volume when one grid point over c is used and finite (though large) volume with two grid points. STIV performs well in terms of variable selection, which translates into less conservative confidence sets based on estimated support. Reducing the strength of the IVs ($\pi = 0.5$) increases the width of the sets but does not yield coverage below the 0.95 level.

Endogenous IVs: Fewer Known Exogenous IVs than Regressors. We take $n = 2000$, $d_X = 6$, $S_I^c = \{1, 5\}$ and $d_Z = 50$. The 45 IVs with indices $S_\perp^c = \{5, 6, ..., 49\}$ are possibly endogenous. We modify the design above by taking $\mathbf{Z}_{i,5} = \sqrt{1 - 0.8^2} e_i + 0.8 \mathbf{U}_i$ with e_i is an independent standard Gaussian. This preserves the variance matrix of $\mathbf{Z}_{i,\cdot}$ as the identity but implies that $\mathbb{E}[\mathbf{Z}_{i,\cdot}^\top \mathbf{U}(\beta^*)_i] = \theta^*$ has one nonzero entry given by $\theta_5^* = 0.8$. We construct SNIV and sensitivity based confidence sets. The sensitivity based confidence sets are based on the C-STIV estimator. Table 5 illustrates the SNIV confidence sets for a single dataset. The confidence sets based on a sparsity certificate fix $s = 4$ for the sparsity of β and vary $\widetilde{s} \in [5]$ for the sparsity of θ. The bounds for the endogenous IV do not include zero, and so SNIV detects it for every sparsity certificate. In contrast, the bounds on the exogenous instrument include zero for all sparsity certificates. Table 6 reports results for the other confidence sets over 1000 replications. The C-STIV estimator detects the endogenous IV, though is downwards biased due to the shrinkage. The confidence sets using a sparsity certificate correctly detect the endogenous IV with frequency 0.74 for $\widetilde{s} = 1$, which decreases as $\widetilde{s}$ increases. The confidence sets based on estimated

TABLE 3. 0.95 confidence sets and bands

$d_Z = 2050, d_X = 50, n = 2000, \pi = 0.8$												
	STIV			SC 4	SC 5	SC 6	SC 10	ES	BC STIV			CB
	p2.5	p50	p97.5	Median width/2					p2.5	p50	p97.5	Width/2
$\beta_1^*(=1)$	0.9	0.95	0.99	0.8	1.02	1.32	6.07	0.33	0.94	1	1.06	0.1
$\beta_2^*(=-2)$	-1.95	-1.9	-1.85	0.58	0.73	0.94	4.55	0.26	-2.04	-1.99	-1.95	0.07
$\beta_3^*(=-0.5)$	-0.45	-0.4	-0.36	0.57	0.73	0.94	4.64	0.26	-0.54	-0.49	-0.45	0.07
$\beta_4^*(=0.25)$	0.11	0.15	0.19	0.57	0.73	0.95	4.65	0.26	0.2	0.24	0.29	0.07
$\beta_5^*(=0)$	0	0	0	0.8	1.02	1.31	6.03	0	-0.05	0	0.06	0.1
$\beta_6^*(=0)$	0	0	0	0.57	0.73	0.95	4.62	0	-0.04	0	0.04	0.07
$S(\widehat{\beta}) \supseteq S(\beta^*)$	1		Cover	1	1	1	1	0.98				0.94
$S(\widehat{\beta}) = S(\beta^*)$	0.98			(0.996,1)	(0.996,1)	(0.996,1)	(0.996,1)	(0.97,0.98)				(0.92,0.95)
$d_Z = 49, d_X = 50, n = 2000, \pi = 0.8$												
$\beta_1^*(=1)$	0.84	0.9	0.96	∞	∞	∞	∞	0.24	0.94	0.99	1.05	0.07
$\beta_2^*(=-2)$	-1.96	-1.91	-1.87	∞	∞	∞	∞	0.2	-2.04	-2	-1.95	0.07
$\beta_3^*(=-0.5)$	-0.47	-0.43	-0.39	∞	∞	∞	∞	0.2	-0.54	-0.5	-0.46	0.07
$\beta_4^*(=0.25)$	0.13	0.18	0.23	∞	∞	∞	∞	0.2	0.2	0.25	0.3	0.07
$\beta_5^*(=0)$	0	0	0	∞	∞	∞	∞	0	-0.03	0.02	0.08	0.1
$\beta_6^*(=0)$	0	0	0	∞	∞	∞	∞	0	-0.04	0	0.04	0.07
$S(\widehat{\beta}) \supseteq S(\beta^*)$	1		Cover	1	1	1	1	1				0.93
$S(\widehat{\beta}) = S(\beta^*)$	0.96			(0.996,1)	(0.996,1)	(0.996,1)	(0.996,1)	(0.996,1)				(0.91,0.95)
$d_Z = 2050, d_X = 50, n = 2000, \pi = 0.5$												
$\beta_1^*(=1)$	0.98	1.02	1.07	1.55	2.31	3.78	∞	0.43	0.92	1	1.07	0.12
$\beta_2^*(=-2)$	-1.95	-1.9	-1.86	0.85	1.26	2.08	∞	0.27	-2.04	-1.99	-1.95	0.07
$\beta_3^*(=-0.5)$	-0.45	-0.4	-0.36	0.86	1.27	2.1	∞	0.27	-0.54	-0.49	-0.45	0.07
$\beta_4^*(=0.25)$	0.11	0.16	0.2	0.85	1.26	2.08	∞	0.26	0.2	0.24	0.29	0.07
$\beta_5^*(=0)$	0	0.03	0.07	1.56	2.32	3.78	∞	0	-0.07	0	0.07	0.12
$\beta_6^*(=0)$	0	0	0	0.85	1.26	2.1	∞	0	-0.04	0	0.04	0.07
$S(\widehat{\beta}) \supseteq S(\beta^*)$	1		Cover	1	1	1	1	0.75				0.95
$S(\widehat{\beta}) = S(\beta^*)$	0.13			(0.996,1)	(0.996,1)	(0.996,1)	(0.996,1)	(0.72,0.77)				(0.93,0.96)
$d_Z = 49, d_X = 50, n = 2000, \pi = 0.5$												
$\beta_1^*(=1)$	1	1.05	1.09	∞	∞	∞	∞	0.31	0.99	1.03	1.08	0.03
$\beta_2^*(=-2)$	-1.97	-1.93	-1.88	∞	∞	∞	∞	0.2	-2.04	-1.99	-1.95	0.07
$\beta_3^*(=-0.5)$	-0.47	-0.42	-0.38	∞	∞	∞	∞	0.2	-0.54	-0.49	-0.45	0.07
$\beta_4^*(=0.25)$	0.13	0.18	0.22	∞	∞	∞	∞	0.2	0.2	0.25	0.29	0.07
$\beta_5^*(=0)$	0	0.05	0.09	∞	∞	∞	∞	0	-0.03	0.03	0.09	0.1
$\beta_6^*(=0)$	0	0	0	∞	∞	∞	∞	0	-0.04	0	0.04	0.07
$S(\widehat{\beta}) \supseteq S(\beta^*)$	1		Cover	1	1	1	1	1				0.50
$S(\widehat{\beta}) = S(\beta^*)$	0.02			(0.996,1)	(0.996,1)	(0.996,1)	(0.996,1)	(0.996,1)				(0.47,0.53)

1000 replications. 'SC s' use sparsity certificate s. 'ES' use estimated support. 'CB' use $\Omega = I$. SC/ES use one grid point for c. For $d_Z = 49$ (resp. 2050) $r_0(n) = 0.074$ (resp. 0.094). 'STIV' uses $c = 0.99/\widehat{r}$. 'BC STIV' is the bias corrected STIV. For SC/ES (resp. CB) 'Cover' reports the frequency with which β^* lies in the bounds defined in (4.9) (resp. (B.20)). 0.95 confidence intervals for the coverage are in parentheses (see [34]).

support detect the endogenous IV in every replication.

Endogenous IVs: As Many Known Exogenous IVs as Regressors. We take $n = 3000$, $d_X = 90$, $S_I^c = \{1, 5\}$ and $d_Z = 100$. There are 10 possibly endogenous IVs with indices $S_\perp^c = \{89, 90, ..., 98\}$. There is one endogenous IV, $\mathbf{Z}_{i,89} = \sqrt{1 - 0.8^2} e_i + 0.8\mathbf{U}_i$ where e_i is an independent standard Gaussian. We apply the NV-STIV estimator. We use STIV for the pilot, taking $r_0(n)$ from class 4 with $\alpha = 0.025$ and $\widehat{r}_1(n) = 1.01 r_0(n)$. We choose one value of c using the rule of thumb above. For the NV-STIV estimator we take $r_2(n)$ from class 4 with $\alpha = 0.025$. As both stages use $\alpha = 0.025$, we construct confidence sets with prescribed coverage probability 0.95. We use sparsity certificates $s = 4$ for β and $\widetilde{s} \in [10]$ for θ. The confidence sets are intersected over a grid of 19 points for $\widetilde{c}$. Table

TABLE 4. 0.95 confidence sets with $d_X > n$

$d_Z = 4100, d_X = 4100, n = 4000, \pi = 0.8$										
	STIV			SC 4	SC 5	SC 6	SC 7	SC* 7	SC 10	ES
	p2.5	p50	p97.5	Median width/2						
$\beta_1^*(=1)$	0.87	0.91	0.94	0.65	1.08	2.66	$\infty^\dagger$	97.7	∞	0.22
$\beta_2^*(=-2)$	-1.96	-1.93	-1.90	0.40	0.59	1.29	$\infty^\dagger$	42.84	∞	0.17
$\beta_3^*(=-0.5)$	-0.46	-0.43	-0.39	0.40	0.60	1.29	$\infty^\dagger$	42.51	∞	0.17
$\beta_4^*(=0.25)$	0.14	0.18	0.21	0.40	0.59	1.29	$\infty^\dagger$	41.89	∞	0.17
$\beta_5^*(=0)$	0	0	0	0.65	1.09	2.68	$\infty^\dagger$	97.37	∞	0
$\beta_6^*(=0)$	0	0	0	0.40	0.60	1.29	$\infty^\dagger$	43.32	∞	0
$S(\widehat{\beta}) \supseteq S(\beta^*)$	1		Cover	1	1	1	1	1	1	0.97
$S(\widehat{\beta}) = S(\beta^*)$	0.99			(0.98,1)	(0.98,1)	(0.98,1)	(0.98,1)	(0.98,1)	(0.98,1)	(0.94,0.99)
$d_Z = 4100, d_X = 4100, n = 4000, \pi = 0.5$										
$\beta_1^*(=1)$	1.02	1.05	1.08	3.52	18.64	∞	∞	∞	∞	0.65
$\beta_2^*(=-2)$	-1.96	-1.93	-1.90	1.56	7.02	∞	∞	∞	∞	0.4
$\beta_3^*(=-0.5)$	-0.46	-0.43	-0.4	1.57	6.97	∞	∞	∞	∞	0.4
$\beta_4^*(=0.25)$	0.15	0.18	0.21	1.55	6.99	∞	∞	∞	∞	0.4
$\beta_5^*(=0)$	0.02	0.05	0.08	3.58	19.18	∞	∞	∞	∞	0
$\beta_6^*(=0)$	0	0	0	1.57	7.05	∞	∞	∞	∞	0
$S(\widehat{\beta}) \supseteq S(\beta^*)$	1		Cover	1	1	1	1	1	1	0.97
$S(\widehat{\beta}) = S(\beta^*)$	0.01			(0.98,1)	(0.98,1)	(0.98,1)	(0.98,1)	(0.98,1)	(0.98,1)	(0.94,0.99)

200 replications. 'SC s' use sparsity certificate s. 'ES' use estimated support. 'CB' use $\Omega = I$. SC/ES use one grid point for c. $r_0(n) = 0.07$.'STIV' uses $c = 0.99/\widehat{r}$. 'Cover' reports the frequency with which β^* lies in the bounds defined in (4.9). †: The frequency of replications with confidence sets of finite width is 0.03. *: Where they are not identical, we report confidence sets using two grid points.

TABLE 5. 0.95 *SNIV* confidence sets for detection of endogenous IVs

$d_Z = 50, d_X = 6, n = 2000, \pi = 0.8$										
	Lower bound					Upper bound				
	SC 4,5	SC 4,4	SC 4,3	SC 4,2	SC 4,1	SC 4,1	SC 4,2	SC 4,3	SC 4,4	SC 4,5
$\theta_5^*(=0.8)$	0.02	0.02	0.06	0.06	0.07	1.31	1.31	1.31	1.36	1.36
$\theta_6^*(=0)$	-1.16	-1.15	-1.12	-1.07	-1.01	0.61	0.61	0.64	0.63	0.68

1 replication. $r_0(n) = 0.07$. 'SC $s, \widetilde{s}$' use sparsity certificates $s, \widetilde{s}$.

TABLE 6. 0.95 C-STIV confidence sets for detection of endogenous IVs

$d_Z = 50, d_X = 6, n = 2000, \pi = 0.8$									
	C-STIV			SC 4,1	SC 4,2	SC 4,3	SC 4,4	SC 4,5	ES
	p2.5	p50	p97.5	Median width/2					
$\theta_5^*(=0.8)$	0.66	0.71	0.77	0.69	0.71	0.73	0.76	0.78	0.25
$\theta_6^*(=0)$	0	0	0	0.69	0.71	0.74	0.76	0.78	0
$S(\widehat{\theta}) \supseteq S(\theta^*)$	1		Power	0.74	0.51	0.26	0.11	0.04	1
$S(\widehat{\theta}) = S(\theta)$	1			(0.71,0.77)	(0.48,0.54)	(0.23,0.29)	(0.09,0.13)	(0.03,0.05)	(0.996,1)

1000 replications. 'SC $s, \widetilde{s}$' use sparsity certificates $s, \widetilde{s}$. 'ES' use estimated support. SC/ES use one grid point for c. $r_0(n) = 0.07$. 'C-STIV' uses $c = 0.99$. 'Power' is the frequency with which the confidence sets do not include $\theta_5 = 0$.

7 reports results. Due to shrinkage, the NV-STIV estimator is centred on 0.6. The endogenous IV is detected with frequency 0.95 for $\widetilde{s} = 1$, decreasing to 0.91 for $\widetilde{s} = 10$.

7.2. **Second-order Approximation of the EASI Demand System.** The EASI demand system of [25] implies the vector of expenditure shares $\mathbf{S}_i \in \mathbb{R}^{d_G}$ for d_G goods consumed by household i satisfies

$$\mathbf{S}_i = \sum_{p=0}^{d_P} b_p \mathbf{T}_i^p + C\mathbf{H}_{i,\cdot}^\top + D\mathbf{H}_{i,\cdot}^\top \mathbf{T}_i + A_0 \mathbf{P}_{i,\cdot}^\top + \sum_{h=1}^{d_H} A_h \mathbf{P}_{i,\cdot}^\top \mathbf{H}_{i,h} + B\mathbf{P}_{i,\cdot}^\top \mathbf{T}_i + \mathbf{W}_i; \tag{7.1}$$

TABLE 7. 0.95 NV-STIV confidence sets for detection of endogenous IVs

$d_Z = 100, d_X = 90, n = 3000, \pi = 0.8$									
	NV-STIV			SC 4,1	SC 4,2	SC 4,3	SC 4,5	SC 4,7	SC 4,10
	p2.5	p50	p97.5	Median width/2					
$\theta_{89}(=0.8)$	0.55	0.6	0.65	0.53	0.53	0.53	0.53	0.54	0.54
$\theta_{90}(=0)$	0	0	0	0.53	0.53	0.53	0.53	0.54	0.54
$S(\widehat{\theta}) \supseteq S(\theta^*)$	1		Power	0.95	0.94	0.94	0.93	0.92	0.91
$S(\widehat{\theta}) = S(\theta^*)$	1			(0.93,0.96)	(0.92,0.95)	(0.92,0.95)	(0.91,0.94)	(0.9,0.93)	(0.89,0.92)

1000 replications. 'SC $s, \widetilde{s}$' use sparsity certificates $s, \widetilde{s}$. SC use one grid point for c. $r_0(n) = 0.07$ is from class 4 with $\alpha = 0.025$, and $\widehat{r}_1(n) = 1.01 r_0(n)$. $r_2(n) = 0.06$ is from class 4 with $\alpha = 0.025$. Confidence sets use a grid of 19 points for $\widetilde{c}$. 'NV-STIV' uses $\widetilde{c} = 0.99$. 'Power' reports the frequency with which the confidence sets do not include $\theta_{89} = 0$.

$$\mathbf{T}_i = \frac{\mathbf{E}_i - \mathbf{P}_{i,\cdot}\mathbf{S}_i + \mathbf{P}_{i,\cdot}(A_0 + \sum_{h=1}^{d_H} A_h \mathbf{H}_{i,h})\mathbf{P}_{i,\cdot}^\top/2}{1 - \mathbf{P}_{i,\cdot} B \mathbf{P}_{i,\cdot}^\top/2}; \tag{7.2}$$

where $\mathbf{E}_i$ is nominal expenditure, $\mathbf{P}_{i,\cdot}$ is a vector of log-prices of size d_G, $\mathbf{H}_{i,\cdot}^\top$ is a vector of household characteristics of size d_H, $\mathbf{W}_i$ is a vector of errors, and $\mathbf{T}_i$ is deflated expenditure. The parameters are $b_p \in \mathbb{R}^{d_G}$ for $p = 0, ..., d_P$, $C, D \in \mathcal{M}_{d_G, d_P}$ and $A_0, ..., A_5, B \in \mathcal{M}_{d_G, d_G}$. Theory imposes restrictions such as (1) expenditure shares sum to one and (2) Slutsky symmetry, hence

$$1^\top b_0 = 1,\ 1^\top C = 1^\top D = 0,\ 1^\top B = 0;\ \forall p \in [d_P],\ 1^\top b_p = 0;$$
$$\forall h \in [d_H],\ 1^\top A_h = 0,\ A_h = A_h^\top;\ B = B^\top;$$
$$A_1, \ldots, A_{d_H} \text{ and } B \text{ are symmetric.}$$

Since $\mathbf{T}_i$ depends on the parameters, the system is nonlinear, and hence cumbersome to estimate. [25] proposes an approximate EASI system in which $\mathbf{T}_i$ is replaced by its first-order in prices approximation $\mathbf{D}_i = \mathbf{E}_i - \mathbf{P}_{i,\cdot}\mathbf{S}_i$. This corresponds to deflating nominal expenditure with a Stone price index. Prices are normalized for a subset of the sample, implying log prices of 0. To reduce approximation error, we consider a second-order approximation. We use (7.2) to obtain, for all $p \in \mathbb{N}$,

$$\mathbf{T}_i^p = \mathbf{D}_i^{p-1}\left(\mathbf{D}_i + \frac{p-1}{2}\mathbf{P}_{i,\cdot}\left(A_0 + \sum_{h=1}^{d_H} A_h \mathbf{H}_{i,h} + B\mathbf{D}_i\right)\mathbf{P}_{i,\cdot}^\top\right) + O(|\mathbf{P}_{i,\cdot}|_2^4) \tag{7.3}$$

and inject (7.3) into (7.1) to obtain a second-order approximation of the expenditure share equation. An approximation error arises due to the second term in (7.3). Due to the normalization, the approximation error is small. Since (7.3) depends on parameters, the second-order approximation depends on products of the parameters, violating linearity. Where this arises, we replace the product with a new parameter and use the parameter restrictions above to obtain restrictions involving the new variables.

We use the Canadian data of [25] for $n = 4847$ rental-tenure single-member households that had positive expenditures on rent recreation, and transportation. The categories of the $d_G = 9$ goods are: (1) food consumed at home, (2) food consumed out of the home, (3) rent, (4) clothing, (5) household operation, (6) household furnishing/equipment, (7) transportation operation, (8) recreation, and (9) personal care. The individual characteristics comprise: (1) age minus 40, (2) gender, (3) a dummy for car non-ownership equal to one if real gasoline expenditures (at 1986 gasoline prices) are less than \$50, (4) a social assistance dummy equal to one if government transfers are greater than 10 percent of gross income, and (5) a time trend equal to the calendar year minus 1986 (that is, equal to zero in 1986). Following [25], we use $d_P = 5$ for the degree of the polynomial in deflated expenditure. Each equation in the second-order approximation has $d_X = 1879$ parameters. It is reasonable to expect that the parameter vector be sparse, particularly for the second-order approximation terms. Prices are

normalized to be 1 for residents of Ontario in 1986, implying log prices of 0. The approximation error from the second order approximation is likely small because $n^{-1}\sum_{i\in[n]}|\mathbf{P}_{\cdot,i}|_2^4 = 0.0008$. In contrast, $n^{-1}\sum_{i\in[n]}|\mathbf{P}_{\cdot,i}|_2^2 = 0.0268$, and the mean budget share for five of the goods is less than 0.1, suggesting a sizeable first order approximation error. This means that, in order to obtain reliable results, a researcher using the first order approximation might need to use a subsample for which the approximation error is not large relative to the budget shares. Since $\mathbf{D}_i = \mathbf{E}_i - \mathbf{P}_{i,\cdot}\mathbf{S}_i$ depends on $\mathbf{W}_i$, every regressor which involves $\mathbf{D}_i$ is endogenous. This implies that 1819 of the 1879 regressors are endogenous. We construct $d_Z = d_X$ IVs by replacing $\mathbf{D}_i$ with $\overline{\mathbf{D}}_i = \mathbf{E}_i - \mathbf{P}_{i,\cdot}\mathbb{E}_n[S]$.

The IVs are strong and $d_Z = d_X$ and so we apply Section 6. We construct uniform 0.9 confidence bands for the Engel curves based on $d_F = 11$ grid points. We use 0.9 for comparability with [25]. In the first step, we apply the *SE-STIV* estimator, adjusting $\widehat{r}$ according to class 5 using $\alpha = 0.05/d_G$, taking $c = 0.99/\widehat{r}$ and $\widehat{v}_g = 1/n$ for all $g \in [d_G]$. We choose S_Q so as to exempt the constant, linear, and quadratic parts of the Engel curves (b_0, b_1, b_2) and the linear price parameters (A_0) from the ℓ_1 penalty, since these form a parsimonious baseline specification for the demand system (see [2]). The SE-STIV estimator permits unrestricted cross-equation correlation in the entries of $\mathbf{W}_i$. For brevity, we do not present the full estimation results, focussing instead on the Engel curves. There are $1879d_G = 16911$ parameters in the system. Among the $1771d_G = 15399$ parameters in S_Q, only 50 are estimated as nonzero, 22 of which are parameters which arise due to the second-order approximation. In the second step we apply the C-STIV estimator in (6.3) using $r_0'(n)$ from class 4 with $\alpha = 0.05$ and $\lambda = 0.99$. Figure 1 depicts the preliminary estimator of the Engel curve for transportation operation, its bias corrected counterpart and 90% confidence bands. The second-order approximation yields a different curve to that of [25], which is downwards sloping and close to linear. The slope is negative and the bias correction large, changing the shape from an inverted U to decreasing and nonlinear. The preliminary estimator lies outside the confidence band, which is wider close to the end points. This is most likely because there is less data at the end points, and is true for all of the goods. Appendix D depicts the Engel curves for the other goods. The curves for food are close to linear and have the expected slopes: negative for food-in and positive for food-out. The bias correction is large for rent, household operation, clothing, and personal care, for which the preliminary estimator does not lie within the confidence bands. The Engel curves are similar to those of [25] apart from household and transportation operation. The confidence bands are marginally wider. This is expected since we construct uniform bands rather than pointwise intervals.

References

[1] Andrews, D. W. K. and Stock, J. H. (2007). Inference with weak instruments. *Advances in Economics and Econometrics Theory and Applications, Ninth World Congress*, Blundell, R., W. K. Newey, and T. Persson, Eds, **3**, 122–174, Cambridge University Press.

[2] Banks, J., Blundell, R. and Lewbel, A. (1997). Quadratic Engel curves and consumer demand. *The Review of Economics and Statistics* **79** 527–539.

[3] Barrenho, E., Gautier, E., Miraldo, M., Propper,C., and Rose, C. (2019). Peer and network effects in medical innovation: the case of laparoscopic surgery in the English NHS. *HEDG Working Papers* **19** 650–659.

[4] Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. (2012). Sparse models and methods for optimal IVs with an application to eminent domain. *Econometrica* **80** 2369–2429.

[5] Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806.

[6] Belloni, A., Chernozhukov, V., and Hansen, C. (2014): Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* **81** 608–650.

[7] Belloni, A., Chernozhukov, V., Hansen, C., and Newey, W. (2018): High-dimensional linear models with many endogenous variables. Preprint 1712.08102.

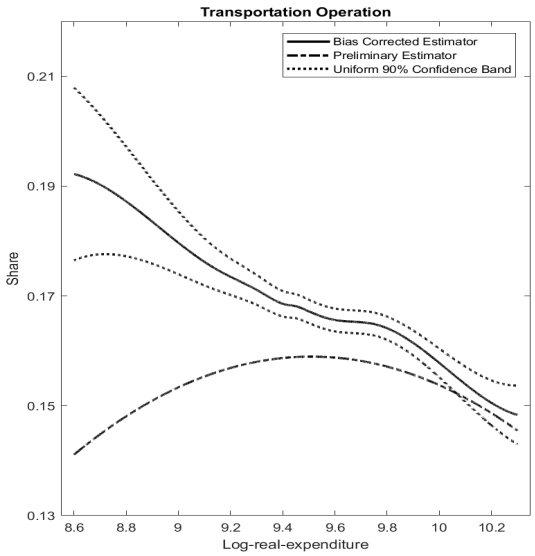


FIGURE 1. Engel Curve for Transportation Operation

[8] BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D., HANSEN, C., and KATO, K. (2018): High-dimensional econometrics and regularized GMM. Preprint 1806.00666.

[9] BICKEL, P., RITOV, Y., and TSYBAKOV, A. (2009). Simultaneous analysis of lasso and Dantzig selector. *Annals of Statistics* **37** 1705–1732.

[10] BREUNIG, C., SIMONI, A., and MAMMEN, E. (2018). Ill-posed estimation in high-dimensional models with instrumental variables. Preprint 1806.00666.

[11] CAI, T., LIU, W. and LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106** 594–607.

[12] CANER, M. (2009). Lasso type GMM estimator. *Econometric Theory* **25** 1–23.

[13] CANER, M. and ZHANG, H. (2014). Adaptive elastic net for generalized methods of moments . *Journal of Business and Economics Statistics* **32** 30–47.

[14] CANNER, M. and KOCK, A. B. (2019). High dimensional GMM. Preprint 1811.08779.

[15] CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W., and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Economics Journal* **21** C1–C68.

[16] CANDÈS, E, and TAO, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics* **35** 2313–2351.

[17] DUFOUR, J.-M. (1997). Impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica* **65** 1365–1387.

[18] FAN, J. and LIAO, Y. (2014). Endogeneity in high dimensions. *Annals of Statistics* **42** 872–917.

[19] GAUTIER, E. and ROSE, C. (2017). Inference on social effects when the network is sparse and unknown. Working paper.

[20] GOLD, D., LEDERER, J., and TAO, J. (2017). Inference for high-dimensional instrumental variables regression. Preprint 1708.05499.

[21] JAVANMARD, A. and MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research* **15** 2869–2909.

[22] KANG, H., ZHANG, A., CAI, T. and SMALL, D. (2016). Instrumental variables estimation with some invalid IVs and its application to mendelian randomization. *Journal of the American Statistical Association* **111** 132–144.

[23] KOLESÁR, M., CHETTY, R., FIEDMAN, J., GLASESER, E., and IMBENS, G. (2015). Identification and inference with many invalid IVs. *Journal of Business & Economic Statistics* **33** 474–484.

[24] LASSERRE, J.-B. (2015). *An Introduction to Polynomial and Semi-Algebraic Optimization*. Cambridge.

[25] LEWBEL, A. and PENDAKUR, K. (2009). Tricks with hicks: the EASI demand system. *American Economic Review* **99** 827–863.

[26] Lounici, K. (2008): Sup-norm convergence rate and sign concentration property of the lasso and Dantzig selector. *Electronic Journal of Statistics* **2** 90–102.
[27] Owen, A. B.. A robust hybrid of lasso and ridge regression. *Contemporary Mathematics* **443** 59–72.
[28] Rose, C. (2016): Identification of spillover effects using panel data. Job Market Paper.
[29] Sargan, J. (1958). The estimation of economic relationships using instrumental variables. *Econometrica* **26** 393–415.
[30] Sun, T. and Zhang, C.-H.. Scaled sparse linear regression. *Biometrika* **99** 879–898.
[31] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* **58** 267–288.
[32] Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics* **7** 1456–1490.
[33] van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models.. *Annals of Statistics* **42** 1166–1202.
[34] Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* **22** 209–212.
[35] Weisser, T., Lasserre, J.B., and Toh, K.C. (2018). Sparse-BSOS: A bounded degree SOS hierarchy for large scale polynomial optimization with sparsity. *Mathematical Programming Computation* **10** 1–32.
[36] Ye, F. and Zhang, C.-H. (2010). Rate minimaxity of the lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research* **11** 3519–3540.
[37] Zhang, C.-H. and Zhang, S. (2014). Confidence intervals for low dimensional parameters in high-dimensional linear models. *Journal of the Royal Statistical Society: Series B* **76** 217–242.
[38] Zhu, Y. (2018). Sparse linear models and ℓ_1 regularized 2SLS with high-dimensional endogenous regressors and IVs. *Journal of Econometrics* **202** 196–213.
[39] Zivot, E., Startz, R. and Nelson, C.R. (1998). Valid confidence intervals and inference in the presence of weak instruments *International Economic Review* 1119–1144.

Toulouse School of Economics, 21 allée de Brienne, 31000 Toulouse, France.

E-mail address: eric.gautier@tse-fr.edu,

School of Economics, University of Queensland, St Lucia, Brisbane, Australia, 4072.

E-mail address: christiern.rose@uq.edu.au,

SUPPLEMENTAL APPENDIX FOR "HIGH-DIMENSIONAL INSTRUMENTAL VARIABLES AND CONFIDENCE SETS"

In this appendix $C_{\mathrm{N}}(m) \triangleq e(2\ln(m)-1)$ for $m \geq 3$ is the Nemirovski constant (see, Theorem 2.2 in [5]).

Appendix A: Proofs and Complements to the Main Text

A.1. **Possible Classes $\mathcal{P}$.** We present 5 baseline classes $\mathcal{P}$ which we further restrict when need be. For example, some confidence sets require very mild assumptions on $\mathcal{P}$ while rates of convergence require working within subsets of these classes.

