High-dimensional instrumental variables regression and confidence sets

Eric Gautier, Alexandre Tsybakov, Christiern Rose

HIGH-DIMENSIONAL INSTRUMENTAL VARIABLES REGRESSION AND CONFIDENCE SETS

ERIC GAUTIER, CHRISTIERN ROSE, AND ALEXANDRE TSYBAKOV

Abstract. This article considers inference in linear models with K regressors, some or many could be endogenous, and L instruments. L can range from less than K to any order smaller than an exponential in the sample size and K is arbitrary. For moderate K, identification robust confidence sets are obtained by solving a hierarchy of semidefinite programs. For larger K, we propose the *STIV* estimator. The analysis of its error uses sensitivity characteristics which are sharper than those in the literature on sparsity. Data-driven bounds on them and robust confidence sets are obtained by solving K linear programs. Results on rates of convergence, variable selection, and confidence sets which "adapt" to the sparsity are given. We generalize our approach to models with approximation errors, systems, endogenous instruments, and two-stage for confidence bands for vectors of linear functionals and functions. The application is to a demand system with many endogenous regressors.

1. Introduction

This article is mainly concerned with inference in the structural model

$$\forall i = 1, \dots, n,\ y_i = x_i^\top \beta + u_i; \tag{1.1}$$

$$\mathbb{E}[z_i u_i] = \mathbf{0}; \tag{1.2}$$

Date: This is the revised and resubmitted version from March 2018 with correction of typographical errors. Due to editorial demand, this revision incorporates results that were aimed to be separate papers (*e.g.*, those of Section 8 have been presented starting in 2014). First version: May 12, 2011, arXiv:1105.2454.

Keywords: Instrumental variables, sparsity, endogeneity, high-dimensional regression, heteroscedasticity, confidence sets, variable selection, unknown variance, partial identification, inference robust to identification, bias correction.

We acknowledge financial support from ERC POEMH, ANR-13-BSH1-0004, and ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047. We thank James Stock, Elie Tamer, and anonymous 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, LSE, TSE, Toulouse Paul Sabatier, UCSD, Zurich, EC2, CIREQ Montreal, Vanderbildt/Cemmap, Asian Meetings, and EcoSta conferences for comments on the results of Section 8.

$$\beta \in \mathcal{R},\ \mathbb{P}(\beta) \in \mathcal{P}; \tag{1.3}$$

where $\mathbb{E}[\cdot]$ denotes the expectation under the distribution $\mathbb{P}$ of the data $\left(y_i, x_i^\top, z_i^\top\right)_{i=1}^n$, x_i and z_i are random vectors of size K and L, and u_i is the mean-zero structural error. The terms in (1.3) are restrictions which we decompose as a set $\mathcal{R}$ for the high-dimensional parametric component and $\mathbb{P}(\beta)$ for the distribution of $\left(x_i^\top, z_i^\top, u_i(\beta)\right)_{i=1}^n$ implied by $\mathbb{P}$, where $u_i(\beta) \triangleq y_i - x_i^\top \beta$, and $\mathcal{P}$ is a nonparametric class. Some or many regressors, called endogenous - as opposed to exogenous - regressors can be correlated with u_i. Endogeneity occurs when a regressor is determined simultaneously with the response variable y_i, when u_i absorbs an unobserved variable which is partially correlated with x_i, or in the errors-in-variables model when the measurement error is independent of the underlying variable. The vector z_i in the moment conditions (1.2) comprises the instruments (also called instrumental variables or IV). Its components are all regressors known to be exogenous and, if available, exogenous variables excluded from the right-hand side of (1.1). The number of instruments L could be of any order smaller than an exponential in n. It could equally be smaller than K. When K is large, possibly much larger than n, we assume that either:

(i) only few coefficients β_k are nonzero (β is *sparse*);

(ii) β can be well approximated by a sparse vector (β is *approximately sparse*).

Estimation under the sparsity scenario when K can be much larger than n is an active and challenging field. The most studied techniques are the Lasso, the Dantzig selector, and aggregation methods. This literature proposes methods that are computationally feasible in high-dimensional settings. For example, the Lasso is a convex program as opposed to the ℓ_0-penalized least squares method, which is *NP*-hard and thus impossible to solve in practice when K is moderately large. Econometrics for high-dimensional sparse models has become an active field as well, to name a few: Belloni and Chernozhukov (2011a) studies the ℓ_1-penalized quantile regression, Belloni, Chen, Chernozhukov et al. (2012) uses Lasso type methods to estimate the optimal instrument and make inference on a low-dimensional structural equation. Caner and Fan (2014), Caner and (2014), Caner and Zhang (2015) are recent contributions to the literature on instruments and large dimensions but do not handle the high-dimensional regime. Fan and Liao (2014) consider a nonconvex approach to instrumental variables estimation. Zhu (2015) studies a type of two-stage least squares approach (henceforth, 2SLS). With regard to inference in high-dimensional models, version 1 of this paper constitutes an early reference. Recently this has been a very active field regarding inference for subvectors. The main references are Zhang and Zhang (2014), Belloni, Chernozhukov, and Hansen (2014), van de Geer, Bühlmann, Ritov, et al. (2014), and Javanmard and Montanari (2014). Recently, using a 2SLS type approach, Gold, Lederer, and Tao (2017) proposes such a solution in high-dimensional model with endogenous regressors with some infeasible choice of the tuning parameters.

We now proceed to a more detailed exposition of our approach. The set

$$\mathcal{I}dent \triangleq \{\beta \in \mathcal{R} :\ \mathbb{P}(\beta) \in \mathcal{P} \text{ and } \forall i = 1, \dots, n,\ \mathbb{E}[z_i u_i(\beta)] = \mathbf{0}\}$$

is the set of vectors compatible with (1.1)-(1.3). The dependence in n allows $\mathbb{E}[z_i y_i]$ and $\mathbb{E}\left[z_i x_i^\top\right]$ to vary with i. $\mathcal{I}dent$ is the usual identified set when they are constant. Assuming there exists a true parameter β^* which would be point identified if additional instruments or restrictions were available, β^* belongs to $\mathcal{I}dent$. If the researcher is willing to specify an upper bound s on the number of nonzero coefficients in a sub-vector of β (which could be β itself), which we call a *sparsity certificate*, it is possible to restrict attention to *sparse* identifiable parameters. Our baseline methods are agnostic on the size of the identified set or set of sparse vectors in the identified set, which would be *NP*-hard to compute if the distribution of the data were known, and on the conditional distribution of the endogenous regressors given the instruments, and hence are fully robust to weak instruments. This is achieved by NOT relying on a model involving the conditional distribution like if we used 2SLS or optimal instruments. Throughout our analysis, we put emphasis on the computational tractability which is of particular importance when K and/or L are large and propose self-tuned, also called pivotal, methods. Our main results are as follows.

We first introduce confidence sets around linear functionals of the vector of coefficients based on self-normalization. They can incorporate prior restrictions and a sparsity certificate s. Coverage is guaranteed uniformly over wide classes of data generating processes. Hence, we can draw nested confidence sets with guaranteed coverage for all values of s. The confidence sets are robust to identification and cover all parameters in the identified set which are compatible with the sparsity certificate s. They do not rely on a pretest nor on selection of the instruments nor on a non standard asymptotic analysis, and the confidence sets sometimes have finite sample validity. For moderate to large K, this approach is feasible whereas test inversion is not. Testing can be performed using the duality between confidence sets and hypothesis testing. The first confidence sets are obtained by solving nonconvex programs, which are *NP*-hard but feasible for moderately large K by solving sequences of semidefinite programs with increasing dimension. For arbitrarily large K, we propose an alternative approach and introduce the *STIV* estimator. We rely on *sensitivity characteristics*, introduced in this paper, to analyze the estimation error for arbitrary loss functions which are homogeneous of degree one. Similar quantities are used in models without endogeneity and we show that, in that case, our constants are sharper. They depend on the unknown parameter and involve solving nonconvex programs. We provide computable bounds for given sparsity certificates using convex relaxation. Convex relaxation is the key to estimation methods in high-dimensions and we use the same idea for inference. The baseline confidence sets are supersets of what one would hope to be able to compute but which is *NP*-hard. These sets cover, as many as we want, homogeneous of degree one functionals of the parameter in the identified set compatible with the sparsity certificate s, for all tuning parameter of the *STIV* estimator. The uniformity in the tuning parameter gives prescribed coverage for the intersection of sets for all values of the tuning parameter. Hence, they do not depend on a smoothing parameter. Computing the basic version of these sets rely on solving K linear programs, which is tractable even for very large K. The confidence sets sometimes have infinite volume. This is unavoidable for confidence sets which are robust to identification in the IV context (see Dufour (1997)).

Under stronger assumptions involving the conditional distribution of the endogenous regressors given the instruments, we obtain rates of convergence of the *STIV* and results on variable selection and confidence sets which are "adaptive" in the sparsity. Motivated by the empirical application in the paper, we consider variants of model (1.1)-(1.3). First, we consider the case in which, the error u_i can be written as $v_i + w_i$ where $\mathbb{E}[z_i w_i] = \mathbf{0}$ and $\mathbb{E}[v_i^2] \le E^2$. This permits the analysis of settings in which the outcome is measured in brackets, nonparametric or partially linear models, and approximations of nonlinear models. Second, we consider the generalization to systems of equations. Third, we allow some instruments to be endogenous, in which case some of the components of $\mathbb{E}[z_i u_i]$ need not be equal to zero, and do not require the knowledge of a subset of exogenous instruments. A variant can be used when such a set is available. We present two two-stage methods which use the *STIV* or one of its variants as a first stage estimator and build confidence bands for G vectors of O functionals using bias correction. Our approach can be used to obtain confidence bands around functions. The first method is demanding in terms of identification and computation, so we also propose a computationally tractable alternative that can handle thousands of regressors and a large number of functionals, and does not require some of the identification assumptions of the first. We propose to combine these bands with the data-driven error bounds on the *STIV* types estimators to obtain a bound on the bias rather than assuming the debiased estimator has negligible bias.

Our methodology is put into practice in a simulation study in which we consider situations such as many instruments, fewer instruments than regressors, and many regressors with a large fraction of endogenous ones. We apply our method to a demand system with many endogenous regressors and produce confidence bands for a system of Engel curves. Proofs and complements are in the appendix.

2. Preliminaries

2.1. **Notations.** For $K \in \mathbb{N}$, $[K]$ denotes $\{1, \dots, K\}$. We write a.s. for almost surely. For random variables $(r_i)_{i=1}^n$, the sample and population means are $\mathbb{E}_n[R] \triangleq \sum_{i=1}^n r_i/n$ and $\mathbb{E}[R] \triangleq \sum_{i=1}^n \mathbb{E}[r_i]/n$. Conditional expectations with non i.d. data are defined like population means. The sample and population means for matrices are defined entrywise. When $(r_i)_{i=1}^n$ are random vectors, $\mathbf{R}$ is the matrix where each row is $r_i^\top$ for $i \in [n]$. We use the diagonal matrices $\widehat{\mathbf{D}}_{\mathbf{X}}$ and $\widehat{\mathbf{D}}_{\mathbf{Z}}$ to rescale the regressors $\mathbf{X}$ and the instruments $\mathbf{Z}$. The diagonal entries of $\widehat{\mathbf{D}}_{\mathbf{X}}$ are $\left(\widehat{\mathbf{D}}_{\mathbf{X}}\right)_{kk} = \mathbb{E}_n\left[X_k^2\right]^{-1/2}$ for $k \in [K]$ and those of $\widehat{\mathbf{D}}_{\mathbf{Z}}$ are $\left(\widehat{\mathbf{D}}_{\mathbf{Z}}\right)_{ll} = \mathbb{E}_n\left[Z_l^2\right]^{-1/2}$ for $l \in [L]$. We write $\mathbf{D}_Z$ and $\mathbf{D}_X$ when sample means are replaced by populations means. For a mean zero random variable A, σ_A denotes $\mathbb{E}[A^2]^{1/2}$. We denote by $\mathcal{R}_D \triangleq \{\beta_1 - \beta_2, \forall \beta_1, \beta_2 \in \mathcal{R}\}$ or a manageable subset. For $\beta \in \mathcal{I}dent$, we denote by $\widehat{Q}(\beta) \triangleq \mathbb{E}_n[U(\beta)^2]$. We refer to the particular case of the linear projection model with random but exogenous regressors by writing "the case where $\mathbf{Z} = \mathbf{X}$". For $\beta \in \mathcal{M}_{O,K}$, let $J(\beta) = \{(o,k) \in [O] \times [K] : \beta_{ok} \neq 0\}$. For $J \subseteq [O] \times [K]$, $|J|$ is its cardinality and J^c its complement. We primarily use these notations when $O = 1$ and we deal with vectors in $\mathbb{R}^K$. The set $I \subseteq [K]$ consists of the indices of the regressors for which the researcher is certain that they are exogenous. The regressors with index in I are used as instruments. The set $P \subseteq [K]$ of size p corresponds to regressors for which we question relevance and is such that P^c has fixed size. The sets $P_\perp$ and $P_{\not\perp}$ are respectively $P \cap I$ and $P \cap I^c$. For $1 \le q \le \infty$, the ℓ_q-norm of a vector Δ is denoted by $|\Delta|_q$. When Δ is a matrix, the ℓ_q-norm is defined entrywise and $|\cdot|_{p,q}$ is the operator norm from ℓ_p to ℓ_q. The operator norms $|\cdot|_{2,\infty}$ and $|\cdot|_{\infty,\infty}$ are, respectively, the maximum ℓ_2 and ℓ_1-norm of the rows. We denote by $(e_k)_{k=1}^K$ the canonical basis of $\mathbb{R}^K$ and $(f_l)_{l=1}^L$ for $\mathbb{R}^L$. For $\Delta \in \mathbb{R}^K$ and $J \subseteq [K]$, we define $\Delta_J \triangleq (\Delta_k \mathbb{1}\{k \in J\})_{k=1}^K$, where $\mathbb{1}\{\cdot\}$ is the indicator function. For a vector $\beta \in \mathbb{R}^K$, we set $\overrightarrow{\text{sign}(\beta)} \triangleq (\text{sign}(\beta_k))_{k=1}^K$, where $\text{sign}(t) \triangleq \mathbb{1}\{t > 0\} - \mathbb{1}\{t < 0\}$. $\mathbf{0}$ and $\mathbf{1}$ are vectors of zeros and ones. Inequality between vectors is understood entrywise and, when M is a symmetric matrix, $M \succcurlyeq 0$ means that M is positive semidefinite. We denote by I_K the $K \times K$ identity matrix, $\mathcal{M}_{L,K}$ the set of $L \times K$ matrices with real entries, and $\mathcal{L}$ the set of functions from $\mathbb{R}^K$ to $[0, \infty)$ which are homogeneous of degree 1 (*i.e.*, $\forall \Delta \in \mathbb{R}^K, a > 0$, $l(a\Delta) = al(\Delta)$) and nondecreasing in its arguments. For a matrix Δ, $\Delta_{o\cdot}$ is the o^{th} row of Δ. For $a \in \mathbb{R}$, we set $a_+ \triangleq \max(0, a)$. We use the conventions $a/0 \triangleq \infty$ for $a > 0$, $0/0 \triangleq 0$, $1/\infty \triangleq 0$, and $\inf \emptyset \triangleq \infty$. Φ is the CDF of the standard normal. $C_{\text{N}}(m) \triangleq e(2\ln(m) - 1)$

for $m \geq 3$ is the Nemirovski constant (see, Theorem 2.2 in Dümbgen, van de Geer, Veraar et al. (2010)). Some results in this paper rely on asymptotic arguments where n goes to infinity in which case, though we do not make the dependence explicit, L, K, and s can increase with n.

2.2. **Motivating Examples.**

2.2.1. *Rich Heterogeneity.* Large datasets have become widely available and can provide rich information on the heterogeneity of the economic agents. It is possible to account for this heterogeneity by including many regressors and interactions.

2.2.2. *Growth.* Sala-i-Martin (1997) and Belloni and Chernozhukov (2011b) give examples from development economics where it is unclear which growth determinant should be included. More than 140 growth determinants have been proposed and usually n is smaller than K. Searching among 2^{140} submodels is simply impossible. Many growth determinants are arguably endogenous.

2.2.3. *Heterogeneous Treatment Effects.* The researcher is interested in determining for which group a policy has an effect in order to implement targeted measures. She interacts the endogenous treatment variable with group dummies. It results in many endogenous regressors when there are many groups. There could be sparsity if the policy has an effect on a few groups only. All treatments are equally important and assessing the effects using independent significance tests is plagued by false discovery (*e.g.*, if the policy has an effect for 10% of 400 groups and we are assessing the significance using a 5% level test, then the policy is found to have an effect for 18 groups for which it has no effect).

2.2.4. *Social Interactions with Unknown Networks.* In social interactions models with n individuals, an individual's outcome is determined simultaneously by the outcomes of their peers, K individual characteristics of her/his own and of her/his peers, and heterogeneity that is common to all individuals (correlated effects). This defines a system of simultaneous equations, in which the outcomes of others are endogenous. Assuming away correlated effects, when peers are not known, everyone could in principle affect everyone, leading to $n - 1$ endogenous regressors and Kn exogenous regressors per equation. Such models could be estimated using panel data (see Rose (2016) who applies this paper to the study of R&D spillovers, see also Gautier and Rose (2015,2017)). In this example, the objects of interest may not be subvectors but rather the whole network structure.

2.2.5. *Partially Linear Model.* One has: $\forall i \in [n],\ y_i = f(\widetilde{x}_i) + \overline{x}_i^\top \gamma + w_i$, where $\gamma \in \mathbb{R}^M$ and $f \in \mathcal{S}$ such that, for a set of functions $(\varphi_k)_{k\in\mathbb{N}}$ and a decaying sequence $(e_N)_{N\in\mathbb{N}}$,

$$\forall N \in \mathbb{N},\ \sup_{g\in\mathcal{S}} \inf_{b\in\mathbb{R}^N} \mathbb{E}\left[\left(g\left(\widetilde{x}_i\right) - \sum_{k=1}^{N} \varphi_k\left(\widetilde{x}_i\right) b_k\right)^2\right] \le e_N^2. \tag{2.1}$$

The class $\mathcal{S}$ is not a sharp class characterizing f but a large enough one so that this is considered a mild assumption It is usually achieved by assuming minimum smoothness. Taking $x_i = (\varphi_1(\widetilde{x}_i), \dots, \varphi_{K-M}(\widetilde{x}_i), w_i^\top)^\top$, $v_i = f(\widetilde{x}_i) - \sum_{k=1}^{K-M} \varphi_k(\widetilde{x}_i)\beta_k$, $(\beta_{K-M+1}, \dots, \beta_K) = \gamma^\top$, (1.1) holds with $u_i = v_i + w_i$ and $\mathbb{E}[v_i^2] \le e_{K-M}^2$. The term v_i is the approximation error made by approximating the function in the high-dimensional space. The constant E is chosen small, usually $n^{-1/2}$ (the parametric rate) or $1/n$, so $\mathbb{E}[v_i^2] \le E^2$ for K large enough. The vector β is not sparse but if the function is smooth there are usually many small coefficients and β is approximately sparse.

Now, if $\mathbb{E}\left[w_i \middle| \widetilde{x}_i, \overline{x}_i\right] \neq 0$, we are in the presence of endogeneity and rely on a vector z_i which is subvector of $z_i^\infty \in \mathbb{R}^\infty$ and contains nonlinear transformations of baseline instruments. When endogeneity is due to $\widetilde{x}_i$, the model has $K - M$ endogenous regressors. A classical situation is when f is an Engle curve. Sometimes elements of $\overline{x}_i$ could be endogenous and $(\varphi_1(\widetilde{x}_i), \dots, \varphi_{K-M}(\widetilde{x}_i))^\top$ are not. This occurs, for example, when one is interested in the parameter γ in $y_i = \overline{x}_i^\top \gamma + \widetilde{u}_i$, some variables in $\overline{x}_i$ are endogenous, and the researcher is willing to use as an instrument z_i for which $\mathbb{E}[z_i\widetilde{u}_i] \neq \mathbf{0}$ but $\mathbb{E}[\widetilde{u}_i|\widetilde{x}_i, z_i] = \mathbb{E}[\widetilde{u}_i|\widetilde{x}_i]$ for a vector of control variables $\widetilde{x}_i$. This yields $y_i = \mathbb{E}[\widetilde{u}_i|\widetilde{x}_i] + \overline{x}_i^\top\gamma + w_i$, where $w_i = \widetilde{u}_i - \mathbb{E}[\widetilde{u}_i|\widetilde{x}_i, z_i]$, $\overline{x}_i$ remains endogenous, but z_i is an instrument.

2.2.6. *Second Order Approximation of the Exact Affine Stone Index Model.* This is the empirical application of this article. The EASI model (Lewbel and Pendakur (2009)) is a model for a cost function which implies that the vector of expenditure shares $y_i \in \mathbb{R}^G$ for G goods consumed by household i satisfies

$$y_i = \sum_{r=0}^{R} b_r t_i^r + Cz + Dz_i t_i + A_0 p_i + \sum_{h=1}^{H} A_h p_i z_{hi} + Bp_i t_i + w_i; \tag{2.2}$$

$$t_i = \frac{x_i - p_i^\top y_i + p_i^\top (A_0 + \sum_{h=1}^H A_h z_{hi}) p_i/2}{1 - p_i^\top B p_i/2}; \tag{2.3}$$

where p_i, and w_i are vectors in $\mathbb{R}^G$ of log-prices, and errors, z_i are vectors of H individual characteristics and time trends. In the application, we have $G = 9$, and $R = H = 5$. The parameters are $b_r \in \mathbb{R}^G$ for $r \in [5]$, the $G \times 5$ matrices C and D, the $G \times G$ matrices $A_0, ..., A_5, B$ which satisfy restrictions to ensure that: (1) expenditure shares sum to one, (2) Slutsky symmetry, (3) monotonicity of cost, and

(4) concavity of the exponential of the cost function. For computational reasons we avoid imposing (3) but verify that it is satisfied post-estimation. This yields, using $\mathbb{S}_z$ to denote the support of z,

$$\mathbf{1}^\top b_0 = 1,\ \mathbf{1}^\top C = \mathbf{1}^\top D = \mathbf{0},\ \mathbf{1}^\top B = \mathbf{0};\ \forall r \in [5],\ \mathbf{1}^\top b_r = 0;\ \forall h \in [5],\ \mathbf{1}^\top A_h = \mathbf{0},\ A_h = A_h^\top;\ B = B^\top;$$

$$\forall z \in \mathbb{S}_Z,\ -\left(A_0 + \sum_{h=1}^{5} A_h z_h + B\right) \succcurlyeq 0.$$

The system is nonlinear in the parameters and can be cumbersome to estimate. To facilitate estimation, Lewbel and Pendakur (2009) proposes an approximate EASI system in which t_i is replaced by its first-order in prices approximation $d_i = x_i - p_i^\top y_i$ which corresponds to using nominal expenditures deflated by a Stone price index. Rather, we consider a second-order in prices approximation. As a starting point, we use (2.3) to obtain, for all $r \in \mathbb{N}$,

$$t_i^r = d_i^r \left(1 + \frac{r}{2} p_i^\top \left(A_0 + \sum_{h=1}^{5} A_h z_{hi} + B\right) p_i\right) + O(|p_i|_2^2). \tag{2.4}$$

Injecting (2.4) into (2.2) yields the second-order approximation

$$\begin{aligned} y_i = &\sum_{r=0}^{5} b_r d_i^r + \sum_{r=1}^{5}\sum_{g=1}^{G} e_g p_i^\top \widetilde{B}_{g,r} p_i d_i^r + \sum_{h=1}^{5}\sum_{r=1}^{5}\sum_{g=1}^{G} e_g p_i^\top \widetilde{A}_{h,g,r} p_i z_{hi} d_i^r + C z_i + D z_i d_i + A_0 p_i + \sum_{h=1}^{5} A_h z_{hi} \\ &+ B p_i d_i + \sum_{h=1}^{5}\sum_{g=1}^{G} e_g p_i^\top \overline{B}_{h,g} p_i z_{hi} d_i + \sum_{h=1}^{5}\sum_{g=1}^{G} e_g (p_i z_i)^\top \overline{A}_{h,g} (p_i z_{hi}) d_i + v_i + w_i, \end{aligned}$$

where v_i is an approximation error, $\widetilde{B}_{h,g,r} = r(b_r)_g(A_h + B)/2$, $\widetilde{A}_{h,g,r} = r(b_r)_g A_h/2$, $\overline{B}_{h,g} = D_{g,h}(A_0 + B)/2$, and $\overline{A}_{h,g} = D_{g,h} A_h/2$. These equality constraints define a nonconvex set, hence we do not impose them directly. Instead we impose the constraints, for all $h, r \in [5]$ and $g \in [G]$,

$$\begin{aligned} &\widetilde{B}_{h,g,r} = \widetilde{B}_{h,g,r}^\top,\ \widetilde{A}_{h,g,r} = \widetilde{A}_{h,g,r}^\top,\ \overline{B}_{h,g} = \overline{B}_{h,g}^\top,\ \overline{A}_{h,g} = \overline{A}_{h,g}^\top; \\ &\sum_{g=1}^{G} \widetilde{B}_{h,g,r} = \sum_{g=1}^{G} \widetilde{A}_{h,g,r} = \sum_{g=1}^{G} \overline{B}_{h,g} = \sum_{g=1}^{G} \overline{A}_{h,g} = \mathbf{0}; \end{aligned}$$

which are implied by the restrictions on the parameters of (2.2). Each equation in (2.2) has $K = 1879$ parameters. This dimensionality stems from the reduction in the approximation error relative to the approximate EASI system. Nonetheless, it is reasonable to expect that the parameter vector is sparse, particularly for the second order approximation terms. In addition, since $d_i = x_i - p_i^\top y_i$ depends on w_i, every regressor which involves d_i is endogenous. This implies that 1819 of the 1879 regressors are endogenous and $L = K$ instruments, where for each regressor involving d_i we replace d_i by $\overline{d}_i = x_i - p_i^\top \mathbb{E}_n[Y]$ (*i.e.*, we replace the individual shares by the mean shares in the sample).

2.3. **Restrictions and Penalization.** The set $\mathcal{R}$, if a proper subset of $\mathbb{R}^K$, accounts for restrictions. The main text considers single equations and typical examples are the whole space $\mathbb{R}^K$, known signs, or bounds on the coefficients. The empirical application in this paper considers a system with cross-equation restrictions. The set P is also relevant for the application. For example, it is known that Engle curves are nonlinear and, if approximated by a polynomial, should include at least a polynomial of degree two in the implicit utility (see, *e.g.*, Banks, Blundell, and Lewbel (1997)) and that the own price has an effect on the demand of the good.

2.4. **s-sparse Identified Set.** Define *s-sparse identifiable parameters* as vectors in

$$\mathcal{B}_s = \mathcal{I}dent \bigcap \{\beta : \ |J(\beta) \cap P| \le s\}.$$

We clearly have, for all $s \le s' \le p$, $\mathcal{B}_s \subseteq \mathcal{B}_{s'} \subseteq \mathcal{B}_p = \mathcal{I}dent$. Consider, for example, the case where $L < K$, $\mathcal{R} = \{\gamma \in \mathbb{R}^K : \ L = R_l \gamma\}$, and $R_l^\top \in \mathcal{M}_{K \times R}$ has full column rank. The situation where $L < K$ occurs when one is uncertain about some exclusion restrictions (see, *e.g.*, Kolesár, Chetty, Friedman, et al. (2015)). Assume that $\mathbb{E}[z_i y_i]$ and $\mathbb{E}\left[z_i x_i^\top\right]$ are constant for all i. If the identity of the zero coefficients were known, we would have $L+R+p-s$ restrictions. When $L+R+p-s > K$ there are $\binom{L+R+p-K}{s}$ overdetermined systems and point identification is achieved if there exists a solution for only one system and it is unique. Testing such an assumption is *NP*-hard. Assuming $P = [K]$ and $R = 0$, another *NP*-hard condition, which is clearly less sharp, is the extension of a condition in Candès and Tao (2007) (page 2320): $\mathcal{B}_s$ is a singleton if every matrix formed by extracting $2s$ columns from $\mathbb{E}\left[z_i x_i^\top\right]$ has rank $2s$ (see Kang, Zhang, Cai, et al. (2016)). Assessing the size of $\mathcal{B}_s$ is also *NP*-hard, hence infeasible when $K - p$ is larger than a few dozens.

2.5. **Uniformity/Honesty.** Working with a class $\mathcal{P}$ of distributions $\mathbb{P}(\beta)$ and with a sparsity certificate s, a honest confidence set for a functional $\varphi(\beta)$ with coverage at least $1-\alpha$ in finite samples is a set $\widehat{S}_l(s)$ which could be computed from the data, s, $\mathcal{R}$, and possibly parameters of $\mathcal{P}$, such that

$$\inf_{\beta, \mathbb{P}: \ \beta \in \mathcal{B}_s} \mathbb{P}\left(\varphi(\beta) \in \widehat{S}_l(s)\right) \ge 1 - \alpha. \tag{2.5}$$

Sets with coverage asymptotically at least $1-\alpha$ satisfy

$$\underline{\lim} \inf_{\beta, \mathbb{P}: \ \beta \in \mathcal{B}_s} \mathbb{P}\left(\varphi(\beta) \in \widehat{S}_l(s)\right) \ge 1 - \alpha. \tag{2.6}$$

The confidence sets cover each parameter in the identified set, hence the true parameter β^*, with coverage probability $1-\alpha$. If this holds asymptotically, the uniformity in $\mathbb{P}$, ensures that for a coverage error ϵ, for n large enough, only depending on ϵ, we have

$$\inf_{\beta,\mathbb{P}:\ \beta\in\mathcal{B}_s} \mathbb{P}\left(\varphi(\beta)\in\widehat{S}_l(s)\right)\geq 1-\alpha-\epsilon. \tag{2.7}$$

Some sets will satisfy the stronger uniformity property, for all countable sets G implying $c>0$:

$$\inf_{\beta,\mathbb{P}:\ \beta\in\mathcal{I}dent} \mathbb{P}\left(\forall s\in[K],\ (c,\varphi)\in G,\ \varphi(\beta)\in\widehat{S}_l(s,c)\right)\geq 1-\alpha-\epsilon, \tag{2.8}$$

for all n and $\epsilon=0$ or for n large enough depending on ϵ, where $\widehat{S}_l(s,c)$ are sets depending on a tuning parameter c. Due to the uniformity in c, it is possible to work with any intersection of sets $\widehat{S}_l(s,c)$ for c on a grid and compare nested confidence sets. The fact that we obtain confidence sets uniform only upon a class of distributions of the observed data is related to the Bahadur and Savage (1956) impossibility result (see also Romano and Wolf (2000)). The joint confidence sets can be used for joint hypothesis testing by duality between confidence sets and hypothesis testing and are particularly important for the models of sections 2.2.3 and 2.2.4.

3. A First Approach to Confidence Sets and the *STIV* Estimator

3.1. **Robustness.** Three features are important in estimation and inference using instrumental variables: (1) for each endogenous regressor there should be an instrument which does not appear as a right-hand side variable in (1.1), (2) instruments should be exogenous (*i.e.*, not correlated with the structural error), (3) they should be relevant (*i.e.*, have sufficiently large predicting power for the endogenous variables). This paper relaxes all three. (1) is the exclusion restriction and it is relaxed throughout the paper if the structural equation is sparse, it is particularly important when $L<K$. (2) is the instrument exogeneity, it is relaxed in Section 7.3 and a method is proposed when one does not know in advance a set of variables which are known to be exogenous. (3) is the relevance of the instruments and is relaxed in our first two inference procedures. The weak instrument or weak identification problem occurs when $\mathbb{E}\left[z_i x_i^\top\right]$ is nearly rank deficient.