A.1.1. *Baseline Classes.* The baseline classes 1-5 are identification robust because they do not restrict the joint distribution of $(\mathbf{Z}, \mathbf{X})$. Classes 1-4 are concerned with independent data but there are also results on moderate deviations for self-normalized sums in the case of dependent data which could be used as well (see, *e.g.*, (v5)). Let us start by presenting classes 1-4. They are the classes under which we work with $\mathcal{G}_0$. The choices of $r_0(n)$ below are based on Theorems A1-A.4 in (v5). Φ is the standard normal CDF.

Class 1: $(\mathbf{Z}_{i,\cdot}^\top \mathbf{U}_i(\beta))_{i\in[n]}$ *are i.i.d. and symmetric and* $d_Z < 9\alpha/\left(4e^3\Phi\left(-\sqrt{n}\right)\right)$*;*

$$r_0(n) = -\frac{1}{\sqrt{n}}\Phi^{-1}\left(\frac{9\alpha}{4d_Z e^3}\right).$$

Class 2: $(\mathbf{Z}_{i,\cdot}^\top \mathbf{U}_i(\beta))_{i\in[n]}$ *are i.i.d., there exists* $\gamma_4 > 0$ *such that* $\max_{l\in[d_Z]} \mathbb{E}[(\mathbf{Z}_{i,l}\mathbf{U}_i)^4](\mathbb{E}[(\mathbf{Z}_{i,l}\mathbf{U}_i)^2])^{-2} \leq \gamma_4$*, and* $d_Z < \alpha\exp\left(n/\gamma_4\right)/(2e+1)$*;*

$$r_0(n) = \sqrt{\frac{2\ln(d_Z(2e+1)/\alpha)}{n - \gamma_4\ln(d_Z(2e+1)/\alpha)}}.$$

Class 3: $(\mathbf{U}_i(\beta))_{i\in[n]}$ *are independent and symmetric conditional on* $\mathbf{Z}$ *or* $(\mathbf{Z}_{i,\cdot}^\top \mathbf{U}_i(\beta))_{i\in[n]}$ *are independent and symmetric;*

$$r_0(n) = \sqrt{\frac{2\ln(2d_Z/\alpha)}{n}}.$$

Class 4: $(\mathbf{Z}_{i,\cdot}^\top \mathbf{U}_i(\beta))_{i\in[n]}$ *are independent, there exists* δ *in* $(0,1]$ *and* $\gamma_{2+\delta} > 0$ *such that*

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

and $d_Z \leq \alpha/\left(2\Phi\left(-n^{1/2-1/(2+\delta)}\gamma_{2+\delta}^{-1/(2+\delta)}\right)\right)$*;*

$$r_0(n) = -\frac{1}{\sqrt{n}}\Phi^{-1}\left(\frac{\alpha}{2d_Z}\right).$$

For classes 1-3 the coverage error $\alpha_B(n)$ is 0. Class 4 yields an asymptotic guarantee with finite sample error $\alpha_B(n) \triangleq \alpha C_1\gamma_{2+\delta}\left(1+\sqrt{n}r_0(n)\right)^{2+\delta} n^{-\delta/2}$, where C_1 is a universal constant. Classes 1 and 3 rely on symmetry. Class 2 relaxes symmetry but requires fourth moments and the upper bound γ_4. When $n - \gamma_4\ln(d_Z(2e+1)/\alpha) \geq n/2$ one can take $r_0(n) = 2\sqrt{\ln(d_Z(2e+1)/\alpha)/n}$. Class 3 allows for dependence in the matrix $\mathbf{Z}$.

For the event $\mathcal{G}$, we can work with classes 1-4. This requires a suitable modification of $r_0(n)$ to obtain $\widehat{r}$. Taking $\widehat{r} = r_0(n)\left|\mathbf{D}_{\mathbf{Z}}\mathbf{Z}^\top\right|_\infty$ yields $\mathcal{G}_0 \subseteq \mathcal{G}$. The following developments are motivated by the fact that choice of $\widehat{r}$ can result in conservative inference.

We can make use of concentration arguments. For example, define the events

$$\mathcal{E}_U \triangleq \left\{\left|(\mathbb{E}_n - \mathbb{E})[U(\beta)^2]\right| \geq \tau(n)\mathbb{E}[U(\beta)^2]\right\}, \quad \beta \in \mathcal{I};$$
$$\mathcal{E}'_Z \triangleq \left\{\min_{l\in[d_Z]} \left(\mathbf{D}_{\mathbf{Z}}^{-1}\right)_{l,l}(D_Z)_{l,l} \leq \sqrt{1-\tau(n)} \text{ or } \max_{l\in[d_Z]} \left(\mathbf{D}_{\mathbf{Z}}^{-1}\right)_{l,l}(D_Z)_{l,l} \geq \sqrt{1+\tau(n)}\right\}.$$

Lemma A.1. *If (C5.ii) holds then* $\mathbb{P}(\mathcal{E}_U) \leq m_4/(n\tau(n)^2)$ *and if (C5.iv) holds then* $\mathbb{P}(\mathcal{E}'_Z) \leq C_{\mathrm{N}}(d_Z)M_Z(d_Z)/(n\tau(n)^2)$.

Proof. The first statement follows from the Chebychev inequality. For the second statement we use

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}'_Z\right) &= \mathbb{P}\left(\left|\left(\sum_{i=1}^n \left(\frac{z_{li}^2}{\mathbb{E}\left[Z_l^2\right]} - 1\right)\right)_{l\in[d_Z]}\right| \geq n\tau(n)\right) \\
&\leq \frac{1}{(n\tau(n))^2}\mathbb{E}\left[\max_{l\in[d_Z]}\left|\sum_{i=1}^n \left(\frac{z_{li}^2}{\mathbb{E}\left[Z_l^2\right]} - 1\right)\right|^2\right] \quad \text{(by the Chebyshev inequality)} \\
&\leq \frac{C_{\mathrm{N}}(d_Z)}{n\tau(n)^2}\mathbb{E}\left[\max_{l\in[d_Z]}\left|\left(\frac{Z_l^2}{\mathbb{E}\left[Z_l^2\right]} - 1\right)\right|^2\right] \quad \text{(by the Nemirovski inequality)} \\
&\leq \frac{C_{\mathrm{N}}(d_Z)M'_Z(d_Z)}{n\tau(n)^2}.
\end{aligned}$$

□

This allows to take $\widehat{r} = r_0(n)\sqrt{1+\tau(n)}/(1-\tau(n))$ based on classes 1-4 if we restrict these classes so that:

Assumption A.1. *Let* $d_X, d_Z \geq 3$, $M'_{ZU}(d_Z) \geq 0$. *For all* $\beta, \mathbb{P}$ *such that* $\beta \in \mathcal{I}$, *(C5.ii) and (C5.iv) below hold, we have* $\mathbf{Z}$ *and* $\mathbf{U}(\beta)$ *are independent, and*

$$\mathbb{E}\left[\left|\left(\left(Z_l U(\beta)\right)^2 / \left(\mathbb{E}\left[Z_l^2\right]\mathbb{E}\left[U(\beta)^2\right]\right) - 1\right)_{l\in[d_Z]}\right|_\infty^2\right] \leq M'_{ZU}(d_Z).$$

Indeed we have $\mathcal{G}_0 \cap \left\{\max_{l\in[d_Z]} \mathbb{E}_n\left[\left(Z_l U(\beta)\right)^2\right] / \left(\mathbb{E}_n[Z_l^2]\mathbb{E}_n[U(\beta)^2]\right) \leq (1+\tau(n))/(1-\tau(n))^2\right\} \subseteq \mathcal{G}$, so

$$\inf_{\beta,\mathbb{P}:\ \beta\in\mathcal{I}} \mathbb{P}(\mathcal{G}) \geq \inf_{\beta,\mathbb{P}:\ \beta\in\mathcal{I}} \mathbb{P}(\mathcal{G}_0) - \alpha_C(n),$$

where $\alpha_C(n) \triangleq \left(m_4 + C_{\mathrm{N}}(d_Z)\left(M'_Z(d_Z) + M'_{ZU}(d_Z)\right)\right)/(n\tau(n)^2)$ by the same arguments as those leading to Lemma A.1. The coverage error $\alpha_B(n)$ in Section 2 corresponds to the one from either of class 1-4 plus $\alpha_C(n)$.

The union bound used for $r_0(n)$ in classes 1-4 does not account for dependence over $l \in [d_Z]$ of $\mathbf{Z}_{i,l}\mathbf{U}_i(\beta)$, and so $r(n)$ can be larger than necessary, even under Assumption A.1. To account for dependence, we consider class 5.

Class 5: $d_Z \geq 3$, $m_4, M_Z(d_Z), M'_Z(d_Z), q_2 > 0$, $B(n) \geq 1$, *and* $(\alpha_E(n))_{n\in\mathbb{N}}$ *is a sequence converging to zero. For all* $n \in \mathbb{N}$,

(C5.*i*) *For all* $i \in [n]$, $\mathbb{E}\left[\mathbf{U}_i(\beta)^2 \middle| \mathbf{Z}_{i,\cdot}\right] = \sigma^2_{U(\beta)}$;

(C5.*ii*) $\mathbb{E}\left[\left(\left(U(\beta)/\sigma_{U(\beta)}\right)^2 - 1\right)^2\right] \leq m_4$;

(C5.*iii*) $\mathbb{E}\left[\left|D_Z\left(ZZ^\top - \mathbb{E}\left[ZZ^\top\right]\right)D_Z\right|_\infty^2\right] \leq M_Z(d_Z)$;

(C5.*iv*) $\mathbb{E}\left[\left|\left(Z_l^2/\mathbb{E}\left[Z_l^2\right] - 1\right)_{l=1}^{d_Z}\right|_\infty^2\right] \leq M'_Z(d_Z)$;

(C5.*v*) $\max\left(\mathbb{E}\left[\left(\left(D_Z\right)_{l,l} Z_l U(\beta)/\sigma_{U(\beta)}\right)^{2+q_1}\right], \mathbb{E}\left[\left(\left(D_Z\right)_{l,l} Z_l \mathbf{E}_i\right)^{2+q_1}\right]\right) \leq B(n)^{q_1}, \forall l \in [d_Z], q_1 \in [2]$;

(C5.*vi*) $\max\left(\mathbb{E}\left[\left(\left|D_Z \mathbf{Z}_{i,\cdot}^\top \mathbf{U}_i(\beta)\right|_\infty / (B(n)\sigma_{U(\beta)})\right)^{q_2}\right], \mathbb{E}\left[\left(\left|D_Z \mathbf{Z}_{i,\cdot}^\top \mathbf{E}_i\right|_\infty / B(n)\right)^{q_2}\right]\right) \leq 2, \ \forall i \in [n]$;

where $\mathbf{E}$ *is a mean zero Gaussian vector with covariance* I*,* $\mathbf{E}$ *and* $\mathbf{Z}$ *are independent.*

Under class 5,

$$\widehat{r} = \frac{1}{\sqrt{n}}\left(q_{G|\mathbf{Z}}(1-\alpha) + 2\zeta(n)\right),$$

where $q_{G|\mathbf{Z}}$ is the quantile function of $G \triangleq |(\sum_{i\in[n]} \mathbf{D_Z} \mathbf{Z}_{i,\cdot}^\top \mathbf{E}_i)/\sqrt{n}|_\infty$ given $\mathbf{Z}$, which is obtained by simulation. $(\zeta(n))_{n\in\mathbb{N}}$ is a sequence such that $\zeta(n) \to 0$ arbitrarily slowly but $\zeta(n)/\tau(n) \to \infty$. For this $\widehat{r}$, for all n, we have $\mathbb{P}(\mathcal{G}) \geq 1 - \alpha - \alpha_B(n)$, where

$$\alpha_B(n) \triangleq 2\zeta_2(n) + \zeta_2'(n) + \varphi(\tau(n)) + \frac{C_{\mathrm{N}}(d_Z(d_Z+1)/2)M_Z(d_Z)}{n\tau(n)^2} + \iota(d_Z, n),$$

$$\zeta_2'(n) \triangleq \mathbb{P}\left(N_0 > \frac{\zeta(1-\tau(n))}{\tau(n)}\right) + \iota(d_Z, n) + \frac{C_{\mathrm{N}}(d_Z)M_Z'(d_Z)}{n\tau(n)^2} + \frac{m_4}{n\tau(n)^2},$$

$$\zeta_2(n)^2 \triangleq \mathbb{P}\left(N_0 > \zeta(n)\left(\frac{1}{\sqrt{1-\tau(n)}} - 1\right)^{-1}\right) + \iota(d_Z, n) + \frac{C_{\mathrm{N}}(d_Z)M_Z'(d_Z)}{n\tau(n)^2},$$

$$\iota(d, n) \triangleq C_2\left(\left(B(n)^2\left(\ln(d_Z n)\right)^7/n\right)^{1/6} + \left(B(n)^2(\ln(d_Z n))^3 n^{-1+2/q_2}\right)^{1/3}\right) \quad \text{for } d \in \mathbb{N},$$

φ is the function $x \in (0,1) \to C_1 x^{1/3}\max\left(1, \ln(2d_Z/x)\right)^{2/3}$, C_1 is a constant and C_2 can depend on q_2, $N_0 \triangleq |\sum_{i\in[n]} (\mathbf{E}_{D_Z Z})_{i,\cdot}|_\infty/\sqrt{n}$ where $(\mathbf{E}_{D_Z Z})_{i,\cdot}$ are independent Gaussian vectors of covariance $\mathbb{E}[D_Z \mathbf{Z}_{i,\cdot}^\top \mathbf{Z}_{i,\cdot} D_Z]$. Also we have $\mathbb{P}(\widehat{r} \leq r(n)) \geq 1 - \alpha_C(n)$, where

$$r(n) \triangleq q_{N_0}\left(1 - \alpha + \zeta_2(n) + \varphi(\tau(n))\right)/\sqrt{n} + 3\zeta(n)/\sqrt{n},$$

$$\alpha_C(n) \triangleq C_{\mathrm{N}}\left(d_Z(d_Z+1)/2\right) M_Z(d_Z)/(n\tau(n)^2) + \zeta_2(n),$$

and q_{N_0} is the quantile function of N_0. (C5.iv) is redundant once (C5.iii) is imposed but we write it for further reference and because we can have $M_Z'(d_Z) \leq M_Z(d_Z)$.

Proof of the statement for Class 5. Let $\mathcal{E}_Z \triangleq \left\{\left|D_Z(\mathbb{E}_n - \mathbb{E})\left[ZZ^\top\right] D_Z\right|_\infty \geq \tau(n)\right\}$. By the same arguments as those leading to Lemma A.1, (C5.iii) yields

$$\mathbb{P}(\mathcal{E}_Z) \leq \frac{C_{\mathrm{N}}(d_Z(d_Z+1)/2)M_Z(d_Z)}{n\tau(n)^2}. \tag{A.1}$$

Let $\beta \in \mathcal{I}$. Define

$$T \triangleq \left|\frac{1}{\sqrt{n}}\sum_{i\in[n]} \mathbf{D_Z}\mathbf{Z}_{i,\cdot}^\top \frac{\mathbf{U}_i(\beta)}{\widehat{\sigma}(\beta)}\right|_\infty, \ T_0 \triangleq \left|\frac{1}{\sqrt{n}}\sum_{i\in[n]} D_Z \mathbf{Z}_{i,\cdot}^\top \frac{\mathbf{U}_i(\beta)}{\sigma_{U(\beta)}}\right|_\infty, \ G_0 \triangleq \left|\frac{1}{\sqrt{n}}\sum_{i\in[n]} D_Z \mathbf{Z}_{i,\cdot}^\top \mathbf{E}_i\right|_\infty.$$

T_0, G_0, and N_0 have same covariance matrix, indeed

$$\mathbb{E}\left[D_Z \mathbf{Z}_{i,\cdot}^\top \mathbf{Z}_{i,\cdot} D_Z \frac{\mathbf{U}_i(\beta)^2}{\sigma_{U(\beta)^2)}}\right] = \mathbb{E}\left[D_Z \mathbf{Z}_{i,\cdot}^\top \mathbf{Z}_{i,\cdot} D_Z \mathbb{E}\left[\frac{\mathbf{U}_i(\beta)^2}{\sigma_{U(\beta)^2)}}\middle| \mathbf{Z}\right]\right] = \mathbb{E}[D_Z \mathbf{Z}_{i,\cdot}^\top \mathbf{Z}_{i,\cdot} D_Z].$$

Let us show that, for all $\alpha \in (0,1)$, $\left|\mathbb{P}\left(T \leq q_{G|\mathbf{Z}}(\alpha)\right) - \alpha\right| \leq \alpha_B(n)$. Using (C5.v), (C5.vi) and Proposition 2.1 in [4], we obtain

$$\max\left(\left|\mathbb{P}(T_0 \leq t) - \mathbb{P}(N_0 \leq t)\right|, \left|\mathbb{P}(G_0 \leq t) - \mathbb{P}(N_0 \leq t)\right|\right) \leq \iota(d_Z, n). \tag{A.2}$$

Indeed, by (C5.i), the law of iterated expectations, and the independence between $\mathbf{E}$ and $\mathbf{Z}$, we have $\mathbb{E}\left[\left((D_Z)_{l,l}\, Z_l U(\beta)/\sigma_{U(\beta)}\right)^2\right] = \mathbb{E}\left[\left((D_Z)_{l,l}\, Z_l E\right)^2\right] = 1$ for all $l \in [d_Z]$, so condition (M.1) in [4] is satisfied. We denote by $q_{G_0|\mathbf{Z}}$ the conditional quantile functions of G_0 given $\boldsymbol{Z}$. By the arguments in the proof of Proposition A.1 and Lemma 3.2 in [3], denoting by $q_{G_0|\mathbf{Z}}(\alpha)$ the conditional α quantile of G_0 given $\mathbf{Z}$, for all $\alpha \in (0,1)$,

$$\begin{aligned}\min\left(\mathbb{P}\left(q_{G_0|\mathbf{Z}}(\alpha) \leq q_{N_0}(\alpha+\varphi(\tau(n)))\right), \mathbb{P}\left(q_{N_0}(\alpha) \leq q_{G_0|\mathbf{Z}}(\alpha+\varphi(\tau(n)))\right)\right)\\ \geq 1 - \frac{C_{\mathrm{N}}(d_Z(d_Z+1)/2)M_Z(d_Z)}{n\tau(n)^2}.\end{aligned} \tag{A.3}$$

For all $\alpha \in (0,1)$, on the event $\{\mathbb{P}\left(T_0 \leq q_{G_0|\mathbf{Z}}(\alpha)\right) - \alpha > 0\}$, we have

$$\begin{aligned}\left|\mathbb{P}\left(T_0 \leq q_{G_0|\mathbf{Z}}(\alpha)\right) - \alpha\right| \leq& \mathbb{P}\left(T_0 \leq q_{N_0}(\alpha+\varphi(\tau(n)))\right) - \alpha + \frac{C_{\mathrm{N}}(d_Z(d_Z+1)/2)M_Z(d_Z)}{n\tau(n)^2} \quad \text{(by (A.3))}\\ \leq& \alpha + \varphi(\tau(n)) - \alpha + \iota(d_Z,n) + \frac{C_{\mathrm{N}}(d_Z(d_Z+1)/2)M_Z(d_Z)}{n\tau(n)^2} \quad \text{(by (A.2))},\end{aligned}$$

on the complement of the event,

$$\begin{aligned}\left|\mathbb{P}\left(T_0 \leq q_{G_0|\mathbf{Z}}(\alpha)\right) - \alpha\right| \leq& \alpha - \mathbb{P}\left(T_0 \leq q_{N_0}(\alpha-\varphi(\tau(n)))\right) + \frac{C_{\mathrm{N}}(d_Z(d_Z+1)/2)M_Z(d_Z)}{n\tau(n)^2} \quad \text{(by (A.3))}\\ \leq& \varphi(\tau(n)) + \iota(d_Z,n) + \frac{C_{\mathrm{N}}(d_Z(d_Z+1)/2)M_Z(d_Z)}{n\tau(n)^2} \quad \text{(by (A.2))},\end{aligned}$$

hence, we always have

$$\left|\mathbb{P}\left(T_0 \leq q_{G_0|\mathbf{Z}}(\alpha)\right) - \alpha\right| \leq \varphi(\tau(n)) + \frac{C_{\mathrm{N}}(d_Z(d_Z+1)/2)M_Z(d_Z)}{n\tau(n)^2} + \iota(d_Z,n). \tag{A.4}$$

On $\mathcal{E}_Z'^c$, we have

$$|G - G_0| \leq \left(\frac{1}{\sqrt{1-\tau(n)}} - 1\right) G_0,$$

hence, by the Markov inequality, the law of iterated expectations, and second bound in (A.2), setting

$$\zeta_2(n)^2 \triangleq \mathbb{P}\left(N_0 > \zeta(n)\left(\frac{1}{\sqrt{1-\tau(n)}} - 1\right)^{-1}\right) + \iota(d_Z,n) + \frac{C_{\mathrm{N}}(d_Z)M_Z'(d_Z)}{n\tau(n)^2},$$

we have

$$\mathbb{P}\left(\mathbb{P}\left(|G - G_0| > \zeta(n) \,|\, \mathbf{Z}\right) > \zeta_2(n)\right) < \zeta_2(n). \tag{A.5}$$

On $\mathcal{E}_Z'^c \cap \mathcal{E}_U^c$, we have

$$|T - T_0| \leq \frac{\tau(n)T_0}{1-\tau(n)}, \tag{A.6}$$

hence, by the first bound in (A.2),

$$\mathbb{P}\left(|T - T_0| > \zeta(n)\right) \leq \mathbb{P}\left(N_0 > \frac{\zeta(n)(1-\tau(n))}{\tau(n)}\right) + \iota(d_Z,n) + \frac{C_{\mathrm{N}}(d_Z)M_Z'(d_Z) + m_4}{n\tau(n)^2} \tag{A.7}$$

and we denote the right-hand side by $\zeta_2'(n)$. Using Lemma 3.3 in [3] and (A.5) in the first display, (A.7) in the second, (A.4) in the third display

$$\mathbb{P}\left(T - 2\zeta(n) \geq q_{G|\mathbf{Z}}(1-\alpha)\right) < \mathbb{P}\left(T - \zeta(n) \geq q_{G_0|\mathbf{Z}}(1-\alpha-\zeta_2(n))\right) + \zeta_2(n)$$

$$\leq \mathbb{P}\left(T_0 \geq q_{G_0|\mathbf{Z}}(1-\alpha-\zeta_2(n))\right) + \zeta_2(n) + \zeta_2'(n)$$
$$\leq \alpha + 2\zeta_2(n) + \zeta_2'(n) + \varphi(\tau(n)) + \frac{C_{\mathrm{N}}(d_Z(d_Z+1)/2)M_Z(d_Z)}{n\tau(n)^2} + \iota(d_Z,n).$$

The result on the deterministic upper bound $r(n)$ on $\widehat{r}$ is obtained using Lemma 3.3 in [3] and (A.3). □

A.1.2. *Deterministic Bounds on Sample Objects.* The purpose of this section is to provide probabilistic conditions under which we can replace random quantities appearing in Section 4.2 by determinist ones. These are: $\widehat{r}$, $\widehat{\sigma}(\beta)$, and the sensitivities. This requires to further restrict the class $\mathcal{P}$ so that:

Assumption A.2. *Let $d_X, d_Z \geq 3$, $\alpha_\infty(n)$, m_4, $M_\Psi(d_Z,d_X)$, $M_X(d_X)$, $M_Z'(d_Z)$, and $B(n,d_Z)$ positive. For all $\beta,\mathbb{P}$ such that $\beta \in \mathcal{I}$, we have*

$$\mathbb{E}\left[\left|D_Z\left(ZX^\top - \mathbb{E}\left[ZX^\top\right]\right)D_X\right|_\infty^2\right] \leq M_\Psi(d_Z,d_X), \tag{A.8}$$

$$\mathbb{E}\left[\left|\left(X_k^2/\mathbb{E}\left[X_k^2\right]-1\right)_{k=1}^{d_X}\right|_\infty^2\right] \leq M_X(d_X). \tag{A.9}$$

For $\mathcal{P}$ from class 1-4, we have either (C5.ii), (C5.iv), $\mathbb{P}\left(\left|D_Z\mathbf{Z}^\top\right|_\infty > B(n,d_Z)\right) \leq \alpha_\infty(n)$ or Assumption A.1.

For classes 1-4 with Assumption A.2 but without Assumption A.1, we set $\alpha_C(n) = \alpha_\infty(n) + (m_4 + C_{\mathrm{N}}(d_Z)M_Z'(d_Z))/(n\tau(n)^2)$ and $r(n) \triangleq r_0(n)B(n,d_Z)/\sqrt{1-\tau(n)}$. For example, if $D_Z\mathbf{Z}^\top$ has sub-Gaussian entries, $B(n,d_Z)$ can be taken proportional to $\sqrt{\ln(Cnd_Z/\alpha_\infty(n))}$, where the constants C and of proportionality depend on tail parameters of the sub-Gaussian distribution. For classes 1-4 with Assumption A.2 and Assumption A.1, $\alpha_C(n)$ has already been introduced and we set $\widehat{r} = r(n)$. For class 5, $\alpha_C(n)$ and $r(n)$ have been defined above. There, $r(n)$ is proportional to $\sqrt{\ln(Cd_Z^2/\alpha_C(n))/n}$ for a constant C depending on the covariance matrix of N_0. The following result relates random quantities to their population counterparts.

Proposition A.1. *Under Assumption A.2, for all $\beta,\mathbb{P}$ such that $\beta \in \mathcal{I}$, on an event $\mathcal{G}\cap\mathcal{G}_\Psi$ such that $\mathbb{P}(\mathcal{G}\cap\mathcal{G}_\Psi) \geq 1-\alpha-\alpha_D(n)$, where*

$$\alpha_D(n) \triangleq \alpha_B(n) + \alpha_C(n) + \left(C_{\mathrm{N}}(d_X)M_X(d_X) + C_{\mathrm{N}}(d_Zd_X)M_\Psi(d_Z,d_X)\right)/(n\tau(n)^2), \tag{A.10}$$

$\mathcal{G}$ holds but also, for all $c>0$, $\widehat{r} \leq r$,

$$\sigma_{U(\beta)}^2(1-\tau(n)) \leq \widehat{\sigma}(\beta)^2 \leq \sigma_{U(\beta)}^2(1+\tau(n)), \tag{A.11}$$

$$\forall b \in \mathbb{R}^{d_X},\ l \in \mathcal{L},\ \sqrt{1-\tau(n)}\,l\left(D_X^{-1}b\right) \leq l\left(\mathbf{D}_\mathbf{X}^{-1}b\right) \leq \sqrt{1+\tau(n)}\,l\left(D_X^{-1}b\right), \tag{A.12}$$

$$\forall S \subseteq [d_X],\ l \in \mathcal{L},\ \widehat{\kappa}_{l,S} \geq \frac{\kappa_{l,S}}{1+\tau(n)}\left(1-\frac{\tau(n)}{\kappa_{1,S}}\right),\quad \forall k \in [d_X],\ \widehat{\kappa}_{e_k,S}^* \geq \frac{\kappa_{e_k,S}^*}{1+\tau(n)}\left(1-\frac{\tau(n)}{\kappa_{1,S}}\right), \tag{A.13}$$

$$\forall S \subseteq [d_X],\ l \in \mathcal{L},\ \widehat{\gamma}_{l,S} \geq \frac{\gamma_{l,S}}{1+\tau(n)}\left(1-\frac{\tau(n)}{\gamma_{1,S}}\right),\quad \forall k \in [d_X],\ \widehat{\gamma}_{e_k,S}^* \geq \frac{\gamma_{e_k,S}^*}{1+\tau(n)}\left(1-\frac{\tau(n)}{\kappa_{1,S}}\right), \tag{A.14}$$

If $|S\cap S_Q| \leq s$,

$$\forall l \in \mathcal{L},\ \widehat{\kappa}_l(s) \geq \overline{\kappa}_l(s),\quad \forall k \in [d_X],\ \widehat{\kappa}_{e_k}^*(s) \geq \overline{\kappa}_{e_k}^*(s),$$

where

$$\overline{\kappa}_l(s) \triangleq \frac{\kappa_l(s)}{1+\tau(n)}\min_{S:|S\cap S_Q|\leq s}\left(1-\frac{\tau(n)}{\kappa_{1,S}}\right),\quad \overline{\kappa}_{e_k}^*(s) \triangleq \frac{\kappa_{e_k}^*(s)}{1+\tau(n)}\min_{S:|S\cap S_Q|\leq s}\left(1-\frac{\tau(n)}{\kappa_{1,S}}\right)$$

and $\kappa_l(s)$ *and* $\kappa^*_{e_k}(s)$ *are the analogue of the lower bounds in* (A.13) *using the population counterparts, and* $\widehat{\theta}_\kappa(s) \geq \overline{\theta}_\kappa(s)$*, where*

$$\overline{\theta}_\kappa(s) \triangleq \left(1 - \frac{r}{\overline{\kappa}_g(s)}\right)_+^{-1}.$$

Proof of Proposition A.1. Define the events

$$\mathcal{E}'_X \triangleq \left\{ \min_{k\in[d_X]} \left(\mathbf{D}_\mathbf{X}^{-1}\right)_{k,k} (D_X)_{k,k} \leq \sqrt{1-\tau(n)} \text{ or } \max_{k\in[d_X]} \left(\mathbf{D}_\mathbf{X}^{-1}\right)_{k,k} (D_X)_{k,k} \geq \sqrt{1+\tau(n)} \right\};$$

$$\mathcal{E}_{ZX^\top} \triangleq \left\{ \left| D_Z(\mathbb{E}_n - \mathbb{E}) \left[ZX^\top\right] D_X \right|_\infty \geq \tau(n) \right\}.$$

The event $\mathcal{E}'_Z$ is such that $\mathcal{E}'_Z \subseteq \mathcal{E}_Z$. Define $\mathcal{G}_\Psi \triangleq \{\widehat{r} \leq r(n)\} \cap \mathcal{E}'^c_Z \cap \mathcal{E}'^c_X \cap \mathcal{E}^c_{ZX^\top} \cap \mathcal{E}^c_U$ for classes 1-4 and $\mathcal{G}_\Psi \triangleq \{\widehat{r} \leq r(n)\} \cap \mathcal{E}^c_Z \cap \mathcal{E}'^c_X \cap \mathcal{E}^c_{ZX^\top} \cap \mathcal{E}^c_U$ for class 5. Recall that $\mathbb{P}(\widehat{r} \leq r(n)) \geq 1 - \alpha_C(n)$. Similarly to Lemma A.1, we have

$$\mathbb{P}\left(\mathcal{E}'_X\right) \leq \frac{C_{\mathrm{N}}(d_X) M_X(d_X)}{n\tau(n)^2}, \ \mathbb{P}\left(\mathcal{E}_{ZX^\top}\right) \leq \frac{C_{\mathrm{N}}(d_Z d_X) M_\Psi(d_Z, d_X)}{n\tau(n)^2}.$$

Clearly, on $\mathcal{E}'^c_X$, (A.12) holds. Assume now that we work on the event $\mathcal{G}_\Psi$.
Let $S \subseteq [d_X]$, $l \in \mathcal{L}$, and $\overline{\Delta} \triangleq D_X^{-1}\mathbf{D}_\mathbf{X}\Delta$. Due to (A.12), we have, for all $l \in \mathcal{L}$, in particular ℓ_1-norms of subvectors, $\sqrt{1-\tau(n)}\, l(\overline{\Delta}) \leq l(\Delta) \leq \sqrt{1+\tau(n)}\, l(\overline{\Delta})$. This, the fact that $\widehat{r} \leq r$, and manipulations on the ℓ_1-norm of subvectors used previously, yield $\overline{\Delta} \in K_S$ if $\Delta \in \widehat{K}_S$ and $\overline{\Delta} \in K_{\gamma,S}$ if $\Delta \in \widehat{K}_{\gamma,S}$. Now, because $\mathcal{G}_\Psi \subseteq \mathcal{E}'^c_Z \cap \mathcal{E}^c_{ZX^\top}$, we obtain

$$\begin{aligned}
\left|\widehat{\Psi}\Delta\right|_\infty &\geq \min_{l\in[d_Z]} \left(\mathbf{D}_\mathbf{Z} D_Z^{-1}\right)_{l,l} \left| D_Z \mathbb{E}_n \left[ZX^\top\right] D_X D_X^{-1} \mathbf{D}_\mathbf{X} \Delta \right|_\infty \\
&\geq \frac{1}{\sqrt{1+\tau(n)}} \left(\left| D_Z \mathbb{E}\left[ZX^\top\right] D_X \overline{\Delta} \right|_\infty - \left| D_Z(\mathbb{E}_n - \mathbb{E}) \left[ZX^\top\right] D_X \overline{\Delta} \right|_\infty \right) \\
&\geq \frac{1}{\sqrt{1+\tau(n)}} \left(\left|\Psi\overline{\Delta}\right|_\infty - \tau(n) \left|\overline{\Delta}\right|_1 \right). \qquad \text{(A.15)}
\end{aligned}$$

Inequalities (A.13) and (A.14) are obtained from the definition of $\kappa_{1,S}$ and $\gamma_{1,S}$ and the fact that, on $\mathcal{G}_\Psi$, $l(\Delta) \leq \sqrt{1+\tau(n)}\, l(\overline{\Delta})$. Finally, we check that $\mathbb{P}(\mathcal{G} \cap \mathcal{G}_\Psi) \geq 1 - \alpha_D(n)$. □

A.2. **Lower Bounds on the Sensitivities.** We use the quantity $\widehat{c}_\kappa(S, S(\widehat{\beta})) \triangleq \min(\widehat{c}_{>,\kappa}(S, S(\widehat{\beta})), \widehat{c}_{<,\kappa}(S, S(\widehat{\beta})))$, where

$$\widehat{c}_{>,\kappa}\left(S, S\left(\widehat{\beta}\right)\right) \triangleq \frac{1}{(1-c\widehat{r})_+} \left(2\left|S \cap S_Q\right| + \left| S^c_Q \cap \left(S \cup S\left(\widehat{\beta}\right)\right)\right| + c(1-\widehat{r}) \left| S^c_I \cap \left(S \cup S\left(\widehat{\beta}\right)\right)\right| \right),$$

$$\widehat{c}_{<,\kappa}\left(S, S\left(\widehat{\beta}\right)\right) \triangleq \frac{1}{(1-c)_+} \left(2|S \cap S_Q| + \left| S^c_Q \cap \left(S \cup S\left(\widehat{\beta}\right)\right)\right| \right).$$

The set $\widehat{S}(S, S(\widehat{\beta}))$ is defined as $(S \cap S_Q) \cup ((S^c_Q \cup S^c_I) \cap (S \cup S(\widehat{\beta})))$, when $1 \leq c < \widehat{r}^{-1}$, and as $S \cup (S^c_Q \cap S(\widehat{\beta}))$, when $c < 1$. The following result relates the sensitivities for various losses. It requires no assumption. Similar bounds can be obtained for the sensitivities based on $\widehat{K}_{\gamma,S}$ (see (v5)).

Proposition A.2. *Let* $S \in [d_X]$ *and* $c > 0$.

(*i*) *Let* $S \subseteq \widehat{S} \subseteq [d_X]$. *Then, for all* $l \in \mathcal{L}$, *we have* $\widehat{\kappa}_{l,S} \geq \widehat{\kappa}_{l,\widehat{S}}$.
(*ii*) *For all* $S_0 \subseteq [d_X]$ *and* $q \in [1, \infty]$, *we have* $\widehat{\kappa}_{q,S_0,S} \geq \widehat{\kappa}_{q,S}$.

(iii) For all $q \in [1,\infty]$ and $S_0 \subseteq [d_X]$,

$$c_\kappa(S, S(\widehat{\beta}))^{-1/q} \widehat{\kappa}_{\infty,S} \leq \widehat{\kappa}_{q,S} \leq \widehat{\kappa}_{\infty,S}, \tag{A.16}$$

$$c_\kappa(S, S(\widehat{\beta}))^{-1} \widehat{\kappa}_{\infty,\widehat{S}(S,S(\widehat{\beta})),S} \leq \widehat{\kappa}_{1,S}, \tag{A.17}$$

$$|S_0|^{-1/q} \widehat{\kappa}_{\infty,S_0,S} \leq \widehat{\kappa}_{q,S_0,S} \leq \widehat{\kappa}_{\infty,S_0,S}. \tag{A.18}$$

(iv) In addition to (A.16)-(A.17)*, we have*

$$\widehat{\kappa}_{1,S} \geq \max\left(\left(\frac{2}{\widehat{\kappa}_{1,S\cap S_Q,S}} + \frac{1}{\widehat{\kappa}_{1,S^c_Q,S}} + \frac{c}{\widehat{\kappa}_{g,S}} \right)^{-1}, (1-c\widehat{r})_+ \left(\frac{2}{\widehat{\kappa}_{1,S\cap S_Q,S}} + \frac{1}{\widehat{\kappa}_{1,S^c_Q,S}} + \frac{c(1-\widehat{r})}{\widehat{\kappa}_{1,S^c_I,S}} \right)^{-1}, \right. \tag{A.19}$$
$$\left. (1-c)_+ \left(\frac{2}{\widehat{\kappa}_{1,S\cap S_Q,S}} + \frac{1}{\widehat{\kappa}_{1,S^c_Q,S}} \right)^{-1} \right).$$

(v) We have

$$\widehat{\kappa}_{g,S} \geq \max\left((1-c\widehat{r})_+ \left(\frac{2}{r(n)\widehat{\kappa}_{1,S\cap S_Q,S}} + \frac{\widehat{r}}{\widehat{\kappa}_{1,S^c_Q,S}} + \frac{1-\widehat{r}}{\widehat{\kappa}_{1,S^c_I,S}} \right)^{-1}, \left(\frac{\widehat{r}}{\widehat{\kappa}_{1,S_I,S}} + \frac{1}{\widehat{\kappa}_{1,S^c_I,S}} \right)^{-1}, \widehat{\kappa}_{1,S} \right). \tag{A.20}$$

(vi) For all $S_0 \subseteq [d_X]$, we have

$$\widehat{\kappa}_{\infty,S_0,S} = \min_{k\in S_0} \widehat{\kappa}^*_{e_k,S} = \min_{k\in S_0} \min_{\Delta\in\widehat{K}_S:\ \Delta_k=1,\ |\Delta|_\infty\leq 1} \left|\widehat{\Psi}\Delta\right|_\infty.$$

Proof of Proposition A.2. We prove the bounds for the sensitivities based on $\widehat{K}_S$, those for the sensitivities based on $\widehat{K}_{\gamma,S}$ are obtained similarly. Parts (i) and (ii) are easy. The upper bound in (A.16) follows from the fact that $|\Delta|_q \geq |\Delta|_\infty$. We obtain the lower bound as follows. Because $|\Delta|_q \leq |\Delta|_1^{1/q} |\Delta|_\infty^{1-1/q}$, we get that, for $\Delta \neq 0$,

$$\frac{\left|\widehat{\Psi}\Delta\right|_\infty}{|\Delta|_q} \geq \frac{\left|\widehat{\Psi}\Delta\right|_\infty}{|\Delta|_\infty} \left(\frac{|\Delta|_\infty}{|\Delta|_1} \right)^{1/q}. \tag{A.21}$$

Furthermore, for $\Delta \in \widehat{K}_S$, by definition of the set, we have

$$|\Delta_{S^c\cap S_Q}|_1 \leq |\Delta_{S\cap S_Q}|_1 + c\widehat{r}|\Delta|_1 + c(1-\widehat{r})|\Delta_{S^c_I}|_1 \tag{A.22}$$

which, by adding $|\Delta_{(S\cap S_Q)\cup S^c_Q}|_1$ on both sides, is equivalent to

$$|\Delta|_1 \leq \frac{1}{(1-c\widehat{r})_+} \left(2|\Delta_{S\cap S_Q}|_1 + |\Delta_{S^c_Q}|_1 + c(1-\widehat{r})|\Delta_{S^c_I}|_1 \right). \tag{A.23}$$

From (A.23) and the fact that $\Delta_{S^c\cap S(\widehat{\beta})^c} = 0$, we deduce

$$|\Delta|_1 \leq \frac{\left|\Delta_{\widehat{S}(S,S(\widehat{\beta}))}\right|_\infty}{(1-c\widehat{r})_+} \left(2|S\cap S_Q| + \left|S^c_Q \cap \left(S \cup S\left(\widehat{\beta}\right)\right)\right| + c(1-\widehat{r}) \left|S^c_I \cap \left(S \cup S\left(\widehat{\beta}\right)\right)\right| \right). \tag{A.24}$$

Let us obtain an alternative lower bound for the case where $c \in (0,1)$. The condition that $\Delta \in \widehat{K}_S$ can also be written as

$$|\Delta_{S^c\cap S_Q}|_1 \leq |\Delta_{S\cap S_Q}|_1 + c(\widehat{r}-1)|\Delta_{S_I}|_1 + c|\Delta|_1$$

which implies

$$|\Delta_{S^c\cap S_Q}|_1 \leq |\Delta_{S\cap S_Q}|_1 + c|\Delta|_1$$

and, by adding $|\Delta_{(S\cap S_Q)\cup S_Q^c}|_1$ on both sides, if $c \in (0,1)$, this is equivalent to

$$|\Delta|_1 \leq \frac{1}{1-c}\left(2|\Delta_{S\cap S_Q}|_1 + |\Delta_{S_Q^c}|_1\right). \tag{A.25}$$

Using $\Delta_{S^c\cap S(\widehat{\beta})^c} = 0$, this yields

$$|\Delta|_1 \leq \frac{2|S\cap S_Q| + \left|S_Q^c \cap \left(S\cup S\left(\widehat{\beta}\right)\right)\right|}{(1-c)_+}\left|\Delta_{\widehat{S}(S,S(\widehat{\beta}))}\right|_\infty. \tag{A.26}$$

We obtain inequality (A.17) and the lower bound in (A.16) using (A.24) and (A.26), that $|\Delta_{S\cup S_Q^c\cup S_I^c}|_\infty \leq |\Delta|_\infty$, $|\Delta_{S\cup S_Q^c}|_\infty \leq |\Delta|_\infty$, and (A.21).
Inequality (A.18) can be proved in a similar manner. The lower bounds follows from the fact that

$$\frac{\left|\widehat{\Psi}\Delta\right|_\infty}{|\Delta_{S_0}|_q} \geq \frac{\left|\widehat{\Psi}\Delta\right|_\infty}{|\Delta_{S_0}|_\infty}\left(\frac{|\Delta_{S_0}|_\infty}{|\Delta_{S_0}|_1}\right)^{1/q}$$

and $|\Delta_{S_0}|_1 \leq |S_0||\Delta_{S_0}|_\infty$. While the upper bound holds because $|\Delta_{S_0}|_q \geq |\Delta_{S_0}|_\infty$.
To prove (A.19) it suffices to note that, by definition of the set $\widehat{K}_S$,

$$|\Delta|_1 \leq \left(\frac{2}{\widehat{\kappa}_{1,S\cap S_Q,S}} + \frac{1}{\widehat{\kappa}_{1,S_Q^c,S}} + \frac{c}{\widehat{\kappa}_{g,S}}\right)\left|\widehat{\Psi}\Delta\right|_\infty, \tag{A.27}$$

by (A.23),

$$|\Delta|_1 \leq \frac{1}{(1-c\widehat{r})_+}\left(\frac{2}{\widehat{\kappa}_{1,S\cap S_Q,S}} + \frac{1}{\widehat{\kappa}_{1,S_Q^c,S}} + \frac{c(1-\widehat{r})}{\widehat{\kappa}_{1,S_I^c,S}}\right)\left|\widehat{\Psi}\Delta\right|_\infty,$$

and, by (A.25),

$$|\Delta|_1 \leq \frac{1}{(1-c)_+}\left(\frac{2}{\widehat{\kappa}_{1,S\cap S_Q,S}} + \frac{1}{\widehat{\kappa}_{1,S_Q^c,S}}\right)\left|\widehat{\Psi}\Delta\right|_\infty.$$

The bound (v) is obtained by rewriting $\Delta \in \widehat{K}_S$ as

$$(1-c\widehat{r})|\Delta_{S_I}|_1 + (1-c)|\Delta_{S_I^c}|_1 \leq 2|\Delta_{S\cap S_Q}|_1 + |\Delta_{S_Q^c}|_1, \tag{A.28}$$

which yields

$$\begin{aligned}\widehat{r}|\Delta_{S_I}|_1 + |\Delta_{S_I^c}|_1 &\leq \frac{\widehat{r}}{(1-c\widehat{r})_+}\left(2|\Delta_{S\cap S_Q}|_1 + |\Delta_{S_Q^c}|_1 + \frac{1-\widehat{r}}{\widehat{r}}|\Delta_{S_I^c}|_1\right)\\ &\leq \frac{\widehat{r}\left|\widehat{\Psi}\Delta\right|_\infty}{(1-c\widehat{r})_+}\left(\frac{2}{\widehat{\kappa}_{1,S\cap S_Q,S}} + \frac{1}{\widehat{\kappa}_{1,S_Q^c,S}} + \frac{1-\widehat{r}}{\widehat{r}\widehat{\kappa}_{1,S_I^c,S}}\right).\end{aligned} \tag{A.29}$$

The second upper bound follows from noticing that, if $\widehat{\kappa}_{g,S} > 0$, we have

$$\frac{1}{\widehat{\kappa}_{g,S}} = \sup_{\Delta\in\widehat{K}_S:\ \left|\widehat{\Psi}\Delta\right|_\infty=1}\left(\widehat{r}\,|\Delta_{S_I}|_1 + \left|\Delta_{S_I^c}\right|_1\right) \leq \sup_{\Delta\in\widehat{S}_S:\ \left|\widehat{\Psi}\Delta\right|_\infty=1}\widehat{r}\,|\Delta_{S_I}|_1 + \sup_{\Delta\in\widehat{K}_S:\ \left|\widehat{\Psi}\Delta\right|_\infty=1}\left|\Delta_{S_I^c}\right|_1.$$

The third upper uses that $\widehat{r}\,|\Delta_{S_I}|_1 + \left|\Delta_{S_I^c}\right|_1 \leq |\Delta|_1$.
Let us now prove (vi). Because for all k in S_0, $|\Delta_{S_0}|_\infty \geq |\Delta_k|$, one obtains that for all k in S_0,

$$\widehat{\kappa}_{\infty,S_0,S} = \min_{\Delta\in\widehat{K}_{S_0}}\frac{\left|\widehat{\Psi}\Delta\right|_\infty}{|\Delta_{S_0}|_\infty} \leq \min_{\Delta\in\widehat{K}_S}\frac{\left|\widehat{\Psi}\Delta\right|_\infty}{|\Delta_k|} = \widehat{\kappa}^*_{e_k,S}.$$

Thus $\widehat{\kappa}_{\infty,S_0,S} \leq \min_{k\in S_0} \widehat{\kappa}^*_{e_k,S}$. But one also has

$$\widehat{\kappa}_{\infty,S_0,S} = \min_{k\in S_0} \min_{\Delta\in\widehat{K}_S:\ |\Delta_k|=|\Delta_{S_0}|_\infty=1} \left|\widehat{\Psi}\Delta\right|_\infty \geq \min_{k\in S_0} \min_{\Delta\in\widehat{K}_S:\ |\Delta_k|=1} \left|\widehat{\Psi}\Delta\right|_\infty . \tag{A.30}$$ □

A.3. Lower Bound on $\kappa_{1,S}$ in Case (IC).

Proposition A.3. *In case (IC), we have*

$$\kappa_{1,S} \geq \frac{1}{u_{\tau(n)}} \max_{m\in[d_X-|S|]} \min_{\substack{S_1\subseteq S^c\\ |S_1|\leq m}} \max_{\lambda:|\lambda|_1\leq 1} \left(\min_{\substack{\Delta:|\Delta|_1=1\\ S(\Delta)\subseteq S\cup S_1}} |\lambda^\top\Psi\Delta| - \frac{u_{\tau(n)}-1}{\sqrt{m}} \max_{\substack{\widetilde{S}\subseteq S^c\\ |\widetilde{S}|\leq m}} |\lambda^\top\Psi_{\cdot,\widetilde{S}}|_2 \right).$$

Proof of Proposition A.3. Take $\lambda\in\mathbb{R}^{d_Z}$ such that $|\lambda|_1\leq 1$ and $\Delta\in K_S$. Define S_1 the set of m largest entries of Δ of index in S^c, S_2 the subsequent m largest in S^c, and so forth, and $S_{01}=S\cup S_1$. By the inverse and direct triangle inequalities, we have

$$\left|\lambda^\top\Psi\Delta_{S_{01}}\right| \leq \sum_{j\geq 2} |\lambda^\top\Psi\Delta_{S_j}| + |\Psi\Delta|_\infty .$$

For $j\geq 2$, we have

$$\begin{aligned} |\lambda^\top\Psi\Delta_{S_j}| &\leq |\lambda^\top\Psi_{\cdot,S_j}|_2 |\Delta_{S_j}|_2 \\ &\leq \frac{1}{\sqrt{m}} |\lambda^\top\Psi_{\cdot,S_j}|_2 |\Delta_{S_{j-1}}|_1 \\ &\leq \frac{1}{\sqrt{m}} \left(\max_{|\widetilde{S}|\leq m,\widetilde{S}\subseteq S^c} |\lambda^\top\Psi_{\cdot,\widetilde{S}}|_2 \right) |\Delta_{S_{j-1}}|_1, \end{aligned}$$

so, using $\Delta\in K_S$ in the last display,

$$\begin{aligned} \sum_{j\geq 2} |\lambda^\top\Psi\Delta_{S_j}| &\leq \frac{1}{\sqrt{m}} \left(\max_{|\widetilde{S}|\leq m,\widetilde{S}\subseteq S^c} |\lambda^\top\Psi_{\cdot,\widetilde{S}}|_2 \right) |\Delta_{S^c}|_1 \\ &\leq \frac{u_{\tau(n)}-1}{\sqrt{m}} \left(\max_{|\widetilde{S}|\leq m,\widetilde{S}\subseteq S^c} |\lambda^\top\Psi_{\cdot,\widetilde{S}}|_2 \right) |\Delta_S|_1. \end{aligned}$$

Moreover,

$$|\Delta_S|_1 \left(\min_{\substack{\Delta:|\Delta|_1=1\\ S(\Delta)\subseteq S\cup S_1}} |\lambda^\top\Psi\Delta| \right) \leq |\Delta_{S_{01}}|_1 \left(\min_{\substack{\Delta:|\Delta|_1=1\\ S(\Delta)\subseteq S\cup S_1}} |\lambda^\top\Psi\Delta| \right) \leq \left|\lambda^\top\Psi\Delta_{S_{01}}\right|$$

and

$$\min_{\substack{\Delta:|\Delta|_1=1\\ S(\Delta)\subseteq S\cup S_1}} |\lambda^\top\Psi\Delta| \geq \frac{1}{\sqrt{m+|S|}} \min_{\substack{\Delta:|\Delta|_2=1\\ S(\Delta)\subseteq S\cup S_1}} |\lambda^\top\Psi\Delta|,$$

so we use that $|\Delta|_1\leq u_{\tau(n)}|\Delta_S|_1$, take the supremum of the lower bound over λ (so λdepends on Δ via S_1) and then the minimum over S_1. Because the computation was made for m arbitrary, the lower bound is the maximum over m of the lower bounds. □

Lower bounds can be obtained, by using $|\lambda|_1 \leq |\lambda|_2\sqrt{|S(\lambda)|}$ and $|\Delta|_1 \leq |\Delta|_2\sqrt{m+|S|}$ when $S(\Delta) \subseteq S \cup S_1$, as follows

$$\kappa_{1,S} \geq \frac{1}{u_{\tau(n)}} \max_{m\in[d_X-|S|]} \min_{\substack{S_1\subseteq S^c \\ |S_1|\leq m}} \max_{\lambda:|\lambda|_2\leq 1} \frac{1}{\sqrt{|S(\lambda)|}} \left(\frac{1}{\sqrt{m+|S|}} \min_{\substack{\Delta:|\Delta|_2=1 \\ S(\Delta)\subseteq S\cup S_1}} |\lambda^\top \Psi \Delta| - \frac{u_{\tau(n)}-1}{\sqrt{m}} \max_{\substack{\widetilde{S}\subseteq S^c \\ |\widetilde{S}|\leq m}} |\lambda^\top \Psi_{\cdot,\widetilde{S}}|_2 \right)$$

and replacing the maximum over λ by

$$\max_{S_2\in[d_Z]} \frac{1}{\sqrt{|S_2|}} \left(\frac{1}{\sqrt{m+|S|}} \max_{\substack{\lambda:|\lambda|_2\leq 1 \\ S(\lambda)\subseteq S_2}} \min_{\substack{\Delta:|\Delta|_2=1 \\ S(\Delta)\subseteq S\cup S_1}} |\lambda^\top \Psi \Delta| - \frac{u_{\tau(n)}-1}{\sqrt{m}} \min_{\substack{\lambda:|\lambda|_2\leq 1 \\ S(\lambda)\subseteq S_2}} \max_{\substack{\Delta:|\Delta|_2\leq 1 \\ S(\Delta)\subseteq S^c \\ |S(\Delta)|\leq m}} |\lambda^\top \Psi \Delta| \right).$$

The argument using the sequence of sets S_j is used in [16] and [9]. A similar result is given in [7]. The main difference is that the negative term above is smaller in absolute value than a maximum singular value and the first term in the bracket above is larger than a minimum singular value. Moreover, this lower bound depends explicitly on S. A disadvantage of this bound (and the one in [7]) is the presence of the set S_1 which implies that, in the lower bound, we can pay a price for a regressor which is not in the model.

A.4. **Proofs of the Results in the Main Text. Proof of Theorem 4.1.** Take $\beta \in \mathcal{I}$ and set $\widehat{\Delta} \triangleq \mathbf{D}_{\mathbf{X}}^{-1}(\widehat{\beta}-\beta)$. By definition of $\widehat{\mathcal{I}}$ and $\widehat{\sigma}(\beta) = \mathbb{E}_n[U(\beta)^2]$, on $\mathcal{G}$, we have $\beta \in \widehat{\mathcal{I}}(\widehat{r}, \widehat{\sigma}(\beta))$. On $\mathcal{G}$, we also have

$$\left|\widehat{\Psi}\widehat{\Delta}\right|_\infty \leq \left|\frac{1}{n}\mathbf{D}_{\mathbf{Z}}\mathbf{Z}^\top\left(\mathbf{Y}-\mathbf{X}\widehat{\beta}\right)\right|_\infty + \left|\frac{1}{n}\mathbf{D}_{\mathbf{Z}}\mathbf{Z}^\top(\mathbf{Y}-\mathbf{X}\beta)\right|_\infty \tag{A.31}$$

$$\leq \widehat{r}\left(\widehat{\sigma}+\widehat{\sigma}(\beta)\right). \tag{A.32}$$

On the other hand, $(\widehat{\beta},\widehat{\sigma})$ minimizes the criterion $\left|\mathbf{D}_{\mathbf{X}}^{-1}\beta\right|_1 + c\sigma$. Thus, on $\mathcal{G}$, we have

$$\left|\mathbf{D}_{\mathbf{X}}^{-1}\widehat{\beta}_{S_Q}\right|_1 + c\widehat{\sigma} \leq |\mathbf{D}_{\mathbf{X}}^{-1}\beta_{S_Q}|_1 + c\widehat{\sigma}(\beta). \tag{A.33}$$

This implies, on $\mathcal{G}$,

$$\begin{aligned} \left|\widehat{\Delta}_{S(\beta)^c\cap S_Q}\right|_1 &= \sum_{k\in S(\beta)^c\cap S_Q} \left|\mathbb{E}_n[X_k^2]^{1/2}\widehat{\beta}_k\right| \\ &\leq \sum_{k\in S(\beta)\cap S_Q} \left(\left|\mathbb{E}_n[X_k^2]^{1/2}\beta_k\right| - \left|\mathbb{E}_n[X_k^2]^{1/2}\widehat{\beta}_k\right|\right) + c\left(\widehat{\sigma}(\beta)-\widehat{\sigma}\right) \\ &\leq \left|\widehat{\Delta}_{S(\beta)\cap S_Q}\right|_1 + c\left(\widehat{\sigma}(\beta)-\widehat{\sigma}\left(\widehat{\beta}\right)\right). \end{aligned} \tag{A.34}$$

The last inequality holds because by construction $\widehat{\sigma}(\widehat{\beta}) \leq \widehat{\sigma}$.
Because $\gamma \to \sqrt{\widehat{\sigma}(\gamma)}$ is convex and

$$w_* \triangleq -\frac{\frac{1}{n}\sum_{i\in[n]} x_i(\mathbf{Y}_i - \mathbf{X}_{i,\cdot}\beta)}{\sqrt{\frac{1}{n}\sum_{i\in[n]}(\mathbf{Y}_i-\mathbf{X}_{i,\cdot}\beta)^2}} \mathbb{1}\left\{\frac{1}{n}\sum_{i\in[n]}(\mathbf{Y}_i-\mathbf{X}_{i,\cdot}\beta)^2 \neq 0\right\} \in \partial\widehat{\sigma}(\cdot)(\beta).$$

we have

$$\widehat{\sigma}(\beta) - \widehat{\sigma}\left(\widehat{\beta}\right) \leq w_*^\top\left(\beta-\widehat{\beta}\right)$$

$$= (\mathbf{D}_{\mathbf{X}} w_*)^\top \mathbf{D}_{\mathbf{X}}^{-1} \left(\beta - \widehat{\beta}\right) = - (\mathbf{D}_{\mathbf{X}} w_*)^\top \widehat{\Delta}.$$

Now, for all $k \in S_I$, we have $|(\mathbf{D}_{\mathbf{X}} w_*)_k| \leq \widehat{r}$ on $\mathcal{G}$. This is because these regressors serve as their own IV and, on $\mathcal{G}$, $\beta \in \widehat{\mathcal{I}}\,(r(n), \widehat{\sigma}(\beta))$. On the other hand, for all row of index k in the set S_I^c, the Cauchy-Schwarz inequality yields

$$|(\mathbf{D}_{\mathbf{X}} w_*)_k| \leq \frac{|\mathbb{E}_n[X_k U(\beta)]|}{\sqrt{\mathbb{E}_n[X_k^2]\mathbb{E}_n[U(\beta)^2]}} \leq 1.$$

Finally, we obtain

$$\widehat{\sigma}(\beta) - \widehat{\sigma}\left(\widehat{\beta}\right) \leq \widehat{r} \left|\widehat{\Delta}_{S_I}\right|_1 + \left|\widehat{\Delta}_{S_I^c}\right|_1. \tag{A.35}$$

Combining (A.35) with (A.34), we find that $\widehat{\Delta} \in \widehat{K}_{S(\beta)}$ on $\mathcal{G}$. Using (A.31) and (A.35), we find

$$\begin{aligned} \left|\widehat{\Psi}\widehat{\Delta}\right|_\infty &\leq \widehat{r}\left(\widehat{\sigma} + \widehat{\sigma}\left(\widehat{\beta}\right) + \widehat{\sigma}(\beta) - \widehat{\sigma}\left(\widehat{\beta}\right)\right) \\ &\leq \widehat{r}\left(2\overline{\sigma} + \widehat{r}\left|\widehat{\Delta}_{S_I}\right|_1 + \left|\widehat{\Delta}_{S_I^c}\right|_1\right). \end{aligned} \tag{A.36}$$

Using the definition of the sensitivities we obtain, on $\mathcal{G}$,

$$\left|\widehat{\Psi}\widehat{\Delta}\right|_\infty \leq \widehat{r}\left(2\overline{\sigma} + \widehat{r}\frac{\left|\widehat{\Psi}\widehat{\Delta}\right|_\infty}{\widehat{\kappa}_{\widehat{g},S(\beta)}}\right),$$

which implies

$$\left|\widehat{\Psi}\widehat{\Delta}\right|_\infty \leq 2\widehat{r}\overline{\sigma}\left(1 - \frac{\widehat{r}}{\widehat{\kappa}_{\widehat{g},S(\beta)}}\right)_+^{-1}. \tag{A.37}$$

(A.37) and the definition of the sensitivities yield the first upper bound.
For the second upper bound we use that, by (A.33) and the definition of $\widehat{\kappa}_{1,S(\beta)\cap S_Q,S(\beta)}$,

$$c\widehat{\sigma} \leq \left|\widehat{\Delta}_{S(\beta)\cap S_Q}\right|_1 + c\widehat{\sigma}(\beta) \leq \frac{\left|\widehat{\Psi}\widehat{\Delta}\right|_\infty}{\widehat{\kappa}_{1,S(\beta)\cap S_Q,S(\beta)}} + c\widehat{\sigma}(\beta), \tag{A.38}$$

and, by adding $c\widehat{\sigma}(\beta)$ to both sides and (A.32),

$$\widehat{\sigma} + \widehat{\sigma}(\beta) \leq 2\widehat{\sigma}(\beta)\left(1 - \frac{r}{c\widehat{\kappa}_{1,S(\beta)\cap S_Q,S(\beta)}}\right)_+^{-1}. \tag{A.39}$$