When L and K are small and $L\geq K$, it is a common practice to proceed in two-stages and estimate a first stage linear projection model of the endogenous regressors on the instruments or nonparametric regressions under the stronger exogeneity assumption $\mathbb{E}[u_i|z_i]=0$. This can be justified using the concept of semiparametric efficiency. In the presence of weak identification and instances such as in sections 2.2.3 and 2.2.4 where the researcher is genuinely interested in the high-dimensional vector, the semi-parametric efficiency framework is not the right one. In the latter case, there did not exist inference procedures until recently[1]. Hence, appealing to a high-dimensional generalization of 2SLS or GMM is not motivated. The approach of this paper does not rely on the estimation of a first stage. For the sake of completeness, we include the simulation results of Section A.8 which show that our baseline confidence sets are sharper than those obtained by a high-dimensional 2SLS procedure even when the first stage is sparse. Also, it is shown in Bekker (1994) that when K is small and L/n converges to 1, the bias of 2SLS and GMM estimators is of the order of the bias of the least squares (henceforth, OLS) estimator and our framework allows for much larger L and K. Inference after selecting instruments is difficult if one wants to account for model selection and is not feasible when this first stage is not sparse (*e.g.*, when all instruments are equally weak).

We now focus on the weak instrument problem. When $\mathbb{E}[z_i z_i^\top] = I$ and all coefficients in the linear projection are equal, their size is at most of the order of $1/\sqrt{L}$ which is much smaller than the regime of weak instruments asymptotic when $L \gg n$ (see Staiger and Stock (1997)). This does not occur if the linear projection is sparse or approximately sparse but in the main part of the paper we posit a sparsity assumption on the reduced form equation only. Also, in the case of nonparametric IV (see also Example O2 in the appendix), the correlation between basis functions of the baseline regressors and baseline instruments can decay very fast to zero. In the presence of weak instruments, the approximations using asymptotic theory for classical estimators such as 2SLS cannot be trusted. This happens even with very large sample sizes and for typical problems of empirical relevance (see Nelson and Startz (1990) and Bound, Baker, and Jaeger (1995)). To deal with the weak instrument problem, the literature (see Andrews and Stock (2007) and the references therein) relies on non-standard asymptotics (weak IV, many IV, or many weak IV) or the inversion, at every parameter value, of tests which are robust to weak instruments. When $\mathcal{I}dent = \{\beta^*\}$ and we take $\beta_{I^c} = \beta^*_{I^c}$, the Anderson-Rubin test is a F-test of $\kappa = 0$ in the model: for all $i \in [n]$, $y_i - x_{I^c i}^\top \beta^*_{I^c} = \widetilde{z}_i^\top \kappa + x_{Ii}^\top \beta_I + u_i$, where $\widetilde{z}_i$ are the instruments other than x_{Ii}. It is robust to arbitrarily weak instruments and the distribution of the test statistic under the null is independent of the parameter Π of the first stage $x_{I^c i} = \Pi \widetilde{z}_i + \Gamma x_{Ii} + v_{2i}$. Similar level α tests have level α for all Π and can be constructed from non-similar tests by using so-called conditional critical values. A common practice is to use a pretest for weak instruments and to use 2SLS if they are not weak and a more complex method if the test fails to reject the null hypothesis. This approach is subject to the criticism related to "uniformity". Inverting tests is only possible on a grid when I^c is small and in the framework of this paper this might not be the case. Also, when $L \gg n$, we can no longer rely on a F-test. Even in low dimensions,

[1] Version 1 of this paper is the first up to our knowledge.

the asymptotic distribution of the Anderson-Rubin test when $|I^c| = 1$ is $\chi^2_{L-|I|}/(L - |I|)$ which can be problematic when $L - |I| > |I^c|$ (*i.e.*, $L > K$). Rather, we make use of a standardized ℓ_∞-norm statistic rather than a weighted squared ℓ_2-norm which mitigates the many instruments problem and is introduced in Section 3.2. Also, rather than inverting a test, we aim directly at the estimation and confidence sets and are based on convex relaxation ideas. Yet, our approach shares in common with the Anderson-Rubin test that the two first proposals of this paper have guaranteed coverage irrespective of identification. This is because the classes $\mathcal{P}$ in Section 3.5 do not restrict the conditional distribution of $\mathbf{X}|\mathbf{Z}$ nor require $L \geq K$. Without model uncertainty, the *STIV* estimator trades-off minimization of a ℓ_∞-norm accounting for the exogeneity of the instruments and OLS (see Remark 3.1).

Importantly, because we allow for a set P of indices of regressors for which we have a doubt whether the coefficient is zero or not and this set could be empty, the paper applies not only to the high-dimensional paradigm but also to more classical situations in Econometrics. It provides algorithmically feasible solutions to robust estimation and inference with weak, many, endogenous, and few IVs for low-dimensional structural equations.

3.2. **A ℓ_∞-norm Statistic.** The sample counterpart of condition (1.2) can be written as

$$\frac{1}{n}\mathbf{Z}^\top(\mathbf{Y} - \mathbf{X}\beta) = \mathbf{0}. \tag{3.1}$$

This is a system of L equations with K unknown parameters, where $\operatorname{rank}(\mathbf{Z}^\top\mathbf{X}) \leq \min(K, L, n)$. Thus, when $K > \min(L, n)$, the matrix cannot have full column rank and, when $L > \min(K, n)$, the system might not have a solution. Furthermore, even if the system had a unique solution, replacing the population equations (1.2) by (3.1) induces a huge error when L, K or both are large and n is not too large. So, looking for the exact solution of (3.1) in high-dimensional settings and finite samples makes no sense. However, we can stabilize the problem by restricting our attention to a suitable "small" candidate set of vectors $\beta \in \mathcal{R}$, for example, to those satisfying the constraint

$$\max_{l\in[L]}\left|\frac{1}{n}\frac{\mathbf{z}_l^\top(\mathbf{Y} - \mathbf{X}\beta)}{\sqrt{\widehat{Q}_l(\beta)}}\right| \leq r_0. \tag{3.2}$$

where, for $l \in [L]$, $\widehat{Q}_l(\beta) \triangleq \mathbb{E}_n\left[(Z_l U(\beta))^2\right]$. We provide values of r_0 for various choices of $\mathcal{P}$ in Section 3.5 that guarantee that, for all $\beta \in \mathcal{I}dent$ and $\mathbb{P}$ such that $\mathbb{P}(\beta) \in \mathcal{P}$, the probability of

$$\mathcal{G}_0 \triangleq \left\{\max_{l\in[L]}\frac{|\mathbb{E}_n[Z_l U(\beta)]|}{\sqrt{\mathbb{E}_n[(Z_l U(\beta))^2]}} \leq r_0\right\},$$

exceeds $1-\alpha$ either in finite samples or asymptotically. A typical "reference" behavior is $r_0 \sim \sqrt{\ln(L)/n}$. This is the usual parametric rate of convergence multiplied by $\ln(L)$. Note that $\ln(L)$ increases very slowly with L. This is the key to handling a very large number of instruments.

3.3. **Self-Normalized IV Confidence Sets: A *NP*-hard Problem.** Assume that $\beta \in \mathcal{R}$ can be written as certain polynomials in the parameters are nonnegative. From the above, the set

$$\widehat{S} \triangleq \left\{ \beta \in \mathcal{R} : \ |J(\beta) \cap P| \le s, \ \max_{l \in [L]} \left| \frac{1}{n} \frac{\mathbf{z}_l^\top (\mathbf{Y} - \mathbf{X}\beta)}{\sqrt{\widehat{Q}_l(\beta)}} \right| \le r_0 \right\} \tag{3.3}$$

satisfies (2.5) or (2.8) depending on the class $\mathcal{P}$. (3.3) is a confidence set for β from which we obtain a confidence set which covers all linear functionals $\lambda^\top \beta$ with $\lambda \in \mathbb{R}^K$ for $\beta \in \mathcal{B}_s$ via

$$\left[\min_{\beta \in \widehat{S}} \lambda^\top \beta, - \min_{\beta \in \widehat{S}} -\lambda^\top \beta \right]. \tag{3.4}$$

For further reference, these are called the *Self-Normalized Instrumental Variables (SNIV)* confidence sets. The minimizations on $\widehat{S}$ are difficult because $\widehat{S}$ is not convex. The constraint involving a maximum in the definition of $\widehat{S}$ can be written: $\forall l \in [L]$, $a_l + b_l^\top \beta + \beta^\top C_l \beta \le 0$, where

$$\begin{aligned}
a_l &= \mathbb{E}_n [Z_l Y]^2 - r_0^2 \mathbb{E}_n \left[Z_l^2 Y^2 \right], \\
b_l &= 2 \left(r_0^2 \mathbb{E}_n \left[Z_l^2 Y X \right] - \mathbb{E}_n [Z_l Y] \mathbb{E}_n [Z_l X] \right), \\
C_l &= \mathbb{E}_n [Z_l X] \mathbb{E}_n [Z_l X]^\top - r_0^2 \mathbb{E}_n \left[Z_l^2 X \right] \mathbb{E}_n \left[Z_l^2 X \right]^\top.
\end{aligned}$$

In general, the matrices C_l need not be positive semidefinite. The cardinality constraint $|J(\beta) \cap P| \le s$ is not convex either. When K is very small, the optimization is feasible by taking β on a grid. When $K+p$ is moderately large, we can rely on the following observation. Using the following reformulation of the cardinality constraint (see Feng, Mitchell, Pang et al. (2013))

$$\widehat{S} = \left\{ \beta \in \mathcal{R} : \begin{array}{l} \epsilon_{P^c} = \mathbf{0}, \ \sum_{k \in P} \epsilon_k \ge p - s, \\ \forall k \in P, \ \epsilon_k \in [0,1], \ \epsilon_k(\epsilon_k - 1) \ge 0, \text{ and } \epsilon_k \beta_k = 0, \\ \forall l \in [L], \ \left(\mathbf{z}_l^\top (\mathbf{Y} - \mathbf{X}\beta) \right)^2 \le (n r_0)^2 \widehat{Q}_l(\beta) \end{array} \right\}.$$

and that, if the polynomials in the definition of $\mathcal{R}$ are of degree strictly greater than two, we can replace them by inequalities with polynomial of degree at most two by increasing the number of variables, the optimization problems defining the upper and lower bounds of the intervals in (3.4) are a type of nonconvex Quadratically Constrained Quadratic Programs (QCQP). If the polynomials in the definition of $\mathcal{R}$ are of degree at most two, the number of variables is $K+p$. These problems are in general *NP*-hard. Various heuristics exist for such problems but, in order to obtain confidence sets with coverage at least $1-\alpha$, it is important to find either a global minimum or a minimizer.

Rather than using methods which can fail to yield a global minimum, a basic idea is to replace $\widehat{S}$ by a larger manageable set. One classical approach to handle quadratic constraints is via Semidefinite Programs (SDP) relaxations, this usually implies that the parameters of the relaxed problem is a square of the number of parameter of the original problem. Rather, the confidence sets in Section 5.1, make use of the fact that if $|J(\beta)| \le s$ then $|\beta|_q \le s^{1/q}|\beta|_\infty$ for all $q \in [1,\infty]$. For moderate K, we can find the minimum of the original QCQP by using Lasserre's hierarchy of SDP relaxations (see Lasserre (2015)). This is illustrated in Section 9.1.1 with $K+p=21$. The Lasserre's hierarchy yields a sequence of lower bounds on the minimum of a QCQP which has theoretical guarantees that it converges to the global minimum. Typically, for an optimization problem in $K+p$ variables, the SDP relaxation of order d in the hierarchy involves $O\left((K+p)^{2d}\right)$ variables and $O\left((K+p)^{d}\right)$ linear matrix inequalities. Note that one does not need to consider the whole sequence of problems until convergence. Indeed, the solution to each problem is a valid lower bound and working with lower bounds produces a confidence set with guaranteed coverage but which could be conservative.

3.4. **The *Self-Tuned IV* estimator.** As this paper is mainly concerned with large K, we avoid dealing with nonconvex constraints using SDP relaxations. First, notice that it makes sense to normalize the matrix $\mathbf{Z}$. This is quite intuitive because, otherwise, an instrument has a larger effect simply because it is on a larger scale. Various choices are possible but in this paper we take $\widehat{\mathbf{D}}_{\mathbf{Z}}$.

Instead of $\widehat{S}$, we can work with the following set with σ in $\mathbb{R}^L$ and $\overline{Q}_l(\beta) \triangleq \left(\widehat{\mathbf{D}}_{\mathbf{Z}}\right)_{ll}\widehat{Q}_l(\beta)$

$$\left\{\beta \in \mathcal{R}:\ \forall l \in [L],\ \left|\frac{1}{n}\left(\widehat{\mathbf{D}}_{\mathbf{Z}}\right)_{ll}\mathbf{z}_l^\top(\mathbf{Y}-\mathbf{X}\beta)\right| \le r_0\sigma_l,\ \overline{Q}_l(\beta) \le \sigma_l^2\right\}, \tag{3.5}$$

and use an objective function that prevents σ to be large and a convex relaxation of the cardinality constraint. For simplicity, we now replace the L constraints $\overline{Q}_l(\beta) \le \sigma_l^2$ by a single one and postpone the discussion of variants of the *STIV* estimator using (3.5) to sections 7.2 and 7.3.

Definition 3.1. *The set of IV-constraints is the set defined, for $\sigma, r > 0$, by*

$$\widehat{\mathcal{I}}(r,\sigma) \triangleq \left\{\beta \in \mathcal{R},\ \left|\frac{1}{n}\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{Z}^\top(\mathbf{Y}-\mathbf{X}\beta)\right|_\infty \le r\sigma,\ \widehat{Q}(\beta) \le \sigma^2\right\}. \tag{3.6}$$

Definition 3.2. *For $c, r > 0$ with $c \in \left(0, r^{-1}\right)$, a* Self-Tuned Instrumental Variables *(STIV) estimator is any solution $\left(\widehat{\beta},\widehat{\sigma}\right)$ of the minimization problem*

$$\min_{\beta \in \widehat{\mathcal{I}}(r,\sigma), \sigma \ge 0}\left(\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_P\right|_1 + c\sigma\right). \tag{3.7}$$

We use the ℓ_1-norm because it is the convex relaxation of the cardinality constraint in the definition of the set $\widehat{S}$. This usually ensures that the solution is sparse. We use the scaling matrix $\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}$ so the procedure is invariant to the scale of the regressors and avoids, for example, that it selects a different model by changing units. Recall that the procedure is also invariant to the scale of the instruments. The product $r\sigma$ plays the role of the penalization parameter in the Lasso. The additonal term $c\sigma$ favors small σ, hence tightening the set $\widehat{\mathcal{I}}(r,\sigma)$. If c were 0, the program would minimize the ℓ_1-norm without constraints hence $\widehat{\beta}=\mathbf{0}$ is the solution. This suggest taking c as large as possible, hence close to r^{-1}. The results of Section 4 allow to analyze the effect of c on the rates and confidence sets. When K is large, the *STIV* estimator is not necessarily unique. This is not a problem since our results hold for all minimizers. If the data is i.i.d., $\widehat{Q}\left(\widehat{\beta}\right)$ and $\widehat{\sigma}^2$ are estimators of the unknown variance of $u_i(\beta)$. For the particular case $\mathbf{Z}=\mathbf{X}$, the *STIV* estimator provides an extension of the Dantzig selector to the setting with unknown variance of the noise. In this particular case, the *STIV* estimator can be related to the Square-root Lasso of Belloni, Chernozhukov, and Wang (2011). The definition of the *STIV* estimator contains the additional constraint involving the instruments not present in the conic program for the Square-root Lasso. A linear programming estimator and confidence sets like in Gautier and Tsybakov (2013) could also be used (see here).

Remark 3.1. *When $P=\emptyset$,* (3.7) *simplifies to*

$$\widehat{\beta}\in \mathrm{argmin}_{\beta\in\mathcal{R}} \max\left(\widehat{Q}(\beta), \frac{1}{r^2}\left|\frac{1}{n}\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{Z}^\top(\mathbf{Y}-\mathbf{X}\beta)\right|_\infty^2\right) \tag{3.8}$$

The first component of the maximum in (3.8) *corresponds to least squares. The second is derived from the exogeneity of the instruments. Hence, without model uncertainty,* STIV *estimators trade off OLS and a ℓ_∞-norm statistic derived from the IVs (see Andrews and Stock (2007) for references to comparisons between 2SLS and least squares).* (3.8) *is a convex program when $\mathcal{R}$ is convex.*

If $\mathcal{R}$ consists of linear equality or inequality constraints, finding a solution $\left(\widehat{\beta},\widehat{\sigma}\right)$ of (3.7) reduces to the following second-order cone program, where $C\triangleq\{(t,v)\in\mathbb{R}\times\mathbb{R}^n:\ t\ge|v|_2\}$ is a cone.

Algorithm 3.1. *Find $\beta\in\mathcal{R}$ and $t>0$ ($\sigma=t/\sqrt{n}$), which achieve the minimum*

$$\min_{(\beta,t,v,w)\in\mathcal{V}}\left(\sum_{k=1}^K w_k + c\frac{t}{\sqrt{n}}\right), \tag{3.9}$$

where $\mathcal{V}$ is the set of (β,t,v,w) satisfying

$$v=\mathbf{Y}-\mathbf{X}\beta,\qquad -rt\mathbf{1}\le\frac{1}{\sqrt{n}}\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{Z}^\top\left(\mathbf{Y}-\mathbf{X}\beta\right)\le rt\mathbf{1},$$

$$-w \le \widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta \le w, \qquad \beta \in \mathcal{R}, \qquad w \ge \mathbf{0}, \qquad (t,v) \in C.$$

Conic programming is a standard tool in optimization. It starts to be difficult when K is of the order of several thousands.

3.5. **Menu of Distributional Assumptions and Choice of r_0 and r.** For $\beta \in \mathcal{I}dent$ and $i \in [n]$, we have $\mathbb{E}[z_i u_i(\beta)] = \mathbf{0}$. We present a menu of classes $\mathcal{P}$ - called scenarii - from which the researcher has to choose. We use the notation $\mathcal{P}_j$ for the class $\mathcal{P}$ defined in Scenario j. From this section and onwards, we assume that the researcher has chosen a Scenario and simply use the notation $\mathcal{P}$. Choosing a confidence level $1-\alpha$, the constants r_0 given below guarantee that

$$\inf_{\beta,\mathbb{P}:\ \mathbb{P}(\beta)\in\mathcal{P}} \mathbb{P}\left(\mathcal{G}_0\right) \ge 1-\alpha-\epsilon$$

holds for all n and $\epsilon = 0$ for scenarii 1-3 and for n large enough depending on ϵ for Scenario 4. We now present four scenarii and the corresponding choice of r_0 obtained using a union bound and moderate deviations for self-normalized sums (see Section A.1).

Scenario 1: *$(z_i u_i(\beta))_{i=1}^n$ are i.i.d. and symmetric and L is such that $L < 9\alpha/\left(4e^3\Phi\left(-\sqrt{n}\right)\right)$.*

Under Scenario 1, taking

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

yields, for all n, $\mathbb{P}\left(\mathcal{G}_0\right) \ge 1-\alpha$.

Scenario 2: *$(z_i u_i(\beta))_{i=1}^n$ are i.i.d., for $\gamma_4 > 0$ $\max_{l\in[L]}\mathbb{E}[(z_{li}u_i)^4](\mathbb{E}[(z_{li}u_i)^2])^{-2} \le \gamma_4$ and $L < \alpha\exp\left(n/\gamma_4\right)/(2e+1)$.*

Under Scenario 2, taking

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

yields, for all n, $\mathbb{P}\left(\mathcal{G}_0\right) \ge 1-\alpha$.

Scenario 3: *$(u_i(\beta))_{i=1}^n$ are independent and symmetric conditional on $\mathbf{Z}$ or $(z_i u_i(\beta))_{i=1}^n$ are independent and symmetric.*

Under Scenario 3, taking

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

yields, for all n, $\mathbb{P}\left(\mathcal{G}_0\right) \ge 1-\alpha$.

For scenarii 4 and 5 below, $\alpha_B(n)$ is a finite sample bound on the coverage error, these bounds depend on the parameters of the classes.

Scenario 4: *$(z_i u_i(\beta))_{i=1}^n$ are independent, $\left|\left(\left(\mathbb{E}\left[|Z_l U(\beta)|^{2+\delta}\right]\right)\left(\mathbb{E}\left[Z_l^2 U(\beta)^2\right]\right)^{-(2+\delta)/2}\right)_{l\in[L]}\right|_\infty \le$*

$\gamma_{2+\delta}$ *for* δ *in* $(0,1]$ *and* $\gamma_{2+\delta}\ge 0$ *and* $L\le \alpha/\left(2\Phi\left(-n^{1/2-1/(2+\delta)}\gamma_{2+\delta}^{-1/(2+\delta)}\right)\right)$.
Under Scenario 4, taking

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

yields, for all n, $\mathbb{P}\left(\mathcal{G}_0\right)\ge 1-\alpha-\alpha_B(n)$, where $\alpha_B(n)\triangleq \alpha A_0\gamma_{2+\delta}\left(1+\sqrt{n}r_0\right)^{2+\delta}n^{-\delta/2}$ and A_0 is a universal constant.

Scenarii 1 and 3 rely on symmetry. This occurs if (1.1) is a first difference between two time periods in a panel data model. Scenario 2 relaxes symmetry but requires fourth moments and the upper bound γ_4. When $n-\gamma_4\ln(L(2e+1)/\alpha)\ge n/2$ one can take $r_0=2\sqrt{\ln(L(2e+1)/\alpha)/n}$. A two-stage approach is used in Bertail, Gauthérat and Harari-Kermadec (2005). One starts by choosing r_0 with an upper bound on γ_4, then one constructs the upper bound from a confidence interval for γ_4 and computes refined confidence sets. Scenarii 3 and 4 allow for conditional heteroscedasticity. Scenario 3 allows for dependence in the matrix $\mathbf{Z}$. Scenario 4 relies on an upper bound $\gamma_{2+\delta}$. The choice of r_0 in Scenario 4 is asymptotic because the moderate deviations result depends on A_0 which is not explicit. The four scenarii require that L does not exceed an exponential in n. Using Chen, Shao, Wu, et al. (2016), one can obtain r_0 for dependent data.

For the analysis of the *STIV* estimator, for $\beta\in\mathcal{I}dent$, we make use of the event

$$\mathcal{G}\triangleq\left\{\max_{l\in[L]}\frac{\left|\mathbb{E}_n[Z_lU(\beta)]\right|}{\sqrt{\mathbb{E}_n[Z_l^2]\mathbb{E}_n[U(\beta)^2]}}\le r\right\}.$$

Note that $\mathcal{G}=\left\{\beta\in\widehat{\mathcal{I}}(r,\sqrt{\widehat{Q}(\beta)})\right\}$ and if we take $r=r_0\left|\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{Z}^\top\right|_\infty$ we have $\mathcal{G}_0\subseteq\mathcal{G}$. Since $\left|\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{Z}^\top\right|_\infty$ can be large, we can further restrict scenarii 1-4 by maintaining Assumption 3.1 below. We use the same denomination scenarii 1-4 and notation $\mathcal{P}_j$ for conciseness. Under Assumption 3.1, taking $r=r_0\sqrt{1+\tau_{\mathcal{G}}}$, where $1+\tau_{\mathcal{G}}=(1+\tau_{ZU})/\left((1-\tau_Z')(1-\tau)\right)$, for arbitrary sequences $\tau,\tau_Z',\tau_{ZU}\in(0,1)$ converging to zero with n and $\alpha_B(n)$ which is nonzero only for Scenario 4, we have

$$\mathcal{G}_0\cap\left\{\max_{l\in[L]}\frac{\mathbb{E}_n\left[\left(Z_lU(\beta)\right)^2\right]}{\mathbb{E}_n[Z_l^2]\mathbb{E}_n[U(\beta)^2]}\le 1+\tau_{\mathcal{G}}\right\}\subseteq\mathcal{G}$$

and $\mathbb{P}\left(\mathcal{G}\right)\ge 1-\alpha-\alpha_B(n)-\alpha_C(n)$, where $\alpha_C(n)=\left(m_4/\tau^2+C_{\mathrm{N}}(L)\left(M_Z'(L)/(\tau_Z')^2+M_{ZU}(L)/\tau_{ZU}^2\right)\right)/n$. This shows that we should take $\tau^2n\to 0$ and $\max\left(\tau_Z',\tau_{ZU}\right)^2n/\log(L)\to 0$. Hence, we can take $\tau_{\mathcal{G}}$ very close to zero (in practice $\sqrt{1+\tau_{\mathcal{G}}}$ can be taken as 1.1 or 1.01 depending on the magnitudes of n and L) and pay the price of a small coverage error.

Assumption 3.1. *$L, K \geq 3$ and for $j \in [4]$, m_4, $M'_Z(L)$, and $M_{ZU}(L)$ positive, for all $\beta, \mathbb{P}$ such that $\beta \in \mathcal{I}dent$, where $\mathcal{P} = \mathcal{P}_j$, we have (S5.ii) and (S5.iv) defined below, $\mathbf{Z}$ and $\mathbf{U}(\beta)$ are independent, and $\mathbb{E}\left[\left|\left((Z_l U(\beta))^2 / \left(\mathbb{E}\left[Z_l^2\right]\mathbb{E}\left[U(\beta)^2\right]\right) - 1\right)_{l=1}^{L}\right|_\infty^2\right] \leq M_{ZU}(L)$.*

The union bound used in scenarii 1-4 does not account for the dependence in the vector $z_i u_i(\beta)$. An alternative which does is the multiplier bootstrap (see Chernozhukov, Chetverikov, and Kato (2013,2017)). Other types could be used (*e.g.*, Zhang and Cheng (2014) for dependent data).

Scenario 5: *Let $L \geq 3$ and positive constants m_4, $M_Z(L)$, $M'_Z(L)$, b, q_2, and $B(n) \geq 1$, where $M_Z(L)$ and $M'_Z(L)$ can depend on L and $B(n)$ can depend on n, and a sequence $(\alpha_E(n))_{n\in\mathbb{N}}$ converging to zero. We have, for all $n \in \mathbb{N}$,*

(S5.*i*) *For all $i \in [n]$, $\mathbb{E}\left[u_i(\beta)^2 \middle| z_i\right] = \sigma^2_{U(\beta)}$;*

(S5.*ii*) $\mathbb{E}\left[\left(\left(U(\beta)/\sigma_{U(\beta)}\right)^2 - 1\right)^2\right] \leq m_4$;

(S5.*iii*) $\mathbb{E}\left[\left|\mathbf{D}_Z\left(ZZ^\top - \mathbb{E}\left[ZZ^\top\right]\right)\mathbf{D}_Z\right|_\infty^2\right] \leq M_Z(L)$;

(S5.*iv*) $\mathbb{E}\left[\left|\left(Z_l^2/\mathbb{E}\left[Z_l^2\right] - 1\right)_{l=1}^{L}\right|_\infty^2\right] \leq M'_Z(L)$;

(S5.*v*) $\max\left(\mathbb{E}\left[\left((\mathbf{D}_Z)_{ll} Z_l U(\beta)/\sigma_{U(\beta)}\right)^{2+q_1}\right], \mathbb{E}\left[\left((\mathbf{D}_Z)_{ll} Z_l e_i\right)^{2+q_1}\right]\right) \leq B(n)^{q_1}$, *$\forall l \in [L]$, $q_1 \in \{1, 2\}$;*

(S5.*vi*) $\max\left(\mathbb{E}\left[\left(\left|\mathbf{D}_Z z_i u_i(\beta)\right|_\infty / (B(n)\sigma_{U(\beta)})\right)^{q_2}\right], \mathbb{E}\left[\left(\left|\mathbf{D}_Z z_i e_i\right|_\infty / B(n)\right)^{q_2}\right]\right) \leq 2$, *$\forall i \in [n]$;*

(S5.*vii*) $\min_{\tau,\tau_Z,\tau'_Z,\zeta_1\in(0,1)}\left(\alpha_B(n,\tau,\tau_Z,\tau'_Z,\zeta_1) + \alpha_C(n,\tau_Z,\tau'_Z,\zeta_1)\right) \triangleq \alpha_B(n) + \alpha_C(n) \leq \alpha_E(n)$;

where e_i for $i \in [n]$ are i.i.d. standard normals independent from $\mathbf{Z}$ and $\alpha_B(n,\tau,\tau_Z,\tau'_Z,\zeta_1)$ and $\alpha_C(n,\tau_Z,\tau'_Z,\zeta_1)$ are defined below.

Under Scenario 5, taking q_W the quantile function of $W \triangleq \left|\left(\sum_{i=1}^n \widehat{\mathbf{D}}_{\mathbf{Z}} z_i e_i\right)/\sqrt{n}\right|_\infty$ treating $\mathbf{Z}$ as fixed,

$$r = \frac{1}{\sqrt{n}} q_W(1-\alpha),$$

yields, for all n and $\alpha \in (0,1)$, $\left|\mathbb{P}(\mathcal{G}) - 1 - \alpha\right| \leq \alpha_B(n,\tau,\tau_Z,\tau'_Z,\zeta_1)$, where, for all $\tau,\tau_Z,\tau'_Z,\zeta_1 \in (0,1)$,

$$\alpha_B(n,\tau,\tau_Z,\tau'_Z,\zeta_1) \triangleq 2C_1\tau_Z^{\frac{1}{3}}\log\left(\frac{L}{\tau_Z}\right)^{\frac{2}{3}} + 2C_2\rho + \frac{2C_{\mathrm{N}}(L^2)M_Z(L)}{n\tau_Z^2} + 2\zeta_2(\zeta_1,\tau'_Z) + \zeta'_2(\zeta_1),$$

$\rho \triangleq C_2\left(\left(B(n)^2\left(\log(Ln)\right)^7/n\right)^{\frac{1}{6}} + \left(B(n)^2(\log(Ln))^3 n^{-1\frac{2}{q_2}}\right)^{\frac{1}{3}}\right)$, C_1 is universal and C_2 a constant which only depends on q_2, and the values of $\zeta_2(\zeta_1,\tau'_Z)$ and $\zeta'_2(\zeta_1)$ are in the appendix. Also, for all $\tau,\tau_Z,\tau'_Z,\zeta_1 \in (0,1)$, we have $\mathbb{P}\left(r \leq \overline{r}(\tau_Z,\tau'_Z,\zeta_1)\right) \geq 1 - \alpha_C(n,\tau_Z,\tau'_Z,\zeta_1)$, where, denoting by q_{N_0} the quantile function of N_0,

$$\alpha_C(n,\tau_Z,\tau'_Z,\zeta_1) \triangleq \frac{C_{\mathrm{N}}\left(L^2\right)M_Z(L)}{n\tau_Z^2} + \zeta_2(\zeta_1,\tau'_Z),$$

$$\overline{r}(\tau_Z,\tau'_Z,\zeta_1) \triangleq \frac{1}{\sqrt{n}} q_{N_0}\left(1-\alpha+\zeta_2(\zeta_1,\tau'_Z)+C_1\tau_Z^{\frac{1}{3}}\log\left(\frac{L}{\tau_Z}\right)^{\frac{2}{3}}\right)+\zeta_1.$$

The conditional quantiles can be computed by simulation. Note that (S5.iv) can be viewed as redundent once (S5.iii) is imposed. We write it for further reference and because we can take $M'_Z(L) \le M_Z(L)$. Chernozhukov, Chetverikov, and Kato (2013) consider heteroscedastic errors in the high-dimensional regression with non stochastic regressors. It uses a plug-in of estimated residuals and relies on a consistent preliminary estimator. We do not consider such an extension because, in the main part of the paper, we are agnostic on features that would deliver a preliminary estimator that converges fast enough, such as strong instruments.