□

Proof of Theorem 4.2. Take $\beta \in \mathcal{I}$ and $S \subseteq [d_X]$. Acting as in (A.34), on $\mathcal{G}$, we get

$$\begin{aligned} &\sum_{k\in S^c\cap S_Q}\left|\mathbb{E}_n[X_k^2]^{1/2}\widehat{\beta}_k\right| + \sum_{k\in S^c\cap S_Q}\left|\mathbb{E}_n[X_k^2]^{1/2}\beta_k\right| \\ &\leq \sum_{k\in S\cap S_Q}\left(\left|\mathbb{E}_n[X_k^2]^{1/2}\beta_k\right| - \left|\mathbb{E}_n[X_k^2]^{1/2}\widehat{\beta}_k\right|\right) + 2\sum_{k\in S^c\cap S_Q}\left|\mathbb{E}_n[X_k^2]^{1/2}\beta_k\right| + c\left(\widehat{\sigma}(\beta) - \widehat{\sigma}\left(\widehat{\beta}\right)\right) \\ &\leq \left|\widehat{\Delta}_{S\cap S_Q}\right|_1 + 2\left|\mathbf{D}_{\mathbf{X}}^{-1}\beta_{S^c\cap S_Q}\right|_1 + c\widehat{r}\left|\widehat{\Delta}_{S_I}\right|_1 + c\left|\widehat{\Delta}_{S_I^c}\right|_1. \end{aligned}$$

This yields

$$\left|\widehat{\Delta}_{S^c\cap S_Q}\right|_1 \leq \left|\widehat{\Delta}_{S\cap S_Q}\right|_1 + 2\left|\mathbf{D}_{\mathbf{X}}^{-1}\beta_{S^c\cap S_Q}\right|_1 + c\widehat{r}\left|\widehat{\Delta}_{S_I}\right|_1 + c\left|\widehat{\Delta}_{S_I^c}\right|_1. \tag{A.40}$$

Let us show the first inequality and consider two cases.
Case 1: $2|(\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta)_{S^c\cap S_Q}|_1 \leq |\widehat{\Delta}_{S\cap S_Q}|_1 + c\widehat{r}|\widehat{\Delta}_{S_I}|_1 + c|\widehat{\Delta}_{S_I^c}|_1 + |\widehat{\Delta}_{S_Q^c}|_1$, then $\widehat{\Delta} \in \widehat{K}_{\gamma,S}$. From this, using the definition of the sensitivity $\widehat{\gamma}_{q,S_0,S}$, we get the upper bound corresponding to the first term in the minimum. Also, we have

$$\begin{aligned}
\widehat{\sigma} &\leq \frac{1}{c}\left(\left|\mathbf{D}_{\mathbf{X}}^{-1}\beta_{S_Q}\right|_1 - \left|\mathbf{D}_{\mathbf{X}}^{-1}\widehat{\beta}_{S_Q}\right|_1\right) + \widehat{\sigma}(\beta) \\
&\leq \frac{1}{c}\min\left(\left|\widehat{\Delta}_{S_Q}\right|_1, \left|\widehat{\Delta}_{S\cap S_Q}\right|_1 + \left|\mathbf{D}_{\mathbf{X}}^{-1}\beta_{S^c\cap S_Q}\right|_1\right) + \widehat{\sigma}(\beta) \\
&\leq \frac{1}{c}\min\left(\left|\widehat{\Delta}_{S_Q}\right|_1, \frac{1}{2}\left(3\left|\widehat{\Delta}_{S\cap S_Q}\right|_1 + c\widehat{r}\left|\widehat{\Delta}_{S_I}\right|_1 + c\left|\widehat{\Delta}_{S_I^c}\right|_1 + \left|\widehat{\Delta}_{S_Q^c}\right|_1\right)\right) + \widehat{\sigma}(\beta) \qquad \text{(A.41)} \\
&\leq \frac{\left|\widehat{\Psi}\widehat{\Delta}\right|_\infty}{c\widehat{\gamma}_{\widehat{h},S}} + \widehat{\sigma}(\beta), \qquad \text{(A.42)}
\end{aligned}$$

which, with (A.32) also yields

$$\widehat{\sigma} + \widehat{\sigma}(\beta) \leq 2\widehat{\sigma}(\beta)\left(1 - \frac{r}{c\widehat{\gamma}_{\widehat{h},S}}\right)_+^{-1}. \qquad \text{(A.43)}$$

Case 2: $2|(\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta)_{S^c\cap S_Q}|_1 > |\widehat{\Delta}_{S\cap S_Q}|_1 + c\widehat{r}|\widehat{\Delta}_{S_I}|_1 + c|\widehat{\Delta}_{S_I^c}|_1 + |\widehat{\Delta}_{S_Q^c}|_1$, then we have

$$\left|\widehat{\Delta}\right|_1 = \left|\widehat{\Delta}_{S^c\cap S_Q}\right|_1 + \left|\widehat{\Delta}_{S\cap S_Q}\right|_1 + \left|\widehat{\Delta}_{S_Q^c}\right|_1 \leq 6\left|\left(\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta\right)_{S^c\cap S_Q}\right|_1.$$

In conclusion, $|\widehat{\Delta}_{S_0}|_q$ is smaller than the maximum of the two bounds. □

Proof of Proposition 4.1. This is a consequence of the definition of the sensitivities, the cones $\widehat{K}_{\widehat{S}}$ and $\widehat{K}_{\gamma,\widehat{S}}$, and the fact that minimizing on a larger set yields lower bounds on the sensitivities. More specifically, we use $|\Delta_{S\cap S_Q}|_1 \leq \min(s, |\widehat{S}\cap S_Q|)|\Delta_{\widehat{S}\cap S_Q}|_\infty$. The last constraint is not convex but the cone is a union of sets involving the linear constraint $|\Delta_{S\cap S_Q}|_1 \leq \min(s, |\widehat{S}\cap S_Q|)|\Delta_j|$, hence the second minimum. □

Proof of Proposition 4.2. Start by proving (vii). Let $\lambda \in \mathbb{R}^{d_Z}$ such that $|\lambda|_1 \leq 1$, $k \in [d_X]$, and $\Delta \in K_S$. By the inverse triangle inequality we have

$$\left|\lambda^\top\Psi\Delta - \lambda^\top\Psi_{\cdot,k}\Delta_k\right| \leq \left(\sum_{k'\neq k}|\Delta_{k'}|\right)\max_{k'\neq k}\left|\lambda^\top\Psi_{\cdot,k'}\right|,$$

which yields

$$\left|\lambda^\top\Psi_{\cdot,k}\right||\Delta_k| \leq \left(\sum_{k'\neq k}|\Delta_{k'}|\right)\max_{k'\neq k}\left|\lambda^\top\Psi_{\cdot,k'}\right| + \left|\lambda^\top\Psi\Delta\right|,$$

hence

$$\begin{aligned}
\left(\left|\lambda^\top\Psi_{\cdot,k}\right| + \max_{k'\neq k}\left|\lambda^\top\Psi_{\cdot,k'}\right|\right)|\Delta_k| &\leq |\Delta|_1\max_{k'\neq k}\left|\lambda^\top\Psi_{\cdot,k'}\right| + |\Psi\Delta|_\infty, \\
&\leq c_\kappa(S)\max_{k'\neq k}\left|\lambda^\top\Psi_{\cdot,k'}\right|\left|\Delta_{\overline{S}}\right|_\infty + |\Psi\Delta|_\infty, \qquad \text{(A.44)}
\end{aligned}$$

$$\leq c_\kappa(S) \max_{k' \neq k} \left|\lambda^\top \Psi_{\cdot,k'}\right| |\Delta_{S_0}|_\infty + |\Psi\Delta|_\infty .$$

(A.44) uses that, by similar arguments as those leading to (A.24) and (A.26), noting that $\widehat{c}_\kappa(S, S(\widehat{\beta})) \geq c_\kappa(S)$ and $\widehat{S}(S, S(\widehat{\beta})) \subseteq \overline{S}$, we have $|\Delta|_1 \leq c_\kappa(S)|\Delta_{\overline{S}}|_\infty$.
For k such that $|\Delta_k| = |\Delta_{S_0}|_\infty$, this yields

$$\left(\left|\lambda^\top \Psi_{\cdot,k}\right| - (c_\kappa(S)-1) \max_{k' \neq k} \left|\lambda^\top \Psi_{\cdot,k'}\right|\right) |\Delta_{S_0}|_\infty \leq |\Psi\Delta|_\infty ,$$

$$\max_{\lambda \in \mathbb{R}^{d_Z}:\ |\lambda|_1 \leq 1} \left(\left|\lambda^\top \Psi_{\cdot,k}\right| - (c_\kappa(S)-1) \max_{k' \neq k} \left|\lambda^\top \Psi_{\cdot,k'}\right|\right) |\Delta_{S_0}|_\infty \leq |\Psi\Delta|_\infty , \tag{A.45}$$

and we conclude by taking the minimum over $k \in S_0$ and then using the definition of the ℓ_∞-S_0 sensitivity. This proves (vii).
Start from (A.45) and use (A.17) in Proposition A.2 to obtain (v).
To prove (viii) we start from (A.44). By definition of the ℓ_∞-S_0 sensitivity,

$$\left(\left|\lambda^\top \Psi_{\cdot,k}\right| + \max_{k' \neq k} \left|\lambda^\top \Psi_{\cdot,k'}\right|\right) |\Delta_k| \leq \left(\frac{c_\kappa(S) \max_{k' \neq k} \left|\lambda^\top \Psi_{\cdot,k'}\right|}{\kappa_{\infty,\overline{S},S}} + 1\right) |\Psi\Delta|_\infty$$

and we conclude by taking $\Delta_k = 1$.
The proof of the other items is very similar to the proof of Proposition A.2. □

Proof of Theorem 4.3. The inequalities in (i) and (iii) follow from the second bounds in theorems 4.1 and 4.2 and Proposition A.1 but also the fact that, on $\mathcal{G} \cap \mathcal{G}_\Psi$,

$$\sqrt{1-\tau(n)}\sigma_{U(\beta)} \left(1 - \frac{2r(n)\Theta_\kappa(S(\beta))}{\kappa_{g,S(\beta)}}\right) \leq \widehat{\sigma}(\widehat{\beta}) \leq \widehat{\sigma} \leq \sqrt{1+\tau(n)}\sigma_{U(\beta)} \left(2\left(1 - \frac{r(n)(1+\tau(n))}{c\kappa_{1,S(\beta)\cap S_Q,S(\beta)}}\left(1 - \frac{\tau(n)}{\kappa_{1,S(\beta)}}\right)^{-1}\right)_+^{-1} - 1\right), \tag{A.46}$$

$$\sqrt{1-\tau(n)} \left(\sigma_{U(\beta)} - \min_{S \subseteq [d_X]} \max \left(\frac{2r(n)(1+\tau(n))\sigma_{U(\beta)}}{\gamma_{g,S}} \left(1 - \frac{\tau(n)}{\kappa_{1,S}} - \frac{r(n)(1+\tau(n))}{c\gamma_{h,S}}\right)_+^{-1}, \frac{2}{c1_n} \left|D_X^{-1}\beta_{S^c \cap S_Q}\right|_1\right)\right) \leq \widehat{\sigma}(\widehat{\beta})$$
$$\leq \widehat{\sigma} \leq \sqrt{1+\tau(n)} \left(\sigma_{U(\beta)} + \frac{1}{c} \min_{S \subseteq [d_X]} \max \left(2\sigma_{U(\beta)} \left(\left(1 - \frac{r(n)(1+\tau(n))}{c\gamma_{h,S}}\left(1 - \frac{\tau(n)}{\gamma_{1,S}}\right)^{-1}\right)_+^{-1} - 1\right), \frac{3}{2}\left|D_X^{-1}\beta_{S^c \cap S_Q}\right|_1\right)\right). \tag{A.47}$$

The right inequality in (A.46) is obtained from (A.39). The left one uses that, from (A.35), (A.32), and (A.39),

$$\widehat{\sigma}\left(\widehat{\beta}\right) \geq \widehat{\sigma}(\beta) - \frac{\widehat{r}\,(\widehat{\sigma} + \widehat{\sigma}(\beta))}{\widehat{\kappa}_{\widehat{g},S}} \geq \widehat{\sigma}(\beta) \left(1 - \frac{2\widehat{r}}{\widehat{\kappa}_{\widehat{g},S}} \left(1 - \frac{r}{c\widehat{\kappa}_{1,S(\beta)\cap S_Q,S(\beta)}}\right)_+^{-1}\right).$$

Similarly, we obtain the right inequality in (A.47) by using (A.43) and that, in case 2, using the second element in the minimum in (A.41), we get

$$\widehat{\sigma} \leq \frac{3}{c}\left|\left(\mathbf{D}_\mathbf{X}^{-1}\beta\right)_{S^c \cap S_Q}\right|_1 + \widehat{\sigma}(\beta).$$

An we obtain the left one, using that, by (A.35),

$$\widehat{\sigma}(\beta) - \frac{2}{c}\left|\left(\mathbf{D}_\mathbf{X}^{-1}\beta\right)_{S^c \cap S_Q}\right|_1 \leq \widehat{\sigma}\left(\widehat{\beta}\right).$$

Part (ii) follows from (i) and (iii) with $l(\Delta) = |e_k^\top \Delta|$ and the fact that the assumption on $|\beta_k|$ implies: $\widehat{\beta}_k \neq 0$ for $k \in S(\beta)$ (resp., S_* as defined at the end of Section 4.4.2). □

Proof of Theorem 4.4. Fix s and β in $\mathcal{I}_s$ and work on $\mathcal{G} \cap \mathcal{G}_\Psi$. Using Theorem 4.3 (i), we obtain $\widehat{\omega}_k(s) \leq \omega_k(s)$. The following two cases can occur. First, if $k \in S(\beta)^c$ (so that $\beta_k = 0$) then, using the bound in (4.8) for l defined by $l(\Delta) = |e_k^\top \Delta|$ we obtain $\sqrt{\mathbb{E}_n[X_k^2]}|\widehat{\beta}_k| \leq \widehat{\omega}_k(s)$, which implies $\widehat{\beta}_k^\omega = 0$. Second, if $k \in S(\beta)$, then using again (4.8) for the same functional, we get $||\widehat{\beta}_k| - |\beta_k|| \leq |\widehat{\beta}_k - \beta_k| \leq \widehat{\omega}_k(s)/\sqrt{(1-\tau(n))\mathbb{E}[X_k^2]} \leq \omega_k(s)/\sqrt{(1-\tau(n))\mathbb{E}[X_k^2]}$. Since $|\beta_k| > 2\omega_k(s)/\sqrt{(1-\tau(n))\mathbb{E}[X_k^2]}$ for $k \in S(\beta)$, we obtain $|\widehat{\beta}_k| > \omega_k(s)/\sqrt{(1-\tau(n))\mathbb{E}[X_k^2]} \geq \widehat{\omega}_k(s)/\sqrt{\mathbb{E}_n[X_k^2]}$, so that $\widehat{\beta}_k^\omega = \widehat{\beta}_k$. □

Proof of Theorem 6.1. The elements relative to assumptions and estimation of Λ are in Section B.3. We make the proof in the case $d_G = 1$. Extension to $d_G > 1$ is straightforward. Take $(\beta, \Lambda) \in \mathcal{I}_{s'_r}$. Let $\widehat{\Delta} \triangleq \mathbf{D}_\mathbf{X}^{-1}(\widehat{\beta} - \beta)$. Due to Assumption B.1 (iv), $\mathbb{P}(\mathcal{E}_{ZV}) \leq C_\mathrm{N}(d_F) M_{ZV}(d_F)/(n\tau(n)^2)$, where

$$\mathcal{E}_{ZV} \triangleq \left\{ \exists f \in [d_F] : \ \mathbb{E}_n\left[(\Lambda_{f,\cdot} ZV(\beta))^2\right] / \mathbb{E}\left[(\Lambda_{f,\cdot} ZV(\beta))^2\right] \geq 1 + \tau(n) \right\}.$$

We work on $\mathcal{E} \triangleq \mathcal{G} \cap \mathcal{G}_0' \cap \mathcal{G}_\Psi \cap \{\widehat{\rho}_{ZX} > \rho_{ZX}(n)\} \cap \mathcal{E}_T'^c \cap \mathcal{E}_{ZW}^c \cap \mathcal{E}_Z^c \cap \mathcal{E}_{ZV}^c$ of probability $1 - \alpha_\beta(n) - \alpha_\Lambda(n) - \mathbb{P}(\mathcal{E}_{ZV})$, where the confidence level associated to $\mathcal{G}$ and $\mathcal{G}_0'$ converge to 1. We use $\max\left(\sqrt{1+\tau(n)} - 1, 1 - \sqrt{1-\tau(n)}\right) = 1 - \sqrt{1-\tau(n)} \leq \tau(n)$ and, for all $a \in \mathbb{R}^{d_Z}$ and $b \in \mathbb{R}^{d_X}$,

$$\begin{aligned}
\frac{1}{n}\sum_{i\in[n]} \left(\left(\mathbf{D}_\mathbf{Z}^{-1} a\right)^\top \left(\mathbf{D}_\mathbf{Z} \mathbf{Z}_{i,\cdot}\right)^\top \mathbf{X}_{i,\cdot} b \right)^2 &= \left(\mathbf{D}_\mathbf{Z}^{-1} a\right)^\top \left(\frac{1}{n}\sum_{i\in[n]} \mathbf{D}_\mathbf{Z} \mathbf{Z}_{i,\cdot}^\top \mathbf{X}_{i,\cdot} b b^\top \mathbf{X}_{i,\cdot}^\top \mathbf{Z}_{i,\cdot} \mathbf{D}_\mathbf{Z} \right) \mathbf{D}_\mathbf{Z}^{-1} a \\
&\leq \left|\mathbf{D}_\mathbf{Z}^{-1} a\right|_1^2 \left| \frac{1}{n}\sum_{i\in[n]} \mathbf{D}_\mathbf{Z} \mathbf{Z}_{i,\cdot}^\top \mathbf{X}_{i,\cdot} b b^\top \mathbf{X}_{i,\cdot}^\top \mathbf{Z}_{i,\cdot} \mathbf{D}_\mathbf{Z} \right|_\infty \\
&\leq \left|\mathbf{D}_\mathbf{Z}^{-1} a\right|_1^2 \left|\mathbf{D}_\mathbf{X}^{-1} b\right|_1^2 \left| \frac{1}{n}\sum_{i\in[n]} \mathbf{D}_\mathbf{X} \mathbf{X}_{i,\cdot}^\top \mathbf{Z}_{i,\cdot} \mathbf{D}_\mathbf{Z}^2 \mathbf{Z}_{i,\cdot}^\top \mathbf{X}_{i,\cdot} \mathbf{D}_\mathbf{X} \right|_\infty \\
&\leq \left|\mathbf{D}_\mathbf{Z}^{-1} a\right|_1^2 \left|\mathbf{D}_\mathbf{X}^{-1} b\right|_1^2 \widehat{\rho}_{ZX}^2.
\end{aligned}$$

For all $f \in [d_F]$, we have

$$\begin{aligned}
&\left(D_{\Lambda ZW(\beta)}\right)_{f,f} \left| \sqrt{\mathbb{E}_n\left[\left(\widehat{\Lambda}_{f,\cdot} ZU(\widehat{\beta})\right)^2\right]} - \sqrt{\mathbb{E}\left[(\Lambda_{f,\cdot} ZW(\beta))^2\right]} \right| \\
&\leq \sqrt{\mathbb{E}_n\left[\left(\left(D_{\Lambda ZW(\beta)}\right)_{f,f} \widehat{\Lambda}_{f,\cdot} Z\right)^2 \left(X^\top \widehat{\Delta}\right)^2\right]} + \sqrt{\mathbb{E}_n\left[\left(\left(D_{\Lambda ZW(\beta)}\right)_{f,f} \widehat{\Lambda}_{f,\cdot} Z\right)^2 V(\beta)^2\right]} \\
&\quad + \left(D_{\Lambda ZW(\beta)}\right)_{f,f} \left(\sqrt{\mathbb{E}_n\left[\left(\left(\widehat{\Lambda}_{f,\cdot} - \Lambda_{f,\cdot}\right) ZW(\beta)\right)^2\right]} + \left| \sqrt{\mathbb{E}_n\left[(\Lambda_{f,\cdot} ZW(\beta))^2\right]} - \sqrt{\mathbb{E}\left[(\Lambda_{f,\cdot} ZW(\beta))^2\right]} \right| \right) \\
&\leq \left| \left(D_{\Lambda ZU(\beta)}\right)_{f,f} \widehat{\Lambda}_{f,\cdot} \mathbf{D}_\mathbf{Z}^{-1} \right|_1 \left(\widehat{\rho}_{ZX} \left|\widehat{\Delta}\right|_1 + \widehat{\rho}_Z \sqrt{1+\tau(n)} v(d_X) \right) + v_{\Lambda,2}(n) + \tau(n) \\
&\leq \sqrt{1+\tau(n)} \left(\left| \left(D_{\Lambda ZU(\beta)}\right)_{f,f} \Lambda_{f,\cdot} D_Z^{-1} \right|_1 + \left| D_{\Lambda ZU(\beta)} \right|_\infty v_{\Lambda,1}(n) \right) \left(\rho_{ZX} \left|\widehat{\Delta}\right|_1 + \rho_Z \sqrt{1+\tau(n)} v(d_X) \right) + v_{\Lambda,2}(n) + \tau(n)
\end{aligned}$$

and

$$\begin{aligned}
&\left(D_{\Lambda ZW(\beta)}\right)_{f,f}\left|\sqrt{\mathbb{E}_n\left[\left(\widehat{\Lambda}_{f,\cdot}ZU(\widehat{\beta})\right)^2\right]}-\sqrt{\mathbb{E}\left[(\Lambda_{f,\cdot}ZW(\beta))^2\right]}\right|\\
&\leq\sqrt{\mathbb{E}_n\left[\left(\left(D_{\Lambda ZW(\beta)}\right)_{f,f}\left(\widehat{\Lambda}_{f,\cdot}-\Lambda_{f,\cdot}\right)ZX^\top\mathbf{D}_{\mathbf{X}}\mathbf{D}_{\mathbf{X}}^{-1}\widehat{\beta}\right)^2\right]}+\left(D_{\Lambda ZW(\beta)}\right)_{f,f}\sqrt{\mathbb{E}_n\left[\left(\Lambda_{f,\cdot}ZX\mathbf{D}_{\mathbf{X}}\widehat{\Delta}\right)^2\right]}\\
&\quad+\left(D_{\Lambda ZW(\beta)}\right)_{f,f}\sqrt{\mathbb{E}_n\left[\left(\Lambda_{f,\cdot}ZV(\beta)\right)^2\right]}+\tau(n)\\
&\leq\sqrt{1+\tau(n)}\left|D_{\Lambda ZU(\beta)}\right|_\infty v_{\Lambda,1}(n)\rho_{ZX}\left(v_{\beta,1}(n)+\sqrt{1+\tau(n)}\left|D_X^{-1}\beta\right|_1\right)+v_{\beta,2}(n)\\
&\quad+\sqrt{1+\tau(n)}\sqrt{\mathbb{E}\left[\left(\left(D_{\Lambda ZW(\beta)}\Lambda\right)_{f,\cdot}ZV(\beta)\right)^2\right]}+\tau(n).
\end{aligned}$$

We have obtained $\left|\mathbf{D}_{\mathbf{U}(\widehat{\beta})\mathbf{Z}\widehat{\Lambda}^\top}D_{\Lambda ZW(\beta)}^{-1}\right|_\infty\leq 1+v_D(n)+\tau(n)$. We now use the decomposition

$$\sqrt{n}\mathbf{D}_{\mathbf{U}(\widehat{\beta})\mathbf{Z}\widehat{\Lambda}^\top}\left(\widehat{\Omega\beta}-\Omega\beta-V(\beta)\right)=R_1+R_2+R_3+\frac{1}{\sqrt{n}}\mathbf{D}_{\mathbf{U}(\widehat{\beta})\mathbf{Z}\widehat{\Lambda}^\top}\widehat{\Lambda}\mathbf{Z}^\top\mathbf{W}(\beta),\tag{A.48}$$

where $R_1\triangleq\sqrt{n}\mathbf{D}_{\mathbf{U}(\widehat{\beta})\mathbf{Z}\widehat{\Lambda}^\top}\left(\Omega-\frac{1}{n}\widehat{\Lambda}\mathbf{Z}^\top\mathbf{X}\right)\mathbf{D}_{\mathbf{X}}\widehat{\Delta}$, $R_2=\mathbf{D}_{\mathbf{U}(\widehat{\beta})\mathbf{Z}\widehat{\Lambda}^\top}\widehat{\Lambda}\mathbf{Z}^\top\mathbf{V}(\beta)/\sqrt{n}$, and $R_3=-\sqrt{n}\mathbf{D}_{\mathbf{U}(\widehat{\beta})\mathbf{Z}\widehat{\Lambda}^\top}\overline{V}(\beta)$. On $\mathcal{E}$, we have

$$\begin{aligned}
&|R_1|_\infty\leq\sqrt{n}(1+v_D(n)+\tau(n))\left|D_{\Lambda ZW(\beta)}\right|_\infty r_0'(n)v_{F(\Lambda)}(n)v_{\beta,1}(n),\\
&|R_2|_\infty\leq\sqrt{n}(1+v_D(n)+\tau(n))\left(v_{\Lambda,3}(n)+(\tau(n)+1)\left|D_{\Lambda ZW(\beta)}D_{\Lambda Z}^{-1}\right|_\infty\right)v(d_X)\sqrt{1+\tau(n)},\\
&|R_3|_\infty\leq\sqrt{n}(1+v_D(n)+\tau(n))\left|D_{\Lambda ZW(\beta)}\right|_\infty|\overline{V}(\beta)|_\infty.
\end{aligned}$$

Define

$$\begin{aligned}
&T_\Omega\triangleq\left|\frac{1}{\sqrt{n}}\sum_{i\in[n]}\mathbf{D}_{\mathbf{U}(\widehat{\beta})\mathbf{Z}\widehat{\Lambda}^\top}\widehat{\Lambda}\mathbf{Z}_{i,\cdot}^\top\mathbf{W}_i(\beta)\right|_\infty,\ T_{\Omega1}=\left|\frac{1}{\sqrt{n}}\sum_{i\in[n]}D_{\Lambda ZW(\beta)}\widehat{\Lambda}\mathbf{Z}_{i,\cdot}^\top\mathbf{W}_i(\beta)\right|_\infty,\\
&T_{\Omega0}\triangleq\left|\frac{1}{\sqrt{n}}\sum_{i\in[n]}D_{\Lambda ZW(\beta)}\Lambda\mathbf{Z}_{i,\cdot}^\top\mathbf{W}_i(\beta)\right|_\infty,\ G_{\Omega1}\triangleq\left|\frac{1}{\sqrt{n}}\sum_{i\in[n]}D_{\Lambda ZW(\beta)}\widehat{\Lambda}\mathbf{Z}_{i,\cdot}^\top\mathbf{U}_i(\widehat{\beta})\mathbf{E}_i\right|_\infty,\\
&G_{\Omega0}\triangleq\left|\frac{1}{\sqrt{n}}\sum_{i\in[n]}D_{\Lambda ZW(\beta)}\Lambda\mathbf{Z}_{i,\cdot}^\top\mathbf{W}_i(\beta)\mathbf{E}_i\right|_\infty,\ N_{\Omega0}\triangleq\left|\frac{1}{\sqrt{n}}\sum_{i\in[n]}\left(\mathbf{E}_{D_{\Lambda ZW(\beta)}\Lambda ZW(\beta)}\right)_{i,\cdot}^\top\right|_\infty,
\end{aligned}$$

where $\left(\mathbf{E}_{D_{\Lambda ZW(\beta)}\Lambda ZW(\beta)}\right)_{i,\cdot}$ are independent Gaussian vectors of covariance $\mathbb{E}[D_{\Lambda ZW(\beta)}\Lambda\mathbf{Z}_{i,\cdot}^\top\mathbf{U}_i(\beta)^2\mathbf{Z}_{i,\cdot}\Lambda^\top D_{\Lambda ZW(\beta)}]$. On $\mathcal{E}$, we have $|T_\Omega-T_{\Omega1}|\leq T_{\Omega1}(v_D(n)+\tau(n))$ and $|T_{\Omega1}-T_{\Omega0}|\leq v_{\Lambda,2}(n)$, so

$$|T_\Omega-T_{\Omega0}|\leq(T_{\Omega0}+v_{\Lambda,2}(n))(v_D(n)+\tau(n))+v_{\Lambda,2}(n).$$

Also, on $\mathcal{E}\cap\{\mathbb{E}_n[E^2]\geq 1+\tau(n)\}$, we have $|G_\Omega-G_{\Omega1}|\leq G_{\Omega1}(v_D(n)+\tau(n))$ and $|G_{\Omega1}-G_{\Omega0}|\leq v_D(n)\sqrt{1+\tau(n)}$, so

$$|G_\Omega-G_{\Omega0}|\leq(G_{\Omega0}+v_D(n)\sqrt{1+\tau(n)})(v_D(n)+\tau(n))+v_D(n)\sqrt{1+\tau(n)}.$$

We can now proceed as for (A.5) and (A.7) and obtain $\mathbb{P}\left(|T_\Omega-T_{\Omega0}|>\zeta(n)\right)\leq\zeta_2'(n)$ and $\mathbb{P}\left(\mathbb{P}\left(|G_\Omega-G_{\Omega0}|>\zeta(n)|\mathbf{U}(\widehat{\beta})\mathbf{Z}\widehat{\Lambda}^\top\right)>\zeta_2(n)\right)<\zeta_2(n)$.

Due to Assumption B.1 (vi), we have $\mathbb{P}(\mathcal{E}_{\Lambda ZW}) \leq C_{\mathrm{N}}(d_F(d_F+1)/2)M_{F,ZW}(d_F)/(n\tau(n)^2)$, where $\mathcal{E}_{\Lambda ZW} \triangleq \left\{\left|D_{\Lambda ZW(\beta)}\Lambda(\mathbb{E}_n - \mathbb{E})\left[ZZ^\top W(\beta)^2\right]\Lambda^\top D_{\Lambda ZW(\beta)}\right|_\infty \geq \tau(n)\right\}$. The same argument as the one leading to (A.4) now gives

$$\left|\mathbb{P}\left(T_{\Omega 0} \leq q_{G_{\Omega 0}|\mathbf{Z}}(\alpha)\right) - \alpha\right| \leq \varphi(\tau(n)) + \frac{C_{\mathrm{N}}(d_F(d_F+1)/2)M_{F,ZW}(d_F)}{n\tau(n)^2} + \iota(d_F, n).$$

Hence, we have

$$\mathbb{P}\left(T \geq q_{G_\Omega|\mathbf{Z}}(1-\alpha) + 2\zeta(n)\right) < \alpha + \alpha_B(n). \qquad \square$$

Appendix B: Details on the Other Moment Models

B.1. **Endogenous IVs.** A more detailed presentation is in (v5), where we allow for the possibility of using additional moments. The C-STIV in the formulation of this paper imposes $c < 1$. The additional moments allow to have $c \geq 1$ when some regressors are exogenous and there are no endogenous IVs (*i.e.*, $d_{EX} = d_Z$). These additional moments are readily assumed when we maintain Assumption A.1.

B.1.1. *Analysis of C-STIV.* The class $\mathcal{P}$ corresponds to any of classes 1-4 where $\mathbf{Z}_{i,l}\mathbf{U}_i(\beta) - \theta_l$ plays the role of $\mathbf{Z}_{i,l}\mathbf{U}_i(\beta)$. To obtain rates of convergence, the class is restricted in a similar manner as in Assumption A.2, replacing $\mathbb{P}\left(\left|D_Z\mathbf{Z}^\top\right|_\infty > B(n, d_Z)\right) \leq \alpha_\infty(n)$ by $\mathbb{P}\left(\widehat{\rho}_{ZX} > \rho_{ZX}(n)\right) \leq \alpha_\infty(n)$, where $\rho_{ZX}(n)$ depends on n via d_Z and d_X, and (C5.ii) by, for $M'_{ZU,NV}(d_Z) > 0$, for all $(\beta,\theta), \mathbb{P}$ such that $(\beta,\theta) \in \mathcal{I}$,

$$\mathbb{E}\left[\left|\left(\left(Z_lU(\beta) - \theta_l\right)^2/\sigma^2_{Z_lU(\beta)-\theta_l} - 1\right)_{l=1}^{d_Z}\right|_\infty\right] \leq M'_{ZU,NV}(d_Z).$$

For simplicity, we still refer to this assumption as Assumption A.2 and use

$$\alpha_C(n) = \alpha_\infty(n) + C_{\mathrm{N}}(d_Z)\left(M'_{ZU,NV}(d_Z) + M'_Z(d_Z)\right)/(n\tau(n)^2)$$
$$\alpha_D(n) = \left(C_{\mathrm{N}}(d_X)M_X(d_X) + C_{\mathrm{N}}(d_Zd_X)M_\Psi(d_Z,d_X)\right)/(n\tau(n)^2) + \alpha_B(n) + \alpha_C(n).$$

For $(\beta,\theta) \in \mathcal{I}$, we work with

$$\mathcal{G} \triangleq \left\{\max_{l\in[d_Z]} \frac{\left|\mathbb{E}_n\left[Z_lU(\beta) - \theta_l\right]\right|}{\sqrt{\mathbb{E}_n\left[\left(Z_lU(\beta) - \theta_l\right)^2\right]}} \leq r_0(n)\right\}$$

and $r_0(n)$ is defined as in Section A.1.1. The cones, for $S \subseteq [d_X]$ and $\widetilde{S} \subseteq [d_Z]$, are given in Table 8. Denote by $m(\tau(n)) \triangleq \sqrt{\min\left(1/(1+\tau(n)), 1-\tau(n)\right)}$, $M(\tau(n)) \triangleq \sqrt{\max\left(1/(1-\tau(n)), 1+\tau(n)\right)}$, κ and γ the population sensitivities and their lower bounds where we replace, in the definitions of $\widehat{\kappa}$ and $\widehat{\gamma}$ and the lower bounds in Proposition 4.1, $\widehat{\Psi}$, $\widehat{K}_S$, $\widehat{K}_{\gamma,S}$, by Ψ, K_S, and $K_{\gamma,S}$. Their lower bounds are computed on the sets of Table 8 and for the deterministic bounds we simply replace $\widehat{\rho}_{ZX}$ by $\rho_{ZX}(n)$. We define similarly $\overline{\theta}_\kappa(s)$ and $\overline{\kappa}^*_{e_k}(s)$. The sensitivities, their population counterparts, and lower bounds are now indexed by two sets or two sparsity certificates. Below, we refer to Assumption A.2 for conciseness, it is indeed the suitable modification based on the elements that we have given. We omit the set coming from $\mathcal{B}$ when we write the cones for conciseness.

Proposition B.1. *Under Assumption A.2, for all $(\beta,\theta), \mathbb{P}$ such that $(\beta,\theta) \in \mathcal{I}$, $n \in \mathbb{N}$ and $c > 0$, we have, on an event $\mathcal{G}_\Psi$ of probability $1 - \alpha_D(n)$,*

$F(\beta,\theta)\sqrt{1-\tau(n)} \leq \widehat{F}(\beta,\theta) \leq F(\beta,\theta)\sqrt{1+\tau(n)}$ *(see Table 8)*;

$\forall \left(b, \widetilde{b}\right) \in \mathbb{R}^{d_X+d_Z},\ l \in \mathcal{L},\ m(\tau(n))\, l\left(D_X^{-1}b, D_Z\widetilde{b}\right) \leq l\left(\mathbf{D}_\mathbf{X}^{-1}b, \mathbf{D}_\mathbf{Z}\widetilde{b}\right) \leq M(\tau(n))l\left(D_X^{-1}b, D_Z\widetilde{b}\right)$;

$$\forall S \subseteq [d_X],\ \forall \widetilde{S} \subseteq [d_Z],\ l \in [d_Z],\ \widehat{\kappa}_{l,S,\widetilde{S}} \geq \frac{\kappa_{l,S,\widetilde{S}}}{\sqrt{1+\tau(n)}m\left(\tau(n)\right)}\left(1-\frac{\tau(n)}{\kappa_{1,[d_X],\varnothing,S,\widetilde{S}}}\right);$$

$$\widehat{\gamma}_{l,S,\widetilde{S}} \geq \frac{\gamma_{l,S,\widetilde{S}}}{\sqrt{1+\tau(n)}m\left(\tau(n)\right)}\left(1-\frac{\tau(n)}{\gamma_{1,[d_X],\varnothing,S,\widetilde{S}}}\right).$$

The lower bounds in Proposition 4.1 involving the sparsity certificates hold if we remove the hats.

The main elements of the proofs are as follows. Take $(\beta,\theta) \in \mathcal{I}$. Set $\widehat{\Delta} \triangleq \mathbf{D}_{\mathbf{X}}^{-1}\left(\widehat{\beta}-\beta\right)$ and $\widetilde{\Delta} \triangleq \mathbf{D}_{\mathbf{Z}}\left(\widehat{\theta}-\theta\right)$. Clearly, on $\mathcal{G}$, (β,θ) belongs to $\widehat{\mathcal{I}}_C\left(r_0(n),\widehat{F}\left(\beta,\theta\right)\right)$. We now work on that event. Following the arguments in the proof of Theorem 4.1, we obtain

$$\left|\widehat{\Psi}\widehat{\Delta}+\widetilde{\Delta}\right|_\infty \leq r_0(n)\left(\widehat{\sigma}+\widehat{F}\left(\beta,\theta\right)\right) \tag{B.1}$$

$$\left|\widehat{\Delta}_{S(\beta)^c\cap S_Q}\right|_1+\left|\widetilde{\Delta}_{S(\theta)^c}\right|_1 \leq \left|\widehat{\Delta}_{S(\beta)\cap S_Q}\right|_1+\left|\widetilde{\Delta}_{S(\theta)}\right|_1+c\left(\widehat{F}\left(\beta,\theta\right)-\widehat{F}\left(\widehat{\beta},\widehat{\theta}\right)\right).$$

Each function $\gamma \in \mathbb{R}^{d_X+d_Z} \to \widehat{\sigma}_l(\gamma)$ is convex and

$$w_{l*} \triangleq -\begin{pmatrix} w_l \\ \widetilde{w}_l \end{pmatrix} \mathbb{1}\Big\{\mathbb{E}_n\left[\left(Z_lU(\beta)-\theta_l\right)^2\right] \neq 0\Big\} \in \partial\widehat{\sigma}_l\left(\beta,\theta\right),$$

where

$$w_l \triangleq \frac{\mathbb{E}_n\left[XZ_l\left(Z_lU(\beta)-\theta_l\right)\right]}{\sqrt{\mathbb{E}_n\left[Z_l^2\right]\mathbb{E}_n\left[\left(Z_lU(\beta)-\theta_l\right)^2\right]}} \quad \text{and} \quad \widetilde{w}_l \triangleq \begin{pmatrix} 0 \\ \frac{\mathbb{E}_n[Z_lU(\beta)-\theta_l]}{\sqrt{\mathbb{E}_n\left[Z_l^2\right]\mathbb{E}_n\left[\left(Z_lU(\beta)-\theta_l\right)^2\right]}} \\ 0 \end{pmatrix}.$$

By the Cauchy-Schwarz inequality, for all $k \in [d_X]$, $(\mathbf{D}_{\mathbf{X}})_{k,k}\left|(w_l)_k\right| \leq \widehat{\rho}_{ZX}$. Taking $w_* = (w^\top,\widetilde{w}^\top)^\top$ as one of the w_{l*} for which $\widehat{\sigma}_l\left(\beta,\theta\right) = \widehat{F}\left(\beta,\theta\right)$ yields an element of $\partial\widehat{F}\left(\beta,\theta\right)$ by Lemma A.1 recalled in (v5). By definition of the subdifferential $\partial\widehat{F}\left(\beta,\theta\right)$, we have

$$\begin{aligned}\widehat{F}\left(\beta,\theta\right)-\widehat{F}\left(\widehat{\beta},\widehat{\theta}\right) &\leq w_*^\top\begin{pmatrix}\beta-\widehat{\beta}\\ \theta-\widehat{\theta}\end{pmatrix}\\ &\leq \left|\mathbf{D}_{\mathbf{X}}w\right|_\infty\left|\widehat{\Delta}\right|_1+\left|\mathbf{D}_{\mathbf{Z}}^{-1}\widetilde{w}\right|_\infty\left|\widetilde{\Delta}_{S_\perp^c}\right|_1\\ &\leq \widehat{\rho}_{ZX}\left|\widehat{\Delta}\right|_1+r_0(n)\left|\widetilde{\Delta}_{S_\perp^c}\right|_1. \end{aligned} \tag{B.2}$$

As a result, we have $(\widehat{\Delta},\widetilde{\Delta}) \in \widehat{K}_{S(\beta),S(\theta)}$. Using (B.1) and (B.2), we find

$$\left|\widehat{\Psi}\widehat{\Delta}+\widetilde{\Delta}\right|_\infty \leq r_0(n)\left(2\overline{\sigma}+\widehat{\rho}_{ZX}\left|\widehat{\Delta}\right|_1+r_0(n)\left|\widetilde{\Delta}_{S_\perp^c}\right|_1\right). \tag{B.3}$$

Using the definition of the sensitivities, we obtain

$$\left|\widehat{\Psi}\widehat{\Delta}+\widetilde{\Delta}\right|_\infty \leq r_0(n)\left(2\overline{\sigma}+r_0(n)^2\frac{\left|\widehat{\Psi}\widehat{\Delta}\right|_\infty}{\widehat{\kappa}_{g,S(\beta),S(\theta)}}\right) \leq 2r_0(n)\overline{\sigma}\left(1-\frac{r_0(n)^2}{\widehat{\kappa}_{g,S(\beta),S(\theta)}}\right)_+^{-1},$$

$$c\widehat{\sigma} \leq |\widehat{\Delta}_{S(\beta)\cap S_Q}|_1+\left|\widetilde{\Delta}_{S(\theta)}\right|_1+c\widehat{F}\left(\beta,\theta\right) \leq \frac{\left|\widehat{\Psi}\widehat{\Delta}+\widetilde{\Delta}\right|_\infty}{\widehat{\kappa}_{1,S(\beta)\cap S_Q,S(\theta),S(\beta),S(\theta)}}+c\widehat{F}\left(\beta,\theta\right).$$

TABLE 8. Table of correspondence for the results on the C-STIV

STIV	C-STIV
$\overline{\sigma},\ \widehat{r},\ r(n)$	$(\widehat{\sigma}+\widehat{F}(\widehat{\beta},\widehat{\theta}))/2,\ r_0(n)$
$\lvert\mathbf{D}_{\mathbf{X}}^{-1}\beta_{S^c\cap S_Q}\rvert_1$	$\lvert\mathbf{D}_{\mathbf{X}}^{-1}\beta_{S^c\cap S_Q}\rvert_1+\lvert\mathbf{D}_{\mathbf{Z}}\theta_{\widetilde{S}^c}\rvert_1$
$\lvert D_X^{-1}\beta_{S^c\cap S_Q}\rvert_1$	$\lvert D_X^{-1}\beta_{S^c\cap S_Q}\rvert_1+\lvert D_Z\theta_{\widetilde{S}^c}\rvert_1/\sqrt{(1-\tau(n))(1+\tau(n))}$
$\lvert D_X^{-1}(\widehat{\beta}-\beta)_{S_0}\rvert_q$	$\lvert D_X^{-1}(\widehat{\beta}-\beta)_{S_0}\rvert_q+\lvert D_Z(\widehat{\theta}-\theta)_{\widetilde{S}_0}\rvert_q/\sqrt{(1-\tau(n))(1+\tau(n))}$
$\widehat{K}_S$	$\widehat{K}_{S,\widetilde{S}}\triangleq\left\{\begin{array}{l}(\Delta,\widetilde{\Delta}):\ \ \Delta_{S^c\cap S(\widehat{\beta})^c}=0,\ \widetilde{\Delta}_{\widetilde{S}^c\cap S(\widehat{\theta})^c}=0,\\ \lvert\Delta_{S^c\cap S_Q}\rvert_1+\lvert\widetilde{\Delta}_{\widetilde{S}^c}\rvert_1\le\lvert\Delta_{S\cap S_Q}\rvert_1+\lvert\widetilde{\Delta}_{\widetilde{S}}\rvert_1+c(\widehat{\rho}_{ZX}\lvert\Delta\rvert_1+r_0(n)\lvert\widetilde{\Delta}_{S_\perp^c}\rvert_1)\end{array}\right\}$
$\widehat{K}_{\gamma,S}$	$\widehat{K}_{\gamma,S,\widetilde{S}}\triangleq\left\{\begin{array}{l}(\Delta,\widetilde{\Delta}):\ \lvert\Delta_{S^c\cap S_Q}\rvert_1+\lvert\widetilde{\Delta}_{\widetilde{S}^c}\rvert_1\\ \le 2(\lvert\Delta_{S\cap S_Q}\rvert_1+\lvert\widetilde{\Delta}_{\widetilde{S}}\rvert_1+c(\widehat{\rho}_{ZX}\lvert\Delta\rvert_1+r_0(n)\lvert\widetilde{\Delta}_{S_\perp^c}\rvert_1))+\lvert\Delta_{S_Q^c}\rvert_1\end{array}\right\}$
K_S	$K_{S,\widetilde{S}}\triangleq\left\{\ (\Delta,\widetilde{\Delta}):\ (1_n-c\rho_{ZX}(n))\lvert\Delta\rvert_1+(1-r_0(n))\lvert\widetilde{\Delta}_{S_\perp^c}\rvert_1\le 2\lvert\Delta_{S\cap S_Q}\rvert_1+\lvert\Delta_{S_Q^c}\rvert_1+2\lvert\widetilde{\Delta}_{\widetilde{S}}\rvert_1\ \right\}$
$K_{\gamma,S}$	$K_{\gamma,S,\widetilde{S}}\triangleq\left\{\ (\Delta,\widetilde{\Delta}):\ (1_n-2c\rho_{ZX}(n))\lvert\Delta\rvert_1+(1-r_0(n))\lvert\widetilde{\Delta}_{S_\perp^c}\rvert_1\le 3\lvert\Delta_{S\cap S_Q}\rvert_1+2\lvert\Delta_{S_Q^c}\rvert_1+3\lvert\widetilde{\Delta}_{\widetilde{S}}\rvert_1\ \right\}$
$\widehat{\kappa}_{q,S_0,S}$	$\widehat{\kappa}_{q,S_0,\widetilde{S}_0,S,\widetilde{S}}\triangleq\min_{(\Delta,\widetilde{\Delta})\in\widehat{K}_{S,\widetilde{S}}:\ \lvert\Delta_{S_0}\rvert_q+\lvert\widetilde{\Delta}_{\widetilde{S}_0}\rvert_q=1}\lvert\widehat{\Psi}\Delta+\widetilde{\Delta}\rvert_\infty$
$\widehat{\kappa}_{\widehat{g},S}$	$\widehat{\kappa}_{\widehat{g},S,\widetilde{S}}\triangleq\min_{(\Delta,\widetilde{\Delta})\in\widehat{K}_{S,\widetilde{S}}:\ \widehat{\rho}_{ZX}\lvert\Delta\rvert_1+\lvert\widetilde{\Delta}_{S_\perp^c}\rvert_1=1}\lvert\widehat{\Psi}\Delta+\widetilde{\Delta}\rvert_\infty$
$\widehat{\gamma}_{\widehat{h},S}$	$\widehat{\gamma}_{\widehat{h},S,\widetilde{S}}\triangleq\min_{(\Delta,\widetilde{\Delta})\in\widehat{K}_{\gamma,S,\widetilde{S}}\ \min(\lvert\Delta_{S_Q}\rvert_1+\lvert\widetilde{\Delta}_{S_\perp^c}\rvert_1,\frac{1}{2}(3\lvert\Delta_{S\cap S_Q}\rvert_1+3\lvert\widetilde{\Delta}_{\widetilde{S}}\rvert_1+c(\widehat{\rho}_{ZX}\lvert\widetilde{\Delta}\rvert_1+r_0(n)\lvert\widetilde{\Delta}_{S_\perp^c}\rvert_1)+\lvert\Delta_{S_Q^c}\rvert_1))=1}\lvert\widehat{\Psi}\Delta+\widetilde{\Delta}\rvert_\infty$
$\widehat{B}(\widehat{S})$	$\widehat{B}(\widehat{S},\widetilde{S})\triangleq\left\{\begin{array}{l}-\mu\le\Delta\le\mu,\ \ -\widetilde{\mu}\le\widetilde{\Delta}\le\widetilde{\mu},\ \mu_{\widehat{S}^c\cap S(\widehat{\beta})^c}=0,\ \mu_{\widetilde{S}^c\cap S(\widehat{\theta})^c}=0,\\ (1-c\widehat{\rho}_{ZX})(\sum_{j\in S_I{}^c}\mu_j)+(1-cr_0(n))(\sum_{l\in S_\perp^c}\widetilde{\mu}_l)\\ \le 2(\sum_{j\in\widehat{S}\cap S_Q}\mu_j+\sum_{l\in\widetilde{S}}\widetilde{\mu}_l)+\sum_{j\in S_Q^c}\mu_j\end{array}\right\}$
$\widehat{B}(k)$	$\widehat{B}(k,l)\triangleq\left\{\begin{array}{l}-\mu\le\Delta\le\mu,\ \ -\widetilde{\mu}\le\widetilde{\Delta}\le\widetilde{\mu}\\ (1-c)(\sum_{j\in S_I}\mu_j)+(1-c\widehat{\rho}_{ZX})(\sum_{j\in S_I^c}\mu_j)+(1-cr_0(n))(\sum_{l\in S_\perp^c}\widetilde{\mu}_l)\\ \le 2(s\mu_k+\widetilde{s}\widetilde{\mu}_l)+\sum_{j\in S_Q^c}\mu_j\end{array}\right\}$
$\widehat{\theta}_\kappa(\widehat{S}),\widehat{\theta}_\kappa(s)$	$\widehat{\theta}_\kappa(\widehat{S},\widetilde{S}),\widehat{\theta}_\kappa(s,\widetilde{s})$
$\Theta_\kappa(S)$	$\Theta_\kappa(S,\widetilde{S})\triangleq\sqrt{1+\tau(n)}\delta(n)\left(1-\frac{\tau(n)}{\kappa_{1,[d_X],\varnothing,S,\widetilde{S}}}-\frac{r_0(n)\delta(n)}{c\kappa_{1,S\cap S_Q,\widetilde{S},S,\widetilde{S}}}\right)_+^{-1}$
$\Theta_\gamma(S)$	$\Theta_\gamma(S)\triangleq\sqrt{1+\tau(n)}\delta(n)\left(1-\frac{\tau(n)}{\gamma_{1,[d_X],\varnothing,S,\widetilde{S}}}-\frac{r_0(n)\delta(n)}{c\gamma_{h,S,\widetilde{S}}}\right)_+^{-1}$
$\sigma_{U(\beta)}$	$F(\beta,\theta)\triangleq\frac{1}{\sqrt{1-\tau(n)}}\max_{l\in[d_Z]}(D_Z)_{l,l}\sigma_{Z_lU(\beta)-\theta_l}$
$\widehat{\beta}^\omega$	$\widehat{\beta}_k^\omega\triangleq\widehat{\beta}_k\mathbb{1}\left\{\lvert\widehat{\beta}_k\rvert>\widehat{\omega}_k(s,\widetilde{s})/\sqrt{\mathbb{E}_n[X_k^2]}\right\},\ \widehat{\theta}_l^\omega\triangleq\widehat{\theta}_l\mathbb{1}\left\{\lvert\widehat{\theta}_l\rvert>\widehat{\omega}_l(s,\widetilde{s})\sqrt{\mathbb{E}_n[Z_l^2]}\right\}$
$\widehat{\omega}_k(s)$	$\widehat{\omega}_k(s,\widetilde{s})\triangleq 2r_0(n)\overline{\sigma}\widehat{\theta}_\kappa(s,\widetilde{s})/\widehat{\kappa}^*_{e_k,0}(s,\widetilde{s}),\ \widehat{\widetilde{\omega}}_l(s,\widetilde{s})\triangleq 2r_0(n)\overline{\sigma}\widehat{\theta}_\kappa(s,\widetilde{s})/\widehat{\kappa}^*_{0,f_l}(s,\widetilde{s})$
$\omega_k(s)$	$\omega_k(s,\widetilde{s})\triangleq\frac{2r_0(n)\sqrt{1+\tau(n)}\overline{\theta}_\kappa(s,\widetilde{s})}{\overline{\kappa}^*_{e_k,0}(s,\widetilde{s})}\sup_{(\beta,\theta)\in\mathcal{I}_s}F(\beta,\theta)\left(2\left(1-\frac{r_0(n)\delta(n)}{c\kappa_{1,S(\beta)\cap S_Q,S(\beta),S(\theta)}}\left(1-\frac{\tau(n)}{\kappa_{1,[d_X],\varnothing,S(\beta),S(\theta)}}\right)_+^{-1}\right)_+^{-1}-1\right)$
	$\widetilde{\omega}_l(s,\widetilde{s})$ obtained by replacing $\overline{\kappa}^*_{e_k,0}(s,\widetilde{s})$ by $\overline{\kappa}^*_{0,f_l}(s,\widetilde{s})$
	$\delta(n)\triangleq\sqrt{1+\tau(n)}m(\tau(n))$.

For nonsparse vectors, $S\subseteq[d_X]$, and $\widetilde{S}\subseteq[d_Z]$, we obtain

$$\left|\widehat{\Delta}_{S^c\cap S_Q}\right|_1+\left|\widetilde{\Delta}_{\widetilde{S}^c}\right|_1\le\left|\widehat{\Delta}_{S\cap S_Q}\right|_1+\left|\widetilde{\Delta}_{\widetilde{S}}\right|_1+c\left(\widehat{\rho}_{ZX}\left|\widehat{\Delta}\right|_1+r_0(n)\left|\widetilde{\Delta}_{S_\perp^c}\right|_1\right)+2\left|\mathbf{D}_{\mathbf{X}}^{-1}\beta_{S^c\cap S_Q}\right|_1+2\left|\mathbf{D}_{\mathbf{Z}}\theta_{\widetilde{S}^c}\right|_1.$$

We again consider two cases.

First, if $2|\mathbf{D}_{\mathbf{X}}^{-1}\beta_{S^c\cap S_Q}|_1+2|\mathbf{D}_{\mathbf{Z}}\theta_{\widetilde{S}^c}|_1\le|\widehat{\Delta}_{S\cap S_Q}|_1+|\widetilde{\Delta}_{\widetilde{S}}|_1+c(\widehat{\rho}_{ZX}|\widehat{\Delta}|_1+r_0(n)|\widetilde{\Delta}_{S_\perp^c}|_1)+|\widehat{\Delta}_{S_Q^c}|_1$, then

$(\widehat{\Delta},\widetilde{\Delta}) \in \widehat{K}_{\gamma,S,\widetilde{S}}$. Also, we have

$$\begin{aligned}\widehat{\sigma} &\leq \frac{1}{c}\left(\left|\mathbf{D}_{\mathbf{X}}^{-1}\beta_{S_Q}\right|_1 - \left|\mathbf{D}_{\mathbf{X}}^{-1}\widehat{\beta}_{S_Q}\right|_1 + \left|\mathbf{D}_{\mathbf{Z}}\theta_{S_\perp^c}\right|_1 - \left|\mathbf{D}_{\mathbf{Z}}\widehat{\theta}_{S_\perp^c}\right|_1\right) + \widehat{F}\left(\beta,\theta\right)\\ &\leq \frac{\left|\widehat{\Psi}\widehat{\Delta}+\widetilde{\Delta}\right|_\infty}{c\widehat{\gamma}_{\widehat{h},S,\widetilde{S}}} + \widehat{F}\left(\beta,\theta\right).\end{aligned}$$

Second, if $2|\mathbf{D}_{\mathbf{X}}^{-1}\beta_{S^c\cap S_Q}|_1 + 2|\mathbf{D}_{\mathbf{Z}}\theta_{\widetilde{S}^c}|_1 > |\widehat{\Delta}_{S\cap S_Q}|_1 + |\widetilde{\Delta}_{\widetilde{S}}|_1 + c(\widehat{\rho}_{ZX}|\widehat{\Delta}|_1 + r_0(n)|\widetilde{\Delta}_{S_\perp^c}|_1) + |\widehat{\Delta}_{S_Q^c}|_1$, then

$$\left|\widehat{\Delta}\right|_1 + \left|\widetilde{\Delta}_{S_\perp^c}\right|_1 = \left|\widehat{\Delta}_{S^c\cap S_Q}\right|_1 + \left|\widehat{\Delta}_{S\cap S_Q}\right|_1 + \left|\widehat{\Delta}_{S_Q^c}\right|_1 + \left|\widetilde{\Delta}_{\widetilde{S}^c}\right|_1 + \left|\widetilde{\Delta}_{S}\right|_1 \leq 6\left(\left|\left(\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta\right)_{S^c\cap S_Q}\right|_1 + \left|\left(\widehat{\mathbf{D}}_{\mathbf{Z}}\theta\right)_{\widetilde{S}^c}\right|_1\right).$$

For the deterministic lower bounds on the sensitivities we use that, on $\mathcal{G}_\Psi$, denoting by $\overline{\widehat{\Delta}} = D_X^{-1}\mathbf{D}_{\mathbf{X}}\widehat{\Delta}$ and $\overline{\widetilde{\Delta}} = D_Z\mathbf{D}_{\mathbf{X}}^{-1}\widetilde{\Delta}$, we have

$$\begin{aligned}\left|\widehat{\Psi}\widehat{\Delta}\right|_\infty &\geq \min_{l\in[d_Z]}\left(\mathbf{D}_{\mathbf{Z}}D_Z^{-1}\right)_{l,l}\left|D_Z\mathbb{E}_n\left[ZX^\top\right]D_X\overline{\widehat{\Delta}} + \overline{\widetilde{\Delta}}\right|_\infty\\ &\geq \frac{1}{\sqrt{1+\tau(n)}}\left(\left|D_Z\mathbb{E}\left[ZX^\top\right]D_X\overline{\widehat{\Delta}} + \overline{\widetilde{\Delta}}\right|_\infty - \left|D_Z(\mathbb{E}_n-\mathbb{E})\left[ZX^\top\right]D_X\overline{\widehat{\Delta}}\right|_\infty\right)\\ &\geq \frac{1}{\sqrt{1+\tau(n)}}\left(\left|\Psi\overline{\widehat{\Delta}} + \overline{\widetilde{\Delta}}\right|_\infty - \tau(n)\left|\overline{\widehat{\Delta}}\right|_1\right).\end{aligned}$$

The rest is easy. □

B.1.2. *The NV-STIV Estimator and Confidence Sets.* The sensitivities for NV-STIV, their lower bounds and population counterparts use $|(\widehat{\Psi}\Delta)_{S_\perp}|_\infty$ instead of $|\widehat{\Psi}\Delta|_\infty$. The one associated to the first bound in (5.1) is

$$\widehat{\kappa}_S^\Psi \triangleq \min_{\Delta\in\widehat{K}_S:\ \left|\left(\widehat{\Psi}\Delta\right)_{S_\perp^c}\right|_\infty = 1}\left|\left(\widehat{\Psi}\Delta\right)_{S_\perp}\right|_\infty.$$

For simplicity, we assume that we use the STIV estimator as a pilot estimator and denote by $\mathcal{G}_1$ the usual event $\mathcal{G}$ with $d_Z - d_{EX}$ moments, and by $\widehat{r}_1$ the constant $\widehat{r}$ under either of classes 1-5 adjusted so that $\mathbb{P}(\mathcal{G}_1) \geq 1-\alpha_1-\alpha_B(n)$. We consider that $\beta \in \mathcal{I}_s$. Using the previous results for sparse vectors, (5.1) holds[1] when $\widehat{b}$ and $\widehat{b}^\sigma$ are

$$\widehat{b} = \frac{2\widehat{r}_1\overline{\sigma}\widehat{\theta}_\kappa(s)}{\widehat{\kappa}^\Psi(s)},\ \widehat{b}^\sigma = \frac{2\widehat{r}_1\widehat{\rho}_{ZX,S_\perp^c}\overline{\sigma}\widehat{\theta}_\kappa(s)}{\widehat{\kappa}_1(s)}.$$

Definition B.1. *For $c>0$, the NV-STIV estimator $(\widehat{\theta},\widehat{\widetilde{\sigma}})$ is any solution of*

$$\min_{\theta\in\widehat{\mathcal{I}}_{NV}(\widetilde{\sigma},r_2(n)),\widetilde{\sigma}\geq 0}\left(\left|\mathbf{D}_{\mathbf{Z}}\theta_{S_\perp^c}\right|_1 + \widetilde{c\sigma}\right), \tag{B.4}$$

where, for a set of restrictions $\widetilde{\Theta}$ on θ including $\theta_{S_\perp}=0$,

$$\widehat{\mathcal{I}}_{NV}(\widetilde{\sigma},r_2(n)) \triangleq \left\{\theta\in\widetilde{\Theta}:\ \left|\mathbf{D}_{\mathbf{Z}}\left(\frac{1}{n}\mathbf{Z}^\top\left(\mathbf{Y}-\mathbf{X}\widehat{\beta}\right)-\theta\right)_{S_\perp^c}\right|_\infty \leq r_2(n)\widetilde{\sigma}+\widehat{b},\ \widehat{F}_2\left(\widehat{\beta},\theta\right)\leq\widetilde{\sigma}+\widehat{b}^\sigma\right\}$$

$$\forall\left(\beta,\theta\right)\in\mathbb{R}^{d_X+d_Z},\ \widehat{F}_2\left(\beta,\theta\right)\triangleq\max_{l\in S_\perp^c}\widehat{\sigma}_l\left(\beta,\theta\right).$$

[1](v5) considers two other cases involving estimated support.