4. Sensitivity Characteristics

In the linear regression in low dimension, when $\mathbf{Z} = \mathbf{X}$ and $\mathbf{X}^\top\mathbf{X}/n$ is positive definite, the minimal eigenvalue of this matrix is an important quantity to obtain error bounds. It is the minimum over all $\beta \in \mathbb{R}^K$ of $\beta^\top\mathbf{X}^\top\mathbf{X}\beta/(n|\beta|_2^2)$. When β is sparse and one uses the ℓ_1-norm like in the Lasso and Dantzig selector, one can consider a subset of $\mathbb{R}^K$ which is a cone. This is typically expressed via the restricted isometry property of Candès and Tao (2007) or the restricted eigenvalue condition of Bickel, Ritov, and Tsybakov (2009). These cannot be used for models with endogenous regressors where we have a rectangular matrix $\mathbf{Z}^\top\mathbf{X}/n$. Due to normalizations, we work with $\widehat{\Psi} \triangleq \widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{Z}^\top\mathbf{X}\widehat{\mathbf{D}}_{\mathbf{X}}/n$.

4.1. Definition and Main Results on the Sensitivities.

We introduce some scalar sensitivity characteristics related to the action of $\widehat{\Psi}$ on vectors in the restricted set

$$\widehat{C}_J \triangleq \left\{\Delta:\ \widehat{\mathbf{D}}_{\mathbf{X}}\Delta \in \mathcal{R}_D, \Delta_{J^c\cap J(\widehat{\beta})^c} = \mathbf{0},\ |\Delta_{J^c\cap P}|_1 \le |\Delta_{J\cap P}|_1 + cr|\Delta_I|_1 + c|\Delta_{I^c}|_1\right\}, \tag{4.1}$$

for $J \subseteq [K]$. The set $\widehat{C}_J$ is a cone when $\mathcal{R}_D$ is a set of the form $\{\gamma \in \mathbb{R}^K:\ R_l\gamma = \mathbf{0}\}$.

Remark 4.1. *When $I = P = [K]$, $\widehat{C}_J$ can be written as*

$$\widehat{C}_J = \left\{\Delta:\ \widehat{\mathbf{D}}_{\mathbf{X}}\Delta \in \mathcal{R}_D,\ \Delta_{J^c\cap J(\widehat{\beta})^c} = \mathbf{0},\ (1-cr)|\Delta_{J^c}|_1 \le (1+cr)|\Delta_J|_1\right\}. \tag{4.2}$$

If the cardinality of J is small, the vectors Δ in $\widehat{C}_J$ have a substantial part of their mass concentrated on a set of small cardinality. The set J that will be used later is the set $J(\beta)$.

Remark 4.2. *If $P^c \subseteq J(\beta)$, the set $\widehat{C}_{J(\beta)}$ obtained by penalizing all coefficients can be written as*

$$\widehat{C}_{J(\beta)} = \left\{\Delta:\ \widehat{\mathbf{D}}_{\mathbf{X}}\Delta \in \mathcal{R}_D,\ \Delta_{J^c\cap J(\widehat{\beta})^c} = \mathbf{0},\ |\Delta_{J(\beta)^c\cap P}|_1 \le |\Delta_{J(\beta)\cap P}|_1 + |\Delta_{J(\beta)\cap P^c}|_1 + cr|\Delta_I|_1 + c|\Delta_{I^c}|_1\right\}.$$

It has the additional term $|\Delta_{J(\beta)\cap P^c}|_1$ on the right-hand side of the inequality, hence is larger.

We show that, for $\beta \in \mathcal{I}dent$, on the event $\mathcal{G}$, for all *STIV* estimator $\left(\widehat{\beta}, \widehat{\sigma}\right)$, $\widehat{\Delta} \triangleq \widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\left(\widehat{\beta}-\beta\right)$, and $\overline{\sigma} \triangleq \left(\widehat{\sigma}+\sqrt{Q\left(\widehat{\beta}\right)}\right)/2$, we have: $\widehat{\Delta} \in \widehat{C}_{J(\beta)}$ and

$$\left|\widehat{\Psi}\widehat{\Delta}\right|_{\infty} \leq r\left(2\overline{\sigma}+r\left|\widehat{\Delta}_{I}\right|_{1}+\left|\widehat{\Delta}_{I^c}\right|_{1}\right). \tag{4.3}$$

Inequality (4.3) includes terms of different nature: $\left|\widehat{\Psi}\widehat{\Delta}\right|_{\infty}$ on one side and the ℓ_1-norms of subvectors on the other. The sensitivities allow one to relate them to each other. More generally, they allow to relate a loss function $l\left(\widehat{\Delta}\right)$, where $l \in \mathcal{L}$, to the quantity $\left|\widehat{\Psi}\widehat{\Delta}\right|_{\infty}$. The choice of the function l is guided by the quantities which appear in the inequalities such as the ℓ_1-norms on the right-hand side of (4.3) and eventually by the researcher who has to specify what feature is more important to her.

We define the *sensitivity*

$$\widehat{\kappa}_{l,J} \triangleq \min_{\Delta \in \widehat{C}_J:\ l(\Delta)=1}\left|\widehat{\Psi}\Delta\right|_{\infty}.$$

It depends on c, r, P, and I but for brevity we do not make the dependence explicit. A function of interest for prediction and nonparametric estimation is

$$l_F(\Delta) \triangleq \left(\sum_{i=1}^{n}\left(x_i^{\top}\widehat{\mathbf{D}}_{\mathbf{X}}\Delta\right)^2/n\right)^{1/2}.$$

If one is interested in building confidence bands by applying Section 8, we use $l(\Delta)=|\Delta|_1$. When l is the ℓ_q-norm of the subvector β_T for $T \subseteq [K]$, we define the ℓ_q-T *block sensitivity* as

$$\widehat{\kappa}_{q,T,J} \triangleq \min_{\Delta \in \widehat{C}_J:\ |\Delta_T|_q=1}\left|\widehat{\Psi}\Delta\right|_{\infty}. \tag{4.4}$$

By convention, we set $\widehat{\kappa}_{q,\varnothing,J}=\infty$ and, when $\widetilde{J}=[K]$, we use the shorthand notation $\widehat{\kappa}_{q,J}$ and call this sensitivity the ℓ_q *sensitivity.* $\widehat{\kappa}^*_{\lambda,J}$ denotes the sensitivity obtained when $l(\Delta)=\left|\lambda^{\top}\Delta\right|$ for some $\lambda \in \mathbb{R}^K$. This is useful for a linear functional of the parameters $\lambda_0^{\top}\beta$ in which case $\lambda=\widehat{\mathbf{D}}_{\mathbf{X}}\lambda_0$. When one is interested in O of them, stacking the vectors $\lambda^{\top}$ as rows of $\Omega \in \mathcal{M}_{O,K}$ and using the diagonal matrix $\widehat{\mathbf{D}}_{\Omega}$ with entries $\left(\widehat{\mathbf{D}}_{\Omega}\right)_{oo}=\left(\left|\Omega_{o\cdot}\widehat{\mathbf{D}}_{\mathbf{X}}\right|_2/K\right)^{-1}$ for $o \in [O]$, one uses $l(\Delta)=\left|\widehat{\mathbf{D}}_{\Omega}\Omega\Delta\right|_{\infty}$ and the sensitivity is $\widehat{\kappa}^*_{\Omega,J}$. In the absence of sparsity, the sensitivities are defined replacing $\widehat{C}_J$ by

$$\widehat{C}_{\gamma,J} \triangleq \left\{\Delta:\ \widehat{\mathbf{D}}_{\mathbf{X}}\Delta \in \mathcal{R}_D,\ |\Delta_{J^c\cap P}|_1 \leq 2\left(|\Delta_{J\cap P}|_1+cr|\Delta_I|_1+c|\Delta_{I^c}|_1\right)+|\Delta_{P^c}|_1\right\}.$$

They are denoted by $\widehat{\gamma}$ instead of $\widehat{\kappa}$. The sensitivities based on $l(\Delta)=|\Delta_I|_1+r^{-1}|\Delta_{I^c}|_1$ (resp., $l(\Delta)=\min\left(|\Delta_P|_1, \frac{1}{2}\left(3|\Delta_{J\cap P}|_1+cr|\Delta_I|_1+c|\Delta_{I^c}|_1+|\Delta_{P^c}|_1\right)\right)$) are denoted by $\widehat{\kappa}_{\sigma,J}$ and $\widehat{\gamma}_{\sigma,J}$ depending on the restricted set (resp., $\widehat{\gamma}_{Q,J}$) and are important for Section 4.2.

To explain the role of sensitivities, let us sketch some elements of our argument. By definition of the sensitivity, one has, for $\beta \in \mathcal{I}dent$, on the event $\mathcal{G}$, for all $\Delta \in \widehat{C}_{J(\beta)}$,

$$l(\Delta) \le \frac{\left|\widehat{\Psi}\Delta\right|_\infty}{\widehat{\kappa}_{l,J(\beta)}}. \tag{4.5}$$

Take $\Delta \in \widehat{C}_{J(\beta)}$. Inequality (4.5) is trivial if $l(\Delta) = 0$ and otherwise follows by homogeneity

$$\frac{\left|\widehat{\Psi}\Delta\right|_\infty}{l(\Delta)} \ge \min_{\widetilde{\Delta}:\ \widetilde{\Delta}\neq 0,\ \widetilde{\Delta}\in\widehat{C}_{J(\beta)}} \frac{\left|\widehat{\Psi}\widetilde{\Delta}\right|_\infty}{l\left(\widetilde{\Delta}\right)}.$$

From (4.3) and (4.5), we obtain, on $\mathcal{G}$,

$$\left|\widehat{\Psi}\Delta\right|_\infty \le 2r\overline{\sigma}\left(1 - \frac{r^2}{\widehat{\kappa}_{\sigma,J(\beta)}}\right)_+^{-1}. \tag{4.6}$$

As is apparent from (4.5) and (4.6), the sensitivities are core elements for error bounds for the performance of the *STIV*. The remaining elements on the right-hand side of (4.6) are: r, of the order of $\sqrt{\log(L)/n}$ (the parametric rate up to a logarithm), and $\widehat{\sigma}$. As shown in Section A.4, the sensitivities are a useful tool to study the performance of the Dantzig selector, but also the Lasso. They provide sharper results than the existing ones for the analysis of Dantzig selector and of the Lasso in classical high-dimensional regression. They are also applicable to non-square non-symmetric matrices. Ye and Zhang (2010) introduced similar quantities as (4.4) which differ in: the definition of $\widehat{C}_J$, the matrix $\widehat{\Psi}$, and involve a scaling by $|J(\beta)|^{1/q}$ which is inadequate when dealing with endogenous regressors. Chernozhukov, Chetverikov, and Kato (2013) proposes a variation with a slightly modified restricted set for the classical high-dimensional regression but for which it seems hard to obtain feasible lower bounds on the sensitivities because we cannot use arguments based on homogeneity.

The sensitivities measure how small $\left|\widehat{\Psi}\Delta\right|_\infty$ can get for Δ in a restricted set. Recall that, in contrast, for the classical low dimension regression without sparsity, the minimum eigenvalue involves a minimum over the whole space. As is apparent from (4.5), the error bounds are decreasing in the sensitivities. Hence, a smaller set means a larger sensitivity and thus a sharper bound. The restricted set accounts for the prior knowledge of the researcher through $\mathcal{R}_D$, P, and I, all of which make them larger. The last elements in the definition of the restricted set are $J(\beta)$ and $J(\beta)^c \cap J\left(\widehat{\beta}\right)^c$. Small $J(\beta)$ implies a small restricted set. However, $J(\beta)$ does not only play a role in the sensitivities via its cardinality. In particular, I and I^c appear in the restricted set and it matters whether the nonzero coefficients within P^c belong to I or I^c. The sensitivities also depend on $\left|\widehat{\Psi}\Delta\right|_\infty$. The ℓ_∞-norm comes from the definition of $\widehat{\mathcal{I}}(r,\sigma)$. It is motivated by computational considerations and to handle possibly very many instruments. From this, we can see that one good instrument is enough to ensure a large sensitivity. Adding instruments can only increase $\left|\widehat{\Psi}\Delta\right|_\infty$ even if they are irrelevant. The price to pay for having many is in r and takes the form of a logarithm in L, hence is very mild. To get the intuition, consider the sensitivity $\widehat{\kappa}^*_{e_k,J(\beta)}$ for $k \in [K]$. When $P = \emptyset$ and $\mathcal{R} = \mathbb{R}^K$, the restricted set becomes $\mathbb{R}^K$, $\widehat{\kappa}^*_{e_k,J(\beta)}$ no longer depends on $J(\beta)$ and can be written

$$\widehat{\kappa}^*_{e_k} = \min_{\Delta \in \mathbb{R}^{K-1}} \left| \widehat{\mathbf{D}}_{\mathbf{Z}} \left(\frac{1}{n} \sum_{i=1}^n z_i x_{ki} - \sum_{m \neq k} \left(\frac{1}{n} \sum_{i=1}^n z_i x_{mi} \right) \Delta_m \right) \right|_\infty .$$

It can be zero if $\sum_{i=1}^n z_i x_{ki}/n$ is in the range of the matrix $\widetilde{\Psi}_{-k}$ obtained from $\sum_{i=1}^n z_i x_{mi}$ by removing the kth column. This is unlikely if L exceeds K, even without sparsity. It is zero only under multicolinearity. However, for large K, this can become very small. When the minimization is carried over a subset, sparsity rules out certain combinations of the columns of $\widetilde{\Psi}_{-k}$.

In Section A.4, we analyze the sensitivities in the classical case of the linear regression without endogeneity and show that they deliver sharper bounds than the ones previously introduced in the literature on sparsity. In Section A.5, we relate the sensitivities for different losses to one another. As is apparent from these inequalities, the cases $c \in (0,1)$ and $P = I = [K]$ are more simple. Else, the situation is complex. In Section 9, we present feasible confidence sets and study their dependence in c. The case where $c \in (0,1)$ gives the sharper sets but usually requires large sample size, overidentification, and strong instruments. In Section 6, we present bounds for specific types of matrices $\widehat{\Psi}$ where we can get more information. The goal of such inequalities for sensitivities is to obtain tractable lower bounds on them since they appear in denominators. This is useful because the sensitivities involve an optimization problem so there are no closed form formulas. Most of them involve the equality of a pseudo-norm to 1. This is not a convex constraint. The ones involving $\widehat{\kappa}_{\infty,J}$ and $\widehat{\kappa}^*_{e_k,J}$ are more tractable from this perspective. Still, they involve a nonconvex constraint due to the inequality with ℓ_1-norms on the right of the inequality in the restricted set. So these are *NP*-hard to compute (see also Dobriban and Fan (2015)). Similar to the fact that for estimation we rely on convex relaxations to overcome the *NP*-hardness, for inference we also rely on relaxations. Else, we have to make strong unverifiable (due to *NP*-hardness) assumptions, in particular regarding the strength of the instruments. The feasible bounds are in Section 5.

4.2. **Basic Bounds.** In this section, we give the basic error bounds for sparse and nonsparse vectors.

Theorem 4.1. *Let $\beta, \mathbb{P}$ such that $\beta \in \mathcal{I}dent$. On $\mathcal{G}$, for all solution $\left(\widehat{\beta}, \widehat{\sigma}\right)$ of (3.7), $l \in \mathcal{L}$, and $c > 0$, we have*

$$l\left(\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\left(\widehat{\beta} - \beta\right)\right) \le \frac{2r}{\widehat{\kappa}_{l,J(\beta)}} \min\left(\overline{\sigma}\left(1 - \frac{r^2}{\widehat{\kappa}_{\sigma,J(\beta)}}\right)_+^{-1}, \sqrt{\widehat{Q}(\beta)}\left(1 - \frac{r}{c\widehat{\kappa}_{1,J(\beta)\cap P,J(\beta)}}\right)_+^{-1}\right).$$

The bound involves a minimum of two terms. The one involving $\overline{\sigma}$ is at the basis of the confidence sets and the one involving $\sqrt{\widehat{Q}(\beta)}$ is used to obtain rates of convergence. The bounds at the basis of the confidence sets depend on

$$\left(1 - \frac{r^2}{\widehat{\kappa}_{\sigma,J(\beta)}}\right)_+^{-1} \tag{4.7}$$

which is infinite on the event $\left\{\widehat{\kappa}_{\sigma,J(\beta)} \le r^2\right\}$. In the classical situation where the researcher knows $J(\beta)$, the error bounds are confidence sets. The fact that these could be infinite is in agreement with Dufour (1997) which shows that confidence sets of infinite volume cannot be avoided for procedures that are robust to weak instruments. Assuming that $\left\{\widehat{\kappa}_{\sigma,J(\beta)} \le r^2\right\}$ has a vanishing probability is a type of strong instrument assumption in our high-dimensional framework. Proposition A.4 allows to obtain sufficient conditions for $\left\{\widehat{\kappa}_{\sigma,J(\beta)} \ge r^2\right\}$. For example, the first upper bound in (A.40) yields

$$\left\{\widehat{\kappa}_{\sigma,J(\beta)} \ge r^2\right\} \supseteq \left\{\frac{2r^2|J(\beta)\cup P^c|}{\widehat{\kappa}_{\infty,J(\beta)\cup P^c,J}} + \frac{r(1-r)|I^c|}{\widehat{\kappa}_{\infty,I^c,J(\beta)}} \le 1 - cr\right\}.$$

Also, using Proposition A.4 and Theorem 4.1 yields

$$l\left(\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\left(\widehat{\beta} - \beta\right)\right) \le \frac{2r}{\widehat{\kappa}_{l,J(\beta)}}\sqrt{\widehat{Q}(\beta)}\left(1 - \frac{r|J(\beta)\cap P|}{c\widehat{\kappa}_{\infty,J(\beta)\cap P,J(\beta)}}\right)_+^{-1},$$

where the right-hand side is finite if $r|J(\beta)\cap P|$ is not too large and $c\widehat{\kappa}_{\infty,J(\beta)\cap P,J(\beta)}$ not too small. The bounds are at the basis to obtain rates of convergence rely on $\sqrt{Q(\beta)}$ and

$$\left(1 - \frac{r}{c\widehat{\kappa}_{1,J(\beta)\cap P,J(\beta)}}\right)_+^{-1} \tag{4.8}$$

instead of (4.7). In the case of i.i.d. data, the terms (4.7) and (4.8) appear because the variances of the structural errors u_i are unknown and we simultaneously estimate it.

Remark 4.3. *In the model where $\mathbf{Z} = \mathbf{X}$, we have $l_F(\Delta)^2 \le \left|\widehat{\Psi}\Delta\right|_\infty |\Delta|_1$ hence $\widehat{\kappa}_{l_F,J} \ge \sqrt{\widehat{\kappa}_{1,J}}$.*

Theorem 4.2. *Let $\beta, \mathbb{P}$ such that $\beta \in \mathcal{I}dent$. On $\mathcal{G}$, for all solution $\left(\widehat{\beta}, \widehat{\sigma}\right)$ of (3.7), $q \in [1, \infty]$, $T, J \subseteq [K]$, and $c > 0$, we have*

$$\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\left(\widehat{\beta} - \beta\right)_T\right|_q \le 2\max\left(\frac{r}{\widehat{\gamma}_{q,T,J}}\min\left(\overline{\sigma}\left(1 - \frac{r^2}{\widehat{\gamma}_{\sigma,J}}\right)_+^{-1}, \sqrt{\widehat{Q}(\beta)}\left(1 - \frac{r}{c\widehat{\gamma}_{Q,J}}\right)_+^{-1}\right), 3\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_{J^c\cap P}\right|_1\right).$$

Remark 4.4. *In the model where* $\mathbf{Z}=\mathbf{X}$*, we have, on* $\mathcal{G}$*,*

$$l_F\left(\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\left(\widehat{\beta}-\beta\right)\right)\le 2\max\left(\frac{r}{\sqrt{\widehat{\gamma}_{1,J}}}\min\left(\overline{\sigma}\left(1-\frac{r^2}{\widehat{\gamma}_{1,J}}\right)_+^{-1},\sqrt{\widehat{Q}(\beta)}\left(1-\frac{r}{c\widehat{\gamma}_{Q,J}}\right)_+^{-1}\right),\right.$$
$$\left.\sqrt{3r\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_{J^c\cap P}\right|_1\left(2\sqrt{\widehat{Q}\left(\beta\right)}+\frac{1}{c}\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_{J^c\cap P}\right|_1\right)}\right).$$

5. Computable Lower Bounds on the Sensitivities and Robust Confidence Sets

5.1. Computationally Efficient Lower Bounds on the Sensitivities.

The only unknown in the upper bound in Theorem 4.1 is $J(\beta)$. We may however have access to $\widehat{J}$ such that $J(\beta)\subseteq\widehat{J}$ and hence to a lower bound on the sensitivities using Proposition A.4 (i). If $\left|\widehat{J}\cup P^c\cup I^c\right|$ is small then we can apply the exact computations in Section A.6. We can also rely on a prior bound $|J(\beta)\cap P|\le s$. We now present easy to compute lower bounds on the sensitivities obtained by minimizing on a larger and manageable set.

Proposition 5.1. *For all* $J\subseteq\widehat{J}\subseteq[K]$ *such that* $|J\cap P|\le s$ *and* $c>0$*, we have*

$$\widehat{\kappa}_{\infty,J}\ge\max\left(\widehat{\kappa}_{\infty}\left(\widehat{J}\right),\widehat{\kappa}_{\infty}(s)\right);\quad \widehat{\kappa}_{\lambda,J}^{*}\ge\max\left(\widehat{\kappa}_{\lambda}^{*}\left(\widehat{J}\right),\widehat{\kappa}_{\lambda}^{*}(s)\right);$$
$$\widehat{\kappa}_{1,J}\ge\max\left(\widehat{\kappa}_{1}\left(\widehat{J}\right),\widehat{\kappa}_{1}(s)\right);\quad \widehat{\kappa}_{J}^{\sigma}\ge\max\left(\widehat{\kappa}_{\sigma}\left(\widehat{J}\right),\widehat{\kappa}_{\sigma}(s)\right);$$
$$\left(1-\frac{r^2}{\widehat{\kappa}_{\sigma,J}}\right)_+^{-1}\le\min\left(\widehat{\theta}_{\kappa}\left(\widehat{J}\right),\widehat{\theta}_{\kappa}(s)\right);$$

where the quantities in the bounds are defined in Table 1. The same holds for the sensitivities based on the restricted set $\widehat{C}_{\gamma,J}$ *using, instead of the sets* B*, the sets* $\widetilde{B}$ *and replacing* $\widehat{\theta}_{\kappa}$ *by* $\widehat{\theta}_{\gamma}$*.*

The lower bounds in Proposition 5.1 involve a maximum. One lower bound is useful when one is given $\widehat{J}\supseteq J$ and the other when one is given a sparsity certificate $|J\cap P|\le s$. The bounds in Table 1 involve two minima. The second one is a linear program. Hence, all lower bounds can be obtained by solving multiple linear programs, for example: $2K$ for $\widehat{\kappa}_{e_k}^{*}\left(\widehat{J}\right)$ or $\widehat{\kappa}_{e_k}^{*}(s)$ and K for $\widehat{\kappa}_{\infty}\left(\widehat{J}\right)$ or $\widehat{\kappa}_{\infty}(s)$. $\widehat{\kappa}_{\infty}(s)$ and $\widehat{\kappa}_{e_k}^{*}(s)$ are at the basis of confidence sets for the whole vector β using only a sparsity certificate. The first approach is less sharp but requires solving only K linear programs while the second requires solving $2K^2$ linear programs. Below, we sometimes refer to lower bounds such as $\widehat{\gamma}_{q,T}(s)$ which are not in Table 1, they are obtained from them and Proposition A.4. The idea behind the construction of the lower bounds can be applied to obtain the sharper bounds by considering subsets of $[K]$ of size m yielding $2^{m-1}\binom{K}{m}$ or $2^{m}\binom{K}{m}$ linear programs. This is feasible for small m if K is not too large. Table 2 considers the case where $m=2$. One can obtain lower bounds which are

TABLE 1. Table of constants

$\widehat{\kappa}_\infty\left(\widehat{J}\right) \triangleq \min_{k\in[K]} \min_{\substack{(\Delta,w)\in\widehat{B}\left(\widehat{J}\right)\\ \Delta_k=1,\ w\leq\mathbf{1}}} \left\lvert\widehat{\Psi}\Delta\right\rvert_\infty$	$\widehat{\kappa}_\infty(s) \triangleq \min_{k\in[K]} \min_{\substack{(\Delta,w)\in\widehat{B}(k)\\ \Delta_k=1,\ w\leq\mathbf{1}}} \left\lvert\widehat{\Psi}\Delta\right\rvert_\infty$
$\widehat{\kappa}^*_\lambda\left(\widehat{J}\right) \triangleq \min_{\substack{k\in[K]\\ \eta=\pm1}} \min_{\substack{(\Delta,w)\in\widehat{B}\left(\widehat{J}\right)\\ \lambda^\top\Delta=\eta,\ w\leq\Delta_k\mathbf{1}}} \left\lvert\widehat{\Psi}\Delta\right\rvert_\infty$	$\widehat{\kappa}^*_\lambda(s) \triangleq \min_{\substack{k\in[K]\\ \eta=\pm1}} \min_{\substack{(\Delta,w)\in\widehat{B}(k)\\ \lambda^\top\Delta=\eta,\ w\leq\Delta_k\mathbf{1}}} \left\lvert\widehat{\Psi}\Delta\right\rvert_\infty$
$\widehat{\kappa}_1\left(\widehat{J}\right) \triangleq \min_{k\in[K]} \min_{\substack{(\Delta,w)\in\widehat{B}\left(\widehat{J}\right)\\ \sum_{j=1,\dots,K} w_j=1,\ w\leq\Delta_k\mathbf{1}}} \left\lvert\widehat{\Psi}\Delta\right\rvert_\infty$	$\widehat{\kappa}_1(s) \triangleq \min_{k\in[K]} \min_{\substack{(\Delta,w)\in\widehat{B}(k)\\ \sum_{j=1,\dots,K} w_j=1,\ w\leq\Delta_k\mathbf{1}}} \left\lvert\widehat{\Psi}\Delta\right\rvert_\infty$
$\widehat{\kappa}_\sigma\left(\widehat{J}\right) \triangleq \min_{k\in[K]} \min_{\substack{(\Delta,w)\in\widehat{B}\left(\widehat{J}\right)\\ \sum_{j\in I} w_j+r^{-1}\sum_{j\in I^c} w_j=1,\ w\leq\Delta_k\mathbf{1}}} \left\lvert\widehat{\Psi}\Delta\right\rvert_\infty$	$\widehat{\kappa}_\sigma(s) \triangleq \min_{k\in[K]} \min_{\substack{(\Delta,w)\in\widehat{B}(k)\\ \sum_{j\in I} w_j+r^{-1}\sum_{j\in I^c} w_j=1,\ w\leq\Delta_k\mathbf{1}}} \left\lvert\widehat{\Psi}\Delta\right\rvert_\infty$
$\widehat{\theta}_\kappa\left(\widehat{J}\right) \triangleq \left(1-\frac{r^2}{\widehat{\kappa}_\sigma\left(\widehat{J}\right)}\right)_+^{-1}$	$\widehat{\theta}_\kappa(s) \triangleq \left(1-\frac{r^2}{\widehat{\kappa}_\sigma(s)}\right)_+^{-1}$
$\widehat{B}\left(\widehat{J}\right) \triangleq \left\{\begin{array}{l}(\Delta,w)\in\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\mathcal{R}_D\times\mathbb{R}^K:\ w\geq\mathbf{0},\ -w\leq\Delta\leq w,\ w_{\widehat{J}^c\cap J\left(\widehat{\beta}\right)^c}=\mathbf{0},\\ (1-cr)\sum_{j\in I} w_j+(1-c)\sum_{j\in I^c} w_j\leq 2\sum_{j\in\widehat{J}\cap P} w_j+\sum_{j\in P^c} w_j\end{array}\right\}$	
$\widehat{B}(k) \triangleq \left\{\begin{array}{l}(\Delta,w)\in\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\mathcal{R}_D\times\mathbb{R}^K:\ w\geq\mathbf{0},\ -w\leq\Delta\leq w,\\ (1-cr)\sum_{j\in I} w_j+(1-c)\sum_{j\in I^c} w_j\leq 2sw_k+\sum_{j\in P^c} w_j\end{array}\right\}$	
$\widehat{B}_\gamma\left(\widehat{J}\right) \triangleq \left\{\begin{array}{l}(\Delta,w)\in\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\mathcal{R}_D\times\mathbb{R}^K:\ w\geq\mathbf{0},\ -w\leq\Delta\leq w,\\ (1-2cr)\sum_{j\in I} w_j+(1-2c)\sum_{j\in I^c} w_j\leq 3\sum_{j\in\widehat{J}\cap P} w_j+2\sum_{j\in P^c} w_j\end{array}\right\}$	
$\widehat{B}_\gamma(k) \triangleq \left\{\begin{array}{l}(\Delta,w)\in\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\mathcal{R}_D\times\mathbb{R}^K:\ w\geq\mathbf{0},\ -w\leq\Delta\leq w,\\ (1-2cr)\sum_{j\in I} w_j+(1-2c)\sum_{j\in I^c} w_j\leq 3sw_k+2\sum_{j\in P^c} w_j\end{array}\right\}$	

TABLE 2. Table of constants for tighter sets ($m=2$)

$\widehat{\kappa}_\infty\left(\widehat{J}\right) \triangleq \min_{\substack{k,l=1,\dots,K\\ l\neq k\\ \epsilon=\pm1}} \min_{\substack{(\Delta,w)\in\widehat{B}\left(\widehat{J}\right)\\ \Delta_k=1\\ w\leq\mathbf{1},\ w_{-k}\leq\epsilon\Delta_l\mathbf{1}}} \left\lvert\widehat{\Psi}\Delta\right\rvert_\infty$	$\widehat{\kappa}_\infty(s) \triangleq \min_{\substack{k,l=1,\dots,K\\ l\neq k\\ \epsilon=\pm1}} \min_{\substack{(\Delta,w)\in B\\ \Delta_k=1\\ w\leq\mathbf{1},\ w_{-k}\leq\epsilon\Delta_l\mathbf{1}}} \left\lvert\widehat{\Psi}\Delta\right\rvert_\infty$
$\widehat{\kappa}^*_\lambda\left(\widehat{J}\right) \triangleq \min_{\substack{k,l=1,\dots,K\\ l\neq k\\ \eta=\pm1\\ \epsilon=\pm1}} \min_{\substack{(\Delta,w)\in\widehat{B}\left(\widehat{J}\right)\\ \lambda^\top\Delta=\eta\\ w\leq\Delta_k\mathbf{1},\ w_{-k}\leq\epsilon\Delta_l\mathbf{1}}} \left\lvert\widehat{\Psi}\Delta\right\rvert_\infty$	$\widehat{\kappa}^*_\lambda(s) \triangleq \min_{\substack{k,l=1,\dots,K\\ l\neq k\\ \eta=\pm1\\ \epsilon=\pm1}} \min_{\substack{(\Delta,w)\in\widehat{B}(k,l)\\ \lambda^\top\Delta=\eta\\ w\leq\Delta_k\mathbf{1},\ w_{-k}\leq\epsilon\Delta_l\mathbf{1}}} \left\lvert\widehat{\Psi}\Delta\right\rvert_\infty$
$\widehat{\kappa}_1\left(\widehat{J}\right) \triangleq \min_{\substack{k,l=1,\dots,K\\ l\neq k\\ \epsilon=\pm1}} \min_{\substack{(\Delta,w)\in\widehat{B}\left(\widehat{J}\right)\\ \sum_{j=1,\dots,K} w_j=1\\ w\leq\Delta_k\mathbf{1},\ w_{-k}\leq\epsilon\Delta_l\mathbf{1}}} \left\lvert\widehat{\Psi}\Delta\right\rvert_\infty$	$\widehat{\kappa}_1(s) \triangleq \min_{\substack{k,l=1,\dots,K\\ l\neq k\\ \epsilon=\pm1}} \min_{\substack{(\Delta,w)\in\widehat{B}(k,l)\\ \sum_{j=1,\dots,K} w_j=1\\ w\leq\Delta_k\mathbf{1},\ w_{-k}\leq\epsilon\Delta_l\mathbf{1}}} \left\lvert\widehat{\Psi}\Delta\right\rvert_\infty$
$\widehat{\kappa}_\sigma\left(\widehat{J}\right) \triangleq \min_{\substack{k,l=1,\dots,K\\ l\neq k\\ \epsilon=\pm1}} \min_{\substack{(\Delta,w)\in\widehat{B}\left(\widehat{J}\right)\\ \sum_{j\in I} w_j+r^{-1}\sum_{j\in I^c} w_j=1\\ w\leq\Delta_k\mathbf{1},\ w_{-k}\leq\epsilon\Delta_l\mathbf{1}}} \left\lvert\widehat{\Psi}\Delta\right\rvert_\infty$	$\widehat{\kappa}_\sigma(s) \triangleq \min_{\substack{k,l=1,\dots,K\\ l\neq k\\ \epsilon=\pm1}} \min_{\substack{(\Delta,w)\in\widehat{B}(k,l)\\ \sum_{j\in I} w_j+r^{-1}\sum_{j\in I^c} w_j=1\\ w\leq\Delta_k\mathbf{1},\ w_{-k}\leq\epsilon\Delta_l\mathbf{1}}} \left\lvert\widehat{\Psi}\Delta\right\rvert_\infty$
w_{-k} is the vector in $\mathbb{R}^{K-1}$ obtained from w by removing the kth row, $\widetilde{B}$ and $\widetilde{B}(k,l)$ are obtained similarly	
$\widehat{B}\left(\widehat{J}\right) \triangleq \left\{\begin{array}{l}(\Delta,w)\in\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\mathcal{R}_D\times\mathbb{R}^K:\ w\geq\mathbf{0},\ -w\leq\Delta\leq w,\ w_{\widehat{J}^c\cap J\left(\widehat{\beta}\right)^c}=\mathbf{0}\\ (1-cr)\sum_{j\in I} w_j+(1-c)\sum_{j\in I^c} w_j\leq 2\sum_{j\in\widehat{J}\cap P} w_j+\sum_{j\in P^c} w_j\end{array}\right\}$	
$\widehat{B}(k,l) \triangleq \left\{\begin{array}{l}(\Delta,w)\in\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\mathcal{R}_D\times\mathbb{R}^K:\ w\geq\mathbf{0},\ -w\leq\Delta\leq w,\\ (1-cr)\sum_{j\in I} w_j+(1-c)\sum_{j\in I^c} w_j\leq s(w_k+w_l)+\sum_{j\in P^c} w_j\end{array}\right\}$	

hybrid between the previous bounds and those of Section A.6. They are obtained by working with vectors of signs $(\epsilon_j)_{j\in S}$, where $S \subseteq [K]$ is small, and imposing that, for $j \in S$, $\epsilon_j = \pm 1$. They are obtained by adding the constraints $w_j = \epsilon_j \Delta_j$ for j in small subsets of $[K]$ (*e.g.*, P^c, I^c, and $\widehat{J} \cap P$), and adding to the first minimum a minimum over all signs.