$r_2(n)$ is obtained using class 4 so that $\mathbb{P}(\mathcal{G}_2) \geq 1 - \alpha_2 - \alpha_B(n)$, where

$$\mathcal{G}_2 \triangleq \left\{ \max_{l \in S_\perp^c} \frac{|\mathbb{E}_n [Z_l U(\beta) - \theta_l]|}{\sqrt{\mathbb{E}_n \left[(Z_l U(\beta) - \theta_l)^2 \right]}} \leq r_2(n) \right\}.$$

We make use of the following notations, for $\widetilde{s} \in [d_Z - d_{EX}]$,

$$\widehat{\omega}(\widetilde{c}, \widetilde{s}) \triangleq 2 \left(1 - \frac{2 r_2(n)^2 \widetilde{s}}{1 - \widetilde{c} r_2(n)} \right)_+^{-1} \left(r_2(n) \widehat{\widetilde{\sigma}} + \widehat{b} + r_2(n) \left(1 + \frac{\widetilde{c} r_2(n)}{1 - \widetilde{c} r_2(n)} \right) \widehat{b}^\sigma \right),$$

$$\omega(\widetilde{c}, \widetilde{s}) \triangleq 2 \left(1 - 2 r_2(n) \widetilde{s} \left(\frac{1}{1 - \widetilde{c} r_2(n)} + \frac{1}{\widetilde{c}} \right) \right)_+^{-1} \left(r_2(n) \sqrt{1 + \tau(n)} F(\beta, \theta) + b_* + r_2(n) \left(1 + \frac{\widetilde{c} r_2(n)}{1 - \widetilde{c} r_2(n)} \right) b_*^\sigma \right),$$

where b_* and b_*^σ are the following deterministic upper bounds on $\widehat{b}$ and $\widehat{b}^\sigma$ on the event $\mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}_\Psi$:

$$b_* = \frac{2 r_1(n) \overline{\theta}_\kappa(s)}{\overline{\kappa}^\Psi(s)} \sup_{\beta \in \mathcal{I}_s} \left(\sigma_{U(\beta)} \Theta_\kappa(S(\beta)) \right), \; b_*^\sigma = \frac{2 r_1(n) \rho_{ZX, S_\perp^c}(n) \overline{\theta}_\kappa(s)}{\overline{\kappa}_1(s)} \sup_{\beta \in \mathcal{I}_s} \left(\sigma_{U(\beta)} \Theta_\kappa(S(\beta)) \right).$$

We continue to use the notation $\mathcal{G}_\Psi$ to denote the event on which we can relate random quantities to deterministic quantities. Its formal definition can be obtained with now obvious modifications. Recall that $\mathbb{P}(\mathcal{G}_\Psi^c)$ appears in the coverage error so we simply choose α_1 and α_2 so that $\alpha_1 + \alpha_2 = \alpha$.

Theorem B.1. *Let $\widetilde{s} \in [d_Z - d_{EX}]$. For all $(\beta, \theta), \mathbb{P}$ such that $(\beta, \theta) \in \mathcal{I}_{d_Q, \widetilde{s}}$ and either of (1)-(3) holds, we have, on $\mathcal{G}_1 \cap \mathcal{G}_2$ in case (1) and $\mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}_\Psi$ in cases (2) or (3), with inequalities holding for all $\widetilde{c} \in \left(0, r_2(n)^{-1}\right)$ (and $c \in \left(0, r_1(n)^{-1}\right)$ in case (1)),*

$$\left| \mathbf{D_Z} \left(\widehat{\theta} - \theta \right) \right|_\infty \leq \widehat{\omega}(\widetilde{c}, \widetilde{s}); \tag{B.5}$$

$$\left| \mathbf{D_Z} \left(\widehat{\theta} - \theta \right) \right|_1 \leq \frac{2 \widetilde{s}}{1 - \widetilde{c} r_2(n)} \widehat{\omega}(\widetilde{c}, \widetilde{s}) + \frac{2 \widetilde{c} \widehat{b}^\sigma}{1 - \widetilde{c} r_2(n)}; \tag{B.6}$$

on $\mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}_\Psi$, for all solution $(\widehat{\theta}, \widehat{\widetilde{\sigma}})$ of (B.4)*, we have, with inequalities holding for all $\widetilde{c} \in \left(0, r_2(n)^{-1}\right)$,*

$$\frac{\left| D_Z \left(\widehat{\theta} - \theta \right) \right|_\infty}{\sqrt{1 + \tau(n)}} \leq \omega(\widetilde{c}, |S(\theta)|); \tag{B.7}$$

$$\frac{\left| D_Z \left(\widehat{\theta} - \theta \right) \right|_1}{\sqrt{1 + \tau(n)}} \leq \frac{2 |S(\theta)|}{1 - \widetilde{c} r_2(n)} \omega(\widetilde{c}, |S(\theta)|) + \frac{2 \widetilde{c} b_*^\sigma}{1 - \widetilde{c} r_2(n)}. \tag{B.8}$$

If $\widetilde{c}$ and c are fixed and we restrict $\mathcal{I}_{d_Q, \widetilde{s}}$ so that $|\theta_l| > \omega(\widetilde{c}, |S(\theta)|) \sqrt{(1 + \tau(n)) \mathbb{E}[Z_l^2]}$, for all $l \in S_\perp^c$, we have, on $\mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}_\Psi$, $S(\theta) \subseteq S(\widehat{\theta})$, while, if we restrict $\mathcal{I}_{d_Q, \widetilde{s}}$ so that, for all $l \in S_\perp^c$, $|\theta_l| > 2\omega(\widetilde{c}, \widetilde{s}) \sqrt{(1 + \tau(n)) \mathbb{E}[Z_l^2]}$, then $\operatorname{sign}(\widehat{\theta}^\omega) = \operatorname{sign}(\theta)$, where $\widehat{\theta}^\omega \triangleq (\widehat{\theta}_l \mathbb{1}\{|\widehat{\theta}_l| > \sqrt{\mathbb{E}_n[Z_l^2]} \widehat{\omega}(\widetilde{c}, \widetilde{s})\})_{l=1}^{d_Z}$.

Inequalities (B.5) and (B.6) are confidence sets and the uniformity in $\widetilde{c}$ and c allows to intersect the sets that various $\widetilde{c}$ and c would produce (in practice to use random values given by a rule of thumb). The last statement yields "adaptive" confidence sets by replacing $\widetilde{s}$ by $|S(\widehat{\theta}^\omega)|$ in (B.5) and (B.6). This theorem is useful when $r_2(n)$ is small (*i.e.*, $n \gg \ln(|S_\perp^c|)$). The first upper bounds are finite if $|S(\theta)| = O\left(1/r_2(n)^2\right) = O\left(n / \ln(|S_\perp^c|)\right)$ is small enough. Bounds for ℓ_q-norms follow by interpolation.

Proof of Theorem B.1. We work on $\mathcal{G}_1 \cap \mathcal{G}_2$. First, we show that $(\theta, \widehat{F}_2(\beta,\theta)) \in \widehat{\mathcal{I}}_{NV}$ by the following computations

$$\begin{aligned}\left|\mathbf{D}_{\mathbf{Z}}\left(\frac{1}{n}\mathbf{Z}^\top\left(\mathbf{Y}-\mathbf{X}\widehat{\beta}\right)-\theta\right)_{S_\perp^c}\right|_\infty &\leq \left|\mathbf{D}_{\mathbf{Z}}\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{U}(\beta)-\theta\right)\right|_\infty + \left|\widehat{\Psi}\mathbf{D}_{\mathbf{X}}^{-1}\left(\widehat{\beta}-\beta\right)_{S_\perp^c}\right|_\infty \\ &\leq r_2(n)\widehat{F}_2\left(\beta,\theta\right)+\widehat{b}.\end{aligned}$$

The second constraint in the definition of $\widehat{\mathcal{I}}_{NV}$ is satisfied because, by the triangle inequality and convexity, $\widehat{F}_2(\widehat{\beta},\theta) \leq \widehat{F}_2\left(\beta,\theta\right)+\widehat{b}^\sigma$. Now, because $(\theta, \widehat{F}_2\left(\beta,\theta\right)) \in \widehat{\mathcal{I}}_{NV}$ and $(\widehat{\theta},\widehat{\widetilde{\sigma}})$ minimizes (B.4), we have

$$\left|\widetilde{\Delta}_{S(\theta)^c}\right|_1 \leq \left|\widetilde{\Delta}_{S(\theta)}\right|_1 + \widetilde{c}\left(\widehat{F}_2\left(\beta,\theta\right)-\widehat{\widetilde{\sigma}}\right). \tag{B.9}$$

Using that $\widehat{F}_2(\widehat{\beta},\theta) \leq \widehat{\widetilde{\sigma}}+\widehat{b}^\sigma$ (by definition of the estimator) and the computations from the proofs of the results of Section B.1, we obtain

$$\widehat{F}_2\left(\beta,\theta\right)-\widehat{\widetilde{\sigma}} \leq r_2(n)\left|\widetilde{\Delta}_{S_\perp^c}\right|_1 + 2\widehat{b}^\sigma. \tag{B.10}$$

This and (B.9) yield

$$\left|\widetilde{\Delta}_{S(\theta)^c}\right|_1 \leq \left|\widetilde{\Delta}_{S(\theta)}\right|_1 + \widetilde{c}r_2(n)\left|\widetilde{\Delta}_{S_\perp^c}\right|_1 + 2\widetilde{c}\widehat{b}^\sigma$$

and, equivalently,

$$\left|\widetilde{\Delta}_{S(\theta)^c}\right|_1 \leq \frac{1+\widetilde{c}r_2(n)}{1-\widetilde{c}r_2(n)}\left|\widetilde{\Delta}_{S(\theta)}\right|_1 + \frac{2\widetilde{c}}{1-\widetilde{c}r_2(n)}\widehat{b}^\sigma. \tag{B.11}$$

Next, using the second constraint in the definition of $(\widehat{\theta},\widehat{\widetilde{\sigma}})$, we find

$$\begin{aligned}\left|\mathbf{D}_{\mathbf{Z}}\left(\widehat{\theta}-\theta\right)\right|_\infty &\leq \left|\widehat{\mathbf{D}}_{\overline{\mathbf{Z}}}\left(\frac{1}{n}\overline{\mathbf{Z}}^\top\left(\mathbf{Y}-\mathbf{X}\widehat{\beta}\right)-\widehat{\theta}\right)_{S_\perp^c}\right|_\infty + \left|\mathbf{D}_{\mathbf{Z}}\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{U}(\beta)-\theta\right)_{S_\perp^c}\right|_\infty + \left|\mathbf{D}_{\mathbf{Z}}\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{X}\left(\widehat{\beta}-\beta\right)\right)_{S_\perp^c}\right|_\infty \\ &\leq r_2(n)\left(\widehat{\widetilde{\sigma}}+\widehat{F}_2\left(\beta,\theta\right)\right)+2\widehat{b}.\end{aligned}$$

This and (B.10) yield

$$\left|\widetilde{\Delta}\right|_\infty \leq r_2(n)\left(2\widehat{\widetilde{\sigma}}+r_2(n)\left|\widetilde{\Delta}\right|_1+2\widehat{b}^\sigma\right)+2\widehat{b}. \tag{B.12}$$

On the other hand, (B.11) implies

$$\begin{aligned}\left|\widetilde{\Delta}\right|_1 &\leq \frac{2}{1-\widetilde{c}r_2(n)}\left|\widetilde{\Delta}_{S(\theta)}\right|_1 + \frac{2\widetilde{c}\widehat{b}^\sigma}{1-\widetilde{c}r_2(n)} \\ &\leq \frac{2\left|S\left(\theta\right)\right|}{1-\widetilde{c}r_2(n)}\left|\widetilde{\Delta}\right|_\infty + \frac{2\widetilde{c}\widehat{b}^\sigma}{1-\widetilde{c}r_2(n)}.\end{aligned} \tag{B.13}$$

Inequalities (B.5) and (B.6) follow by simple manipulations of (B.12)-(B.13).
As before, we obtain

$$\widehat{\widetilde{\sigma}} \leq \frac{\left|\widetilde{\Delta}_{S(\theta)}\right|_1}{\widetilde{c}} + \sqrt{1+\tau(n)}\widehat{F}_2\left(\beta,\theta\right) \leq \frac{\left|S\left(\theta\right)\right|\left|\widetilde{\Delta}\right|_\infty}{\widetilde{c}} + \sqrt{1+\tau(n)}\widehat{F}_2\left(\beta,\theta\right), \tag{B.14}$$

which, together with (B.12)-(B.13), yield

$$\left|\mathbf{D}_{\mathbf{Z}}\left(\widehat{\theta}-\theta\right)\right|_\infty \leq 2\left(1-2r_2(n)\left|S\left(\theta\right)\right|\left(\frac{1}{1-\widetilde{c}r_2(n)}+\frac{1}{\widetilde{c}}\right)\right)_+^{-1}\left(r_2(n)\sqrt{1+\tau(n)}F\left(\beta,\theta\right)+\widehat{b}+r_2(n)\left(1+\frac{\widetilde{c}r_2(n)}{1-\widetilde{c}r_2(n)}\right)\widehat{b}^\sigma\right).$$

The rest is as before. □

B.2. **Systems of Equations with Approximation Errors.** To allow for approximation error, we modify $\mathcal{P}$, so that $\mathbf{W}(\beta)$ plays the role of $\mathbf{U}(\beta)$. For simplicity we only consider classes 1-4. We modify $\mathcal{G}$ accordingly and $\mathcal{G}_\Psi$ is defined as in the proof of Proposition A.1 replacing $\mathcal{E}_U^c$ by $\mathcal{E}_V^c \cap \mathcal{E}_W^c$, where both are defined like $\mathcal{E}_U$, and the probability $m_4(n\tau(n)^2)^{-1}$ in the definition of $\alpha_C(n)$ is replaced by $2d_G m_4(n\tau(n)^2)^{-1}$. We choose $r(n)$ and $r_0(n)$ as in Section A.1.1 replacing α by α/d_G and use the event

$$\mathcal{G} \triangleq \left\{ \max_{\substack{g\in[d_G]\\ l\in[d_Z]}} \frac{|\mathbb{E}_n\left[Z_l W_g(\beta)\right]|}{\sqrt{\mathbb{E}_n[Z_l^2]\mathbb{E}_n\left[W_g(\beta)^2\right]}} \leq r(n) \right\}.$$

We rely on a union bound because there could be dependence between the errors in each equation and we do not model them. The number of equations d_G can depend on n. The population sensitivities are obtained replacing $|\Psi\Delta|_\infty$ by $\sum_{g=1}^{d_G} |\Psi\Delta_{\cdot,g}|_\infty$, $\widehat{K}_S$ and $\widehat{K}_{\gamma,S}$ by

$$K_S \triangleq \left\{\Delta \in \mathbb{R}^{d_X} : 1_n \left|\Delta_{S^c\cap S_Q}\right|_1 \leq \left|\Delta_{S\cap S_Q}\right|_1 + cg(\Delta)\right\},$$
$$K_{\gamma,S} \triangleq \left\{\Delta \in \mathbb{R}^{d_X} : 1_n \left|\Delta_{S^c\cap S_Q}\right|_1 \leq 2\left(\left|\Delta_{S\cap S_Q}\right|_1 + cg(\Delta)\right) + \left|\Delta_{S_Q^c}\right|_1\right\},$$

where $g(\Delta) \triangleq r(\beta,n)|\Delta_{S_I}|_1 + |\Delta_{S_I^c}|_1$ and $r(\beta,n) \triangleq \max_{g\in[d_G]} \min(r(n) + (r(n)+1)(1_n\sigma_{W_g(\beta)}/v_g(d_X) - 1)_+^{-1}, 1)$ is used instead of $r(n)$ in the definition of the sensitivities. We also denote by $\Psi_X = D_X\mathbb{E}[XX^\top]D_X$ and $\mathcal{E}_X$ is defined like $\mathcal{E}_Z$ replacing Z by X of probability $C_{\mathrm{N}}(d_X(d_X+1)/2)M_X(d_X)/(n\tau(n)^2)$.

Theorem B.2. *For all $\beta, \mathbb{P}$ such that $\beta \in \mathcal{I}$, assuming as well $\mathbb{E}_n[v_g(d_X)^2] \leq \widehat{v}_g^2$ on $\mathcal{G}_\Psi$ and all solution $\left(\widehat{\beta}, \widehat{\sigma}\right)$ of (5.4), the following hold on $\mathcal{G}\cap\mathcal{G}_\Psi$ (on $\mathcal{G}\cap\mathcal{G}_\Psi\cap\mathcal{E}_X'^c$ for the second inequality of (ii))*

(i) For a sparse matrix β, for all $l \in \mathcal{L}$, we have

$$1_n l\left(D_X^{-1}\left(\widehat{\beta}-\beta\right)\right) \leq \frac{2r(n)\sum_{g=1}^{d_G}\left(\sigma_{W_g(\beta)} + (r(n)+2)\,v_g(d_X)\right)}{\kappa_{l,S(\beta)}}\Theta_\kappa(S(\beta)),$$
$$\mathbb{E}_n\left[\left(X^\top\left(\widehat{\beta}_{\cdot,g}-\beta_{\cdot,g}\right)\right)^2\right] \leq \frac{2r(n)\sum_{g=1}^{d_G}\left(\sigma_{W_g(\beta)} + (r(n)+2)\,v_g(d_X)\right)}{1_n\sqrt{\kappa_{1,[d_X]\times\{g\},S(\beta)}}}\Theta_\kappa(S(\beta));$$

(ii) For all $S, S_0 \in [d_X]^{d_G}$, and $q \in [1,\infty]$, we have

$$1_n\left|D_X^{-1}\left(\widehat{\beta}-\beta\right)_{S_0}\right|_q \leq 2\max\left(\frac{r(n)\sum_{g=1}^{d_G}\left(\sigma_{W_g(\beta)} + (r(n)+2)\,v_g(d_X)\right)}{\gamma_{q,S_0,S}}\Theta_\gamma(S), 3\left|D_X^{-1}\beta_{S^c\cap S_Q}\right|_1\right)$$
$$\mathbb{E}_n\left[\left(X^\top\left(\widehat{\beta}_{\cdot,g}-\beta_{\cdot,g}\right)\right)^2\right]^{1/2} \leq \left|D_X^{-1}\left(\widehat{\beta}_{\cdot,g}-\beta_{\cdot,g}\right)\right|_1\sqrt{|\Psi_X|_\infty + \tau(n)};$$

In all cases, we have

$$|\widehat{\sigma}_g(\widehat{\beta}) - \sigma_{W_g(\beta)}| \leq \mathbb{E}_n\left[\left(X^\top\left(\widehat{\beta}_{\cdot,g}-\beta_{\cdot,g}\right)\right)^2\right]^{1/2} + \sigma_{W_g(\beta)}\tau(n) + \sqrt{1+\tau(n)}v_g(d_X). \tag{B.15}$$

Proof of Theorem B.2. Take β in $\mathcal{I}$, set $\widehat{\Delta} \triangleq \mathbf{D}_\mathbf{X}^{-1}\left(\widehat{\beta}-\beta\right)$, and work on $\mathcal{G}\cap\mathcal{G}_\Psi$. We have, for $g \in [d_G]$, using the triangle inequality in the second and fourth display, and the definition of $\mathcal{G}$ and the Cauchy-Schwartz inequality in the third,

$$\left|\frac{1}{n}\mathbf{D}_\mathbf{Z}\mathbf{Z}^\top(\mathbf{Y}_{\cdot,g} - \mathbf{X}\beta_{\cdot,g})\right|_\infty \leq \left|\frac{1}{n}\mathbf{D}_\mathbf{Z}\mathbf{Z}^\top\mathbf{W}_{\cdot,g}(\beta)\right|_\infty + \left|\frac{1}{n}\mathbf{D}_\mathbf{Z}\mathbf{Z}^\top\mathbf{V}_{\cdot,g}(\beta)\right|_\infty$$

$$\leq r(n)\sqrt{\mathbb{E}_n[W_g(\beta)^2]} + \sqrt{\mathbb{E}_n[V_g(\beta)^2]}$$
$$\leq r(n)\widehat{\sigma}_g(\beta) + (r(n)+1)\sqrt{\mathbb{E}_n[V_g(\beta)^2]}$$
$$\leq r(n)\widehat{\sigma}_g(\beta) + (r(n)+1)\widehat{v}_g.$$

Hence, $\beta \in \widehat{\mathcal{I}}_E\left(r(n), \widehat{\sigma}(\beta)\right)$ and

$$\left|\widehat{\Psi}\widehat{\Delta}_{\cdot,g}\right|_\infty \leq r(n)\left(\widehat{\sigma}_g + \widehat{\sigma}_g(\beta)\right) + 2(r(n)+1)\sqrt{1+\tau(n)}v_g(d_X). \tag{B.16}$$

Moreover, by the inverse triangle inequality, we have

$$\widehat{\sigma}_g(\beta) \geq \sqrt{\mathbb{E}_n[W_g(\beta)^2]} - \sqrt{\mathbb{E}_n[V_g(\beta)^2]} \geq \sqrt{1-\tau(n)}\sigma_{W_g(\beta)} - \sqrt{1+\tau(n)}v_g(d_X).$$

Hence, by convexity, we have

$$\widehat{\sigma}_g(\beta) - \widehat{\sigma}_g\left(\widehat{\beta}\right) \leq \min\left(r(n) + \frac{(r(n)+1)\sqrt{1+\tau(n)}v_g(d_X)}{\left(\sqrt{1-\tau(n)}\sigma_{W_g(\beta)} - \sqrt{1+\tau(n)}v_g(d_X)\right)_+}, 1\right)|\widehat{\Delta}_{S_I}|_1 + \left|\widehat{\Delta}_{S_I^c}\right|_1$$
$$\leq \min\left(r(n) + (r(n)+1)\left(1_n\frac{\sigma_{W_g(\beta)}}{v_g(d_X)} - 1\right)_+^{-1}, 1\right)\left|\left(\widehat{\Delta}_{S_I}\right)_{\cdot,g}\right|_1 + \left|\left(\widehat{\Delta}_{S_I^c}\right)_{\cdot,g}\right|_1$$
$$\leq r(\beta,n)\left|\left(\widehat{\Delta}_{S_I}\right)_{\cdot,g}\right|_1 + \left|\left(\widehat{\Delta}_{S_I^c}\right)_{\cdot,g}\right|_1. \tag{B.17}$$

Hence we obtain the first inequality in (i).
Denoting by $\widehat{\Psi}_X \triangleq D_X\mathbb{E}_n[XX^\top]D_X$, the second inequality comes from

$$\mathbb{E}_n\left[\left(X^\top\left(\widehat{\beta}_{\cdot,g} - \beta_{\cdot,g}\right)\right)^2\right] \leq \left|\widehat{\Psi}_X\Delta_g\right|_\infty |\Delta_g|_1.$$

The third inequality comes from

$$|\widehat{\sigma}_g(\widehat{\beta}) - \sigma_{W_g(\beta)}| \leq \mathbb{E}_n\left[\left(X^\top\left(\widehat{\beta}_{\cdot,g} - \beta_{\cdot,g}\right)\right)^2\right]^{1/2} + |\sigma_{W_g(\beta)} - \widehat{\sigma}_g(\beta)|$$

and $\max\left(\sqrt{1+\tau(n)} - 1, 1 - \sqrt{1-\tau(n)}\right) \leq \tau(n)$.
Now, by definition of the estimator, we have

$$\left|\widehat{\Delta}_{S^c\cap S_Q}\right|_1 \leq \left|\widehat{\Delta}_{S\cap S_Q}\right|_1 + 2\left|\mathbf{D}_{\mathbf{X}}^{-1}\beta_{S^c\cap S_Q}\right|_1$$
$$+ c\sum_{g\in[d_G]}\left(\min\left(r(n) + (r(n)+1)\left(\frac{1_n\sigma_{W_g(\beta)}}{v_g(d_X)} - 1\right)_+^{-1}, 1\right)\left|\left(\widehat{\Delta}_{S_I}\right)_{\cdot,g}\right|_1 + \left|\left(\widehat{\Delta}_{S_I^c}\right)_{\cdot,g}\right|_1\right)$$
$$\leq \left|\widehat{\Delta}_{S\cap S_Q}\right|_1 + 2\left|\mathbf{D}_{\mathbf{X}}^{-1}\beta_{S^c\cap S_Q}\right|_1 + c\left(r(\beta,n)\left|\widehat{\Delta}_{S_I}\right|_1 + \left|\widehat{\Delta}_{S_I^c}\right|_1\right)$$

and

$$\sum_{g=1}^{d_G}\left|\widehat{\Psi}\widehat{\Delta}_{\cdot,g}\right|_\infty \leq r(n)\sum_{g=1}^{d_G}\left(\widehat{\sigma}_g + \widehat{\sigma}_g(\beta)\right) + 2(r(n)+1)\sqrt{1+\tau(n)}\sum_{g=1}^{d_G}v_g(d_X)$$
$$\leq \frac{r(n)}{c}\left(\left|\mathbf{D}_{\mathbf{X}}^{-1}\beta_{S_Q}\right|_1 - \left|\mathbf{D}_{\mathbf{X}}^{-1}\widehat{\beta}_{S_Q}\right|_1\right) + 2r(n)\sum_{g=1}^{d_G}\widehat{\sigma}_g(\beta) + 2(r(n)+1)\sqrt{1+\tau(n)}\sum_{g=1}^{d_G}v_g(d_X).$$

The second inequality from (ii) is obtained in a similar manner as in the proof of Theorem B.3. The last statement is obtained using that, by similar arguments as those leading to (A.1) and (A.15),

$$\mathbb{E}_n\left[\left(X^\top\left(\widehat{\beta}_{\cdot,g}-\beta_{\cdot,g}\right)\right)^2\right]\leq\left|D_X^{-1}\left(\widehat{\beta}_{\cdot,g}-\beta_{\cdot,g}\right)\right|_1^2|\widehat{\Psi}_X|_\infty. \tag{B.18}$$ □

The second inequalities in both items allow to bound the loss in a system of nonparametric IV equations. In a model where all the $\mathbf{V}_{i,g}(\beta)$ are zero, we take $\widehat{v}_g=0$ and can derive the same results as for the STIV estimator, including the confidence sets.

B.3. **C-STIV Estimation of Λ.** Assumption B.1 gives the class $\mathcal{P}'$ which we consider in Section 6, it allows for approximation errors as in sections 5.2 and B.2. We denote by

$$\widehat{\rho}_Z\triangleq\max_{l,l'\in[d_Z]}\left(\mathbf{D_Z}\right)_{l,l}\left(\mathbf{D_Z}\right)_{l',l'}\sqrt{\mathbb{E}_n\left[Z_l^2Z_{l'}^2\right]},$$

$$\mathcal{G}_0'\triangleq\left\{\max_{f\in[d_F],k\in[d_X]}\frac{\left|\mathbb{E}_n\left[\left(T(\Lambda)\right)_{f,k}\right]\right|}{\sqrt{\mathbb{E}_n\left[\left(T(\Lambda)_{f,k}\right)^2\right]}}\leq r_0'(n)\right\}.$$

Assumption B.1. *Let $d_X,d_Z,d_F\geq 3$. There exists $M_T(d_F,d_X),M_{ZV}(d_F),M_{ZW}(d_Z),M_{F,ZW}(d_F),q_2>0$, $B(n)\geq 1$, positive sequence $\left(\alpha_\infty(n)\right)_{n\in\mathbb{N}}$ decaying to zero, such that, for all $\mathbb{P},(\beta,\Lambda)\in\mathcal{I}_\Omega$ such that $\mathbb{P}(\beta,\Lambda)\in\mathcal{P}'$, positive sequences $\left(\alpha_\beta(n)\right)_{n\in\mathbb{N}}$, and $\left(v_{\beta,1}(n)\right)_{n\in\mathbb{N}}$ and $\left(v_{\beta,2}(n)\right)_{n\in\mathbb{N}}$ which can depend on β, for all $n\in\mathbb{N}$,*

(i) $\mathcal{P}$ is class 4, replacing α with α/d_G in the definition of $r_0(n)$, restricted in a similar manner as in Assumption A.2, assuming as well $\mathbb{P}\left(\{\widehat{\rho}_{ZX}>\rho_{ZX}(n)\}\cup\{\widehat{\rho}_Z>\rho_Z(n)\}\right)\leq\alpha_\infty(n)$;
The estimator $\widehat{\beta}$ is such that, with probability $1-\alpha_\beta(n)$ (on $\mathcal{G}\cap\mathcal{G}_\Psi$),

$$\max_{g\in[d_G]}\left|\mathbf{D}_\mathbf{X}^{-1}\left(\widehat{\beta}_{\cdot,g}-\beta_{\cdot,g}\right)\right|_1\leq v_{\beta,1}(n),\ \max_{g\in[d_G]}\sqrt{\mathbb{E}_n\left[\left(\left(D_{\Lambda ZW(\beta)}\Lambda\right)_{f,\cdot}ZX\left(\widehat{\beta}_{\cdot,g}-\beta_{\cdot,g}\right)\right)^2\right]}\leq v_{\beta,2}(n);$$

(ii) For all $(f,k)\in[d_F]\times[d_X]$, the distribution of $\left(\left(\mathbf{T}_i\left(\Lambda\right)\right)_{f,k}\right)_{i\in[n]}$ belongs to class 4;

(iii) $\mathbb{E}\left[\left|\left(\left(T(\Lambda)\right)_{f,k}^2/\sigma^2_{\left(T(\Lambda)\right)_{f,k}}-1\right)_{(f,k)\in[d_F]\times[d_X]}\right|_\infty^2\right]\leq M_T(d_F,d_X)$;

(iv) $\mathbb{E}\left[\left|\left(\left(\Lambda_{f,\cdot}ZV(\beta)\right)^2/\mathbb{E}\left[\left(\Lambda_{f,\cdot}ZV(\beta)\right)^2\right]-1\right)_{f\in[d_F]}\right|_\infty^2\right]\leq M_{ZV}(d_F)$;

(v) $\mathbb{E}\left[\left|D_{ZW(\beta)}\left(ZZ^\top W(\beta)^2-\mathbb{E}\left[ZZ^\top W(\beta)^2\right]\right)D_{ZW(\beta)}\right|_\infty^2\right]\leq M_{ZW}(d_Z)$;

(vi) $\mathbb{E}\left[\left|D_{\Lambda ZW(\beta)}\Lambda\left(ZZ^\top W(\beta)^2-\mathbb{E}\left[ZZ^\top W(\beta)^2\right]\right)\Lambda^\top D_{\Lambda ZW(\beta)}\right|_\infty^2\right]\leq M_{F,ZW}(d_F)$;

(vii) $\max\left(\mathbb{E}\left[\left(\left(D_{\Lambda ZW(\beta)}\Lambda\right)_{f,\cdot}ZW(\beta)\right)^{2+q_1}\right],\mathbb{E}\left[\left(\left(D_{\Lambda ZW(\beta)}\Lambda\right)_{f,\cdot}ZW(\beta)E\right)^{2+q_1}\right]\right)\leq B(n)^{q_1},\forall f\in[d_F],q_1\in[2]$;

(viii) $\max\left(\mathbb{E}\left[\left(\left|D_{\Lambda ZW(\beta)}\Lambda\mathbf{Z}_{i,\cdot}^\top\mathbf{W}_i(\beta)\right|_\infty/B(n)\right)^{q_2}\right],\mathbb{E}\left[\left(\left|D_{\Lambda ZW(\beta)}\Lambda\mathbf{Z}_{i,\cdot}^\top\mathbf{W}_i(\beta)\mathbf{E}_i\right|_\infty/B(n)\right)^{q_2}\right]\right)\leq 2,\forall i\in[n]$;

The cones for the population sensitivities used to establish the rate of convergence of $\widehat{\Lambda}$ are

$$K_S'\triangleq\left\{\Delta'\in\mathcal{M}_{d_F,d_Z}:1_n\left|\Delta'_{S^c}\right|_1\leq\frac{1+\lambda}{1-\lambda}\left|\Delta'_S\right|_1\right\},\ K_{\gamma,S}'\triangleq\left\{\Delta'\in\mathcal{M}_{d_F,d_Z}:1_n\left|\Delta'_{S^c}\right|_1\leq\frac{2+\lambda}{1-\lambda}\left|\Delta'_S\right|_1\right\}.$$

We use κ',γ' to denote the population sensitivities using the cones above, which are defined identically to κ,γ, replacing $|\Psi\Delta|_\infty$ with $|\Delta'\Psi^\top|_\infty$. Since $\widehat{\Lambda}$ can have more than one column, its analysis requires

the use of mixed norms. We use $|\cdot|_{p,q}$ norm to denote the operator norm from ℓ_p to ℓ_q. Note that $|\cdot|_{2,\infty}$ and $|\cdot|_{\infty,\infty}$ are, respectively, the maximum ℓ_2 and ℓ_1-norm of the rows. We denote the population sensitivities for mixed $\ell_p - \ell_q$ loss by $\kappa'_{(p,q),S}, \gamma'_{(p,q),S}$. We also define, for $l_{F,\infty}(\Delta') \triangleq \max_{f\in[d_F]}(D_{\Lambda ZW(\beta)})_{f,f}|\Delta'_{f,\cdot}|_1$,

$$\Theta'_\kappa(S) \triangleq (1+\tau(n))\left(1 - \frac{\tau(n)}{\kappa'_{(\infty,\infty),S}} - \frac{r'_0(n)\rho_{ZX}(n)(1+\tau(n))}{\lambda\kappa'_{1,S,S}}\right)_+^{-1}$$

and $\Theta'_\gamma(S)$ which is obtained by replacing $\kappa'_{(\infty,\infty),S}$ and $\kappa'_{1,S,S}$ by $\gamma'_{(\infty,\infty),S}$ and $\gamma'_{h,S}$. The event $\mathcal{E}'_T$ is defined like $\mathcal{E}'_X$ for $\mathbb{E}_n[T(\Lambda)]D_X$ and has probability $C_{\mathrm{N}}(d_F d_X)M_T(d_F,d_X)/W(\beta)(n\tau(n)^2)$ and $\mathcal{E}_{ZW}$ is defined like $\mathcal{E}_Z$ replacing Z by $ZW(\beta)$ of probability $C_{\mathrm{N}}(d_Z(d_Z+1)/2)M_{ZU}(d_Z)/(n\tau(n)^2)$, and we define $F(L) \triangleq |(\sigma_{(T(L))_{f,k}}(D_X)_{k,k})_{(f,k)\in[d_F]\times[d_X]}|_\infty$ for $L \in \mathcal{M}_{d_F,d_Z}$, $\Psi_Z \triangleq D_Z\mathbb{E}[ZZ^\top]D_Z$, and $\Psi_{ZW(\beta)} \triangleq D_Z\mathbb{E}[ZZ^\top W(\beta)^2]D_Z$.

Theorem B.3. *Under Assumption B.1* (ii), (iii), (v), *and (C5.iii), for all* $(\beta,\Lambda),\mathbb{P}$ *such that* $(\beta,\Lambda) \in \mathcal{I}_\Omega$ *and all solution* $(\widehat{\Lambda},\widehat{\nu})$ *of* (6.3) *with* $\lambda \in (0,1)$, *we have, on* $\mathcal{G}'_0\cap\mathcal{G}_\Psi\cap\{\widehat{\rho}_{ZX} > \rho_{ZX}(n)\}\cap\mathcal{E}'^c_T\cap\mathcal{E}^c_{ZW}\cap\mathcal{E}^c_Z$,

(*i*) *If* Λ *is sparse, for all* $l \in \mathcal{L}$,

$$l\left(\left(\widehat{\Lambda}-\Lambda\right)D_Z^{-1}\right) \le \frac{2r'_0(n)\sqrt{1+\tau(n)}F(\Lambda)\,\Theta'_\kappa(S(\Lambda))}{1_n\kappa'_{l,S(\Lambda)}},$$

$$\widehat{\nu} \le (1+\tau(n))F(\Lambda)\left(1 + \frac{2r'_0(n)\rho_{ZX}(n)\Theta'_\kappa(S(\Lambda))}{\lambda\kappa'_{1,S(\Lambda),S(\Lambda)}}\right),$$

$$\max_{f\in[d_F]}\left(D_{\Lambda ZW(\beta)}\right)_{f,f}\mathbb{E}_n\left[\left(\left(\widehat{\Lambda}_{f,\cdot}-\Lambda_{f,\cdot}\right)ZW(\beta)\right)^2\right]^{1/2} \le 2r'_0(n)F(\Lambda)\frac{\Theta'_\kappa(S(\Lambda))}{\kappa'_{l_{F,\infty},S(\Lambda)}}\sqrt{|\Psi_{ZW(\beta)}|_\infty+\tau(n)};$$

$$\max_{f\in[d_F]}\left(D_{\Lambda ZW(\beta)}\right)_{f,f}\mathbb{E}_n\left[\left(\left(\widehat{\Lambda}_{f,\cdot}-\Lambda_{f,\cdot}\right)Z\right)^2\right]^{1/2} \le 2r'_0(n)F(\Lambda)\frac{\Theta'_\kappa(S(\Lambda))}{\kappa'_{l_{F,\infty},S(\Lambda)}}\sqrt{|\Psi_Z|_\infty+\tau(n)};$$

(*ii*) *Else,*

$$\left|\left(\widehat{\Lambda}-\Lambda\right)D_Z^{-1}\right|_1 \le \frac{2}{1_n}\min_{S\subseteq[d_F]\times[d_Z]}\max\left(\frac{r'_0(n)\sqrt{1+\tau(n)}F(\Lambda)\,\Theta'_\gamma(S)}{\gamma'_{1,S}}, \frac{3+\lambda}{1-\lambda}\left|\Lambda_{S^c}D_Z^{-1}\right|_1\right),$$

$$\widehat{\nu} \le (1+\tau(n))\left(F(\Lambda) + \frac{\rho_{ZX}(n)}{\lambda}\min_{S\subseteq[d_X]}\max\left(2F(\Lambda)\left(\left(1 - \frac{r'_0(n)\rho_{ZX}(n)(1+\tau(n))}{\lambda\gamma'_{h,S}}\left(1-\frac{\tau(n)}{\gamma'_{1,S}}\right)^{-1}\right)_+^{-1} - 1\right),\right.\right.$$
$$\left.\left.\frac{3}{2\sqrt{1+\tau(n)}}\left|\Lambda_{S^c}D_Z^{-1}\right|_1\right)\right),$$

$$\max_{f\in[d_F]}\left(D_{\Lambda ZW(\beta)}\right)_{f,f}\mathbb{E}_n\left[\left(\left(\widehat{\Lambda}_{f,\cdot}-\Lambda_{f,\cdot}\right)ZW(\beta)\right)^2\right]^{1/2}$$
$$\le \sqrt{|\Psi_{ZW(\beta)}|_\infty+\tau(n)}2\min_{S\subseteq[d_F]\times[d_Z]}\max\left(\frac{r'_0(n)\sqrt{1+\tau(n)}F(\Lambda)\,\Theta'_\gamma(S)}{\gamma'_{l_{F,\infty},S}}, \frac{3+\lambda}{1-\lambda}\left|D_{\Lambda ZW(\beta)}\right|_\infty\left|\Lambda_{S^c}D_Z^{-1}\right|_1\right),$$

and with obvious modifications a bound on $\max_{f\in[d_F]}\left(D_{\Lambda ZW(\beta)}\right)_{f,f}\mathbb{E}_n\left[\left(\left(\widehat{\Lambda}_{f,\cdot}-\Lambda_{f,\cdot}\right)Z\right)^2\right]^{1/2}$.

Proof of Theorem B.3. Take $(\beta,\Lambda)\in\mathcal{I}_\Omega$. Set $\widehat{\Delta}'\triangleq(\widehat{\Lambda}-\Lambda)\mathbf{D}_{\mathbf{Z}}^{-1}$, and $\overline{\widehat{\Delta}'}\triangleq\widehat{\Delta}'\mathbf{D}_{\mathbf{Z}}D_Z^{-1}$. Clearly, on $\mathcal{G}_0'$, Λ belongs to $\widehat{\mathcal{I}}_C\left(r_0'(n),\widehat{F}(\Lambda\right)$. We now work on the event in the statement of the theorem.
We start by proving (i). The arguments in the proof of Theorem 4.1 yield

$$\begin{aligned}\left|\widehat{\Delta}'\widehat{\Psi}^\top\right|_\infty &\le r_0'(n)\left(\widehat{\nu}+\widehat{F}\left(\Lambda\right)\right)\\ \left|\widehat{\Delta}'_{S(\Lambda)^c}\right|_1 &\le \left|\widehat{\Delta}'_{S(\Lambda)}\right|_1+\frac{\lambda}{\widehat{\rho}_{ZX}}\left(\widehat{F}\left(\Lambda\right)-\widehat{F}\left(\widehat{\Lambda}\right)\right)\end{aligned}\tag{B.19}$$

and, by those of the proof results of Section B.1, $\widehat{F}(\Lambda)-\widehat{F}(\widehat{\Lambda})\le\widehat{\rho}_{ZX}|\widehat{\Delta}'|_1$.
As a result, $\overline{\widehat{\Delta}'}\in K'_{S(\Lambda)}\subseteq\widehat{K}'_{S(\Lambda)}$ and, using the definition of $\widehat{\kappa}'_{1,S(\Lambda),S(\Lambda)}$ and of the objective function in (6.3) in the first display and (B.19) in the third display,

$$\begin{aligned}\widehat{\nu}&\le\frac{\widehat{\rho}_{ZX}\left|\widehat{\Delta}'\widehat{\Psi}^\top\right|_\infty}{\lambda\widehat{\kappa}'_{1,S(\Lambda),S(\Lambda)}}+\widehat{F}\left(\Lambda\right);\\ \widehat{\nu}+\widehat{F}\left(\Lambda\right)&\le 2\widehat{F}\left(\Lambda\right)\left(1-\frac{r_0'(n)\widehat{\rho}_{ZX}}{\lambda\widehat{\kappa}'_{1,S(\Lambda),S(\Lambda)}}\right)_+^{-1};\\ \left|\widehat{\Delta}'\widehat{\Psi}^\top\right|_\infty&\le 2r_0'(n)\widehat{F}\left(\Lambda\right)\left(1-\frac{r_0'(n)\widehat{\rho}_{ZX}}{\lambda\widehat{\kappa}'_{1,S(\Lambda),S(\Lambda)}}\right)_+^{-1}.\end{aligned}$$

Let us now show the results of item (ii). Take $S\subseteq[d_F]\times[d_Z]$. We have

$$|\widehat{\Delta}'_{S^c}|_1\le\left|\widehat{\Delta}'_S\right|_1+2\left|\Lambda_{S^c}\mathbf{D}_{\mathbf{Z}}^{-1}\right|_1+\lambda\left|\widehat{\Delta}'\right|_1$$

and distinguish the two cases:
Case 1: $2|\Lambda_{S^c}\mathbf{D}_{\mathbf{Z}}^{-1}|_1\le|\widehat{\Delta}'_S|_1$ and the rest is usual.
Case 2: $2|\Lambda_{S^c}\mathbf{D}_{\mathbf{Z}}^{-1}|_1>|\widehat{\Delta}'_S|_1$. In that case, we have

$$\left|\widehat{\Delta}'\right|_1=\left|\widehat{\Delta}'_{S^c}\right|_1+\left|\widehat{\Delta}'_S\right|_1\le 2\frac{3+\lambda}{1-\lambda}\left|\Lambda_{S^c}\mathbf{D}_{\mathbf{Z}}^{-1}\right|_1,$$

hence

$$\left|\overline{\widehat{\Delta}'}\right|_1\le 2\frac{3+\lambda}{1-\lambda}\frac{1}{1_n}\left|\Lambda_{S^c}D_Z^{-1}\right|_1.\qquad\square$$

Due to item(ii), even if multiple matrices Λ satisfy (6.1), $\widehat{\Lambda}$ can converge to a sparse solution. We denote by $\alpha_\Lambda(n)$ the probability of the complement of the event in Theorem B.3, it is independent of Λ. We also denote by $v_{\Lambda,1}(n)$, $v_{\Lambda,2}(n)$, $v_{\Lambda,3}(n)$, and $v_{F(\Lambda)}(n)$ the first (taking for l the norm $|\cdot|_{\infty,\infty}$), third, fourth, and second upper bounds on the right of (i) and (ii), they can depend on Λ such that $(\beta,\Lambda)\in\mathcal{I}_\Omega$. Define the following sequences

$$\begin{aligned}v_D(n)\triangleq\min\Big(&\sqrt{1+\tau(n)}\left(\left|D_{\Lambda ZW(\beta)}\Lambda D_Z^{-1}\right|_{\infty,\infty}+\left|D_{\Lambda ZW(\beta)}\right|_\infty v_{\Lambda,1}(n)\right)(\rho_{ZX}(n)v_{\beta,1}(n)\\ &+\rho_Z\sqrt{1+\tau(n)}v(d_X))+v_{\Lambda,2}(n),\\ &\sqrt{1+\tau(n)}\left|D_{\Lambda ZU(\beta)}\right|_\infty v_{\Lambda,1}(n)\rho_{ZX}\left(v_{\beta,1}(n)+\sqrt{1+\tau(n)}\left|D_X^{-1}\beta\right|_1\right)\\ &+\sqrt{1+\tau(n)}\max_{f\in[d_F]}\sqrt{\mathbb{E}\left[\left(\left(D_{\Lambda ZW(\beta)}\Lambda\right)_{f,\cdot}ZV(\beta)\right)^2\right]}+v_{\beta,2}(n)\Big),\end{aligned}$$

$$\overline{v}_0(n) \triangleq \sqrt{n}(1+v_D(n)+\tau(n))\left(\left|D_{\Lambda ZW(\beta)}\right|_\infty r_0'(n)v_{F(\Lambda)}(n)v_{\beta,1}(n) + \left|D_{\Lambda ZW(\beta)}\right|_\infty |V(\beta)|_\infty \right.$$
$$\left. + \left(v_{\Lambda,3}(n) + (\tau(n)+1)\left|D_{\Lambda ZW(\beta)}D_{\Lambda Z}^{-1}\right|_\infty\right)|v(d_X)|_\infty\sqrt{1+\tau(n)}\right),$$

$$\zeta_2(n)^2 \triangleq \mathbb{P}\left(N_{\Omega 0} > \frac{\zeta(n)-v_D(n)\sqrt{1+\tau(n)}}{v_D(n)+\tau(n)} - v_D(n)\sqrt{1+\tau(n)}\right) + \alpha_\beta(n)+\alpha_\Lambda(n)+e^{-n\tau(n)^2/8}$$
$$+ C_{\mathrm{N}}(d_F)M_{ZV}(d_F)/(n\tau(n)^2)+\iota(d_F,n),$$

$$\zeta_2'(n) \triangleq \mathbb{P}\left(N_{\Omega 0} > \frac{\zeta(n)-v_{\Lambda,2}(n)}{v_D(n)+\tau(n)} - v_{\Lambda,2}(n)\right) + \alpha_\beta(n)+\alpha_\Lambda(n)++C_{\mathrm{N}}(d_F)M_{ZV}(d_F)/(n\tau(n)^2)+\iota(d_F,n),$$

$$\alpha_B(n) \triangleq 2\zeta_2(n)+\zeta_2'(n)+\varphi(\tau(n))+\frac{C_{\mathrm{N}}(d_F(d_F+1)/2)M_{F,ZW}(d_F)}{n\tau(n)^2}+\iota(d_F,n),$$

$$\alpha_\Omega(n) \triangleq \alpha_B(n)+\alpha_\beta(n)+C_{\mathrm{N}}(d_F)M_{ZV}(d_F)/(n\tau(n)^2)+\alpha_\Lambda(n).$$

To obtain coverage guarantees for the confidence bands, we use:

Assumption B.2. *For all $(\beta,\Lambda)\in\mathcal{I}_\Omega$, we have*

(i) $\zeta(n)/v_D(n) - \max(\sqrt{2}v_D(n), v_{\Lambda,2}(n)) \to \infty$;
(ii) $\alpha_\Omega(n)\to 0$;
(iii) $\forall n\in\mathbb{N},\ \overline{v}(n)\geq\overline{v}_0(n)$, *and* $\overline{v}(n)\to 0$.

The use of Assumption B.1 (iv) and $v_{\beta,2}(n)$ is not necessary. Using them allows the second term in the minimum in the definition of $v_D(n)$.

Let us now consider alternative confidence bands when we maintain (C5.i), defined as

$$\widehat{\underline{C}}_{\Omega,g} \triangleq \widehat{\Omega\beta}_{\cdot,g} - \widehat{q},\ ,\widehat{\overline{C}}_{\Omega,g} \triangleq \widehat{\Omega\beta}_{\cdot,g} + \widehat{q},\ \widehat{C}_{\Omega,g} \triangleq \left[\widehat{\underline{C}}_{\Omega,g}, \widehat{\overline{C}}_{\Omega,g}\right], \tag{B.20}$$
$$\widehat{q} \triangleq \frac{q_{G_\Omega|Z\widehat{\Lambda}^\top}(1-\alpha)+2\zeta(n)}{\sqrt{n}}\widehat{\sigma}_g\left(\widehat{\beta}\right)\mathbf{D}^{-1}_{\mathbf{Z}\widehat{\Lambda}^\top}1+\frac{\overline{v}(n)}{\sqrt{n}},$$

where $q_{G_\Omega|Z\widehat{\Lambda}^\top}(1-\alpha)$ is the $1-\alpha$ quantile of $G_\Omega = |\mathbf{D}_{\mathbf{Z}\widehat{\Lambda}^\top}\widehat{\Lambda}\mathbf{Z}^\top\mathbf{E}|_\infty/\sqrt{n}$ given $\mathbf{Z}\widehat{\Lambda}^\top$.
For the analysis, we add to Assumption B.1 (i)

$$\max_{g\in[d_G]}|\widehat{\sigma}_g(\widehat{\beta})-\sigma_{W_g(\beta)}| \leq v_{\sigma(W_g(\beta))}(n),$$

which is obtained from (B.15) and yields, for all $g\in[d_G]$,

$$\left|\frac{1}{\widehat{\sigma}_g(\widehat{\beta})}-\frac{1}{\sigma_{W_g(\beta)}}\right| \leq \frac{v_{\sigma(W_g(\beta))}(n)}{\sigma_{W_g(\beta)}(\sigma_{W_g(\beta)}-v_{\sigma(W_g(\beta))}(n))_+} \triangleq v_{\sigma,g}(n).$$

We also replace (v)-(viii) in Assumption B.1 by (C5.iii) and

(v') $\mathbb{E}\left[\left|D_{\Lambda Z}\Lambda\left(ZZ^\top-\mathbb{E}\left[ZZ^\top\right]\right)\Lambda^\top D_{\Lambda Z}\right|_\infty^2\right] \leq M_{F,Z}(d_F)$;
(vi') $\max\left(\mathbb{E}\left[\left(\left(D_{\Lambda Z}\Lambda\right)_{f,\cdot}ZW_g(\beta)/\sigma(W_g(\beta))\right)^{2+q_1}\right], \mathbb{E}\left[\left(\left(D_{\Lambda Z}\Lambda\right)_{f,\cdot}ZW_g(\beta)E/\sigma(W_g(\beta))\right)^{2+q_1}\right]\right) \leq B(n)^{q_1}, \forall f\in[d_F],\ q_1\in[2],\ g\in[d_G]$;
(vii') $\max\left(\mathbb{E}\left[\left(\left|D_{\Lambda Z}\Lambda\mathbf{Z}_{i,\cdot}^\top\mathbf{W}_{i,g}(\beta)\right|_\infty/(B(n)\sigma(W_g(\beta)))\right)^{q_2}\right], \mathbb{E}\left[\left(\left|D_{\Lambda Z}\Lambda\mathbf{Z}_{i,\cdot}^\top\mathbf{W}_{i,g}(\beta)\mathbf{E}_i\right|_\infty/(B(n)\sigma(W_g(\beta))\right)^{q_2}\right]\right) \leq 2,\ \forall i\in[n],\ g\in[d_G]$.

The loss $l_{F,\infty}$ is replaced by $l_{F,\infty}(\Delta') \triangleq \max_{f\in[d_F]}(D_{\Lambda Z})_{f,f}|\Delta'_{f,\cdot}|_1$ and Theorem B.3 (but the third inequalities of both items which we do not use) holds replacing $D_{\Lambda ZU(\beta)}$ by $D_{\Lambda Z}$. The coverage is guaranteed if we assume:

Assumption B.3. *For all $(\beta,\Lambda)\in\mathcal{I}_\Omega$, we have*

(i) $\tau(n)/v_{\Lambda,3}(n)\to 0$, $\zeta(n)/v_{\Lambda,3}(n)-\max(\sqrt{2}v_{\Lambda,3}(n),v_{\sigma,g}(n)(1+1/v_{\Lambda,3}(n)))\to\infty$;
(ii) $\alpha_\Omega(n)\to 0$;
(iii) $\forall n\in\mathbb{N}$, $\overline{v}(n)\geq\overline{v}_0(n)$, *and* $\overline{v}(n)\to 0$.

Item (iii) is weaker than for the previous confidence bands. Indeed, the condition of item 1 in the discussion after Theorem 6.1 simply becomes $v_{\Lambda,3}(n)=O(1)$ and does not involve any norm of Λ.

Analysis of the Bands Under (C5.i). On the event $\mathcal{E}\triangleq\mathcal{G}_0'\cap\mathcal{G}_\Psi\cap\{\widehat{\rho}_{ZX}>\rho_{ZX}(n)\}\cap\mathcal{E}_T'^c\cap\mathcal{E}_Z^c$ of probability $1-\alpha_\Lambda(n)$, we have

$$\begin{aligned}&(D_{\Lambda Z})_{f,f}\left|\sqrt{\mathbb{E}_n\left[\left(\widehat{\Lambda}_{f,\cdot}Z\right)^2\right]}-\sqrt{\mathbb{E}\left[(\Lambda_{f,\cdot}Z)^2\right]}\right|\\&\leq(D_{\Lambda Z})_{f,f}\left(\sqrt{\mathbb{E}_n\left[\left(\left(\widehat{\Lambda}_{f,\cdot}-\Lambda_{f,\cdot}\right)Z\right)^2\right]}+\left|\sqrt{\mathbb{E}_n\left[(\Lambda_{f,\cdot}Z)^2\right]}-\sqrt{\mathbb{E}\left[(\Lambda_{f,\cdot}Z)^2\right]}\right|\right)\end{aligned}$$

so $\left|\mathbf{D}_{\mathbf{Z}\widehat{\Lambda}^\top}D_{\Lambda Z}^{-1}\right|_\infty\leq 1+v_{\Lambda,3}(n)+\tau(n)$. We now use the decomposition

$$\sqrt{n}\mathbf{D}_{\mathbf{Z}\widehat{\Lambda}^\top}\left(\widehat{\Omega\beta}-\Omega\beta-V(\beta)\right)=R_1+R_2+R_3+\frac{1}{\sqrt{n}}\mathbf{D}_{\mathbf{Z}\widehat{\Lambda}^\top}\widehat{\Lambda}\mathbf{Z}^\top\mathbf{W}(\beta),\tag{B.21}$$

where $R_1\triangleq\sqrt{n}\mathbf{D}_{\mathbf{Z}\widehat{\Lambda}^\top}\left(\Omega-\frac{1}{n}\widehat{\Lambda}\mathbf{Z}^\top\mathbf{X}\right)\mathbf{D}_{\mathbf{X}}\widehat{\Delta}$, $R_2=\mathbf{D}_{\mathbf{Z}\widehat{\Lambda}^\top}\widehat{\Lambda}\mathbf{Z}^\top\mathbf{V}(\beta)/\sqrt{n}$, and $R_3=-\sqrt{n}\mathbf{D}_{\mathbf{Z}\widehat{\Lambda}^\top}V(\beta)$. On $\mathcal{E}\cap\mathcal{G}$, we have

$$\begin{aligned}|R_1|_\infty&\leq\sqrt{n}(1+v_{\Lambda,3}(n)+\tau(n))\left|D_{\Lambda ZW(\beta)}\right|_\infty r_0'(n)v_{F(\Lambda)}(n)v_{\beta,1}(n),\\|R_2|_\infty&\leq\sqrt{n}(1+v_{\Lambda,3}(n)+\tau(n))\left(v_{\Lambda,3}(n)+\tau(n)+1\right)v(d_X)\sqrt{1+\tau(n)},\\|R_3|_\infty&\leq\sqrt{n}(1+v_{\Lambda,3}(n)+\tau(n))\left|D_{\Lambda ZW(\beta)}\right|_\infty|V(\beta)|_\infty.\end{aligned}$$

Define

$$T_\Omega\triangleq\left|\frac{1}{\sqrt{n}}\sum_{i\in[n]}\mathbf{D}_{\mathbf{Z}\widehat{\Lambda}^\top}\widehat{\Lambda}\mathbf{Z}_{i,\cdot}^\top\frac{\mathbf{W}_i(\beta)}{\widehat{\sigma}(\widehat{\beta})}\right|_\infty,\ T_{\Omega1}=\left|\frac{1}{\sqrt{n}}\sum_{i\in[n]}D_{\Lambda Z}\widehat{\Lambda}\mathbf{Z}_{i,\cdot}^\top\frac{\mathbf{W}_i(\beta)}{\widehat{\sigma}(\widehat{\beta})}\right|_\infty,\ T_{\Omega0}\triangleq\left|\frac{1}{\sqrt{n}}\sum_{i\in[n]}D_{\Lambda Z}\Lambda\mathbf{Z}_{i,\cdot}^\top\frac{\mathbf{W}_i(\beta)}{\sigma_{W(\beta)}}\right|_\infty,$$

$$G_{\Omega1}\triangleq\left|\frac{1}{\sqrt{n}}\sum_{i\in[n]}D_{\Lambda Z}\widehat{\Lambda}\mathbf{Z}_{i,\cdot}^\top\mathbf{E}_i\right|_\infty,\ G_{\Omega0}\triangleq\left|\frac{1}{\sqrt{n}}\sum_{i\in[n]}D_{\Lambda Z}\Lambda\mathbf{Z}_{i,\cdot}^\top\mathbf{E}_i\right|_\infty,\ N_{\Omega0}\triangleq\left|\frac{1}{\sqrt{n}}\sum_{i\in[n]}\left(\mathbf{E}_{D_{\Lambda Z}\Lambda Z}\right)_{i,\cdot}^\top\right|_\infty,$$

where $\left(\mathbf{E}_{D_{\Lambda Z}\Lambda Z}\right)_{i,\cdot}$ are independent Gaussian vectors of covariance $\mathbb{E}[D_{\Lambda Z}\Lambda\mathbf{Z}_{i,\cdot}^\top\mathbf{Z}_{i,\cdot}\Lambda^\top D_{\Lambda Z}]$.
On $\mathcal{E}$, $|T_\Omega-T_{\Omega1}|\leq T_{\Omega1}(v_{\Lambda,3}(n)+\tau(n))$ and $|T_{\Omega1}-T_{\Omega0}|\leq v_{\sigma,g}(n)$, so

$$|T_\Omega-T_{\Omega0}|\leq(T_{\Omega0}+v_{\sigma,g}(n))(v_{\Lambda,3}(n)+\tau(n))+v_{\sigma,g}(n).$$