5.2. **Feasible Confidence Sets Under a Sparsity Certificate.** The confidence sets in this section are valid if the Scenario chosen from those of Section 3.5 is correct. They do not restrict the joint distribution of $(\mathbf{Z}, \mathbf{X})$. Hence, the sets are robust to identification and many instruments.

Theorem 5.1. *For all $\beta, \mathbb{P}$ such that $\beta \in \mathcal{I}dent$, on $\mathcal{G}$, for all solution $\left(\widehat{\beta}, \widehat{\sigma}\right)$ of* (3.7)*, $s \in [K-p]$, $l \in \mathcal{L}$, $T \subseteq [K]$, $q \in [1, \infty]$, and $c > 0$, we have, if $\beta \in \mathcal{B}_s$,*

$$l\left(\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\left(\widehat{\beta} - \beta\right)\right) \le \frac{2r\overline{\sigma}\widehat{\theta}_{\kappa}(s)}{\widehat{\kappa}_l(s)}. \tag{5.1}$$

These sets are all obtained on the same event $\mathcal{G}$ and when the researcher is uncertain about a sparsity certificate she can draw nested sets for increasing values of s. The parameter c appears in the definitions of the *STIV* estimator and in the bound in Theorem 5.1 . Choosing small c leads to a smaller restricted set but implies that we penalize less σ in (3.7), which tends to increase the resulting $\widehat{\sigma}$ and $\overline{\sigma}$. Overall, there might be some optimal c. However, the dependency of the bounds on c does not have a tractable form. Importantly, just like for s, the result of Theorem 5.1 is uniform in c and the event does not depend on it. Because the procedure is fast to implement, it is possible to vary c on a grid, intersect the obtained sets, and obtain a confidence set of coverage $1 - \alpha$.

6. Rates, Model Selection, and Refined Confidence Sets

6.1. **Deterministic Bounds.** Denote by $\Psi \triangleq \mathbf{D}_Z \mathbb{E}[ZX^\top]\mathbf{D}_X$. We consider subclasses $\widetilde{\mathcal{P}}_j$ of $\mathcal{P}_j$ for $j \in [5]$ defined below. For asymptotic statements c can grow with n.

Assumption 6.1. *$L, K \ge 3$ and for $j \in [5]$, $\alpha_\infty(n)$, m_4, $M(L,K)$, $M_X(K)$, $M_Z'(L)$, $M_{ZU}(L)$, and $B(n,L)$ positive, for all $\beta, \mathbb{P}$ such that $\beta \in \mathcal{I}dent$, where $\mathcal{P} = \widetilde{\mathcal{P}}_j \subseteq \mathcal{P}_j$, we have, if $j \in [4]$, (S5.ii), (S5.iv), and $\mathbb{P}\left(\left|\mathbf{D}_Z \mathbf{Z}^\top\right|_\infty > B(n,L)\right) \le \alpha_\infty(n)$ or Assumption 3.1 holds, and, for all $j \in [5]$,*

$$\mathbb{E}\left[\left|\mathbf{D}_Z\left(ZX^\top - \mathbb{E}\left[ZX^\top\right]\right)\mathbf{D}_X\right|_\infty^2\right] \le M(L,K), \tag{6.1}$$

$$\mathbb{E}\left[\left|\left(\frac{X_k^2}{\mathbb{E}\left[X_k^2\right]} - 1\right)_{k=1}^K\right|_\infty^2\right] \le M_X(K). \tag{6.2}$$

For scenarii 1-4 without Assumption 3.1, we use $\alpha_C(n)=\alpha_\infty(n)+\left(m_4/\tau^2+C_{\mathrm{N}}(L)M'_Z(L)/(\tau'_Z)^2\right)/n$ and $\overline{r}\triangleq r_0B(n,L)/\sqrt{1-\tau'_Z}$ as a deterministic upper bound on r. For scenarii 1-4 with Assumption 3.1, we use $\overline{r}\triangleq r_0\sqrt{1+\tau_{\mathcal{G}}}$. For Scenario 5, we use $\overline{r}\triangleq\overline{r}(\tau_Z,\tau'_Z,\zeta_1)$. We use $\alpha_D(n)\triangleq\alpha_B(n)+\alpha_C(n)+\left(C_{\mathrm{N}}(K)M_X(K)/\tau_X^2+C_{\mathrm{N}}(LK)M(L,K)/r_\Psi^2\right)/n$ for all scenarii. $\tau,\tau_Z,\tau'_Z,\tau_{ZU},\zeta_1,\tau_X,r_\Psi\in(0,1)$ are arbitrary sequences decaying to zero with n and such that $\overline{r}\in(0,1)$. In the rest of the section, we consider that $\mathcal{P}=\widetilde{\mathcal{P}}_j$ for some $j\in[5]$ chosen by the researcher. The results are stated on the event $\mathcal{G}\cap\mathcal{G}_\Psi$ which has probability at least $1-\alpha-\alpha_D(n)$. The restricted sets for the population sensitivities are defined as

$$C_J\triangleq\left\{\Delta:\ \mathbf{D}_X\Delta\in\mathcal{R}_D,\ \left(\sqrt{\frac{1-\tau_X}{1+\tau_X}}-c\overline{r}\right)|\Delta_I|_1+\left(\sqrt{\frac{1-\tau_X}{1+\tau_X}}-c\right)|\Delta_{I^c}|_1\le 2|\Delta_{J\cap P}|_1+|\Delta_{P^c}|_1\right\}$$

$$C_J^\gamma\triangleq\left\{\Delta:\ \mathbf{D}_X\Delta\in\mathcal{R}_D,\ \left(\sqrt{\frac{1-\tau_X}{1+\tau_X}}-2c\overline{r}\right)|\Delta_I|_1+\left(\sqrt{\frac{1-\tau_X}{1+\tau_X}}-2c\right)|\Delta_{I^c}|_1\le 3|\Delta_{J\cap P}|_1+2|\Delta_{P^c}|_1\right\}.$$

We denote by 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 5.1, $\widehat{\Psi}$, $\widehat{C}_J$, $\widehat{C}_J^\gamma$, and r, by Ψ, C_J, C_J^γ, and $\overline{r}$. We define similarly $\theta_\kappa(s)$ and $\theta_\gamma(s)$. These restricted sets are almost identical to the ones for the sensitivities and the results of Proposition A.4 hold with minor modifications for the population sensitivities. The following result relates random quantities to their population counterparts.

Proposition 6.1. *Under Assumption 6.1, for all $\beta,\mathbb{P}$ such that $\beta\in\mathcal{I}dent$, on an event $\mathcal{G}_\Psi$ of probability $1-\alpha_D(n)$ such that $\mathbb{P}\left(\mathcal{G}\cap\mathcal{G}_\Psi\right)\ge 1-\alpha-\alpha_D(n)$, we have, for all $c>0$, $r\le\overline{r}$ and*

$$\sigma^2_{U(\beta)}(1-\tau)\le\widehat{Q}(\beta)\le\sigma^2_{U(\beta)}(1+\tau);$$

$$\forall b\in\mathbb{R}^K,\ l\in\mathcal{L},\ \sqrt{1-\tau_X}l\left(\mathbf{D}_X^{-1}b\right)\le l\left(\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}b\right)\le\sqrt{1+\tau_X}l\left(\mathbf{D}_X^{-1}b\right);\tag{6.3}$$

$$\forall J\subseteq[K],\ l\in\mathcal{L},\ \widehat{\kappa}_{l,J}\ge\frac{\kappa_{l,J}}{\sqrt{(1+\tau_Z)(1+\tau_X)}}\left(1-\frac{r_\Psi}{\kappa_{1,J}}\right);\tag{6.4}$$

$$\widehat{\gamma}_{l,J}\ge\frac{\gamma_{l,J}}{\sqrt{(1+\tau_Z)(1+\tau_X)}}\left(1-\frac{r_\Psi}{\gamma_{1,J}}\right).\tag{6.5}$$

The lower bounds in Proposition 5.1 involving the sparsity certificates hold if we remove the hats.

For the lower bound (6.4) to be meaningful, we should have $\kappa_{1,J}>r_\Psi$ and r_Ψ could be taken of the order of $\sqrt{\ln(LK)M(L,K)\log(n)/n}$ and the same for (6.5) and $\gamma_{1,J}$. The following proposition follows from the definition of the population sensitivities. The condition, when holding for all J such that $|J|\le s$, is a type of *null-space property with order s*. It can be satisfied even if $\mathrm{rank}\,(\Psi)<K$.

Proposition 6.2. *Let $J, P \subseteq [K]$. The population sensitivities based on C_J, respectively on $C_{\gamma,J}$, are positive if and only if, for all $x \in \ker(\Psi) \setminus \{0\}$, we have $x \in C_J^c$, respectively $x \in C_{\gamma,J}^c$.*

Lower bounds on the population sensitivities in benchmark cases are given in Section A.7. These include a situation often encountered in nonparametric instrumental variables and one where there are less instruments than potential regressors.

6.2. **Rates of Convergence.** In this section, we derive rates of convergence of *STIV* estimators. The argument is based on replacing the random right-hand sides in Theorems 4.1 and 4.2 by deterministic upper bounds. We make use of the following notations, for all $J \subset [K]$,

$$\Theta_\kappa(J) \triangleq \sqrt{(1+\tau_Z)(1+\tau_X)}\left(1 - \frac{r_\Psi}{\kappa_{1,J}} - \frac{\overline{r}\sqrt{(1+\tau_Z)(1+\tau_X)}}{c\kappa_{1,J\cap P,J}}\right)_+^{-1};$$

$$\Theta_\gamma(J) \triangleq \sqrt{(1+\tau_Z)(1+\tau_X)}\left(1 - \frac{r_\Psi}{\gamma_{1,J}} - \frac{\overline{r}\sqrt{(1+\tau_Z)(1+\tau_X)}}{c\gamma_{Q,J}}\right)_+^{-1}.$$

Theorem 6.1. *Under Assumption 6.1, for all $\beta, \mathbb{P}$ such that $\beta \in \mathcal{I}dent$, on $\mathcal{G} \cap \mathcal{G}_\Psi$, for all solution $\left(\widehat{\beta}, \widehat{\sigma}\right)$ of* (3.7) *and $c \in \left(0, \overline{r}^{-1}\sqrt{(1-\tau_X)/(1+\tau_X)}\right)$, we have*

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

$$\sqrt{1-\tau_X}\, l\left(\mathbf{D}_X^{-1}\left(\widehat{\beta} - \beta\right)\right) \le \frac{2\overline{r}\sigma_{U(\beta)}\sqrt{1+\tau}}{\kappa_{l,J(\beta)}}\Theta_\kappa(J(\beta));$$

$$\sigma_{U(\beta)}\left(\sqrt{1-\tau} - \frac{2\overline{r}^2\sqrt{1+\tau}\,\Theta_\kappa\left(J(\beta)\right)}{\kappa_{\sigma,J(\beta)}}\right) \le \sqrt{\widehat{Q}\left(\widehat{\beta}\right)} \le \widehat{\sigma} \le \sigma_{U(\beta)}\sqrt{1+\tau}\left(1 + \frac{2\overline{r}\,\Theta_\kappa\left(J(\beta)\right)}{c\kappa_{1,J(\beta)\cap P,J(\beta)}}\right);$$

(ii) *For all $q \in [1, \infty]$ and $T \subseteq [K]$,*

$$\sqrt{1-\tau_X}\left|\mathbf{D}_X^{-1}\left(\widehat{\beta} - \beta\right)_T\right|_q \le 2\min_{J\subseteq[K]}\max\left(\frac{\overline{r}\sigma_{U(\beta)}\sqrt{1+\tau}}{\gamma_{q,T,J}}\Theta_\gamma(J), 3\sqrt{1+\tau_X}\left|\mathbf{D}_X^{-1}\beta_{J^c\cap P}\right|_1\right);$$

$$\sqrt{1-\tau}\left(\sigma_{U(\beta)} - \min_{J\subseteq[K]}\max\left(\sqrt{\frac{1+\tau}{1-\tau}}\frac{2\overline{r}^2\sigma_{U(\beta)}\Theta_\gamma\left(J\right)}{\gamma_{\sigma,J}}, \sqrt{\frac{1+\tau_X}{1-\tau}}\frac{2}{c}\left|\mathbf{D}_X^{-1}\beta_{J^c\cap P}\right|_1\right)\right)$$

$$\le \sqrt{\widehat{Q}\left(\widehat{\beta}\right)} \le \widehat{\sigma} \le \sqrt{1+\tau}\left(\sigma_{U(\beta)} + \frac{1}{c}\min_{J\subseteq[K]}\max\left(\frac{2\overline{r}\sigma_{U(\beta)}\Theta_\gamma\left(J\right)}{\gamma_{Q,J}}, 3\sqrt{\frac{1+\tau_X}{1+\tau}}\left|\mathbf{D}_X^{-1}\beta_{J^c\cap P}\right|_1\right)\right);$$

(iii) *If we add to the definition of $\mathcal{I}dent$ the restriction*

$$\forall k \in J(\beta),\ \sqrt{(1-\tau_X)\mathbb{E}[X_k^2]}\,|\beta_k| > \omega_k \triangleq \frac{2\overline{r}\sigma_{U(\beta)}\sqrt{1+\tau}}{\kappa^*_{e_k,J(\beta)}}\Theta_\kappa(J(\beta)) \tag{6.6}$$

then $J(\beta) \subseteq J\left(\widehat{\beta}\right)$ and the inequalities of item (i) *hold when we work with the sharper population sensitivities where we add $\Delta_{J(\beta)^c} = \mathbf{0}$ in the restricted sets;*

If we compute the right-hand side of the inequality in Theorem 4.2 at $J = J\left(\widehat{\beta}\right)$*, the second term is smaller than* $\left|J(\beta)\cap P\setminus J\left(\widehat{\beta}\right)\right|\frac{2\overline{r}\sigma_{U(\beta)}\sqrt{1+\tau}}{\kappa^*_{e_k,J(\beta)}}\Theta_\kappa(J(\beta))$.

The second inequalities of item (i) and (ii) allow to sandwich $\widehat{\sigma}$ and $\sqrt{\widehat{Q}\left(\widehat{\beta}\right)}$ by expressions in $\sigma_{U(\beta)}$, hence both are consistent estimators when $n \to \infty$, L and/or K could increase with n, and $\tau \to 0$ provided the other terms converge to zero. Alternatively, one can use the first inequalities and the fact that, by the inverse triangle inequality, $\left|\sqrt{\widehat{Q}(\beta)} - \sqrt{\widehat{Q}\left(\widehat{\beta}\right)}\right| \le l_F\left(\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\left(\widehat{\beta}-\beta\right)\right)$. The first inequality in item (ii) implies that our estimator adapts to the unknown β, *i.e.*, it performs as well as if we knew β and the optimal set $J = J_*$ such that

$$\left|\mathbf{D}_X^{-1}\left(\widehat{\beta}-\beta\right)\right|_1 \le \frac{2\overline{r}\sigma_{U(\beta)}}{\gamma_{1,J_*}}\sqrt{\frac{1+\tau}{1-\tau_X}}\Theta_\gamma(J_*). \tag{6.7}$$

This allows to define formally approximately sparse vectors as vectors which are sufficiently well approximated by a sparse vector so that the right-hand side of (6.7) is small. For large enough sample size ($n \gg \ln(L)$), r and r_Ψ are small, and $\Theta_\kappa(J(\beta))$ is approaching 1 as $r, r_\Psi \to 0$. Also, (A.37) holds for the population sensitivities and, when $I^c \cup P^c = \emptyset$ and $\mathbf{Z} = \mathbf{X}$, based on (4.1) and Proposition A.4, one can take c of the order of $\overline{r}^{-1}$ to estimate well the noise level. Then, the bound (i) taking as l the ℓ_q-norm is of the order $O(\overline{r}|J(\beta)|^{1/q})$ and $O(\overline{r})$ respectively. These are the same rates, in terms of the sparsity $|J(\beta)|$, the dimension L, and the sample size n, that were proved for the Lasso and Dantzig selector in high-dimensional regression with Gaussian errors, fixed regressors, and without endogenous variables in Candès and Tao (2007), Bickel, Ritov and Tsybakov (2009), and Lounici (2008). In the presence of endogenous regressors, there is not such a clear dependence in c and the sparsity. We refer to Proposition A.4 and the examples for various possible regimes depending on the strength of instruments, the number of endogenous regressors, the identity of the coefficients which are nonzero, which all play a role in a complicated way and through an optimization program. The advantage of the feasible sets in Section 5 is to discipline the choice of c in practice. We provide a rule of thumb which works well in practice in Section 9.1.3.

Condition (6.6) and (6.10) below are *beta-min* conditions. They restrict the joint distribution of $(\mathbf{Z}, \mathbf{X})$ which affects the rates of convergence via the sensitivities. Hence, unlike those based on sparsity certificates, the confidence sets relying on the beta-min conditions are not robust to identification.

6.3. **Confidence Sets Based on an Estimated Superset of the Important Regressors.** Based on Proposition 5.1 and Theorem 6.1, we obtain the following confidence sets.

Theorem 6.2. *Let $c \in \left(0, \overline{r}^{-1}\sqrt{(1-\tau_X)/(1+\tau_X)}\right)$. Under Assumption 6.1, for all $\beta, \mathbb{P}$ such that $\beta \in \mathcal{I}dent$ satisfies* (6.6)*, on $\mathcal{G} \cap \mathcal{G}_\Psi$, for all solution $\left(\widehat{\beta}, \widehat{\sigma}\right)$ of* (3.7)*, $\widehat{J} = J\left(\widehat{\beta}\right)$, and $l \in \mathcal{L}$, we have*

$$\max\left(\sqrt{1-\tau_X}\, l\left(\mathbf{D}_X^{-1}\left(\widehat{\beta}-\beta\right)\right), l\left(\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\left(\widehat{\beta}-\beta\right)\right)\right) \le \frac{2r\overline{\sigma}\widehat{\theta}_\kappa\left(\widehat{J}\right)}{\widehat{\kappa}_l\left(\widehat{J}\right)}. \tag{6.8}$$

When we do not restrict $\mathcal{I}dent$, the right-hand side of (6.8) *becomes*

$$\frac{2r\overline{\sigma}\widehat{\theta}_\gamma\left(\widehat{J}\right)}{\widehat{\gamma}_l\left(\widehat{J}\right)} + \left|J(\beta) \cap P \setminus J\left(\widehat{\beta}\right)\right| \frac{2\overline{r}\sigma_{U(\beta)}\sqrt{1+\tau}}{\kappa^*_{e_k, J(\beta)}}\Theta_\kappa(J(\beta)).$$

6.4. **Selection of Variables.** Theorem 6.1 (iii) provides an upper estimate on the set of nonzero or important components of β. Exact selection of variables can be performed as well. For this purpose, we use the thresholded *STIV* estimator $\widehat{\beta}^\omega$ which coordinates are defined, for $k \in [K]$, by

$$\widehat{\beta}^\omega_k \triangleq \widehat{\beta}_k \mathbb{1}\left\{\sqrt{\mathbb{E}_n[X_k^2]}\left|\widehat{\beta}_k\right| > \widehat{\omega}_k(s)\right\},\ \widehat{\omega}_k(s) \triangleq \frac{2r\overline{\sigma}\widehat{\theta}_\kappa(s)}{\widehat{\kappa}^*_{e_k}(s)}. \tag{6.9}$$

Denote by

$$\theta_\kappa(s) \triangleq \sqrt{(1+\tau_Z)(1+\tau_X)}\left(1 - \frac{r_\Psi}{\kappa_1(s)} - \frac{\overline{r}\sqrt{(1+\tau_Z)(1+\tau_X)}}{\kappa^\sigma(s)}\right)_+^{-1};$$

$$\Theta^\sigma_\kappa(s) \triangleq \left(1 - \frac{r_\Psi}{\kappa_1(s)} + \frac{\overline{r}s\sqrt{(1+\tau_Z)(1+\tau_X)}}{c\kappa_\infty(s)}\right)\left(1 - \frac{r_\Psi}{\kappa_1(s)} - \frac{\overline{r}s\sqrt{(1+\tau_Z)(1+\tau_X)}}{c\kappa_\infty(s)}\right)_+^{-1};$$

$$\omega_k(s) \triangleq \frac{\overline{r}\sigma_{U(\beta)}\sqrt{1+\tau}(\Theta^\sigma_\kappa(s)+1)\theta_\kappa(s)}{\kappa^*_{e_k}(s)}.$$

The following theorem shows that, based on thresholding a *STIV* estimator, on $\mathcal{G} \cap \mathcal{G}_\Psi$, we achieve sign consistency, hence $J\left(\widehat{\beta}^\omega\right) = J(\beta)$.

Theorem 6.3. *Let $c \in \left(0, \overline{r}^{-1}\sqrt{(1-\tau_X)/(1+\tau_X)}\right)$, and $s \in [p]$. Under Assumption 6.1, for all $\beta, \mathbb{P}$ such that $\beta \in \mathcal{B}_s$ satisfies*

$$\forall k \in J(\beta),\ \sqrt{(1-\tau_X)\mathbb{E}[X_k^2]}|\beta_k| > 2\omega_k(s), \tag{6.10}$$

on $\mathcal{G} \cap \mathcal{G}_\Psi$, for all solution $\left(\widehat{\beta}, \widehat{\sigma}\right)$ of (3.7) *and $\widehat{\beta}^\omega$ defined in* (6.9)*, we have $\overrightarrow{\operatorname{sign}\left(\widehat{\beta}^\omega\right)} = \overrightarrow{\operatorname{sign}(\beta)}$ and the inequalities of Theorem 6.1 item* (i) *hold when we work with the sharper population sensitivities where we add $\Delta_{J(\beta)^c} = \mathbf{0}$ in the restricted sets.*

6.5. **Adaptive Confidence Sets.**

Theorem 6.4. *Let* $c \in \left(0, \overline{r}^{-1}\sqrt{(1-\tau_X)/(1+\tau_X)}\right)$ *and* $s \in [p]$. *Under Assumption 6.1, for all* $\beta, \mathbb{P}$ *such that* $\beta \in \mathcal{B}_s$ *satisfies* (6.10), *on* $\mathcal{G} \cap \mathcal{G}_\Psi$, *for all solution* $\left(\widehat{\beta}, \widehat{\sigma}\right)$ *of* (3.7), $\widehat{J} = J\left(\widehat{\beta}^\omega\right)$ *where* $\widehat{\beta}^\omega$ *is defined in* (6.9), *the conclusion of Theorem 6.2 holds.*

The sparsity certificate s can be taken large, possibly K. If the beta-min assumption holds for that s, the width of the confidence sets matches the upper bound in the error bounds which depends on the unknown and set of smaller cardinality $J(\beta)$, hence the terminology adaptive here.

Remark 6.1. *It is impossible to obtain honest confidence sets which diameter corresponds to the optimal rate of estimation in high-dimensional regression (see Nickl and Van de Geer (2013) for the* ℓ_2*-norm and Gaussian errors independent of the regressors). It becomes possible if one removes from the parameter space vectors which are too close to* $|J(\beta)|$*-sparse vectors.* (6.10) *is a* ℓ_∞*-norm analogue.*

7. Generalizations of the *STIV*

7.1. **Model with Approximation Errors.** We consider the model

$$\forall i \in [n],\ y_i = x_i^\top \beta + v_i + w_i; \tag{7.1}$$

$$\mathbb{E}[z_i w_i] = \mathbf{0}; \tag{7.2}$$

$$\beta \in \mathcal{R},\ \mathbb{P}(\beta) \in \widetilde{\mathcal{P}}; \tag{7.3}$$

where $\widetilde{\mathcal{P}}$ is either of $\widetilde{\mathcal{P}}_j$ for $j \in [5]$ and we modify $\widetilde{\mathcal{P}}_j$ so that $u_i(\beta) = v_i(\beta) + w_i(\beta)$, where $w_i(\beta)$ plays the role of $u_i(\beta)$ in the previous definition of $\widetilde{\mathcal{P}}_j$, $\mathcal{G}$ is modified accordingly, and (S5.ii) also holds for $v_i(\beta)$ with $\sigma_{V(\beta)} \le E$. E is a parameter of the class $\widetilde{\mathcal{P}}$. $\mathcal{G}_\Psi$ is defined like before replacing $\mathcal{E}_U^c$ by $\mathcal{E}_V^c \cap \mathcal{E}_W^c$, where both are defined like $\mathcal{E}_U$ (*c.f.*, the analysis of Scenario 5 in the appendix), and the probability $m_4(\tau^2 n)^{-1}$ by $2m_4(\tau^2 n)^{-1}$. The identified set for this model is

$$\mathcal{I}dent \triangleq \left\{\beta \in \mathcal{R}:\ \mathbb{P}(\beta) \in \widetilde{\mathcal{P}} \text{ and } \forall i \in [n],\ \mathbb{E}[z_i w_i(\beta)] = \mathbf{0}\right\}.$$

Formulation (7.1) allows to handle the model in Section 2.2.5. It also allows to handle the case where, for i in $B \subseteq [n]$, the outcomes are observed in brackets. Indeed, we can assume that, for $i \in B^c$, y_i is the observed outcome and $v_i = 0$, while, for $i \in B$, y_i is the midpoint of the bracket, $y_i - v_i$ is the unobserved outcome. There, one has $|v_i| \le e_i$, where e_i are half-widths of the brackets.

Definition 7.1. *For $c>0$ and $\rho_E \geq \mathbb{E}_n[V^2]^{1/2}$, the* E-STIV *estimator $\left(\widehat{\beta}, \widehat{\sigma}\right)$ is any solution of*

$$\min_{\substack{\beta \in \widehat{\mathcal{I}}_E(r,\sigma) \\ \sigma \geq 0}} \left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1} \beta_P\right|_1 + c\sigma,$$

where

$$\widehat{\mathcal{I}}_E(r,\sigma) \triangleq \left\{\beta \in \mathcal{R},\ \left|\frac{1}{n}\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{Z}^\top(\mathbf{Y}-\mathbf{X}\beta)\right|_\infty \leq r\sigma + (r+1)\rho_E,\ \widehat{Q}(\beta) \leq \sigma^2\right\}.$$

This is a second-order cone program. For bracketed outcomes, we take $\rho_E = n^{-1}\sum_{i \in B} e_i^2$ and $E^2 = \mathbb{E}_n\mathbb{E}\left[\mathbb{1}\{i \in B\}e_i^2\right]$. For the partially linear model, we take $\rho_E = \sqrt{1+\tau}E$ so that $\mathbb{E}_n[V^2] \leq \rho_E^2$ on $\mathcal{G}_\Psi$. The following theorem, where $\overline{r}(\beta) \triangleq \min\left(\overline{r} + (\overline{r}+1)\left(\sqrt{\frac{1-\tau}{1+\tau}}\frac{\sigma_W(\beta)}{E} - 1\right)_+^{-1}, 1\right)$ and $\sigma(\beta) \triangleq \sigma_W(\beta) + (\overline{r}+2)\, E$, holds.

Theorem 7.1. *All results of Theorem 6.1 hold for the* E-STIV *estimator, replacing $\overline{r}$ by $\overline{r}(\beta)$ in the definition of the population sensitivities and the condition $c \in \left(0, \overline{r}^{-1}\sqrt{(1-\tau_X)/(1+\tau_X)}\right)$, but the bounds for $\widehat{\sigma}$ and $\sqrt{\widehat{Q}(\widehat{\beta})}$ which become*

$$\begin{aligned}
&\sqrt{1-\tau}\left(\sigma_{W(\beta)} - \sqrt{\frac{1+\tau}{1-\tau}}\left(E + \frac{2\overline{r}\ \overline{r}(\beta)\sigma(\beta)\Theta_\kappa\left(J(\beta)\right)}{\kappa_{\sigma,J(\beta)}}\right)\right) \\
&\leq \sqrt{\widehat{Q}\left(\widehat{\beta}\right)} \leq \widehat{\sigma} \leq \sqrt{1+\tau}\left(\sigma_W(\beta) + E + \frac{2\overline{r}\sigma(\beta)\Theta_\kappa\left(J(\beta)\right)}{c\kappa_{1,J(\beta)\cap P,J(\beta)}}\right)
\end{aligned}$$

for sparse vectors and, else,

$$\begin{aligned}
&\sqrt{1-\tau}\left(\sigma_{W(\beta)} - \sqrt{\frac{1+\tau}{1-\tau}}\left(E + \min_{J \subseteq [K]}\max\left(\sqrt{\frac{1+\tau}{1-\tau}}\frac{2\overline{r}\overline{r}(\beta)\sigma(\beta)\Theta_\gamma\left(J\right)}{\gamma_{\sigma,J}}, \sqrt{\frac{1+\tau_X}{1-\tau}}\frac{2}{c}\left|\mathbf{D}_X^{-1}\beta_{J^c\cap P}\right|_1\right)\right)\right) \\
&\leq \sqrt{\widehat{Q}\left(\widehat{\beta}\right)} \leq \widehat{\sigma} \leq \sqrt{1+\tau}\left(\sigma_{W(\beta)} + E + \frac{1}{c}\min_{J \subseteq [K]}\max\left(\frac{2\overline{r}\sigma(\beta)\Theta_\gamma\left(J\right)}{\gamma_{Q,J}}, 3\sqrt{\frac{1+\tau_X}{1+\tau}}\left|\mathbf{D}_X^{-1}\beta_{J^c\cap P}\right|_1\right)\right).
\end{aligned}$$

7.2. **Systems with Approximation Errors.** Though, *E-STIV* estimators can be obtained for each equation separately, we present a method which allows to handle cross-equation restrictions. Consider

$$\forall g \in [G],\ i \in [n],\ y_{g,i} = x_i^\top\beta_g + v_{g,i} + w_{g,i};$$

$$\mathbb{E}[z_i v_{g,i}] = \mathbf{0};$$

$$\beta \triangleq (\beta_1, \ldots, \beta_G) \in \mathcal{R},\ \mathbb{P}(\beta) \in \widetilde{\mathcal{P}};$$

where $\mathbb{P}(\beta)$ is the distribution of $\left(x_i^\top, z_i^\top, u_{1,i}(\beta_1), \ldots, u_{G,i}(\beta_G)\right)_{i=1}^n$, we still denote by $\widetilde{\mathcal{P}}_j$ the classes for each Scenario $j \in [5]$, for $g \in [G]$, $u_{g,i}(\beta_g) \triangleq y_{g,i} - x_i^\top\beta_g$ can be decomposed as $v_{g,i}(\beta_g) + w_{g,i}(\beta_g)$. We maintain the distributional assumptions of Section 7.1 with G errors and $\sigma_{V_g(\beta_g)} \leq E_g$ for $g \in [G]$, where E is a vector in $\mathbb{R}^G$ of small constants which is a parameter of the class, so we replace $2m_4(\tau^2 n)^{-1}$ from the *E-STIV* by $2Gm_4(\tau^2 n)^{-1}$, choose r as in Section 3.5 replacing α by α/G and use the event

$$\mathcal{G} \triangleq \left\{ \max_{\substack{g\in[G]\\ l\in[L]}} \frac{|\mathbb{E}_n[Z_l V_g(\beta_g)]|}{\sqrt{\mathbb{E}_n[Z_l^2]\mathbb{E}_n[V_g(\beta_g)^2]}} \le r \right\}.$$

We rely on a union bound because there could be dependence between the errors in each equation and we do not model them or rely on a multi stages approach. The identified set for this model is

$$\mathcal{I}dent \triangleq \left\{\beta \in \mathcal{R} :\ \mathbb{P}(\beta) \in \widetilde{\mathcal{P}} \text{ and } \forall g \in [G],\ i \in [n],\ \mathbb{E}[z_i w_{g,i}(\beta)] = \mathbf{0}\right\}.$$

We make use of the following notations. For $J_1, \dots, J_G \subseteq [K]$, we write $J \triangleq \prod_{g=1}^G J_g$ and $J^c \triangleq \prod_{g=1}^G J_g^c$. We also write $J(\beta) \triangleq \prod_{g=1}^G J(\beta_g)$, $\mathbf{I} \triangleq I \times \dots \times I$, and $\mathbf{I}^c \triangleq I^c \times \dots \times I^c$. For two products of subsets of $[K]$ J and P, we denote by $J \cap P \triangleq \prod_{g=1}^G J_g \cap P_g$. For Δ in $\mathcal{M}_{K,G}$ and $J = \prod_{g=1}^G J_g$, where $J_1, \dots, J_G \subseteq [K]$, we denote by $\Delta_J = \left((\Delta_1)_{J_1}, \dots, (\Delta_G)_{J_G}\right)$. The sparse identified set is defined similarly. Sparsity can take various forms such as total sparsity of the matrix β, row, or column sparsity. To avoid burdensome notations, when later we use s as a sparsity certificate, we refer to total sparsity.

Definition 7.2. *For $c > 0$ and $\rho_{g,E} \ge \mathbb{E}_n[V_g^2]^{1/2}$ for all $g \in [G]$, the* SE-STIV *estimator $\left(\widehat{\beta}, \widehat{\sigma}\right)$ is any solution of*

$$\min_{\beta \in \widehat{\mathcal{I}}_{SE}(r,\sigma), \sigma_1 \ge 0, \dots, \sigma_G \ge 0} \left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_P\right|_1 + c|\sigma|_1,$$

where

$$\widehat{\mathcal{I}}_{SE}(r,\sigma) \triangleq \left\{\beta \in \mathcal{R},\ \forall g \in [G],\ \left|\frac{1}{n}\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{Z}^\top(\mathbf{Y}_g - \mathbf{X}\beta_g)\right|_\infty \le r\sigma_g + (r+1)\rho_{g,E},\ \widehat{Q}(\beta_g) \le \sigma_g^2\right\}.$$

Remark 7.1. *An alternative is to rely on a scalar parameter σ like the* C-STIV *in Section 7.3.*

This has the advantage that we have a single conic constraint. The population sensitivities are obtained replacing $|\Psi\Delta|_\infty$ by $\sum_{g=1}^G |\Psi\Delta_g|_\infty$, C_J and $C_{\gamma,J}$ by

$$C_J \triangleq \left\{ \begin{array}{l} \Delta :\ \mathbf{D}_X\Delta \in \mathcal{R}_D,\ \Delta_{J^c \cap J\left(\widehat{\beta}\right)^c} = \mathbf{0}, \\ \left(\sqrt{\frac{1-\tau_X}{1+\tau_X}} - c\overline{r}(\beta)\right)|\Delta_{\mathbf{I}}|_1 + \left(\sqrt{\frac{1-\tau_X}{1+\tau_X}} - c\right)|\Delta_{\mathbf{I}^c}|_1 \le 2\,|\Delta_{J\cap P}|_1 + |\Delta_{P^c}|_1 \end{array} \right\},$$

$$C_{\gamma,J} \triangleq \left\{\Delta :\ \mathbf{D}_X\Delta \in \mathcal{R}_D,\ \left(\sqrt{\frac{1-\tau_X}{1+\tau_X}} - c\overline{r}(\beta)\right)|\Delta_{\mathbf{I}}|_1 + \left(\sqrt{\frac{1-\tau_X}{1+\tau_X}} - c\right)|\Delta_{\mathbf{I}^c}|_1 \le 3\,|\Delta_{J\cap P}|_1 + 2\,|\Delta_{P^c}|_1\right\},$$

where

$$\overline{r}(\beta) \triangleq \max_{g\in[G]} \min\left(\overline{r} + (\overline{r}+1)\left(\sqrt{\frac{1-\tau}{1+\tau}}\frac{\sigma_{W_g(\beta_g)}}{E_g} - 1\right)_+^{-1}, 1\right),$$

and using $\overline{r}(\beta)$ instead of $\overline{r}$ in the definition of the sensitivities. We use the same notations for them as well as for $\Theta_\kappa(J)$ and $\Theta_\gamma(J)$ and denote by $\overline{\kappa}_{\sigma,J(\beta)}$ and $\overline{\overline{\gamma}}_{\sigma,J}$ the sensitivities for $l(\Delta)=\max_{g\in[G]} l_F\left(\Delta_g\right)$ and the set C_J and for $l(\Delta)=\max_{g\in[G]}\left(\left|(\Delta_g)_I\right|_1+\overline{r}(\beta)^{-1}\left|(\Delta_g)_{I^c}\right|_1\right)$ and the set $C_{\gamma,J}$.

Theorem 7.2. *For all $\beta,\mathbb{P}$ such that $\beta\in\mathcal{I}dent$, the following hold on $\mathcal{G}\cap\mathcal{G}_\Psi$ for all solution $\left(\widehat{\beta},\widehat{\sigma}\right)$ of* (3.7) *and $c\in\left(0,\overline{r}(\beta)^{-1}\sqrt{(1-\tau_X)/(1+\tau_X)}\right)$:*

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

$$l\left(\mathbf{D}_X^{-1}\left(\widehat{\beta}-\beta\right)\right)\le\frac{2\overline{r}\left(\sum_{g=1}^G\sigma_{W_g(\beta_g)}+(\overline{r}+2)\,E_g\right)}{\kappa_{l,J(\beta)}}\sqrt{\frac{1+\tau}{1-\tau_X}}\Theta_\kappa(J(\beta));$$

moreover, for all $g\in[G]$,

$$\sqrt{1-\tau}\left(\sigma_{W_g(\beta_g)}-\sqrt{\frac{1+\tau}{1-\tau}}\left(E_g+\frac{2\overline{r}\left(\sum_{g=1}^G\sigma_{W_g(\beta_g)}+(\overline{r}+2)\,E_g\right)\Theta_\kappa\left(J(\beta)\right)}{\overline{\kappa}_{\sigma,J(\beta)}\sqrt{1-\tau_X}}\right)\right)$$

$$\le\sqrt{\widehat{Q}\left(\widehat{\beta}_g\right)}\le\sqrt{1+\tau}\left(\sigma_{W_g}(\beta_g)+E_g+\frac{2\overline{r}\left(\sum_{g=1}^G\sigma_{W_g(\beta_g)}+(\overline{r}+2)\,E_g\right)\Theta_\kappa\left(J(\beta)\right)}{\overline{\kappa}_{\sigma,J(\beta)}\sqrt{1-\tau_X}}\right);$$

(ii) *For all $J=\prod_{g=1}^G J_g$ and $T=\prod_{g=1}^G T_g$, where $J_1,\dots,J_G\subseteq[K]$ and $T_1,\dots,T_G\subseteq[K]$, and $q\in[1,\infty]$, we have*

$$\left|\mathbf{D}_X^{-1}\left(\widehat{\beta}-\beta\right)_T\right|_q\le 2\max\left(\frac{\overline{r}\left(\sum_{g=1}^G\sigma_{W_g(\beta_g)}+(\overline{r}+2)\,E_g\right)}{\gamma_{q,T,J}}\sqrt{\frac{1+\tau}{1-\tau_X}}\Theta_\gamma(J),3\sqrt{\frac{1+\tau_X}{1-\tau_X}}\left|\mathbf{D}_X^{-1}\beta_{J^c\cap P}\right|_1\right);$$

in particular

$$\max_{g\in[G]}\left|\mathbf{D}_X^{-1}\left(\widehat{\beta}_g-\beta_g\right)\right|_1\le 2\max\left(\frac{\overline{r}\left(\sum_{g=1}^G\sigma_{W_g(\beta_g)}+(\overline{r}+2)\,E_g\right)}{\min_{g\in[G]}\gamma_{1,[K]\times\{g\},J}}\sqrt{\frac{1+\tau}{1-\tau_X}}\Theta_\gamma(J),3\sqrt{\frac{1+\tau_X}{1-\tau_X}}\left|\mathbf{D}_X^{-1}\beta_{J^c\cap P}\right|_1\right);$$

moreover, for all $g\in[G]$,

$$\sigma_{W_g(\beta_g)}-\sqrt{\frac{1+\tau}{1-\tau}}\left(E_g+\min_{J\subseteq[K]}\max\left(\sqrt{\frac{1+\tau}{1-\tau}}\frac{2\overline{r}\overline{r}(\beta)\sigma(\beta)\Theta_\gamma\left(J\right)}{\overline{\overline{\gamma}}_{\sigma,J}},3\sqrt{\frac{1+\tau_X}{1-\tau}}\left|\mathbf{D}_X^{-1}\beta_{J^c\cap P}\right|_1\right)\right)\le\sqrt{\frac{\widehat{Q}\left(\widehat{\beta}_g\right)}{1-\tau}}.$$

In a model where the $v_{g,i}$s are zero, we take $\rho_{g,E}=0$ and can derive the same results as for the *STIV* estimator, including the confidence sets.

7.3. **Endogenous Instruments.**

7.3.1. *The* C-STIV *estimator and confidence sets.* We present a simple extension of the *STIV* that we call *C-STIV* to emphasize that, like in Section 3.3, we rely on multiple conic constraints. The problem of checking instrument exogeneity when there is overidentification is a classical problem studied in Sargan (1958) and Basmann (1960) for the linear IV model, and in Hansen (1982) for GMM (see also Andrews (1999), Cheng and Liao (2015)). The procedure in this paper handles high-dimensions in the number of instruments and allows for fewer instruments than regressors, high-dimensions in the number of regressors, and is robust to arbitrarily weak instruments. We start this section by presenting a version that could be used without overidentification. We replace (1.2) by

$$\mathbb{E}\left[z_i u_i - \widetilde{\beta}\right] = \mathbf{0}; \tag{7.4}$$

$$\mathbb{E}\left[x_{Ii} z_{li}\left(z_{li} u_i - \widetilde{\beta}_l\right)\right] = \mathbf{0}; \tag{7.5}$$

and (1.3) by $\left(\beta, \widetilde{\beta}\right) \in \mathcal{R}$ and $\mathbb{P}\left(\beta, \widetilde{\beta}\right) \in \widetilde{\mathcal{P}}$, where $\mathbb{P}\left(\beta, \widetilde{\beta}\right)$ is the distribution of

$$\left(x_i^\top, z_i^\top, u_i(\beta), z_i u_i(\beta) - \widetilde{\beta}, x_{Ii} z_{li}\left(z_{li} u_i(\beta) - \widetilde{\beta}_l\right)\right)_{i=1}^n$$

implied by $\mathbb{P}$. Equations (7.4)-(7.5) allow endogeneity of some of the instruments (*i.e.*, $\mathbb{E}[z_{li}u_i] \neq 0$). Equation (7.5) holds if, for example, we replace (7.4) by $\mathbb{E}\left[\left. z_{li} u_i - \widetilde{\beta}_l \right| z_i\right] = 0$. The set $\mathcal{R}$ is a subset of $\mathbb{R}^K \times \mathbb{R}^L$ which accounts for the restrictions on $\left(\beta, \widetilde{\beta}\right)$. A first restriction is $\widetilde{\beta}_{\widetilde{P}^c} = \mathbf{0}$. This means that the indices in $\widetilde{P}^c$ correspond to instruments which the researcher knows are exogenous. When $\widetilde{P} = \emptyset$, the researcher knows that there are no endogenous instruments. The researcher could also know the sign of the correlation between an endogenous regressor and the structural error or that she has imperfect instruments which have smaller correlation with the structural error than the endogenous regressor (see Nevo and Rosen (2012)). The class $\widetilde{\mathcal{P}}$ corresponds to either of Scenario 1 to 4 where $z_{li}u_i - \widetilde{\beta}_l$ (resp., 1 and $x_{Ii}z_{li}$), for $l \in [L]$, play the role of u_i (resp., z_i) and we again use the notation $\widetilde{\mathcal{P}}_j$. Denote, for $s \in [p]$, $\widetilde{s} \in [\widetilde{p}]$, and $P \subseteq [K]$, by

$$\mathcal{B}_{s,\widetilde{s}} = \left\{ \left(\beta, \widetilde{\beta}\right) \in \mathcal{R} : \left\{ \begin{array}{l} \mathbb{P}\left(\beta, \widetilde{\beta}\right) \in \widetilde{\mathcal{P}},\ |J(\beta) \cap P| \leq s,\ \left|J\left(\widetilde{\beta}\right) \cap \widetilde{P}\right| \leq \widetilde{s}, \\ \forall i \in [n],\ \mathbb{E}\left[z_i u_i(\beta) - \widetilde{\beta}\right] = \mathbf{0},\ \forall i \in [n],\ l \in [L],\ \mathbb{E}\left[x_{Ii} z_{li}\left(z_{li} u_i(\beta) - \widetilde{\beta}_l\right)\right] = 0 \end{array} \right. \right\}$$

and by $\mathcal{I}dent$ the set $\mathcal{B}_{p,\widetilde{p}}$. For $\left(\beta, \widetilde{\beta}\right) \in \mathcal{I}dent$, the following event plays the role of $\mathcal{G}$ before and for conciseness we use the same notation

$$\mathcal{G} \triangleq \left\{ \max\left(\max_{l \in [L]} \frac{\left|\mathbb{E}_n\left[Z_l U(\beta) - \widetilde{\beta}_l\right]\right|}{\sqrt{\mathbb{E}_n\left[\left(Z_l U(\beta) - \widetilde{\beta}_l\right)^2\right]}}, \max_{l \in [L], k \in I} \frac{\left|\mathbb{E}_n\left[X_k Z_l\left(Z_l U(\beta) - \widetilde{\beta}_l\right)\right]\right|}{\sqrt{\mathbb{E}_n\left[\left(X_k Z_l\left(Z_l U(\beta) - \widetilde{\beta}_l\right)\right)^2\right]}} \right) \leq \overline{r}_0 \right\}$$

and $\overline{r}_0$ is obtained like in Section 3.5 ($\overline{r}_0$ plays the role of r_0 and $L(|I|+1)$, or $L|I|$ when the model has an intercept, plays the role of L).

Definition 7.3. *For $c \in (0, \overline{r}_0^{-1})$, the* C-STIV *estimator* $\left(\widehat{\beta}, \widehat{\widetilde{\beta}}, \widehat{\sigma}\right)$ *is any solution of*

$$\min_{\left(\beta,\widetilde{\beta}\right)\in\widehat{\mathcal{I}}_C(\overline{r}_0,\sigma),\sigma\geq 0} \left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_P\right|_1 + \left|\widehat{\mathbf{D}}_{\mathbf{Z}}\widetilde{\beta}_{\widetilde{P}}\right|_1 + c\sigma, \tag{7.6}$$

where

$$\widehat{\mathcal{I}}_C(\overline{r}_0,\sigma) \triangleq \left\{\left(\beta,\widetilde{\beta}\right)\in\mathcal{R}:\ \left|\widehat{\mathbf{D}}_{\mathbf{Z}}\left(\frac{1}{n}\mathbf{Z}^\top\mathbf{U}(\beta)-\widetilde{\beta}\right)\right|_\infty \leq \overline{r}_0\sigma,\ \widehat{F}\left(\beta,\widetilde{\beta}\right)\leq\sigma\right\};$$

$$\forall\left(\beta,\widetilde{\beta}\right)\in\mathbb{R}^{K+L},\ \widehat{F}\left(\beta,\widetilde{\beta}\right)\triangleq\max_{l\in[L]}\sqrt{\overline{Q}_l\left(\beta,\widetilde{\beta}\right)},\ \textit{and}\ \overline{Q}_l\left(\beta,\widetilde{\beta}\right)\triangleq\left(\widehat{\mathbf{D}}_{\mathbf{Z}}\right)_{ll}^2\mathbb{E}_n\left[\left(Z_lU(\beta)-\widetilde{\beta}_l\right)^2\right].$$

Remark 7.2. *When we do not assume* (7.5)*, we remove the second term in the maximum in the definition of $\mathcal{G}$ and replace $\widehat{\rho}_I$ and $\widehat{\rho}_{I^c}$ in the appendix by a single $\widehat{\rho}\triangleq\max_{l\in[L],\ k\in I}\left(\widehat{\mathbf{D}}_{\mathbf{Z}}\right)_{ll}\left(\widehat{\mathbf{D}}_{\mathbf{X}}\right)_{kk}\sqrt{\mathbb{E}_n\left[(X_kZ_l)^2\right]}$. The cones are then larger hence the bounds are less tight.*

The analogue of the results of sections 4.2, 5, and 6 are obtained with the correspondence of Table 17 in the appendix where we also give the main results and proofs. Computing the *C-STIV* estimator can carry a high computational cost for large L due to the L conic constraints.

7.3.2. *The* NV-STIV *estimator and confidence sets.* Instead of using the possibly endogenous instruments, it is possible to only use the known to be exogenous instruments to obtain, using the *STIV* or *C-STIV* estimator (see Remark A.1 in the appendix for its simple form in this case[2]), $\widehat{\beta}$ and the upper bounds $\widehat{b}$ and $\widehat{b}^\sigma$ such that, on an event of probability close to 1,

$$\left|\left(\widehat{\Psi}\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}(\widehat{\beta}-\beta)\right)_{\widetilde{P}}\right|_\infty\leq\widehat{b},\quad \widehat{\rho}_I\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}(\widehat{\beta}-\beta)_I\right|_1+\widehat{\rho}_{I^c}\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}(\widehat{\beta}-\beta)_{I^c}\right|_1\leq\widehat{b}^\sigma, \tag{7.7}$$

where $\widehat{\rho}_I$ and $\widehat{\rho}_{I^c}$ are defined in the appendix. The sensitivities, their lower bounds and population counterparts use $\left|\left(\widehat{\Psi}\Delta\right)_{\widetilde{P}^c}\right|_\infty$ instead of $\left|\widehat{\Psi}\Delta\right|_\infty$. The one associated to the first bound in (7.7) is

$$\widehat{\kappa}_J^\Psi\triangleq\min_{\Delta\in\widehat{C}_J:\ \left|\left(\widehat{\Psi}\Delta\right)_{\widetilde{P}}\right|_\infty=1}\left|\left(\widehat{\Psi}\Delta\right)_{\widetilde{P}^c}\right|_\infty.$$

Let us assume that we use the *STIV* estimator as a pilot estimator and denote by $\mathcal{G}_1$ the usual event $\mathcal{G}$ with $\widetilde{p}$ moments, and by r_1 the constant r under either of scenarii 1-5 adjusted so that $\mathcal{G}_1$ has probability $1-\alpha_1$ up to the usual coverage error for scenarii 4 and 5.

[2] This was the *STIV* estimator used throughout the paper in the revisions between 2012 and 2014, see *e.g.* 2012.

We distinguish 3 cases: (1) $\beta \in \mathcal{B}_s$, (2) $\beta \in \mathcal{I}dent$ and satisfies (6.6), and (3) $\beta \in \mathcal{I}dent$ and satisfies (6.10). Using the previous results for sparse vectors, (7.7) holds when $\widehat{b}$ and $\widehat{b}^\sigma$ are obtained as

$$\widehat{b} = \frac{2r_1\overline{\sigma}\widehat{\theta}_\kappa(s)}{\widehat{\kappa}^\Psi(s)},\ \widehat{b}^\sigma = \frac{2r_1^2\overline{\sigma}\widehat{\theta}_\kappa(s)}{\widehat{\kappa}^\sigma(s)} \text{ in case (1)} \quad \text{or} \quad \widehat{b} = \frac{2r_1\overline{\sigma}\widehat{\theta}_\kappa\left(\widehat{J}\right)}{\widehat{\kappa}^\Psi\left(\widehat{J}\right)},\ \widehat{b}^\sigma = \frac{2r_1^2\overline{\sigma}\widehat{\theta}_\kappa\left(\widehat{J}\right)}{\widehat{\kappa}^\sigma\left(\widehat{J}\right)} \text{ in case (2) or (3),}$$

where, in case (2), $\widehat{J} = J\left(\widehat{\beta}\right)$ and, in case (3), $\widehat{J}$ is the support of the thresholded estimator.

The *NV-STIV* estimator $\left(\widehat{\widetilde{\beta}}, \widehat{\widetilde{\sigma}}\right)$ is now simply, for $\widetilde{c} \in \left(0, r_2^{-1}\right)$, any solution of

$$\min_{\widetilde{\beta}\in\widehat{\mathcal{I}}_{NV}(\widetilde{\sigma},r_2),\widetilde{\sigma}\geq 0} \left|\widehat{\mathbf{D}}_{\mathbf{Z}}\widetilde{\beta}_{\widetilde{P}}\right|_1 + \widetilde{c}\widetilde{\sigma}, \tag{7.8}$$

where, for a set of restrictions $\widetilde{\mathcal{R}}$ on $\widetilde{\beta}$ including $\widetilde{\beta}_{\widetilde{P}^c} = \mathbf{0}$,

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

$$\widehat{F}_2\left(\beta, \widetilde{\beta}\right) \triangleq \max_{l\in\widetilde{P}} \sqrt{\overline{Q}_l\left(\beta, \widetilde{\beta}\right)} \quad \forall\left(\beta, \widetilde{\beta}\right) \in \mathbb{R}^{K+L},$$

where r_2 is obtained using Scenario 4 so that the probability of the event

$$\mathcal{G}_2 \triangleq \left\{\max\left(\max_{l\in\widetilde{P}} \frac{\left|\mathbb{E}_n\left[Z_lU(\beta) - \widetilde{\beta}_l\right]\right|}{\sqrt{\mathbb{E}_n\left[\left(Z_lU(\beta) - \widetilde{\beta}_l\right)^2\right]}}, \max_{l\in\widetilde{P},k\in I} \frac{\left|\mathbb{E}_n\left[X_kZ_l\left(Z_lU(\beta) - \widetilde{\beta}_l\right)\right]\right|}{\sqrt{\mathbb{E}_n\left[\left(X_kZ_l\left(Z_lU(\beta) - \widetilde{\beta}_l\right)\right)^2\right]}}\right) \leq r_2\right\},$$

up to the usual coverage error, is $1 - \alpha_2$. Note that Remark 7.2 still applies.

We make use of the following notations, for $\widetilde{s} \in [\widetilde{p}]$,

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

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

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

$$b_* = \frac{\overline{\theta}(\beta)}{\kappa^\Psi(s)},\ \widehat{b}^\sigma = \frac{\overline{r}_1\overline{\theta}(\beta)}{\kappa^\sigma(s)} \text{ in case (1)} \quad \text{or} \quad b_*^\sigma = \frac{\overline{\theta}(\beta)}{\kappa^\Psi\left(J(\beta)\right)},\ \widehat{b}^\sigma = \frac{r_1\overline{\theta}(\beta)}{\kappa^\sigma\left(J(\beta)\right)} \text{ in case (3)}$$

$$\overline{\theta}(\beta) \triangleq 2\overline{r}_1\sigma_{U(\beta)}\sqrt{1 + \tau}\left(1 + \frac{2\overline{r}_1\Theta_\kappa\left(J(\beta)\right)}{c\kappa_{1,J(\beta)\cap P,J(\beta)}}\right)\theta_\kappa(s)\left(1 - \frac{r_\Psi}{\kappa_{1,J(\beta)}}\right)_+^{-1}\sqrt{(1 + \tau_Z)(1 + \tau_X)}.$$

We still 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 $\mathbb{P}\left(\mathcal{G}_\Psi^c\right)$ appears in the coverage error so we simply choose α_1 and α_2 so that $\alpha_1 + \alpha_2 = \alpha$.

Theorem 7.3. *Let $\widetilde{s} \in [\widetilde{p}]$. For all vector $\widetilde{\beta}$ such that $\left(\beta, \widetilde{\beta}\right) \in \mathcal{B}_{p,\widetilde{s}}$ 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 $\widetilde{c} \in \left(0, r_2^{-1}\right)$ (and $c \in \left(0, \overline{r}_1^{-1}\right)$ in case (1)),*

$$\left|\widehat{\mathbf{D}}_{\mathbf{Z}}\left(\widehat{\widetilde{\beta}} - \widetilde{\beta}\right)\right|_\infty \le \widehat{\omega}\left(\widetilde{c}, \widetilde{s}\right); \tag{7.9}$$

$$\left|\widehat{\mathbf{D}}_{\mathbf{Z}}\left(\widehat{\widetilde{\beta}} - \widetilde{\beta}\right)\right|_1 \le \frac{2\widetilde{s}}{1 - \widetilde{c}r_2}\widehat{\omega}\left(\widetilde{c}, \widetilde{s}\right) + \frac{2\widetilde{c}\widehat{b}^\sigma}{1 - \widetilde{c}r_2}; \tag{7.10}$$

on $\mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}_\Psi$, in case (1) or (3) for any solution $\left(\widehat{\widetilde{\beta}}, \widehat{\widetilde{\sigma}}\right)$ of (7.8)*, we have, with inequalities holding for all $\widetilde{c} \in \left(0, r_2^{-1}\right)$ (and $c \in \left(0, \overline{r}_1^{-1}\right)$ in case (1)),*

$$\frac{\left|\mathbf{D}_Z\left(\widehat{\widetilde{\beta}} - \widetilde{\beta}\right)\right|_\infty}{\sqrt{1 + \tau_Z}} \le \omega\left(\widetilde{c}, \left|J\left(\widetilde{\beta}\right)\right|\right); \tag{7.11}$$

$$\frac{\left|\mathbf{D}_Z\left(\widehat{\widetilde{\beta}} - \widetilde{\beta}\right)\right|_1}{\sqrt{1 + \tau_Z}} \le \frac{2\left|J\left(\widetilde{\beta}\right)\right|}{1 - \widetilde{c}r_2}\omega\left(\widetilde{c}, \left|J\left(\widetilde{\beta}\right)\right|\right) + \frac{2\widetilde{c}b_*^\sigma}{1 - \widetilde{c}r_2}. \tag{7.12}$$

If $\widetilde{c}$ and c are fixed and we restrict $\mathcal{B}_{p,\widetilde{s}}$ so that $\left|\widetilde{\beta}_l\right| > \omega\left(\widetilde{c}, \left|J\left(\widetilde{\beta}\right)\right|\right)\sqrt{(1 + \tau_Z)\mathbb{E}\left[Z_l^2\right]}$, for all $l \in \widetilde{P}$, we have, on $\mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}_\Psi$, $J\left(\widetilde{\beta}\right) \subseteq J\left(\widehat{\widetilde{\beta}}\right)$, while, if we restrict $\mathcal{B}_{p,\widetilde{s}}$ so that, for all $l \in \widetilde{P}$, $\left|\widetilde{\beta}_l\right| > 2\omega\left(\widetilde{c}, \widetilde{s}\right)\sqrt{(1 + \tau_Z)\mathbb{E}\left[Z_l^2\right]}$, then $\overrightarrow{\mathrm{sign}\left(\widehat{\widetilde{\beta}}^\omega\right)} = \overrightarrow{\mathrm{sign}\left(\widetilde{\beta}\right)}$*, where* $\widehat{\widetilde{\beta}}^\omega \triangleq \left(\widehat{\widetilde{\beta}}_l \mathbb{1}\left\{\left|\widehat{\widetilde{\beta}}_l\right| > \sqrt{\mathbb{E}_n[Z_l^2]}\widehat{\omega}\left(\widetilde{c}, \widetilde{s}\right)\right\}\right)_{l=1}^L$.

Inequalities (7.9) and (7.10) are confidence sets and the uniformity in $\widetilde{c}$ and sometimes c allows to intersect the sets that various parameters would produce. The last statement yields "adaptive" confidence sets by replacing $\widetilde{s}$ by $\left|J\left(\widehat{\widetilde{\beta}}^\omega\right)\right|$ in (7.9) and (7.10). This theorem is usefull when r_2 is small (*i.e.*, $n \gg \ln(L|I|)$ if the model includes a constant). The first upper bounds are finite if $\left|J\left(\widetilde{\beta}\right)\right| = O\left(1/r_2^2\right) = O\left(n/\ln(L\widetilde{p})\right)$ is small enough. Bounds for ℓ_q-norms follow by interpolation.