Also, on $\mathcal{E}\cap\{\mathbb{E}_n[E^2]\geq 1+\tau(n)\}$, we have $|G_\Omega-G_{\Omega1}|\leq G_{\Omega1}(v_{\Lambda,3}(n)+\tau(n))$ and $|G_{\Omega1}-G_{\Omega0}|\leq v_{\Lambda,3}(n)\sqrt{1+\tau(n)}$, so

$$|G_\Omega-G_{\Omega0}|\leq(G_{\Omega0}+v_{\Lambda,3}(n)\sqrt{1+\tau(n)})(v_{\Lambda,3}(n)+\tau(n))+v_{\Lambda,3}(n)\sqrt{1+\tau(n)}.$$

We obtain $\mathbb{P}\left(|T_\Omega-T_{\Omega0}|>\zeta(n)\right)\leq\zeta_2'(n)$ and $\mathbb{P}\left(\mathbb{P}\left(|G_\Omega-G_{\Omega0}|>\zeta(n)|\mathbf{Z}\widehat{\Lambda}^\top\right)>\zeta_2(n)\right)<\zeta_2(n)$, where

$$\zeta_2(n)^2\triangleq\mathbb{P}\left(N_{\Omega0}>\frac{\zeta(n)-v_{\Lambda,3}(n)\sqrt{1+\tau(n)}}{v_{\Lambda,3}(n)+\tau(n)}-v_{\Lambda,3}(n)\sqrt{1+\tau(n)}\right)+\alpha_\Lambda(n)+e^{-n\tau(n)^2/8}+\iota(d_F,n),$$

$$\zeta_2'(n) \triangleq \mathbb{P}\left(N_{\Omega 0} > \frac{\zeta(n) - v_{\sigma,g}(n)}{v_{\Lambda,3}(n) + \tau(n)} - v_{\sigma,g}(n)\right) + \alpha_\Lambda(n) + \iota(d_F, n),$$

$$\alpha_B(n) \triangleq 2\zeta_2(n) + \zeta_2'(n) + \varphi(\tau(n)) + \frac{C_{\mathrm{N}}(d_F(d_F+1)/2)M_{F,Z}(d_F)}{n\tau(n)^2} + \iota(d_F, n),$$

$$\alpha_\Omega(n) \triangleq \alpha_B(n) + \alpha_\beta(n) + \alpha_\Lambda(n),$$

and $\mathbb{P}\left(T \geq q_{G_\Omega|\mathbf{Z}}(1-\alpha) + 2\zeta(n)\right) < \alpha + \alpha_B(n)$.
Finally $\overline{v}_0(n)$ is replaced by

$$\overline{v}_0(n) \triangleq \sqrt{n}(1 + v_{\Lambda,3}(n) + \tau(n)) \Big(\left|D_{\Lambda Z W(\beta)}\right|_\infty \left(r_0'(n) v_{F(\Lambda)}(n) v_{\beta,1}(n) + |V(\beta)|_\infty\right) + (v_{\Lambda,3}(n) + \tau(n) + 1) \left|v(d_X)\right|_\infty \sqrt{1+\tau(n)} \Big).$$

Appendix C: Computational details

C.1. **SNIV Confidence Sets and Sparse BSOS.** If $\mathcal{B} = \mathbb{R}^{d_X}$ or $\mathcal{B}$ comprises polynomial (in)equality restrictions (*e.g.*, linear restrictions), the SNIV confidence set $\widehat{C}(s)$ in (3.1) is characterized by polynomial inequalities. To model the sparsity constraint we introduce $\eta \in \{0,1\}^{d_X}$ as indicators for the nonzero entries of β (see [6]), yielding the quadratic constraints

$$|S(b) \cap S_Q| \leq s \Leftrightarrow \exists \eta \in \mathbb{R}^{d_X} : \forall k \in S_Q, (1-\eta_k) b_k = 0, \eta_k(1-\eta_k) = 0, \quad \sum_{k \in S_Q} \eta_k \leq s \tag{C.1}$$

Squaring both sides, the remaining constraints in (3.1) can be rewritten as the quadratic inequality

$$\max_{l \in [d_Z]} \left(\mathbf{Z}_{\cdot,l}^\top \mathbf{U}(b)\right)^2 \leq n^2 r_0(n)^2 \widetilde{\sigma}_l(b)^2. \tag{C.2}$$

This implies that if $\varphi(b)$ is a polynomial (or more generally, a rational function), (3.3) comprises polynomial optimization problems. With possibly endogenous IVs, the SNIV confidence set $\widehat{C}(s, \widetilde{s})$ in Section 5 is characterized by replacing (C.1)-(C.2) by

$$\forall k \in S_Q, (1-\eta_k) b_k = 0, \eta_k(1-\eta_k) = 0, \quad \sum_{k \in S_Q} \eta_k \leq s; \tag{C.3}$$

$$\forall l \in S_\perp^c, (1-\delta_l) t_l = 0, \delta_l(1-\delta_l) = 0, \quad \sum_{l \in S_\perp^c} \delta_l \leq \widetilde{s}; \tag{C.4}$$

$$\max_{l \in [d_Z]} \left(\mathbf{Z}_{\cdot,l}^\top \mathbf{U}(b) - n t_l\right)^2 \leq n^2 r_0(n)^2 \widetilde{\sigma}_l(b,t)^2, \tag{C.5}$$

We now describe a general approach to convex relaxation in polynomial optimization. The main idea is to introduce new decision variables to replace monomials in the original decision variables of degree two or larger. For example, $p(x) = x_1^2 + x_1 x_2 + x_1 - x_2 - 1$ is replaced by $p(x,y) = y_{20} + y_{11} + x_1 - x_2 - 1$, where y are new decision variables with subscripts referring to the exponents of the monomials they replace. One then imposes additional constraints on x and y. There are two means of doing this. First, notice that if the constraints of the original problem are $0 \leq g_1(x) \leq 1, 0 \leq g_2(x) \leq 1 ... 0 \leq g_m(x) \leq 1$ then $\prod_{i=1}^m g_i(x)^{a_i}(1-g_i(x))^{b_i} \geq 0$ for all $a_i \in \{0,1,...d\} \forall i \in [m], b_i \in \{0,1,...d\} \forall i \in [m], \sum_{i=1}^m a_i + b_i \leq d$ and $d \in \{1,2,...\}$. In this way we can obtain additional constraints through polynomial multiplication. Equality constraints can be represented by two inequalities of opposing direction. This yields linear constraints in the decision variables x, y. For given d, the linear

constraints are given by the vector

$$h^d(x) \triangleq \left(\prod_{i=1}^{m} g_i(x)^{a_i}(1-g_i(x))^{b_i} \right)_{\substack{a_i \in \{0,1,...d\} \forall i \in [m], \\ b_i \in \{0,1,...d\} \forall i \in [m], \\ \sum_{i=1}^m a_i + b_i \leq d}} \geq 0$$

Second, one can use semidefinite constraints. If $y_{20} = x_1^2, y_{11} = x_1 x_2, y_{02} = x_2^2$ then

$$\mathbf{S}^1(x) = (1, x_1, x_2)^\top (1, x_1, x_2) = \mathbf{S}^1(x, y) = \begin{pmatrix} 1 & x_1 & x_2 \\ x_1 & y_{20} & y_{11} \\ x_2 & y_{11} & y_{02} \end{pmatrix}$$

has rank 1. The set of rank 1 matrices is not convex, and so we replace this with the set of positive semi-definite matrices. Notice that $\mathbf{S}^1(x)$ is the outer product of the vector of monomials in x of degree less than or equal to 1. More generally, we can use

$$\mathbf{S}^k(x) = (1, x_1, x_2, x_1^2, x_1 x_2, x_2^2, ..., x_2^k)^\top (1, x_1, x_2, x_1^2, x_1 x_2, x_2^2, ..., x_2^k),$$

for $k \in \mathbb{N}_+$. If the polynomial to be minimized is $f(x)$, for given k and d the optimization problem is:

$$\begin{aligned} \min_{x,y} \quad & f(x,y) \\ & h^d(x,y) \geq 0 \\ & \mathbf{S}^k(x,y) \text{ is positive semi-definite} \end{aligned}$$

A solution x^*, y^* of the level d hierarchy is a solution of the original optimization problem if $2k$ is larger than the maximum degree of the polynomials $f(x), g_1(x), ...g_m(x)$ and $\mathbf{S}^k(x^*, y^*)$ is rank 1.

To represent the SNIV confidence set we apply the Sparse BSOS hierarchy of [7] to solve the optimization problems in (3.3). Sparse BSOS is a variant of the Sum Of Squares (SOS) hierarchy of semidefinite optimization problems. It is a bounded degree hierarchy, meaning that the sequence of optimization problems is defined by holding k fixed (and small in practice) and taking $d \in \{1, 2, ...\}$. Increasing d increases the computational burden, but also provides tighter bounds. [7] show that the bounds bind as $d \to \infty$. Fixing k is computationally advantageous, since adding linear constraints is less computationally intensive than increasing the dimension of semidefinite matrices. Sparse BSOS is also a sparse hierarchy. This is not related to sparsity in the sense of few nonzero entries in a parameter, but refers to the idea that each of the polynomials $f(x), g_1(x), g_2(x), ...g_m(x)$ depends on only a few decision variables. The advantage of sparsity is that one need only impose a subset of the linear constraints $h^d(x) \geq 0$ and it permits a large semi-definite matrix $\mathbf{S}^k(x)$ to be replaced by multiple smaller semi-definite matrices. This makes the optimization problems more computationally tractable. This idea is stated precisely in the Running Intersection Property (RIP) of [7].

In our simulations, for SNIV without possibly endogeous IVs we use $f(b, \eta) = \varphi(b) = \pm b_k$ for $k \in [d_X]$ and the constraints are in (C.1)-(C.2). To make the constraints take the form $0 \leq g_1(b, \eta) \leq 1, 0 \leq g_2(b, \eta) \leq 1, ...0 \leq g_m(b, \eta) \leq 1$ we augment $\widehat{C}(s)$ to include $|b| \leq \overline{b}$, where $\overline{b}$ is a large positive constant. Since $|\eta| \leq 1$, we can then rescale the coefficients of $g(b, \eta) \geq 0$ to guarantee $0 \leq g(b, \eta) \leq 1$. We fix $k = 2$, which yields the semidefinite matrices:

$$\mathbf{S}^{2,b} = (1, b_1, b_2, ..., b_{d_X}^2)^\top (1, b_1, b_2, ..., b_{d_X}^2), \quad \mathbf{S}^{2,\eta} = (1, \eta_1, \eta_2, ..., \eta_{d_X}^2)^\top (1, \eta_1, \eta_2, ..., \eta_{d_X}^2),$$

$$\mathbf{S}^{2,b_k,\eta_k} = (1, b_k, \eta_k, b_k^2, b_k\eta_k, \eta_k^2)^\top (1, b_k, \eta_k, b_k^2, b_k\eta_k, \eta_k^2) \quad \forall k \in [d_X]$$

The implementation with possibly endogenous IVs is similar, replacing $\mathbf{S}^{2,b}$ with

$$\mathbf{S}^{2,(b,t)} = (1, b_1, b_2, ..., b_{d_X}, t_1, ..., t_{d_Z}, ..., t_{d_Z}^2)^\top (1, b_1, b_2, ..., b_{d_X}, t_1, ..., t_{d_Z}, ..., t_{d_Z}^2),$$

$$\mathbf{S}^{2,t_l,\delta_l} = (1, t_l, \delta_l, t_l^2, t_l\delta_l, \delta_l^2)^\top (1, t_l, \delta_l, t_l^2, t_l\delta_l, \delta_l^2) \quad \forall l \in [d_Z], \quad \mathbf{S}^{2,\delta} = (1, \delta_1, \delta_2, ..., \delta_{d_Z}^2)^\top (1, \delta_1, \delta_2, ..., \delta_{d_Z}^2)$$

All of our reported results are for level $d = 1$ of the hierarchy, which uses only the original constraints. To compute the *SNIV* confidence sets we use SuperSCS 1.3.2. Our implementation satisfies the RIP condition.

C.2. **Computable Lower Bounds on Sensitivities.** In Table 9 we present the most computationally tractable lower bounds on the sensitivities, which are applicable even if d_X and d_Z are large. We omit the sets coming from $\mathcal{B}$ for conciseness (see (v5)). If $\mathcal{B} = \mathbb{R}^{d_X}$ or comprises linear (in)equality constraints, the programs are LPs. The set $\widehat{S}$ below is such that $S \subseteq \widehat{S} \subseteq [d_X]$. To tighten the bounds in Table 9, one can specify a small set $U \subseteq [d_X]$ and include the additional constraint $\mu_j = \eta_j \Delta_j$, $\forall j \in U$ in the LPs of Table 9, where $\eta_j = \pm 1$ is the sign of Δ_j. Since the signs are unknown, one replaces $\min_{k\in[d_X]}$ with $\min_{k\in[d_X],\eta_j=\pm 1 \forall j\in U}$ in Table 9. This augments the number of linear programs by a factor of $2^{|U|}$. In our simulations we take $U = S_I^c$ to construct lower bounds based on a sparsity certificate. The design is such that $|U| = 2$. If constructing lower bounds using $\widehat{S}$ of small cardinality, we use $U = \widehat{S}$. Further details can be found in sections 5.1 and A.6 in (v5).

One can also compute the lower bounds below on $\widehat{\kappa}^*_{e_k}$ and obtain the lower bounds on the other sensitivities from them using Proposition A.2. In the cases where $S \subseteq \widehat{S} \subseteq S(\widehat{\beta})$ which we consider, we can use $\widehat{c}_\kappa(S, S(\widehat{\beta})) \leq \widehat{c}_\kappa(\widehat{S}, S(\widehat{\beta}))$ (a lot of simplifications occur) and $\widehat{S}(S, S(\widehat{\beta})) \subseteq (\widehat{S} \cap S_Q) \cup ((S_Q^c \cup S_I^c) \cap S(\widehat{\beta}))$, when $1 \leq c < \widehat{r}^{-1}$, and $\widehat{S}(S, S(\widehat{\beta})) \subseteq (\widehat{S} \cap S_Q) \cup (S_Q^c \cap S(\widehat{\beta}))$, when $c < 1$. When we assume $|S \cap S_Q| \leq s$, we have $\widehat{c}_\kappa(S, S(\widehat{\beta})) \leq \widehat{c}_\kappa(s) \triangleq \min(\widehat{c}_{>,\kappa}(s), c_{<,\kappa}(s))$, where

$$\widehat{c}_{>,\kappa}(s) \triangleq \frac{1}{(1 - c\widehat{r})_+} \left(2s + \left|S_Q^c\right| + c(1 - \widehat{r}) \left(\left|S_I^c \cap S_Q^c\right| + \min\left(\left|S_I^c \cap S_Q\right|, s + \left|S_I^c \cap S_Q \cap S\left(\widehat{\beta}\right)\right|\right)\right)\right)$$

and $c_{<,\kappa}(s) \triangleq (2s + |S_Q^c|)/(1 - c)_+$ and $\widehat{S}(S, S(\widehat{\beta})) \subseteq \overline{S}$. For example, to compute an alternative lower bound on $\widehat{\kappa}_1(\widehat{S})$, one can rely on (vi) in Proposition A.2 to obtain a lower bound on $\widehat{\kappa}_{\infty,\widehat{S}(\widehat{S},S(\widehat{\beta})),\widehat{S}}$ and multiply it by $\widehat{c}_\kappa(\widehat{S}, S(\widehat{\beta}))^{-1}$. To compute a lower bound on $\widehat{\kappa}_1(s)$, one can use $\widehat{c}_\kappa(s)^{-1}\widehat{\kappa}_\infty(s)$.

C.3. **FISTA with Partial Smoothing.** The C-STIV estimator $(\widehat{\beta}, \widehat{\theta}, \widehat{\sigma})$ is a solution to a conic program with d_Z cones. The C-STIV estimator $(\widehat{\Lambda}, \widehat{v})$ defined in (6.3) is a solution of a conic program with $d_X d_Z$ cones. If d_Z is large, interior-point based methods are not computationally tractable. For this purpose, we apply an iterative procedure. We present the algorithm for the C-STIV estimator of $(\widehat{\beta}, \widehat{\theta}, \widehat{\sigma})$, though it can be applied for $(\widehat{\Lambda}, \widehat{v})$ with minor modifications. Because $u = \min_{\sigma>0} \{\sigma + u^2/\sigma\}$, the C-STIV estimator can alternatively be obtained as a solution to a similar program as (4.4), where the minimum is over $(b, t, \sigma) \in \mathcal{B} \times (0, \infty)$, $\left|\mathbf{D}_\mathbf{X}^{-1} b_{S_Q}\right|_1$ is replaced by $|\mathbf{D}_\mathbf{X}^{-1} b_{S_Q}|_1 + |\mathbf{D}_\mathbf{Z} t_{S_\perp^c}|_1$, and the loss becomes

$$\mathcal{O}(b, t)^2 \triangleq \max_{l\in[d_Z]} (\mathbf{D}_\mathbf{Z})_{l,l}^2 \max\left(\widehat{\sigma}_l (b, t)^2, \frac{\left(\mathbb{E}_n [Z_l U(b)] - t_l\right)^2}{r_0(n)^2}\right).$$

The objective function is convex because $f(x, y) = x^2/y$ is convex on $\mathbb{R} \times (0, \infty)$. As a result when the restrictions defining $\mathcal{B}$ are a product of restrictions for β and restrictions for θ, a solution of C-STIV can be obtained by iteratively minimizing over (β, θ) and over σ.

Algorithm C.1. *Initialize at* $(\widehat{\beta}^{(0)}, \widehat{\theta}^{(0)}, \widehat{\sigma}^{(0)})$. *At iteration s, solve*

$$(\widehat{\beta}^{(s)}, \widehat{\theta}^{(s)}) \in \operatorname{argmin}_{(b,t)\in\mathcal{B}} \left(\frac{2\widehat{\sigma}^{(s-1)}}{c} \left(\left|\mathbf{D}_\mathbf{X}^{-1} b_{S_Q}\right|_1 + \left|\mathbf{D}_\mathbf{Z} t_{S_\perp^c}\right|_1\right) + \mathcal{O}(b, t)^2\right),$$

$$\widehat{\sigma}^{(s)} = \mathcal{L}\left(\widehat{\beta}^{(s)}, \widehat{\theta}^{(s)}\right),$$

TABLE 9. Lower bounds on sensitivities

$\widehat{\kappa}_\infty(\widehat{S}) \triangleq \min_{j\in[d_X]} \min_{\substack{(\Delta,\mu)\in\widehat{B}(\widehat{S})\\ \Delta_j=1,\ \mu\leq 1}} \lvert\widehat{\Psi}\Delta\rvert_\infty$	$\widehat{\kappa}_\infty(s) \triangleq \min_{j\in[d_X]} \min_{\substack{(\Delta,\mu)\in\widehat{B}(j)\\ \Delta_j=1,\ \mu\leq 1}} \lvert\widehat{\Psi}\Delta\rvert_\infty$
$\widehat{\kappa}^*_\omega(\widehat{S}) \triangleq \min_{\substack{j\in[d_X]\\ \eta=\pm 1}} \min_{\substack{(\Delta,\mu)\in\widehat{B}(\widehat{S})\\ \omega^\top\mu=1,\ \mu\leq\eta\Delta_j}} \lvert\widehat{\Psi}\Delta\rvert_\infty$	$\widehat{\kappa}^*_\omega(s) \triangleq \min_{\substack{j\in[d_X]\\ \eta=\pm 1}} \min_{\substack{(\Delta,\mu)\in\widehat{B}(j)\\ \omega^\top\mu=1,\ \mu\leq\eta\Delta_j}} \lvert\widehat{\Psi}\Delta\rvert_\infty$
$\widehat{\kappa}_1(\widehat{S}) \triangleq \min_{\substack{j\in[d_X]\\ \eta=\pm 1}} \min_{\substack{(\Delta,\mu)\in\widehat{B}(\widehat{S})\\ \sum_{k\in[d_X]}\mu_k=1,\ \mu\leq\eta\Delta_j}} \lvert\widehat{\Psi}\Delta\rvert_\infty$	$\widehat{\kappa}_1(s) \triangleq \min_{\substack{j\in[d_X]\\ \eta=\pm 1}} \min_{\substack{(\Delta,\mu)\in\widehat{B}(j)\\ \sum_{k\in[d_X]}\mu_k=1,\ \mu\leq\eta\Delta_j}} \lvert\widehat{\Psi}\Delta\rvert_\infty$
$\widehat{\kappa}_{\widehat{g}}(\widehat{S}) \triangleq \min_{\substack{j\in[d_X]\\ \eta=\pm 1}} \min_{\substack{(\Delta,\mu)\in\widehat{B}(\widehat{S})\\ \sum_{k\in S_I}\widehat{r}\mu_k+\sum_{k\in S_{I^c}}\mu_k=1,\ \mu\leq\eta\Delta_j}} \lvert\widehat{\Psi}\Delta\rvert_\infty$	$\widehat{\kappa}_{\widehat{g}}(s) \triangleq \min_{\substack{j\in[d_X]\\ \eta=\pm 1}} \min_{\substack{(\Delta,\mu)\in\widehat{B}(j)\\ \sum_{k\in S_I}\widehat{r}\mu_k+\sum_{k\in S_{I^c}}\mu_k=1,\ \mu\leq\eta\Delta_j}} \lvert\widehat{\Psi}\Delta\rvert_\infty$
$\widehat{\theta}_\kappa(\widehat{S}) \triangleq \left(1-\widehat{r}/\widehat{\kappa}_{\widehat{g}}(\widehat{S})\right)_+^{-1}$	$\widehat{\theta}_\kappa(s) \triangleq (1-\widehat{r}/\widehat{\kappa}_{\widehat{g}}(s))_+^{-1}$
$\widehat{B}(\widehat{S}) \triangleq \left\{ \begin{array}{l} -\mu\leq\Delta\leq\mu,\ \mu_{\widehat{S}^c\cap S(\widehat{\beta})^c}=0,\\ (1-c\widehat{r})\sum_{k\in S_I}\mu_k+(1-c)\sum_{k\in S_{I^c}}\mu_k\leq 2\sum_{k\in\widehat{S}\cap S_Q}\mu_k+\sum_{k\in S_Q^c}\mu_k \end{array} \right\}$	
$\widehat{B}(j) \triangleq \left\{ \begin{array}{l} -\mu\leq\Delta\leq\mu,\\ (1-c\widehat{r})\sum_{k\in S_I}\mu_k+(1-c)\sum_{k\in S_{I^c}}\mu_k\leq 2s\mu_j+\sum_{k\in S_Q^c}\mu_k \end{array} \right\}$	
$\widehat{B}_\gamma(\widehat{S}) \triangleq \left\{ \begin{array}{l} -\mu\leq\Delta\leq\mu,\\ (1-2c\widehat{r})\sum_{k\in S_I}\mu_k+(1-2c)\sum_{k\in S_{I^c}}\mu_k\leq 3\sum_{k\in\widehat{S}\cap S_Q}\mu_k+2\sum_{k\in S_Q^c}\mu_k \end{array} \right\}$	
$\widehat{B}_\gamma(j) \triangleq \left\{ \begin{array}{l} -\mu\leq\Delta\leq\mu,\\ (1-2c\widehat{r})\sum_{k\in S_I}\mu_k+(1-2c)\sum_{k\in S_{I^c}}\mu_k\leq 3s\mu_j+2\sum_{k\in S_Q^c}\mu_k \end{array} \right\}$	

then replace s by $s+1$, and iterate until convergence.

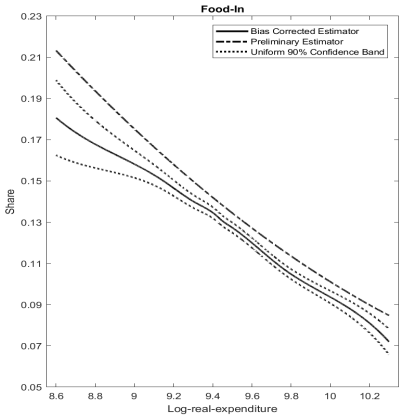
Algorithm C.1 iterates between (b,t) and σ. Since the program is convex, the iterative algorithm converges to a global minimum. Step 1 can be computationally intensive, whereas Step 2 is trivial. To solve Step 1, we use FISTA with partial smoothing ([1, 2]). Both terms in the minimization problem are convex but non-smooth; the first involves an ℓ_1-norm, and the second a maximum. The smoothing is partial because, following [2], we smooth only the maximum, for which we use log-sum-exp smoothing, replacing it with

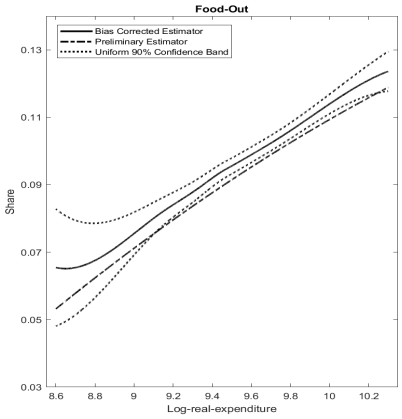
$$g_\mu(b,t) \triangleq \mu \ln\left(\sum_{l\in[d_Z]} \exp\left(\frac{\widehat{\sigma}_l^2(b,t)}{\mu}\right) + \exp\left(\frac{1}{\mu r_0(n)^2}\left((\mathbf{D}_\mathbf{Z})_{l,l}\left(\frac{1}{n}\mathbf{Z}_{\cdot,l}^\top\mathbf{U}(b)-t_l\right)\right)^2\right)\right).$$

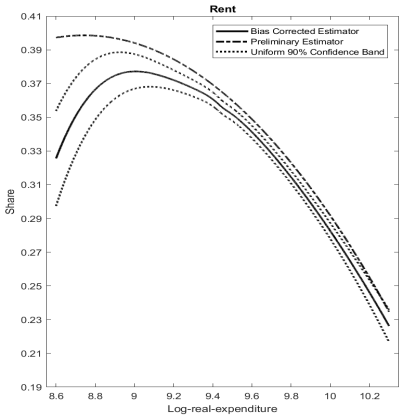
Based on Proposition 4.1 and Theorem 3.1 of [2], in practice we take $\mu=\epsilon/(2\ln 2d_Z)$ and $\epsilon=0.1$. Smaller values of ϵ improve the approximation of the maximum but increase the computational burden. After smoothing we are left with the sum of an ℓ_1-norm and a smooth function, to which we apply FISTA ([2]).

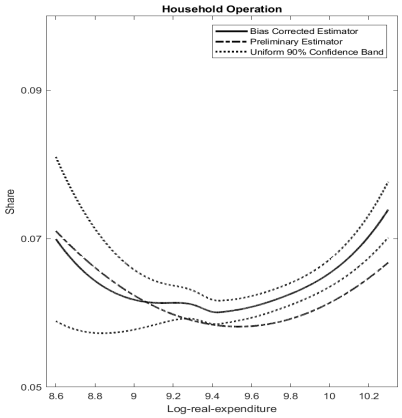
C.4. **Software.** We implement our methods in MATLAB. To compute the STIV, C-STIV and lower bounds on the sensitivities we use MOSEK version 9 (https://www.mosek.com). To compute for C-STIV estimator of Λ we use FISTA with partial smoothing ([2], https://github.com/tiepvupsu/FISTA). To compute the *SNIV* confidence sets we use SuperSCS 1.3.2 (https://kul-forbes.github.io/scs).

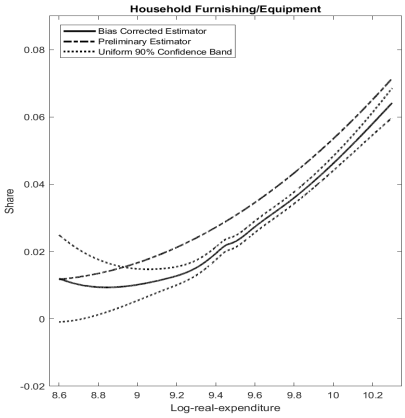
Appendix D: Engle Curves for the Remaining Goods

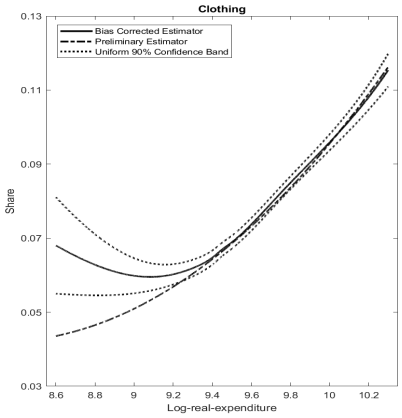


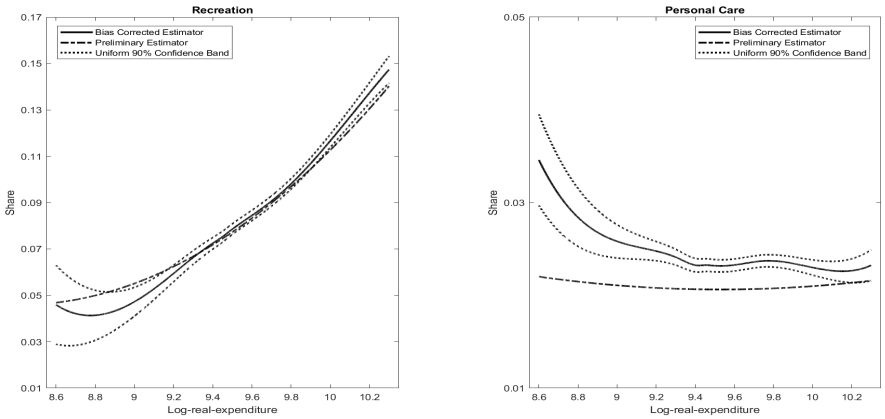












References

[1] Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* **2** 183–202.

[2] Beck, A. and Teboulle, M. (2012): Smoothing and first-order methods: a unified framework. *SIAM Journal on Optimization* **22** 557–580.

[3] Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Annals of Statistics* **41** 2786–2819.

[4] Chernozhukov, V., Chetverikov, D., and Kato, K. (2017). Central limit theorems and bootstrap in high dimensions. *Annals of Probability* **45** 2309–2352.

[5] Dümbgen, L., van de Geer, S., Veraar, M., and Wellner, J. (2010). Nemirovski's inequalities revisited. *American Mathematical Monthly* **117** 138–160.

[6] Feng, M., Mitchell, M., Pang, J.-S., Shen, X., and Wächter, A. (2014). Complementarity formulation of ℓ_0-norm optimization problems. Technical report, Northwestern University.

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