8. Confidence Bands by a Two-Stage Procedure and Bias Correction

We now consider the construction of confidence bands using a bias correction device for the models of the previous sections. To cover all cases, β is a matrix with G columns. When there is a single equation, we have $G = 1$ and $\beta_g = \beta$ and, in the absence of approximation error, $\sigma_{W_g(\beta_g)} = \sigma_{U_g(\beta_g)}$. For $s \in [Gp]$, $\Omega \in \mathcal{M}_{O,K}$, and $\beta \in \mathcal{B}_s$, the confidence bands are around either:

(1) Each $\Omega\beta_g$ for $g \in [G]$, with or without approximation errors;
(2) Each $\Omega\beta_g + V_g(\beta_g)$ for $g \in [G]$, when there are approximation errors.

Case (2) corresponds to estimation of G functions at grid points and $V_g(\beta_g)$ are the approximation errors at the grid points. Indeed, consider a system of equations like in Section 2.2.5 with $\gamma = 0$ and suppose that the researcher wants to produce a band around $\mathbf{f}_g \triangleq (f_g(t_1), \dots, f_g(t_O))^\top$ where $(t_o)_{o=1}^O$ are the same grid points for all $g \in [G]$. With the notations of Section 7.1, each $\mathbf{f}_g$ can be written in the form (2) with $\Omega = (\varphi_k(t_o))_{o\in[O],k\in[K]}$ and $V_g(\beta_g) \in \mathbb{R}^O$ is the approximation error at every point of the grid. Any of the previous estimator from the *STIV* family can be used as a preliminary estimator to obtain $\Omega\widehat{\beta}_g$. We rely on the assumption

$$\exists \Lambda \in \mathcal{M}_{O,L}: \ \mathbb{E}\left[\Lambda Z X^\top\right] = \Omega \tag{8.1}$$

and construct a suitable approximation $\widehat{\Lambda}$. This is the core ingredient for a "bias correction" which is added to $\Omega\widehat{\beta}_g$ as follows

$$\widehat{\Omega\beta_g} \triangleq \Omega\widehat{\beta}_g + \frac{1}{n}\widehat{\Lambda}\mathbf{Z}^\top\left(\mathbf{Y}_g - \mathbf{X}\widehat{\beta}_g\right) \quad \forall g \in [G].$$

Because the multiplication by Λ appears on the left in (8.1), this amounts to estimating linear combinations of the instruments and interacting them with the estimated residuals from the first stage. Because the set of matrices Λ which satisfy (8.1) is an affine space, Λ might not be point identified unless one maintains a sparsity assumption on it. We denote, for $s \in [Gp]$, $s' \in [OL]$, and $s_r \in [L]$, by $\mathbb{P}(\beta, \Lambda)$ the distribution of $\left(x_i, z_i, (u_{g,i}(\beta))_{g=1}^G, t_i(\Lambda), \mathbf{D}_{\Lambda Z}\Lambda z_i\right)_{i=1}^n$, where $t_i(\Lambda) = \Omega - \Lambda z_i x_i^\top$, $\mathbf{D}_{\Lambda Z}$ and later $\widehat{\mathbf{D}}_{\mathbf{Z}\widehat{\Lambda}^\top}$ are defined like $\mathbf{D}_Z$ and $\widehat{\mathbf{D}}_{\mathbf{Z}}$, and $\mathcal{P}$ is a nonparametric class for it, and by

$$\mathcal{B}_{s,s',s_r} = \left\{\beta \in \mathcal{B}_s, \Lambda \in \mathcal{M}_{O,L} : \mathbb{P}(\beta,\Lambda) \in \mathcal{P}; |J(\Lambda_{o\cdot})| \le s'; \max_{o\in[O]} |J(\Lambda_{o\cdot})| \le s_r; \forall i \in [n], \Omega = \Lambda\mathbb{E}\left[z_i x_i^\top\right]\right\}.$$

Nonidentification of Λ is a concern when $L > K$. For this reason, we present two approaches. In the first approach, we require assumptions so that $\widehat{\Lambda}$ is consistent and the rate of convergence is small enough. In the second approach, we do not assume identification of a sparse Λ but rely instead on a sample splitting argument.

8.1. **Confidence Bands Under Consistent Estimation of Λ.** We obtain $\widehat{\Lambda}$ by solving

$$\min_{\Lambda \in \widehat{\mathcal{A}}(r_0', \nu), \nu > 0} \left|\Lambda \widehat{\mathbf{D}}_{\mathbf{Z}}^{-1}\right|_1 + \frac{\lambda\nu}{\widehat{\rho}}, \tag{8.2}$$

where $\lambda \in (0,1)$ and

$$\widehat{\rho} = \max_{k\in[K],l\in[L]} \left(\widehat{\mathbf{D}}_{\mathbf{X}}\right)_{kk}\left(\widehat{\mathbf{D}}_{\mathbf{Z}}\right)_{ll}\sqrt{\mathbb{E}_n\left[(X_k Z_l)^2\right]};$$

$$\widehat{\mathcal{A}}(r_0', \nu) \triangleq \left\{\Lambda \in \mathcal{M}_{O,L} : \ \left|\left(\Omega - \frac{1}{n}\Lambda\mathbf{Z}^\top\mathbf{X}\right)\widehat{\mathbf{D}}_{\mathbf{X}}\right|_\infty \le r_0'\nu, \ \widehat{F}(\Lambda) \le \nu\right\};$$

$$\forall \Lambda \in \mathcal{M}_{O,L},\ \widehat{F}\left(\Lambda\right) \triangleq \max_{o\in[O],k\in[K]} \sqrt{Q'_{ok}\left(\Lambda\right)},\ \text{ and } Q'_{ok}\left(\Lambda\right) \triangleq \left(\widehat{\mathbf{D}}_{\mathbf{X}}\right)^2_{kk} \mathbb{E}_n\left[\left(\Omega_{ok} - \Lambda_{o\cdot} Z X_k\right)^2\right];$$

and r_0' is chosen like r_0 from Scenario 4 in Section 3.5, taking for α a sequence converging to zero, replacing L by OK and $\mathcal{G}_0$ by

$$\mathcal{G}_0' \triangleq \left\{ \max_{o\in[O],k\in[K]} \frac{\left|\mathbb{E}_n\left[\Omega_{ok} - \Lambda_{o\cdot} Z X_k\right]\right|}{\sqrt{\mathbb{E}_n\left[\left(\Omega_{ok} - \Lambda_{o\cdot} Z X_k\right)^2\right]}} \le r_0' \right\}.$$

This is the *C-STIV* estimator for a system of O equations under the premises of Remark 7.2. We do not allow for additional moments (for faster rates or to allow for larger values of λ and hence less sparsity enhancing penalities like in (7.5)) because they do not make sense here. We simply handle the system as in Remark 7.1, replace $c\sigma$ by $\lambda\nu/\widehat{\rho}$, $\mathbf{X}$ by $\mathbf{Z}$, $\mathbf{U}\left(\beta\right)$ by $\mathbf{U}\left(\Lambda_{o\cdot}\right) \triangleq \mathbf{Z}\Lambda_{o\cdot}^\top$, and note that

$$\left|\left(\Omega - \frac{1}{n}\Lambda\mathbf{Z}^\top\mathbf{X}\right)\widehat{\mathbf{D}}_{\mathbf{X}}\right|_\infty = \max_{o\in[O]}\left|\widehat{\mathbf{D}}_{\mathbf{X}}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{U}\left(\Lambda_{o\cdot}\right) - \Omega_{o\cdot}^\top\right)\right|_\infty.$$

When $\Omega = I_K$ and $\mathbf{Z} = \mathbf{X}$, $\widehat{\Lambda}$ is an approximate inverse of the matrix $\mathbf{X}^\top\mathbf{X}/n$ and $\widehat{\Lambda}$ is a “self-tuned” version of the CLIME estimator of Cai, Liu, and Luo (2011).

For all $g \in [G]$, the confidence bands, that we denote by $\widehat{\mathcal{C}}_g$, are defined as: for all $o \in [O]$,

$$\left[\left(\widehat{\Omega\beta_g}\right)_o - \frac{q_{W_\Lambda}(1-\alpha)}{\sqrt{n}}\sqrt{\mathbb{E}_n\left[\left(\widehat{\Lambda}_{o\cdot}Z\right)^2\right]\widehat{Q}\left(\widehat{\beta}_g\right)}, \left(\widehat{\Omega\beta}\right)_o + \frac{q_{W_\Lambda}(1-\alpha)}{\sqrt{n}}\sqrt{\mathbb{E}_n\left[\left(\widehat{\Lambda}_{o\cdot}Z\right)^2\right]\widehat{Q}\left(\widehat{\beta}_g\right)}\right],$$

where $q_{W_\Lambda}(1-\alpha)$ is the $1-\alpha$ quantile of $W_\Lambda \triangleq \left|\widehat{\mathbf{D}}_{\mathbf{Z}\widehat{\Lambda}^\top}\widehat{\Lambda}\mathbf{Z}^\top\mathbf{E}\right|_\infty/\sqrt{n}$ acting as if $\mathbf{Z}$ and $\widehat{\Lambda}$ were not random and $\mathbf{E}$ is a standard Gaussian random vector in $\mathbb{R}^n$. This can be obtained by a Monte-Carlo method. The bands $\widehat{\mathcal{C}}_g$ need not be unique because $\widehat{\beta}_g$ and $\widehat{\Lambda}$ are not necessarily unique.

The next assumption introduces a preliminary version of the class $\mathcal{P}$ that we consider.

Assumption 8.1. *For $K, L, O \ge 3$, $s \in [p]$, $s' \in [Os_r]$, $s_r \in [L]$, parameters of the class $\mathcal{P}_5$ and $\mathcal{P}_4$, positive b, $M_{\Lambda Z}(O)$, $M_{\Lambda,1}(O,K)$, $M_{\Lambda,2}(O)$, positive sequences $\left(\alpha_\beta(n)\right)_{n\in\mathbb{N}}$, $\left(v_\beta(n)\right)_{n\in\mathbb{N}}$, and $\left(v_\sigma(n)\right)_{n\in\mathbb{N}}$ decaying to zero, we have, for all $\mathbb{P}, (\beta,\Lambda) \in \mathcal{B}_{p,s',s_r}$ such that $\mathbb{P}(\beta,\Lambda) \in \mathcal{P}$, for all $n \in \mathbb{N}$,*

(DGP.1) $\mathbb{P}(\beta) \in \widetilde{\mathcal{P}}_5$;

(DGP.2) For all $(o,k) \in [O]\times[K]$, the distribution of $\left(\left(t_i\left(\Lambda\right)\right)_{ok}\right)_{i=1}^n$ belongs to $\mathcal{P}_4$;

(DGP.3) $\mathbb{E}\left[\left|\left(\left(T(\Lambda)\right)^2_{ok}/\sigma^2_{\left(T(\Lambda)\right)_{ok}} - 1\right)_{o\in[O],k\in[K]}\right|_\infty^2\right] \le M_{\Lambda,1}(O,K)$ *and*

$$\mathbb{E}\left[\left|\left(\left(\Lambda_{o\cdot}Z\right)^2/\mathbb{E}\left[\left(\Lambda_{o\cdot}Z\right)^2\right] - 1\right)_{o\in[O],k\in[K]}\right|_\infty^2\right] \le M_{\Lambda,2}(O);$$

(DGP.4) $\mathbb{E}\left[\left|\mathbf{D}_{\Lambda Z}\Lambda\left(ZX^\top-\mathbb{E}\left[ZX^\top\right]\right)\Lambda^\top\mathbf{D}_{\Lambda Z}\right|_\infty^2\right]\le M_{\Lambda Z}(O)$, $\mathbb{E}\left[(v'\mathbf{D}_Z Z)^2\right]\ge b$ *for all* $v\in\mathbb{R}^L$ *such that* $|v|_2=1$ *and* $|J(v)|\le s_r$*;*

(DGP.5) The preliminary estimator is such that, on an event of probability $1-\alpha_\beta(n)$*,*

$$\max_{g\in[G]}\left|\mathbf{D}_X^{-1}\left(\widehat{\beta}_g-\beta_g\right)\right|_1\le v_\beta(n),\ \max_{g\in[G]}\left|\sqrt{\widehat{Q}\left(\widehat{\beta}_g\right)}-\sigma_{W_g(\beta_g)}\right|\le v_\sigma(n).$$

We use the class $\widetilde{\mathcal{P}}_5$ in the assumption above but an analogue holds for the other scenarii and enables to use a preliminary *C-STIV* estimator. Like τ_Z, τ_X, and r_Ψ, the constants $c,r,\overline{r}$ and α in the definition of the preliminary estimators can also vary with n, the last three converging to zero. The restricted sets for the definition of the sensitivities for the rates of convergence of $\left(\widehat{\Lambda},\widehat{\nu}\right)$ are denoted

$$C'_J\triangleq\left\{\Delta'\in\mathcal{M}_{O,L}:\left|\Delta'_{J^c}\right|_1\le\frac{1+\lambda}{1-\lambda}\sqrt{\frac{1+\tau_Z}{1-\tau_Z}}\left|\Delta'_J\right|_1\right\},\ C'_{\gamma,J}\triangleq\left\{\Delta'\in\mathcal{M}_{O,L}:\left|\Delta'_{J^c}\right|_1\le\frac{2+\lambda}{1-\lambda}\sqrt{\frac{1+\tau_Z}{1-\tau_Z}}\left|\Delta'_J\right|_1\right\}.$$

The population sensitivities are denoted with the letters κ' and γ'. They are defined in the same way replacing $|\Psi\Delta|_\infty$ by $|\Delta'\Psi^\top|_\infty$. We use the notation $\kappa'_{(\infty,\infty),J}$ and $\gamma'_{(\infty,\infty),J}$ for the sensitivities for $l(\Delta')=|\Delta'|_{\infty,\infty}$, $\kappa'_{l_{F,\infty},J}$ and $\gamma'_{l_{F,\infty},J}$ for the sensitivities for $l_{F,\infty}(\Delta')\triangleq\max_{o\in[O]}\left(\sum_{i=1}^n\left(\Delta'_{o\cdot}\mathbf{D}_Z z_i\right)^2/n\right)^{1/2}$,

$$\Theta'_\kappa(J)\triangleq\sqrt{(1+\tau_Z)(1+\tau_X)}\left(1-\frac{r_\Psi}{\kappa'_{(\infty,\infty),J}}-\frac{r'_0\rho\sqrt{(1+\tau_Z)(1+\tau_X)}}{\lambda\kappa'_{1,J,J}}\right)_+^{-1};$$

$$\Theta'_\gamma(J)\triangleq\sqrt{(1+\tau_Z)(1+\tau_X)}\left(1-\frac{r_\Psi}{\gamma'_{(\infty,\infty),J}}-\frac{r'_0\rho\sqrt{(1+\tau_Z)(1+\tau_X)}}{\lambda\gamma'_{Q,J}}\right)_+^{-1}.$$

The event $\mathcal{E}'_Z$ is defined in the appendix. The event $\mathcal{E}_T$ is defined like $\mathcal{E}_X$ in the appendix for the random matrix $T(\Lambda)$ using τ_T instead of τ_X and we define $F(\Lambda)\triangleq\left|\left((\mathbf{D}_X)_{kk}\,\sigma_{(T(\Lambda))_{ok}}\right)_{o\in[O],k\in[K]}\right|_\infty$.

Theorem 8.1. *Under Assumption 8.1, for all* $(\beta,\Lambda),\mathbb{P}$ *such that* $(\beta,\Lambda)\in\mathcal{B}_{s,OL,L}$*, on* $\mathcal{G}'_0\cap\mathcal{G}_\Psi\cap\mathcal{E}^c_{Z'}\cap\mathcal{E}^c_T$*, for all solution* $\left(\widehat{\Lambda},\widehat{\nu}\right)$ *of* (8.2) *with* $\lambda\in(0,1)$*, we have*

(i) *If* Λ *is sparse,*

$$\left|\left(\widehat{\Lambda}-\Lambda\right)\mathbf{D}_Z^{-1}\right|_{\infty,\infty}\le\frac{2r'_0F(\Lambda)}{\kappa'_{(\infty,\infty),J(\Lambda)}}\sqrt{\frac{1+\tau_T}{(1-\tau_{Z'})(1-\tau_X)}}\Theta'_\kappa(J(\Lambda));$$

$$\widehat{\nu}\le F(\Lambda)\sqrt{\frac{1+\tau_T}{1-\tau_X}}\left(1+\frac{2r'_0\rho\Theta'_\kappa\left(J(\Lambda)\right)}{\lambda\kappa'_{1,J(\Lambda),J(\Lambda)}}\right);$$

$$\max_{o\in[O]}\mathbb{E}_n\left[\left(\left(\widehat{\Lambda}_{o\cdot}-\Lambda_{o\cdot}\right)Z\right)^2\right]^{1/2}\le\frac{2r'_0F(\Lambda)}{\kappa'_{l_{F,\infty},J(\Lambda)}}\sqrt{\frac{1+\tau_T}{(1-\tau_{Z'})(1-\tau_X)}}\Theta'_\kappa(J(\Lambda));$$

(ii) *Else,*

$$\left|\left(\widehat{\Lambda}-\Lambda\right)\mathbf{D}_Z^{-1}\right|_{\infty,\infty}\le 2\min_{J\subseteq[O]\times[L]}\max\left(\frac{r_0'F\left(\Lambda\right)}{\min_{o\in[O]}\gamma'_{1,\{o\}\times[L],J}}\sqrt{\frac{1+\tau_T}{(1-\tau_{Z'})(1-\tau_X)}}\Theta'_\gamma(J),\frac{3+\lambda}{1-\lambda}\sqrt{\frac{1+\tau_{Z'}}{1-\tau_{Z'}}}\left|\Lambda_{J^c}\mathbf{D}_Z^{-1}\right|_1\right)$$

$$\widehat{\nu}\le\min_{J\subseteq[O]\times[L]}\max\left(F\left(\Lambda\right)\sqrt{\frac{1+\tau_T}{1-\tau_X}}\left(1+\frac{2r_0'\rho\Theta'_\gamma\left(J\right)}{\lambda\gamma'_{Q,J}}\right),\frac{3\rho\sqrt{1+\tau_{Z'}}}{\lambda}\left|\Lambda_{J^c}\mathbf{D}_Z^{-1}\right|_1+F\left(\Lambda\right)\sqrt{\frac{1+\tau_T}{1-\tau_X}}\right).$$

We denote by $\alpha_\Lambda(n)$ the probability of the complement of the event in Theorem 8.1, by $v_\Lambda(n)$ and $v_\nu(n)$ the two first upper bounds on the right of (i) and (ii), and by $v_{\Lambda,2}(n)$ the last upper bound of (i). This last bound is used to handle the scaling by $\widehat{\mathbf{D}}_{\mathbf{Z}\widehat{\Lambda}^\top}$ of the leading stochastic term of the expansion of the bias corrected estimator. Without this scaling, we can handle both sparse and approximately sparse matrices with minor modifications. We now only consider the case of sparse matrices where we can make use of all three inequalities of (i). For further use of this result for inference after bias correction, the sensitivities appearing on the right-hand side of the inequalities need to be well behaved so that, on an event of probability converging to 1, the first right-hand side goes to zero and the second remains bounded. By working under enough sparsity, $\mathcal{B}_{s,s',s_r}$ could be a singleton. This is important when $L>K$. The sensitivities depend again on Ψ through $\Psi^\top$ and the fact that $\Theta'_\kappa(J(\Lambda))$ is close to one is a type of strong instruments assumption.

We obtain the following result where we denote by $\widehat{\mathcal{C}}_{g,v(n)}$ a $v(n)$-neighborhood of $\widehat{\mathcal{C}}_g$.

Theorem 8.2. *Let $(\alpha_{\mathcal{C}}(n))_{n\in\mathbb{N}}$, $\left(\alpha_E^B(n)\right)_{n\in\mathbb{N}}$, $(v(n))_{n\in\mathbb{N}}$, and $(v_v(n))_{n\in\mathbb{N}}$ be positive sequences converging to zero and further reduce the class $\mathcal{P}$ defined in Assumption 8.1 so that, for all $n\in\mathbb{N}$,*

$$\begin{aligned}&\alpha_\beta(n)+\alpha_\Lambda(n)+\alpha_E^B(n)\le\alpha_{\mathcal{C}}(n);\\ &\sqrt{n}r_0'v_\beta(n)v_\nu(n)\sqrt{1+\tau_X}\le v_v(n);\end{aligned}\tag{8.3}$$

where $\alpha_E^B(n)$ is an upper bound on the coverage error of the bootstrap for the bands. Then, for all $\mathbb{P},(\beta,\Lambda)$ such that $(\beta,\Lambda)\in\mathcal{B}_{s,s',s_r}$ and $n\in\mathbb{N}$, every collection $\left(\widehat{\mathcal{C}}_g\right)_{g=1}^G$ satisfies

$$\mathbb{P}\left(\Omega\beta_g\in\widehat{\mathcal{C}}_{g,v(n)}\right)\ge 1-\alpha-\alpha_{\mathcal{C}}(n)\quad\forall g\in[G].$$

For bands around functions evaluated on a grid, assume that $\mathcal{P}$ is such that all approximation errors are bounded in absolute value by $v_v(n)/\sqrt{n}$ and there exists $\tau_\Lambda,\tau_{\Lambda Z}>0$ such that, for all $n\in\mathbb{N}$,

$$\alpha_\beta(n)+\alpha_\Lambda(n)+\alpha_E^B(n)+C_{\mathrm{N}}(O)M_{\Lambda,2}(O)/(n\tau_\Lambda^2)+C_{\mathrm{N}}(OL)M_{\Lambda Z}(O)/(n\tau_{\Lambda Z}^2)\le\alpha_{\mathcal{C}}(n);$$

$$\sqrt{n}r_0'v_\beta(n)v_\nu(n)\sqrt{1+\tau_X}+v_v(n)\left(\left(\sqrt{1+\tau_{\Lambda Z}}-1\right)+\sqrt{1+\tau_\Lambda}\left|(\sigma_{\Lambda_o\cdot Z})_{o\in[O]}\right|_\infty+v_{\Lambda,2}(n)\right)+v_v(n)\le v(n);$$

then, for all $\mathbb{P}, (\beta, \Lambda)$ *such that* $(\beta, \Lambda) \in \mathcal{B}_{s,s',s_r}$ *and* $n \in \mathbb{N}$, *every collection* $\left(\widehat{\mathcal{C}}_g\right)_{g=1}^{G}$ *satisfies*

$$\mathbb{P}\left(\Omega\beta_g + V_g(\beta_g) \in \widehat{\mathcal{C}}_{g,v(n)}\right) \geq 1 - \alpha - \alpha_{\mathcal{C}}(n) \quad \forall g \in [G].$$

The bound $\alpha_E^B(n)$ can be computed like for Scenario 5 and depends as well on s_r, $M_{\Lambda Z}(O)$, $M_{\Lambda,2}(O)$, and $v_\sigma(n)$. Importantly, based on condition 8.3, we need to have $\sqrt{n} r_0' v_\beta(n) \to 0$. This is flexible enough so that one rate can be fast only. It is not needed that both are faster than $n^{1/4}$.

8.2. **Confidence Bands Using Sample Splitting.** We now allow for non identification of Λ, propose a more numerically efficient method to obtain $\widehat{\Lambda}$, and do not require that $\widehat{\Lambda}$ converges. To simplify the arguments, we rely on a sample splitting argument. The main reason for splitting the sample is that the leading stochastic term in the expansion of the debiased estimator is $\widehat{\Lambda}\mathbf{Z}^\top\mathbf{W}(\beta)/\sqrt{n}$ and using all the sample implies that $\widehat{\Lambda}$ depends on $\mathbf{X}$ thus on $\mathbf{W}(\beta)$ because of endogeneity. This is problematic if $\widehat{\Lambda}$ does not converge. Constructing $\widehat{\Lambda}$ using a different sample avoids this concern.

We split the sample in two and use the index - for the first sample of size n_- and the index + for the second sample of size n_+. The subsample sizes n_+ and n_- increase to infinity with n and we assume that we use a deterministic rule to obtain (n_-, n_+) from n. The sample size which matters for the width of the bands is n_+ so, provided the bias remains small, we can avoid paying a price for splitting the sample asymptotically if we take $n_+/n \to 1$. The observations in the first sample have indices in $-[n_-]$ and those in the second sample in $[n_+]$. The sigma-field generated by $(\mathbf{Z}, \mathbf{X}_-, \mathbf{W}_{1-}(\beta_1), \ldots, \mathbf{W}_{G-}(\beta_G))$ is denoted by $\mathcal{F}_n$ and $\mathcal{F}_\infty$ is the smallest sigma-field containing $\bigcup_{n\in\mathbb{N}} \mathcal{F}_n$. The first sample is used to obtain $\widehat{\Lambda}$ as

$$\widehat{\Lambda} \in \operatorname{argmin}_{\Lambda \in \mathcal{M}_{O,L}} \left(\left| \left(\Omega - \frac{1}{n_-}\Lambda \mathbf{Z}_-^\top \mathbf{X}_- \right) \widehat{\mathbf{D}}_{\mathbf{X}_-} \right|_\infty + \lambda_1 \left| \Lambda \widehat{\mathbf{D}}_{\mathbf{Z}_-}^{-1} \right|_{\infty,\infty} + \frac{\lambda_2}{\sqrt{n_+}} \left| \Lambda \mathbf{Z}_+^\top \right|_{2,\infty} \right), \tag{8.4}$$

where λ_1 and λ_2 are nonnegative parameters. The role of the first term in the objective function is to minimize the bias. The second term is introduced to minimize an additional term induced by sample splitting and enhances row sparsity of $\widehat{\Lambda}$. It is used to obtain conditions under which the bias could be neglected asymptotically. With the approach of Section 8.3, we can take $\lambda_1 = 0$. The last term is the maximum of the inverse of the diagonal elements of the scaling matrix $\widehat{\mathbf{D}}_{\mathbf{Z}_+\widehat{\Lambda}^\top}$. It controls their dispersion and prevents the confidence bands from being too U-shaped. This is useful for bands around functions for which there is little data close to the end points of the support. It plays a role in reducing the bias when there are approximation errors. When $\lambda_2 = 0$, (8.4) is a linear program. When $\lambda_2 > 0$, (8.4) is a conic program with O cones only. In contrast, the method of the previous section involves OK conic constraints which quickly becomes prohibitive. The second sample is used to obtain preliminary estimators $\widehat{\beta}_{g+}$ and apply the bias corrections.

The debiased estimators are defined as

$$\widehat{\Omega\beta_g} \triangleq \Omega\widehat{\beta}_{g+} + \frac{1}{n_+}\widehat{\Lambda}\mathbf{Z}_+^\top\left(\mathbf{Y}_+ - \mathbf{X}_+\widehat{\beta}_{g+}\right).$$

For all $g \in [G]$, the confidence bands, that we denote by $\widehat{\mathcal{C}}_g$, are defined as: for all $o \in [O]$,

$$\left[\left(\widehat{\Omega\beta_g}\right)_o - r_+^\Lambda\sqrt{\mathbb{E}_{n_+}\left[\left(\widehat{\Lambda}_{o\cdot}Z\right)^2\right]\widehat{Q}\left(\widehat{\beta}_{g+}\right)}, \left(\widehat{\Omega\beta_g}\right)_o + r_+^\Lambda\sqrt{\mathbb{E}_{n_+}\left[\left(\widehat{\Lambda}_{o\cdot}Z\right)^2\right]\widehat{Q}\left(\widehat{\beta}_{g+}\right)}\right],$$

where $\mathbb{E}_{n_+}$ is the mean over the second sample. We adjust r_+^Λ using one of scenarii 1-4 replacing $\mathbf{Z}$ with $\mathbf{Z}_+\widehat{\mathbf{\Lambda}}$, L with O and n with n_+. The scenario has to hold conditional on $\mathcal{F}_\infty$. This is possible, for example, when $\mathbf{U}_+(\beta)$ is independent of $\mathcal{F}_\infty$ and symmetric. An alternative approach is to assume that $\mathbf{Z}$ and $\mathbf{U}(\beta)$ are independent and that $u_i(\beta)$ for $i \in [n]$ are i.i.d. normally distributed. In this case, r_+^Λ can be replaced by the quantile $q_{W_+}(1-\alpha)/\sqrt{n_+}$ where $W_{\Lambda+} \triangleq \left|\widehat{\mathbf{D}}_{\mathbf{Z}_+\widehat{\Lambda}^\top}\widehat{\Lambda}\mathbf{Z}_+^\top\mathbf{E}/(\sqrt{n_+}(1-\epsilon))\right|_\infty$ computed holding fixed $\mathbf{Z}_+$ and $\widehat{\Lambda}$, in which ϵ is an arbitrarily small positive constant, and $\mathbf{E}$ is a standard Gaussian random vector in $\mathbb{R}^{n_+}$ independent of the observed data. This conditional quantile can be obtained by a Monte-Carlo method. The constant r when the preliminary estimator is a *STIV* or *E-STIV* estimator can be taken simply as a quantile of $\sqrt{1-\tau(n_+)}\left|\widehat{\mathbf{D}}_{\mathbf{Z}_+}\mathbf{Z}_+^\top\mathbf{E}/\sqrt{n_+}\right|_\infty$ conditional on $\mathbf{Z}$, where $\tau(n_+) \to 0$. In which case, the coverage error is less than $2/(n_+\tau(n_+)^2)$ for a *STIV* or *E-STIV* estimator. The coverage errors are the same because, like in the previous section, we maintain the assumption that all approximation errors are bounded in absolute value by $v_v(n)/\sqrt{n}$ for some sequence $(v_v(n))_{n\in\mathbb{N}}$ converging to zero. For a *SE-STIV* estimator, when we allow for arbitrary dependence between the errors in different equations, one has to replace $1-\alpha$ quantiles by $1-\alpha/G$ quantiles and the coverage error is less than $2G/(n_+\tau(n_+)^2)$.

Let $r_{\Psi+}$, $r_{\Psi-}$, τ_{X-}, and τ'_{Z-}, play the role of r_Ψ, τ_X, and τ'_Z before and taken as positive sequences indexed by either n_+ or n_- depending on the sample and which converge to 0. We define r'_{0-} is defined on the first sample like r'_0 for the full sample taking for α a sequence $\alpha_{\mathcal{G}'_0-}$ which converges to zero. Define, for additional positive sequences τ_{T-} and $\tau_{\Lambda+}$ converging to zero in the same way as the other slackness parameters,

$$\mathcal{S}_{\lambda_1,\lambda_2} \triangleq \operatorname{argmin}_{\Lambda:\Lambda\mathbb{E}[ZX^\top]=\Omega}\left(r'_{0-}\sqrt{\frac{1+\tau_{T-}}{1-\tau_{X-}}}F(\Lambda) + \lambda_1\sqrt{1+\tau'_{Z-}}\left|\Lambda\mathbf{D}_Z^{-1}\right|_{\infty,\infty} + \lambda_2\sqrt{1+\tau_{\Lambda+}}\left|(\sigma_{\Lambda_o\cdot Z})_{o\in[O]}\right|_\infty\right)$$

and denote the minimum by m_{λ_1,λ_2}.

Theorem 8.3. *Let $\lambda_1, \lambda_2 \geq 0$ and $\mathcal{P}$ be defined via Assumption 8.1 maintaining only (DGP.1), (DGP.2), (DGP.3), and replacing (DGP.5) by*

$$\max_{g\in[G]} \left|\mathbf{D}_X^{-1}\left(\widehat{\beta}_g - \beta_g\right)\right|_1 \leq v_\beta(n), \quad \text{and, } \forall g \in [G],\ \sqrt{\widehat{Q}\left(\widehat{\beta}_g\right)} \geq \sigma_{W_g(\beta_g)}\left(1 - v_\sigma(n)\right),$$

for all $\Lambda \in \mathcal{S}_{\lambda_1,\lambda_2}$, either of Scenario 1-4 holds for the distribution of $\left(\widehat{\Lambda} z_i u_i(\beta)_+\right)_{i=1}^{n_+}$ given $\mathcal{F}_\infty$ or z_i and $u_i(\beta)$ are independent for all $i \in [n]$ and $u_i(\beta)$ for $i \in [n]$ are i.i.d. normally distributed. Let $\left(\alpha_{\mathcal{C}}(n)\right)_{n\in\mathbb{N}}$, $\left(v_1(n)\right)_{n\in\mathbb{N}}$ and $\left(v_2(n)\right)_{n\in\mathbb{N}}$ be positive sequences converging to zero, and further reduce the class $\mathcal{P}$ so that, for all $\mathbb{P}, (\beta, \Lambda)$ such that $(\beta,\Lambda) \in \mathcal{B}_{s,s',s_r}$ and $n \in \mathbb{N}$,

$$\begin{aligned}&\alpha_\beta(n_+) + \alpha_B(n_+) + \frac{C_{\mathrm{N}}(KL)M(L,K)}{n_- r_{\Psi-}^2} + \frac{C_{\mathrm{N}}(K)M_X(K)}{n_-\tau_{X-}^2} + \frac{C_{\mathrm{N}}(L)M_Z'(L)}{n_-(\tau_{Z-}')^2}\\ &+ C_{\mathrm{N}}(O)\left(\frac{M_{\Lambda,1}(O)}{n_-\tau_{T-}^2} + \frac{M_{\Lambda,2}(O)}{n_+\tau_{\Lambda+}^2}\right) + \alpha_{\mathcal{G}_0'-} + \alpha_B(n_-) \leq \alpha_{\mathcal{C}}(n);\end{aligned}$$

$$\sqrt{n_+}v_\beta(n_+)M(\lambda_1)m_{\lambda_1,\lambda_2} \leq v_1(n); \tag{8.5}$$

$$v_\sigma(n_+) \leq v_2(n); \tag{8.6}$$

where $\alpha_B(n_+)$ is nonzero only if we maintain Scenario 4 for the distribution of $\left(\widehat{\Lambda} z_i u_i(\beta)_+\right)_{i=1}^{n_+}$ and

$$M(\lambda_1) \triangleq \sqrt{1+\tau_{X-}}\max\left(1, \frac{r_{\Psi-} + r_{\Psi+}}{\lambda_1\sqrt{(1+\tau_{X-})(1-\tau_{Z-}')}}\right).$$

Then[3]*, for all $\mathbb{P}, (\beta,\Lambda)$ such that $(\beta,\Lambda) \in \mathcal{B}_{s,s',s_r}$ and $n \in \mathbb{N}$*[4]*, every collection $\left(\widehat{\mathcal{C}}_g\right)_{g=1}^G$ satisfies*

$$\mathbb{P}\left(\Omega\beta_g \in \widehat{\mathcal{C}}_{g,v_1(n)}\right) \geq 1 - \alpha - \alpha_\beta(n_+) - \alpha_B(n_+) \quad \forall g \in [G].$$

For bands around functions evaluated on a grid, assume that $\mathcal{P}$ is such that all approximation errors are bounded in absolute value by $v_v(n)/\sqrt{n}$ and we replace (8.5) by

$$\sqrt{n_+}\left(v_\beta(n_+)M(\lambda_1) + \frac{v_v(n)}{\lambda_2}\sqrt{\frac{n_+}{n}}\right)m_{\lambda_1,\lambda_2} \leq v_1(n), \tag{8.7}$$

then[5]*, for all $\mathbb{P}, (\beta,\Lambda)$ such that $(\beta,\Lambda) \in \mathcal{B}_{s,s',s_r}$ and $n \in \mathbb{N}$*[6]*, every collection $\left(\widehat{\mathcal{C}}_g\right)_{g=1}^G$ satisfies*

$$\mathbb{P}\left(\Omega\beta_g + V_g(\beta_g) \in \widehat{\mathcal{C}}_{v_1(n),g}\right) \geq 1 - \alpha - \alpha_\beta(n_+) - \alpha_B(n_+) \quad \forall g \in [G].$$

[3] Add here for normal errors: for all $\epsilon > 0$, there exists n_0 such that.

[4] Replace for normal errors by: $n_+ \geq n_0$.

[5] Add here for normal errors: for all $\epsilon > 0$, there exists n_0 such that.

[6] Replace for normal errors by: $n_+ \geq n_0$.

A natural choice for λ_1 to ensure a small bias is to take $M(\lambda_1) = 1$. In practice, this often puts too much emphasis on sparsity. We describe the empirical rule in the next Section 9.1.3. In the absence of approximation error, the bias is smallest for $\lambda_2 = 0$. When there are approximation errors, the penalty involving λ_2 plays a role on the bias and appears in the denominator of (8.7) so it should not be too small. The penalty involving λ_2 can be viewed as providing robustness to approximation errors. In the definition of $\mathcal{S}_{\lambda_1,\lambda_2}$, the first and last terms could be comparable in terms of magnitude so that λ_2 should not exceed r'_{0-}.

8.3. **Confidence Bands with a Data-Driven Upper Bound on the Bias.** In the previous sections, we have made assumptions to guarantee that the bias of the bias-corrected estimator is negligible. This might not hold in practice, especially for small sample sizes or in the presence of weak identification or non identification without sparsity. One way to restore coverage is to combine the approach with the one-stage confidence sets. It also allows to remove entirely (8.3).

For the confidence bands of Section 8.2, we use that the ℓ_∞-norm of the bias term is less than

$$\sqrt{n_+}\left|\left(\Omega - \frac{1}{n_+}\widehat{\Lambda}\mathbf{Z}_+^\top\mathbf{X}_+\right)\widehat{\mathbf{D}}_{\mathbf{X}_+}\right|_\infty \left|\widehat{\mathbf{D}}_{\mathbf{X}_+}^{-1}\left(\widehat{\beta}_{g+} - \beta_g\right)\right|_1 + v_v(n)\frac{1}{\sqrt{n}}\left|\widehat{\Lambda}\mathbf{Z}_+^\top\right|_{2,\infty} + v_v(n), \tag{8.8}$$

where $v_v(n) = 0$ when there are no approximation errors. For the confidence bands of Section 8.1, we drop the index + above. We can now make use of sections (5.2), (6.3), (6.5) to have a data-driven bound on $\left|\widehat{\mathbf{D}}_{\mathbf{X}_+}^{-1}\left(\widehat{\beta}_{g+} - \beta_g\right)\right|_1$ holding on the same event as the one used to obtain the bands which ignore the bias. In sum, we add to the upper bound and substract to the lower bound of the confidence bands of the previous sections this upper bound on the bias.

The additional advantage of enlarging the confidence bands to account for the bias is that all confidence bands are valid, whatever the value of the parameters λ, λ_1, and λ_2. These only play a role in the estimation of $\widehat{\Lambda}$. We see now clearly that the penalization involving λ_2 allows to have a small second term in (8.8), hence that it is particularly important in the presence of approximation errors. Moreover, we see that λ_1 plays no role in the data-driven upper bound on the bias. This is important because we can take $\lambda_1 = 0$ and $\widehat{\Lambda}$ can be nonsparse and not even approximately sparse. A similar feature occurs in Javanmard and Montanari (2014) with the important difference that they ignore the bias while we account for it. The penalty involving λ_2 is similar in spirit to the objective function in Javanmard and Montanari (2014). Ours extends theirs and allows for confidence bands.

9. Inference Put Into Practice

9.1. **Simulation Study.** We consider the model

$$y_i = \sum_{k=1}^{K} x_{ki}\beta_k + u_i,$$

$$x_{ki} = \sum_{l=1}^{L} z_{li}\zeta_{kl} + v_{ki} \quad \text{for } k \in I^c, \quad x_{ki} = e_{ki} \quad \text{otherwise}$$

For v_i the $|I^c|$ dimensional vector stacking v_{ki} for $k \in I^c$, the data $(y_i, x_i^\top, z_i^\top, u_i, v_i^\top)$ are i.i.d. and the random vector $(u_i, v_i^\top)$ follows a mean zero normal distribution, $\mathbb{E}[u_i^2] = \sigma_{\text{str}}^2$, $\mathbb{E}[v_{ki}^2] = \sigma_{\text{end}}^2$, $\mathbb{E}[v_{ki}u_i] = \rho|I^c|^{-1/2}$ for $k \in I^c$, and e_{ki} are i.i.d. draws from the standard normal distribution truncated to the interval $[-5, 5]$. We take $\sigma_{\text{str}} = \sigma_{\text{end}} = 1$ and $\rho = 0.3$. All of the remaining covariance terms are equal to zero. We take $J(\beta) = \{1, 2, 3, 4, 5\}$ and $\beta_{J(\beta)} = (1, -2, -0.5, 0.25, -1)^\top$.

All of the inference is at the $\alpha = 0.05$ level. We use the MATLAB freeware GloptiPoly 3[7] (see Henrion, Lasserre, and Lofberg (2009)) in Section 9.1.1 and the CVX package (see Grant and Boyd (2013)) with the solver Mosek for the subsequent optimization routines. We report a coefficient of zero whenever the estimated value is smaller than the tolerance of the solver, which is 10^{-8}. We simulate the estimators both for $K > n$ and $K < n$. For the confidence sets, we deal only with $K < n$ but take K so large that BIC is not feasible. We use the results of Section A.6 to calculate the lower bounds on the sensitivities when they are based on an estimated support $\widehat{J}$. We use the notation p 2.5 for the 2.5 percentile and similarly for p 50 and p 97.5.

9.1.1. *The* SNIV *Confidence Set.* We illustrate the confidence sets from Sections 3.3 with $n = 1000$, $K = 12$, $I^c = \{1\}$, $L = 11$, and $\zeta_{kl} = 0.3$. We suppose that it is known that x_{1i}, x_{2i}, and x_{3i} are included in the true model but there is uncertainty regarding $x_{4i}, ..., x_{12i}$. That is, we take $P = \{4, 5, ..., 12\}$ in the definition of the s-sparse identified set $\mathcal{B}_s$. We choose r based on Scenario 3, hence the confidence sets have coverage at least $1-\alpha$ in finite samples. We compute nested confidence sets, taking the sparsity certificate s from 2 to 8. The true sparsity is 2, since β_4 and β_5 are nonzero.

Table 3 displays these nested confidence sets. The number in each cell is obtained by solving Lasserre's relaxations of order two for one SCQP problem. Cells in gray correspond to cases in which it is known that the iteration has reached the global minimum. Stopping at order 2, whether or not the hierarchy has converged, could produce non-nested sets. This did not occur.

For comparison, Table 4 presents the infimum and supremum of each of the parameters in the sparse identified set $\mathcal{B}_s$ under the various sparsity certificates. This is obtained by solving a QCQP.

[7] www.laas.fr/∼henrion/software/gloptipoly3

TABLE 3. Confidence sets based on $\widehat{S}$ with fewer instruments than regressors

	$\beta_{l,8}$	$\beta_{l,7}$	$\beta_{l,6}$	$\beta_{l,5}$	$\beta_{l,4}$	$\beta_{l,3}$	$\beta_{l,2}$	β^*	$\beta_{u,2}$	$\beta_{u,3}$	$\beta_{u,4}$	$\beta_{u,5}$	$\beta_{u,6}$	$\beta_{u,7}$	$\beta_{u,8}$
β_1	-9.043	0.285	0.349	0.397	0.445	0.507	0.589	1	1.247	1.264	1.363	1.410	1.460	1.551	2.343
β_2	-2.558	-2.297	-2.264	-2.249	-2.223	-2.203	-2.175	-2	-1.807	-1.729	-1.693	-1.684	-1.655	-1.641	2.489
β_3	-1.009	-0.794	-0.7547	-0.732	-0.713	-0.699	-0.687	-0.5	-0.316	-0.296	-0.275	-0.258	-0.243	-0.206	2.815
β_4	-0.045	-0.001	0	0.028	0.049	0.077	0.101	0.25	0.483	0.533	0.580	0.590	0.602	0.640	4.806
β_5	-1.509	-1.236	-1.210	-1.192	-1.173	-1.156	-1.141	-1	-0.801	-0.778	-0.759	-0.729	-0.712	-0.671	0
β_6	-0.557	-0.276	-0.251	-0.213	-0.203	-0.186	0	0	0	0.207	0.240	0.262	0.294	0.330	3.366

Here: $r = 0.097$

TABLE 4. The sparse identified set with fewer instruments than regressors

	$\beta_{l,8}$	$\beta_{l,7}$	$\beta_{l,6}$	$\beta_{l,5}$	$\beta_{l,4}$	$\beta_{l,3}$	$\beta_{l,2}$	β^*	$\beta_{u,2}$	$\beta_{u,3}$	$\beta_{u,4}$	$\beta_{u,5}$	$\beta_{u,6}$	$\beta_{u,7}$	$\beta_{u,8}$
β_1	-2.333	1	...	...	...	...	1	1	1	...	...	...	...	1	1.8333
β_2	-2.25	-2	...	...	...	...	-2	-2	-2	...	...	...	...	-2	-1
β_3	-0.75	-0.5	...	...	...	...	-0.5	-0.5	-0.5	...	...	...	...	-0.5	0.5
β_4	0	0.25	...	...	...	...	0.25	0.25	0.25	...	...	...	...	0.25	1.25
β_5	-1.25	-1	...	...	...	...	-1	-1	-1	...	...	...	...	-1	0
β_6	-0.25	0	...	...	...	...	0	0	0	...	...	...	...	0	1

This is a rare instance in which we can compute the sparse identified sets and is due to the fact that K is not large. Point identification is obtained under sparsity certificates ranging from 2 to 7, whilst partial identification is obtained for a sparsity certificate of 8 or larger. When the sparsity certificate is 8, β is the solution of either of 9 linear systems with 11 equations and 11 parameters. In contrast, if it is 7, β is the solution of either of 36 linear systems with 11 equations and 10 parameters.

9.1.2. *Estimation when $K > n$, $K \gg L$.* We take $n = 500$, $L = 30$, $K = 600$, and $I^c = \{1, 552, ..., 600\}$. Thus, we are under the high-dimensional regime. There are many more regressors than variables known to be exogenous and used as instruments. There are 50 endogenous regressors, one of the indices of which is in $J(\beta)$. We set $z_{li} = x_{l'i}$ for $l' = l + 1$ and $l \in [L]$. We take $\zeta_{kl} = 0.3$ for $k \in I^c$ and $l \in [L]$ and vary c such that $0 < cr < 1$. We adjust r using Scenario 5.

The results are summarized in tables 5 and 6. Table 5 studies the performance of the estimator at $cr = 0.95$, which corresponds to the least shrinkage to zero. The parameter vector β is sparse since $|J(\beta)| = 5$ and $K = 600$. Sparsity provides exclusion restrictions and possible overidentification for each submodel. The coefficients β_3 and β_4 are smaller than the detection level (see Theorem 6.3). Indeed, not even accounting for the sensitivities, we have $\frac{4\sigma_{\text{end}} r}{\mathbb{E}_n[X_k^2]^{1/2}} > 0.5$ for all k. We expect from

Table 5. Monte-Carlo study (1000 replications, $n = 600$)

	p 2.5	p 50	p 97.5		p 2.5	p 50	p 97.5
$\widehat{\beta}_1$	0.818	0.883	0.940	$\widehat{\beta}_6$	0	0	0
$\widehat{\beta}_2$	-1.906	-1.814	-1.713	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\widehat{\beta}_3$	-0.407	-0.309	-0.212	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\widehat{\beta}_4$	0.022	0.121	0.212	$\widehat{\beta}_{600}$	0	0	0
$\widehat{\beta}_5$	-0.904	-0.809	-0.709	$\widehat{\sigma}$	0.997	1.063	1.132

Here: $cr = 0.95$.

Theorem 6.1 that β_3 and β_4 are impossible to distinguish from 0. The results highlighted in the grey box correspond to the nonzero coefficients of the estimator. The estimator performs well in terms of selecting the nonzero components of β, though the point estimates are biased towards zero due to shrinkage. The shrinkage also results in a slight upwards bias of the estimator of the variance. The estimator also performs well for the components with true parameter equal to zero.

Table 6 summarizes the performance of the estimator over various choices of c such that cr ranges from 0.05 to 0.95. For $c = 0.05/r$ we have $c \approx 1/3$. This is because $r \approx 0.15$. For larger values of c, the estimation error is smaller for the nonzero elements of β. However, it is also the case that some of the zero coefficients are estimated to be nonzero. The converse is true for smaller values of c: the estimation error is larger for the nonzero coefficients but smaller for the zero coefficients. This is because there is more shrinkage as c becomes smaller.

9.1.3. *Choice of c for Confidence Sets and λ, λ_1, λ_2 for Confidence bands.* In this simulation study, we do not intersect the confidence sets based on a sparsity certificate for different values of c. This means that the confidence sets are more conservative than they could have been. We rather started by estimating each model with $c = r^{-1}$, which corresponds to the least shrinkage of $\widehat{\beta}$, and compared the estimates obtained for decreasing c. For sufficiently large sample size, the estimators remain almost unchanged when c decreases, until the point at which $\widehat{\sigma}$ starts to increase. We chose c at that value. We apply the same strategy to choose λ.

We follow an identical heuristic to choose λ_1. We start by solving (8.4) for λ_1 close to zero, which corresponds to the least shrinkage of $\widehat{\Lambda}$. As λ_1 increases, $\widehat{\Lambda}$ remains almost unchanged until a point at which the first term in the objective function in (8.4) starts to increase. We chose λ_1 at that value. Throughout this section, we do not include approximation errors and take $\lambda_2 = 0$.

TABLE 6. Monte-Carlo study at different values of c, (1000 replications, $n = 600$)

	p 2.5	p 50	p 97.5	p 2.5	p 50	p 97.5	p 2.5	p 50	p 97.5
$\lvert\Delta_{J(\beta)\cap I}\rvert_\infty$	0.163	0.230	0.313	0.206	0.274	0.359	0.296	0.385	0.491
$\lvert\Delta_{J(\beta)\cap I^c}\rvert_\infty$	0.122	0.223	0.345	0.159	0.264	0.386	0.278	0.393	0.521
$\lvert\Delta_{J^c(\beta)\cap I}\rvert_\infty$	0	0	0.045	0	0	0.012	0	0	0
$\lvert\Delta_{J^c(\beta)\cap I^c}\rvert_\infty$	0	0	0.093	0	0	0.066	0	0	0.015
$\widehat{\sigma}$	0.997	1.066	1.1294	1.026	1.097	1.174	1.122	1.209	1.289
	cr=0.95			**cr=0.80**			**cr=0.60**		
	p 2.5	p 50	p 97.5	p 2.5	p 50	p 97.5	p 2.5	p 50	p 97.5
$\lvert\Delta_{J(\beta)\cap I}\rvert_\infty$	0.573	0.703	0.875	1.870	2.000	2.121	1.865	2.001	2.120
$\lvert\Delta_{J(\beta)\cap I^c}\rvert_\infty$	0.611	0.779	0.965	1.792	1.923	2.046	1.805	1.918	2.048
$\lvert\Delta_{J^c(\beta)\cap I}\rvert_\infty$	0	0	0	0	0	0	0	0	0
$\lvert\Delta_{J^c(\beta)\cap I^c}\rvert_\infty$	0	0	0	0	0	0	0	0	0
$\widehat{\sigma}$	1.448	1.567	1.710	2.684	2.855	3.042	2.677	2.848	3.036
	cr=0.40			**cr=0.20**			**cr=0.05**		

Here: $\Delta = \mathbf{D_X}^{-1}(\widehat{\beta} - \beta)$.

9.1.4. *Confidence Sets: Fewer Instruments than Potential Regressors.* Under the simulation design of the previous section, the confidence sets are infinite. Consider now the following modification in which $K = 50$, $I^c = \{1, 27, 28, ..., 50\}$, and $\zeta_{kl} = 1$ if $(k, l) = (1, 6), (27, 7), (28, 8), ..., (49, 29), (50, 30)$ and 0.1 otherwise. This implies that there is one strong instrument for each endogenous regressor, and the remaining 24 instruments are "weak". We maintain $K \gg L$, $n = 500$ and set $z_{li} = x_{l'i}$ for $l' = l + 1$ and $l \in [25]$, and $z_{li} = e_{li}$ for $l \in [26, ..., L]$, where e_{li} is drawn from the standard normal truncated to the interval $[-5, 5]$. We adjust r using Scenario 5. We illustrate here the confidence sets for one dataset. The 5 true nonzero coefficients are detected to be nonzero based on the *STIV* estimated values. All of the remaining 45 estimated coefficients are zero. The confidence sets based on sparsity certificate with $s = 5$ are infinite. However, those based on the estimated support $\widehat{J} = J\left(\widehat{\beta}\right)$ are finite. They are presented in Table 7. The parameter β_1 corresponds to an endogenous regressor. The associated sensitivity is smaller than those of the exogenous regressors.

9.1.5. *Confidence Sets: One Strong and Many Weak Instruments.* Here we consider the case of many "weak" instruments. We take $L = 155$, $K = 150$, $I^c = \{1, 149, 150\}$, $\zeta_{kl} = 1$ for $(k, l) = (1, 153), (149, 154), (150, 155)$ and $\zeta_{kl} = 0.1$ otherwise. Thus, there is one strong instrument for each endogenous regressor and the remaining 154 are "weak". We set $z_{li} = x_{l'i}$ for $l' = l + 1$ and $l \in [149]$, and $z_{li} = e_{li}$ for $l \in [150, ..., L]$. We adjust r using Scenario 5. This model cannot be estimated using

Table 7. Fewer instruments than regressors (estimated support, $n = 500$)

	β_l	$\widehat{\beta}$	β_u	$\widehat{\kappa}^*_{e_k}$
β_1	0.385	0.892	1.398	0.586
β_2	-2.408	-1.778	-1.148	0.730
β_3	-0.854	-0.349	0.156	0.904
β_4	-0.469	0.082	0.634	0.817
β_5	-1.430	-0.847	-0.265	0.801

Here: $r = 0.140$, $c = 0.854$ and $\widehat{\sigma} = 1.091$.

Table 8. One strong and many weak instruments (estimated support , $n = 500$)

	β_l	$\widehat{\beta}$	β_u	$\widehat{\kappa}^*_{e_k}$
β_1	0.254	0.913	1.573	0.492
β_2	-2.670	-1.860	-1.049	0.773
β_3	-0.953	-0.232	0.488	0.897
β_4	-0.618	0.168	0.953	0.784
β_5	-1.664	-0.812	0.040	0.726

Here: $r = 0.159$, $c = 0.950$ and $\widehat{\sigma} = 1.098$.

Table 9. One strong and many weak instruments (sparsity certificate, $n = 4000$)

	$\beta_{l,10}$	$\beta_{l,9}$	$\beta_{l,8}$	$\beta_{l,7}$	$\beta_{l,6}$	$\beta_{l,5}$	$\widehat{\beta}$	$\beta_{u,5}$	$\beta_{u,6}$	$\beta_{u,7}$	$\beta_{u,8}$	$\beta_{u,9}$	$\beta_{u,10}$
β_1	$-\infty$	-8.641	-1.881	-0.603	-0.043	0.254	0.934	1.614	1.910	2.471	3.749	10.508	∞
β_2	$-\infty$	-8.728	-3.943	-3.0480	-2.662	-2.460	-1.944	-1.428	-1.227	-0.841	0.055	4.840	∞
β_3	$-\infty$	-7.235	-2.443	-1.545	-1.155	-0.947	-0.426	0.095	0.302	0.692	1.591	6.383	∞
β_4	$-\infty$	-6.690	-1.808	-0.899	-0.505	-0.296	0.219	0.734	0.943	1.336	2.246	7.128	∞
β_5	$-\infty$	-8.812	-3.231	-2.192	-1.747	-1.514	-0.931	-0.348	-0.115	0.3300	1.369	6.951	∞
β_6	$-\infty$	-7.389	-2.178	-1.204	-0.777	-0.561	0	0.561	0.777	1.204	2.178	7.389	∞
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_{150}	$-\infty$	-8.454	-2.491	-1.366	-0.875	-0.614	0	0.614	0.875	1.365	2.490	8.451	∞

Here: $r = 0.0569$ and $c = 0.3505$ and $\widehat{\sigma} = 1.0129$

BIC since the first stage would require solving around 10^{45} least squares problems. The estimated values of the nonzero entries of β are clearly distinct from zero. We present in Table 8 the confidence sets based on the estimated support with $\widehat{J} = J\left(\widehat{\beta}\right)$.

For $n = 500$ the confidence sets based on a sparsity certificate are infinite. We consider instead $n = 4000$ and obtain the sets in Table 9. We use the notation $\beta_{l,s}$ and $\beta_{u,s}$ for the lower and upper bounds of the nested confidence sets computed using the sparsity certificate for various degrees of sparsity s. The thresholded estimator depends on a sparsity certificate. When 0 does not lie in the projected confidence intervals, the thresholded *STIV* estimate corresponds to the *STIV* estimate. For comparison, the confidence sets based on $\widehat{J}$ are in Table 10. These are tighter than the confidence sets under the sparsity certificate $s = 5$.

Table 10. One strong and many weak instruments (estimated support , $n = 4000$)

	β_l	$\widehat{\beta}$	β_u	$\widehat{\kappa}^*_{e_k}$
β_1	0.791	0.934	1.076	0.506
β_2	-2.089	-1.944	-1.799	0.924
β_3	-0.579	-0.426	0.2734	0.870
β_4	0.070	0.219	0.368	0.909
β_5	-1.096	-0.931	-0.766	0.832

Here: $r = 0.0569$, $c = 0.3518$, and $\widehat{\sigma} = 1.0129$.

Table 11. Many endogenous regressors and instruments, (estimated support, n=750)

	β_l	$\widehat{\beta}$	β_u	$\widehat{\kappa}^*_{e_k}$
β_1	0.407	0.855	1.302	0.483
β_2	-2.430	-1.843	-1.345	0.897
β_3	-0.907	-0.331	0.245	0.802
β_4	-0.440	0.101	0.641	0.821
β_5	-1.301	-0.772	-0.235	0.861

Here: $r = 0.134$, $c = 0.849$, and $\widehat{\sigma} = 1.103$.

9.1.6. *Many Endogenous Regressors and Many Instruments.* We now consider many endogenous regressors and many instruments. We take $n = 750$, $L = 205$, $K = 200$, $I^c = \{1, 102, 103, ..., 200\}$, and $\zeta_{kl} = 1$ for $(k, l) = (1, 106), (102, 107), (103, 108), ..., (200, 205)$ and 0.1 otherwise. Thus there are 100 endogenous regressors and there is one strong instrument and 99 "weak" instruments for each. We set $z_{li} = x_{l'i}$ for $l' = l + 1$ and $l \in [100]$, and $z_{li} = e_{li}$ for $l \in [101, ..., L]$. We adjust r using Scenario 5. The estimated values of the nonzero components of β are clearly distinct from zero. The remaining entries are 0. Table 11 summarizes the confidence sets based on $\widehat{J} = J\left(\widehat{\beta}\right)$. The confidence sets based on a sparsity certificate are infinite.

9.1.7. *Endogenous Instruments.* Here we consider endogenous instruments as in Section 7.3. We first study case (1), in which we do not rely on beta-min assumptions. We take $n = 4000$, $K = 60$, $L = 100$, $I^c = \{1\}$ and $\zeta_{1l} = 1$ if $l = 90$ and 0.1 otherwise. We consider the case where there is uncertainty regarding the exogeneity of the instruments and $\widetilde{P} = \{91, 92, ..., 100\}$. We set $z_{li} = x_{l'i}$ for $l' = l + 1$ and $l \in [59]$, $z_{li} = e_{li}$ for $l \in [60, 61, ..., 99]$, and $z_{li} = e_{li} + 0.9u_i$ otherwise, where e_{li} are drawn from the standard normal truncated to the interval $[-5, 5]$. Consequently instrument 100 is endogenous and $\widetilde{\beta}_{100} = 0.9$, whilst all other entries are equal to zero. We use the *STIV* for the first stage estimator using only the instruments in $\widetilde{P}^c$, taking $\alpha_1 = 0.025$, using Scenario 5 for $r_1 = 0.057$ and choosing $c = 0.358$ as described in Section 9.1.3. We compute $\widehat{b}$ and $\widehat{b}^\sigma$ with a sparsity certificate $s = 5$. We choose sparsity certificate 5 since the first stage estimator has only 5 entries larger than 10^{-12} in magnitude. This yields $\widehat{b} = 0.104$ and $\widehat{b}^\sigma = 0.657$. For the second stage we use the *NV-STIV* estimator with $\alpha_2 = 0.025$ and compute $r_2 = 0.065$ based on Scenario 4. We choose $\widetilde{c} = 0.087$ as described in Section 9.1.3 and compute nested confidence sets for $\widetilde{\beta}$ using sparsity $\widetilde{s}$ from 5 to 10. Since $\alpha_1 + \alpha_2 = 0.05$, the sets are at the 0.05 level.

Table 12. Detection of endogenous instruments (sparsity certificate, $n = 4000$)

	$\widetilde{\beta}_{l,10}$	$\widetilde{\beta}_{l,9}$	$\widetilde{\beta}_{l,8}$	$\widetilde{\beta}_{l,7}$	$\widetilde{\beta}_{l,6}$	$\widetilde{\beta}_{l,5}$	$\widehat{\widetilde{\beta}}$	$\widetilde{\beta}_{u,5}$	$\widetilde{\beta}_{u,6}$	$\widetilde{\beta}_{u,7}$	$\widetilde{\beta}_{u,8}$	$\widetilde{\beta}_{u,9}$	$\widetilde{\beta}_{u,10}$
$\widetilde{\beta}_{91}$	-0.427	-0.423	-0.419	-0.415	-0.412	-0.408	0	0.408	0.412	0.415	0.419	0.423	0.427
$\widetilde{\beta}_{92}$	-0.432	-0.428	-0.424	-0.421	-0.417	-0.413	0	0.413	0.417	0.421	0.424	0.428	0.432
$\widetilde{\beta}_{93}$	-0.423	-0.419	-0.415	-0.411	-0.408	-0.404	0	0.404	0.408	0.411	0.415	0.419	0.423
$\widetilde{\beta}_{94}$	-0.424	-0.4200	-0.416	-0.412	-0.409	-0.405	0	0.405	0.409	0.412	0.416	0.420	0.424
$\widetilde{\beta}_{95}$	-0.420	-0.416	-0.413	-0.409	-0.405	-0.402	0	0.402	0.405	0.409	0.413	0.416	0.420
$\widetilde{\beta}_{96}$	-0.420	-0.417	-0.413	-0.409	-0.405	-0.402	0	0.402	0.405	0.409	0.413	0.417	0.420
$\widetilde{\beta}_{97}$	-0.430	-0.426	-0.422	-0.418	-0.415	-0.411	0	0.411	0.415	0.418	0.422	0.426	0.430
$\widetilde{\beta}_{98}$	-0.429	-0.425	-0.421	-0.417	-0.413	-0.410	0	0.410	0.413	0.417	0.421	0.425	0.429
$\widetilde{\beta}_{99}$	-0.431	-0.427	-0.423	-0.419	-0.415	-0.412	0	0.412	0.415	0.419	0.423	0.427	0.431
$\widetilde{\beta}_{100}$	0.331	0.335	0.339	0.342	0.346	0.350	0.756	1.162	1.166	1.170	1.173	1.177	1.181

Here: $r_1 = 0.0574$, $r_2 = 0.0648$, $\widetilde{c} = 0.3579$ and $\widehat{\widetilde{\sigma}} = 0.7360$

The results are presented in Table 12. Each of the exogenous instruments is estimated as such. The endogenous instrument is, however, detected. The point estimate of $\widetilde{\beta}_{100}$ is 0.756 compared to a true value of 0.9. The point estimate is biased towards zero due to the shrinkage. The projected confidence intervals for $\widetilde{\beta}_{100}$ does not include 0 for any level of the sparsity in the table, and always includes 0.9. This means that the thresholded *NV-STIV* estimate of this coefficient corresponds to the *NV-STIV* estimate.

Next we consider case (2), which requires beta-min assumptions. In this setting we consider a more demanding data generating process. We take the data generating process of Section 9.1.5 with $n = 1000$, and suppose that there is uncertainty regarding the exogeneity of the instruments and $\widetilde{P} = \{141, 142, ..., 150\}$. We take $\widetilde{\beta}_{150} = 0.9$ and all other entries equal to zero. We choose r_1, r_2, c, $\widetilde{c}$ in the same way as for case (1). The first stage *STIV* estimator uses only the 140 instruments which are known to be exogenous, which is smaller than K. We compute $\widehat{b}$ and $\widehat{b}^{\sigma}$ using the estimated support $\widehat{J} = J\left(\widehat{\beta}\right)$, which is equal to $J(\beta)$. The entries of $\widehat{\beta}$ with coordinates in $\widehat{J}^c$ are effectively 0. This yields $\widehat{b} = 0.120$ and $\widehat{b}^{\sigma} = 0.931$. We use the second stage *NV-STIV* estimator to compute nested confidence sets for $\widetilde{\beta}$ using sparsity $\widetilde{s}$ from 5 to 10.

The results are displayed in Table 13. Each of the exogenous instruments is estimated as such. The endogenous instrument is, however, detected. The point estimate of $\widetilde{\beta}_{150}$ is 0.666 compared to a true value of 0.9. The confidence set for $\widetilde{\beta}_{150}$ does not include 0 for $\widetilde{s} \leq 9$.

Table 13. Detection of endogenous instruments, (estimated support, $n = 1000$)

	$\widetilde{\beta}_{l,10}$	$\widetilde{\beta}_{l,9}$	$\widetilde{\beta}_{l,8}$	$\widetilde{\beta}_{l,7}$	$\widetilde{\beta}_{l,6}$	$\widetilde{\beta}_{l,5}$	$\widehat{\widehat{\beta}}$	$\widetilde{\beta}_{u,5}$	$\widetilde{\beta}_{u,6}$	$\widetilde{\beta}_{u,7}$	$\widetilde{\beta}_{u,8}$	$\widetilde{\beta}_{u,9}$	$\widetilde{\beta}_{u,10}$
$\widetilde{\beta}_{141}$	-0.722	-0.689	-0.658	-0.630	-0.6044	-0.581	0	0.581	0.604	0.630	0.658	0.689	0.722
$\widetilde{\beta}_{142}$	-0.732	-0.698	-0.667	-0.639	-0.6123	-0.588	0	0.588	0.612	0.638	0.667	0.698	0.732
$\widetilde{\beta}_{143}$	-0.724	-0.690	-0.660	-0.631	-0.6056	-0.582	0	0.582	0.606	0.631	0.659	0.690	0.724
$\widetilde{\beta}_{144}$	-0.703	-0.670	-0.641	-0.613	-0.5881	-0.566	0	0.565	0.588	0.613	0.641	0.670	0.703
$\widetilde{\beta}_{145}$	-0.727	-0.693	-0.662	-0.634	-0.6081	-0.584	0	0.584	0.608	0.634	0.662	0.693	0.727
$\widetilde{\beta}_{146}$	-0.730	-0.696	-0.665	-0.637	-0.6105	-0.587	0	0.587	0.611	0.637	0.665	0.696	0.7300
$\widetilde{\beta}_{147}$	-0.727	-0.693	-0.662	-0.634	-0.6083	-0.584	0	0.584	0.609	0.634	0.662	0.693	0.727
$\widetilde{\beta}_{148}$	-0.737	-0.702	-0.671	-0.643	-0.6163	-0.592	0	0.592	0.616	0.643	0.671	0.702	0.737
$\widetilde{\beta}_{149}$	-0.722	-0.688	-0.658	-0.630	-0.6039	-0.580	0	0.580	0.604	0.6300	0.658	0.688	0.722
$\widetilde{\beta}_{150}$	-0.010	0.021	0.050	0.076	0.1002	0.122	0.666	1.209	1.231	1.256	1.282	1.310	1.342

Here: $r_1 = 0.1188$, $r_2 = 0.1361$, $\widetilde{c} = 0.1444$ and $\widehat{\widehat{\sigma}} = 0.5460$

9.1.8. *Two-Stage Confidence Bands.* We illustrate the confidence bands of Section 8 and compute confidence bands for the nonzero components of β. We consider the challenging data generating process of Section 9.1.4 with $n = 4000$. We take Ω as the first five rows of I_K. For the sets of Section 8.1, we take $\alpha_\beta(n) = \alpha_\Lambda(n) = 0.01$, $\alpha = 0.03$, and $\lambda = 0.95$, which yield asymptotic coverage 0.95 under the premises of Theorem 8.2. The first stage estimator is the *STIV* adjusting r using Scenario 5. For the sets of Section 8.2, we take $n_+ = 3200$ and $n_- = 800$. We also adjust our sets to account for the possibility of non-negligible bias, as explained in Section 8.3, based on an estimated support $\widehat{J} = J\left(\widehat{\beta}\right)$. In most cases the estimated support is equal to $J(\beta)$, though it is sometimes a superset.

The results for the method of Section 8.1 are depicted in Table 14. Columns 2-4 summarize the distribution of the preliminary estimator $\Omega\widehat{\beta}$. The shrinkage of the *STIV* estimator leads to a mild bias towards zero for each element of $\beta_{J(\beta)}$. Columns 5-7 summarize the distribution of the bias corrected estimator $\widehat{\Omega\beta}$. The bias correction increases the magnitude of the parameters relative to the preliminary estimator, which become centered around their true values. The difference between the 97.5 and 2.5 percentiles is around 0.06 in all cases. Columns 8-10 summarize the width for each of the parameters and coverage of the confidence bands. That is, the width of the interval around $(\Omega\beta)_o$ for each o. These sets have coverage 0.93, which is marginally below the desired level. The width of the bands is around 0.08 in all cases. This is larger than the difference between the 97.5 and 2.5 percentiles of the bias corrected estimator (0.06). This is because we have constructed a band, as opposed to individual intervals for each parameter. Columns 11-13 summarize the confidence bands of Section 8.3, which account for the remaining bias by computing an upper bound. These sets are

Table 14. Two-stage confidence bands through consistent estimation of Λ (7300 replications)

$\Omega\beta$	Preliminary ($\Omega\widehat{\beta}$)			Debiased ($\widehat{\Omega\beta}$)			CB width (basic)			CB width (bound on bias)		
	p 2.5	p 50	p 97.5	p 2.5	p 50	p 97.5	p 2.5	p 50	p 97.5	p 2.5	p 50	p 97.5
β_1	0.961	0.983	1.006	0.971	1.003	1.031	0.071	0.075	0.081	0.243	0.263	0.392
β_2	-1.980	-1.945	-1.912	-2.038	-2.002	-1.969	0.076	0.080	0.083	0.248	0.266	0.396
β_3	-0.469	-0.438	-0.409	-0.526	-0.495	-0.466	0.076	0.080	0.084	0.248	0.266	0.397
β_4	0.159	0.189	0.220	0.216	0.246	0.277	0.076	0.080	0.083	0.247	0.266	0.396
β_5	-0.971	-0.942	-0.907	-1.028	-0.999	-0.965	0.076	0.080	0.083	0.247	0.267	0.396
							Cover	0.928		Cover	1	

The accuracy of the coverage probabilities is $\pm 1.96\sqrt{0.95(0.05)/7300} = 0.005$ with probability 95%.

conservative. This is because the upper bound on the bias is based on convex relaxation. Despite their conservative nature, the bands are still sufficiently narrow so as to be informative, for example, regarding the signs of the parameters.

The results for the method of Section 8.2 are depicted in Table 15. The bias correction increases the magnitude of the preliminary estimator and is centred around the true values. The width of the confidence bands under the assumption of normality of the error is around 0.1, which is slightly larger than in Table 14, and the coverage is close to the desired 0.95. We also compute confidence bands based on r_+^Λ using Scenario 4, which are wider but still sufficiently narrow as to be informative, for example, with regards to the signs of the parameters. The conservative coverage stems from the use of $r_+^\Lambda = r_{0+}^\Lambda \left|\mathbf{Z}_+\widehat{\Lambda}\mathbf{D}_{\mathbf{Z}_+\widehat{\Lambda}}\right|_\infty$, which is a conservative choice. Assumption 3.1 cannot be applied to replace the second term with something close to 1 because $\widehat{\Lambda}$ might not converge to a nonrandom matrix.

The bands from Section 8.2 take around one second to compute, whereas those from Section 8.1 can take at least twenty times as long. In additional simulations we found instances in which the coverage of the bands which ignore the bias of the debiased estimator was incorrect. This was in cases where n is small and/or instruments are weak for which the rest of the paper proposes solutions.

We now consider a data generating process closer to the empirical application, in which we have $n \approx 5000$, $K = L \approx 1900$ and around 1800 endogenous regressors. These dimensions are too large to simulate a sufficient number of datasets in order to accurately compute the coverage probability. Instead, we reduce n, K and L by a factor 0.5 yielding $n = 2500$, $K = L = 950$ and we set $I^c = \{1, 52, ..., 950\}$. We use $\zeta_{kl} = 1$ for $(k, l) = (1, 51), (51, 2), ..., (899, 949), (900, 950)$ and $\zeta_{kl} = 0.1$ otherwise. Thus, there is one strong instrument for each endogenous regressor and the remaining 949 are "weak". This mirrors the empirical application, in which the instrument for any

Table 15. Two-stage confidence bands through sample splitting (7300 replications)

$\Omega\beta$	Preliminary $(\Omega\widehat{\beta})$			Debiased $(\widehat{\Omega\beta})$			CB width (Normal)			CB width (Scenario 4)		
	p 2.5	p 50	p 97.5	p 2.5	p 50	p 97.5	p 2.5	p 50	p 97.5	p 2.5	p 50	p 97.5
β_1	0.955	0.980	1.001	0.962	1.003	1.043	0.092	0.105	0.120	0.341	0.407	0.506
β_2	-1.974	-1.939	-1.902	-2.035	-1.999	-1.959	0.088	0.099	0.110	0.330	0.385	0.464
β_3	-0.479	-0.437	-0.404	-0.539	-0.497	-0.464	0.089	0.099	0.111	0.336	0.386	0.467
β_4	0.154	0.192	0.227	0.213	0.253	0.291	0.089	0.098	0.111	0.327	0.385	0.468
β_5	-0.973	-0.940	-0.903	-1.034	-0.999	-0.960	0.089	0.100	0.111	0.333	0.386	0.465
							Cover	0.952		Cover	≈ 1	

Table 16. Two-stage confidence bands for a DGP similar to the application, 7300 replications

$\Omega\beta$	Preliminary $(\Omega\widehat{\beta})$			Debiased $(\widehat{\Omega\beta})$			CB width (Normal)			CB width (Scenario 4)		
	p 2.5	p 50	p 97.5	p 2.5	p 50	p 97.5	p 2.5	p 50	p 97.5	p 2.5	p 50	p 97.5
β_1	0.923	0.954	0.975	0.951	0.991	1.039	0.095	0.114	0.136	0.346	0.424	0.544
β_2	-1.937	-1.889	-1.845	-2.043	-1.996	-1.948	0.108	0.125	0.143	0.381	0.467	0.563
β_3	-0.432	-0.388	-0.342	-0.542	-0.493	-0.446	0.107	0.124	0.143	0.379	0.462	0.573
β_4	0.103	0.142	0.192	0.202	0.247	0.299	0.107	0.125	0.145	0.387	0.462	0.579
β_5	-0.934	-0.889	-0.847	-1.043	-0.993	-0.946	0.106	0.124	0.143	0.376	0.462	0.570
							Cover	0.933		Cover	≈ 1	

regressor which is a function of the expenditure share takes exactly the same form but replaces the expenditure share with the average expenditure share in the sample. In the application the first stage R^2 for each reduced form equation exceeds 0.9. The data generating process we consider here yields values of 0.7. To complete the data generating process we set $z_{li} = x_{l'i}$ for $l' = l + 1$ and $l \in [49]$, and $z_{li} = e_{li}$ for $l \in [50, ..., L]$. We adjust r using Scenario 5. We do not compute the confidence bands of Section 8.1, as it is computationally infeasible for the dimensions considered here. We focus instead on the feasible bands of Section 8.2. We use $n_+ = 2000$ and $n_- = 500$.

Table 16 presents the results. The bias corrected estimates of the coefficients become centered on their true values. The difference between the 97.5 and 2.5 percentiles of the bias corrected estimator is around 0.09 for each of the parameters. Under the assumption of normally distributed errors, the width of the bands is around 0.12. This exceeds 0.09 since we have constructed bands for the 5 coefficients, as opposed to invidiual confidence intervals. The coverage is around 0.93, which is marginally below the desired level. The bands based on r_+^{Λ} using Scenario 4 are wider, but still sufficiently narrow so as to be informative, particularly regarding the signs of the parameters.

9.2. **Application to the Second Order Approximation of the EASI Model.** We use the Canadian demand data of Lewbel and Pendakur (2009) for $n = 4,847$ rental-tenure single-member households that had positive expenditures on rent recreation, and transportation. The categories of goods considered 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) the individual's age minus 40, (2) the gender dummy equal to one for men, (3) a car-nonowner dummy 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 variable equal to the calendar year minus 1986 (that is, equal to zero in 1986). In this application, the instruments are strong and $L = K$ so it makes sense to proceed in two-stages, using the confidence bands of Section 8. The exposition of the results focuses on Engel curves. We construct uniform confidence bands at the 0.95 level for the Engel curves based on a grid of $O = 25$ points, and apply the sample splitting method of Section 8.2 with $n_+ = 4000$, $n_- = 847$ and $\alpha_\beta(n_+) = \alpha_\Lambda(n_-) = \alpha = 0.05/3$. We use sample splitting for computational reasons, since the the program to compute the bands based on consistent estimation of Λ involves $OK = 46,975$ conic constraints, and is not computationally tractable.

In the first step, we apply the *SE-STIV* estimator to sample $+$, adjusting r based on Scenario 5 using $\alpha_\beta(n_+)/G$, $c = 0.99/r$, and taking $\rho_{g,E} = 1/n_+$ for each good $g \in [G]$. By appealing to the union bound, we allow for unrestricted correlation between ϵ_{gi} and ϵ_{hi} for each pair of goods g and h. We choose the sets $P_1, P_2, ..., P_G$ so as to exempt the constant 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 Banks, Blundell, and Lewbel (1997)). We impose all of the restrictions in Section 2.2.6 apart from monotonicity of cost, which we verify is satisfied by the estimated parameters. For the negative semidefinite restriction, we construct a grid over the characteristics z_i using the minimum and maximum values observed in the sample. This leads to 64 semidefinite restrictions. For brevity, we do not present the full estimation results here, focussing instead on the Engel curves which we discuss below. We note that of the $1771G = 15399$ parameters out of $1879G = 16911$ on which we allow sparsity, only 50 are estimated as nonzero, 22 of which are parameters which arise due to the second order approximation of the Exact Affine Stone Index.

In the second step, we compute $\widehat{\Lambda}$, choosing λ_1 according to the method discussed in Section 9.1.3 and setting $\lambda_2 = \lambda_1$, which ensures $\lambda_2 < r'_{0-}$. Figure 1 depicts the preliminary estimator of the Engel curves, its bias corrected counterpart and 95% confidence bands for food-in and food-out. The curves are close to linear and have the expected slopes: negative for food-in and positive for food-out. The bias correction for food-in is smaller than for food-out and the band is also tighter, indicating less uncertainty. The band is wider close to the end points of the support of the functions. This is most likely because there is less data at the end points than in the centre, and is true for all of the goods. Figures 2-5 depict the Engel curves for the other goods. The bias correction is large for rent, transportation-operation, and personal care, for which the bands are also wider than the other goods.

The Engel curves are similar to those of Lewbel and Pandakur (2009) apart from clothing and transportation-operation. The confidence bands are marginally wider. This is expected since we present uniform bands rather than pointwise intervals and consider a coverage of 95% compared to 90% in Lewbel and Pandakur (2009). Our estimated Engel-curve for clothing follows a less pronounced U-shape with minimum at expenditure $4,450 than the estimated Engel curve of Lewbel and Pandakur (2009) with minimum at expenditure $8,100. Lewbel and Pandakur (2009) find a downwards sloping Engel curve for transport operation, whereas we find evidence of an inverted U. Our Engel curve for transportation operation suggests that those with low expenditure increase the expenditure share on transportation as their expenditure rises. This may be due to substitution between modes of transport which become more affordable with rising expenditure.

References

[1] Andrews, D. (1999): "Consistent Moment Selection Procedures for Generalized Method of Moments Estimation". *Econometrica,* 67, 543–564.

[2] Andrews, D., and J. Stock (2007): "Inference with Weak Instruments", in: *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.

[3] Bahadur, R., and L. Savage (1956): "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems". *Annals of Mathematical Statistics,* 27, 1115–1122.

[4] Banks, J., R. Blundell, and A. Lewbel (1997): "Quadratic Engel Curves and Consumer Demand". *The Review of Economics and Statistics,* 79, 527–539.

[5] Basmann, R. (1960): "On Finite Sample Distributions of Generalized Classical Linear Identifiability Test Statistics". *Journal of the American Statistical Association,* 55, 650–659.

[6] Bekker, P.A. (1994): "Alternative approximations to the distributions of instrumental variable estimators". *Econometrica,* 62, 657–681.

[7] Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain". *Econometrica,* 80, 2369–2429.

[8] Belloni, A., V. Chernozhukov, and L. Wang (2011): "Square-Root Lasso: Pivotal Recovery of Sparse Signals Via Conic Programming". *Biometrika,* 98, 791–806.

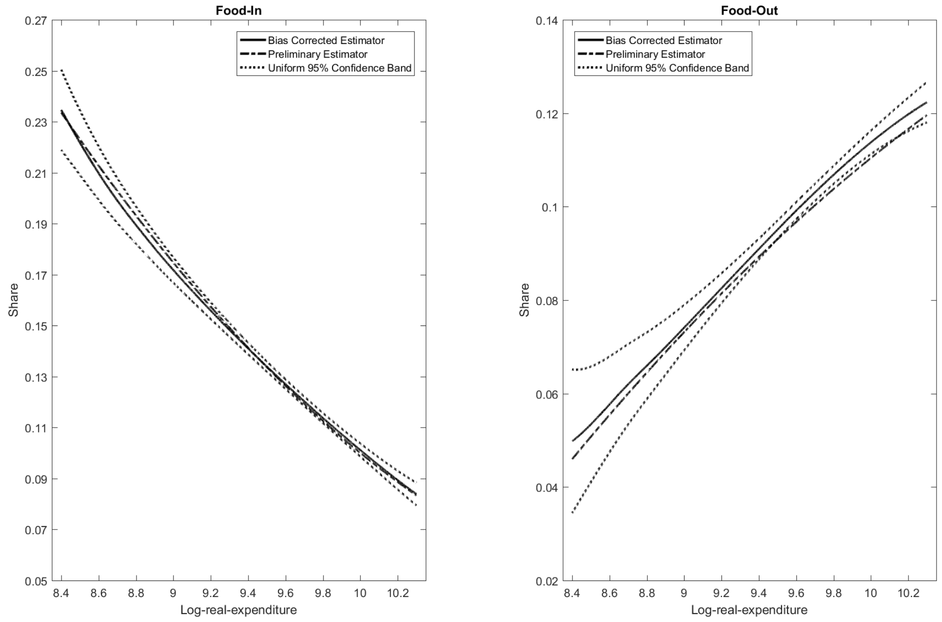

Figure 1. Engel curves for Food-in and Food-out

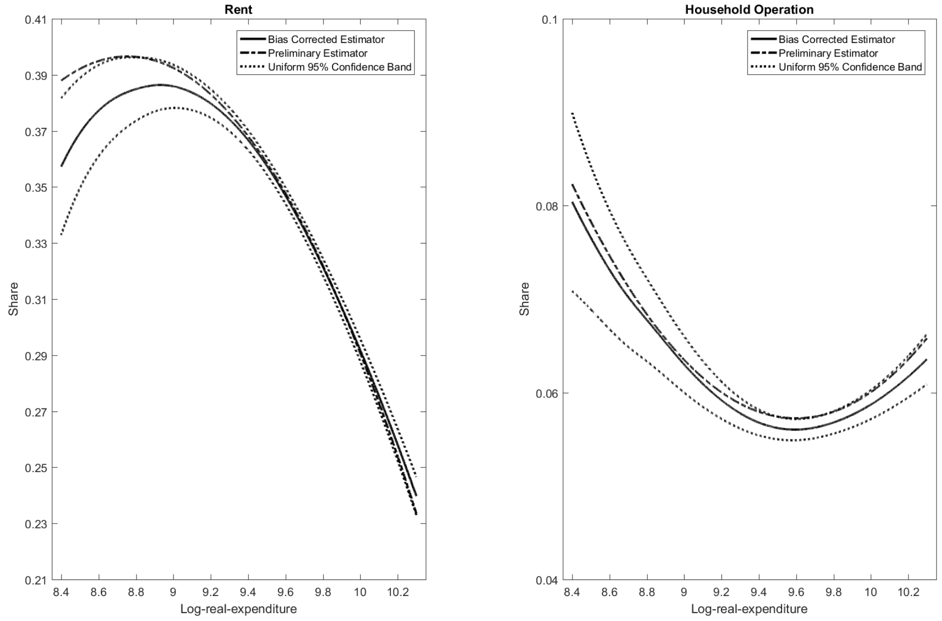

Figure 2. Engel curves for Rent and Household operation

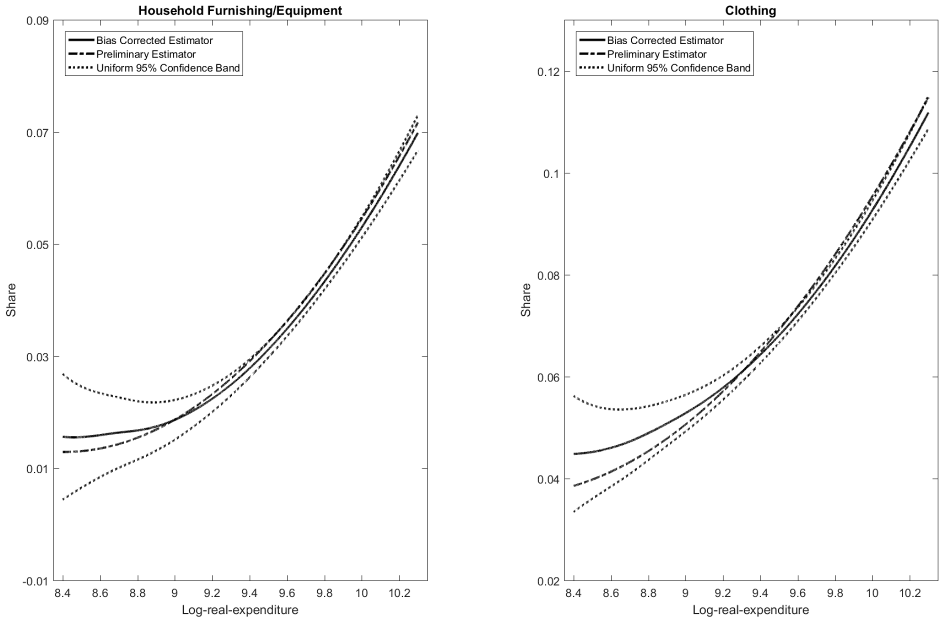

Figure 3. Engel curves for Household furnishing and Clothing

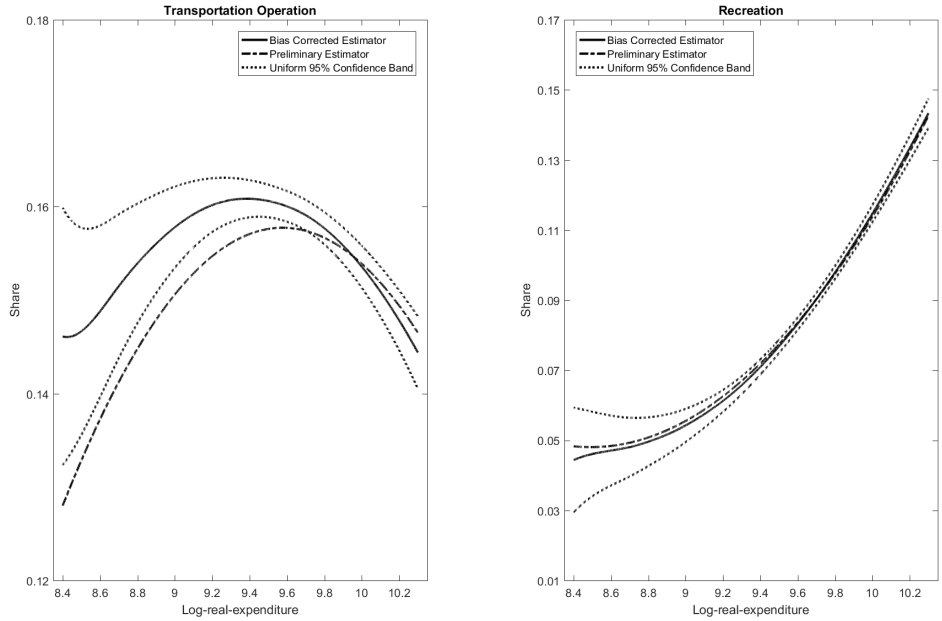

Figure 4. Engel curves for Transport operation and Recreation

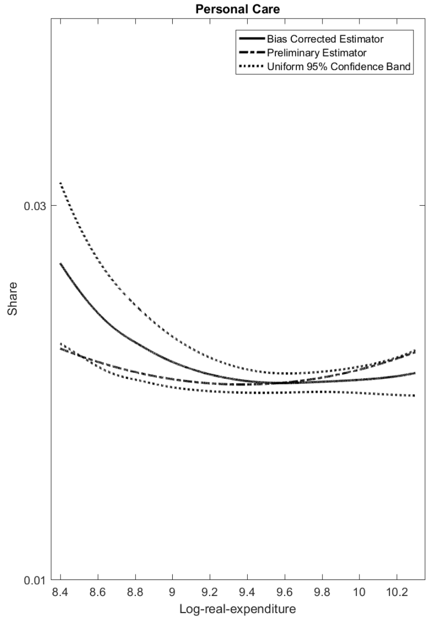

Figure 5. Engel curves for Personal care

[9] Belloni, A., and V. Chernozhukov (2011a): "L1-Penalized Quantile Regression in High-Dimensional Sparse Models". *Annals of Statistics,* 39, 82–130.

[10] Belloni, A., and V. Chernozhukov (2011b): "High-dimensional Sparse Econometric Models: an Introduction", in: *Inverse Problems and High Dimensional Estimation, Stats in the Château 2009*, Alquier, P., E. Gautier, and G. Stoltz, Eds., *Lecture Notes in Statistics*, 203, 127–162, Springer.

[11] Belloni, A., and V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls". *The Review of Economic Studies,* 81, 608-650

[12] Bertail, P. , E. Gauthérat, and H. Harari-Kermadec (2005): "Empirical-Discrepancies and Quasi-Empirical Likelihood : Exponential Bounds". Preprint CREST 2005-34.

[13] Bickel, P., J. Ritov, and A. Tsybakov (2009): "Simultaneous Analysis of Lasso and Dantzig Selector". *Annals of Statistics,* 37, 1705–1732.

[14] Bound, J., D. Jaeger, and R. Baker (1995): "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak,". *Journal of the American Statistical Association,* 90, 443–450.

[15] Cai, T., W. Liu, and X. Luo (2011): "A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation". *Journal of the American Statistical Association,* 106, 594–607.

[16] Caner M., and H. Zhang (2014): "Adaptive Elastic Net for Generalized Methods of Moments ". *Journal of Business and Economics Statistics,* 32, 30–47.

[17] Caner M., and Q. Zhang (2015): "Hybrid Generalized Empirical Likelihood Estimators: Instrument Selection with Adaptive Lasso". *Journal of Econometrics,* 187, 256–274.

[18] Caner M. (2014): "Near Exogeneity and Weak Identification in Generalized Empirical Likelihood Estimators: Many Moment Asymptotics". *Journal of Econometrics,* 182, 247–268.

[19] Candès, E., and T. Tao (2007): "The Dantzig Selector: Statistical Estimation when p is Much Larger Than n". *Annals of Statistics,* 35, 2313–2351.

[20] Chen, X., Q.-M. Shao, W. B. Wu, and L. Xu (2016): "Self-normalized cramér-type moderate deviations under dependence". *Annals of Statistics,* 44, 1593–1617.

[21] Cheng, X., and Z. Liao (2015): "Select the Valid and Relevant Moments: An Information-based LASSO for GMM with Many Moments". *Journal of Econometrics,* 186, 443–464.

[22] Chernozhukov, V., D. Chetverikov, and K. Kato (2013): "Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-Dimensional Random Vectors". *Annals of Statistics,* 41, 2786–2819.

[23] Chernozhukov, V., D. Chetverikov, and K. Kato (2017): "Central Limit Theorems and Bootstrap in High Dimensions". *Annals of Probability,* 45, 2309–2352.

[24] Dobriban, E., and J. Fan: "Regularity Properties of High-dimensional Covariate Matrices". Preprint 1305.5198v2.

[25] Dufour, J.-M. (1997): "Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models". *Econometrica,* 65, 1365–1387.

[26] Dümbgen, L., S. van de Geer, M. Veraar, and J. Wellner (2010): "Nemirovski's Inequalities Revisited". *American Mathematical Monthly,* 117, 138–160.

[27] Fan, J., and Liao, Y. (2014): "Endogeneity in High dimensions". *Annals of Statistics,* 42, 872–917.

[28] Feng, M., J. Mitchell, J.-S. Pang, X. Shen, and A. Wächter (2014): "Complementarity Formulation of ℓ_0-norm optimization problems". Technical Report, Northwestern University.

[29] Gautier, E., and C. Rose (2015,17): "Inference on Social Effects when the Network is Sparse and Unknown", Working Paper.

[30] Gautier, E., and A. Tsybakov (2013): "Pivotal Estimation in High-Dimensional Regression via Linear Programming", in: *Empirical Inference, Festschrift in Honor of Vladimir N. Vapnik*, Springer.

[31] Gold, D., J. Lederer, and J. Tao (2017): "Inference for High-dimensional Instrumental Variables Regression". Preprint 1708.05499v2.

[32] Henrion, D., J.-B. Lasserre, and J. Lofberg (2009): "GloptiPoly 3: Moments, Optimization and Semidefinite Programming". *Optimization Methods and Software* 24, 761—779.

[33] Hansen, L.-P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators". *Econometrica*, 50, 1029–1054.

[34] Javanmard, A., and A. Montanari (2014): "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression". *Journal of Machine Learning Research*, 15, 2869–2909.

[35] Kang, H., A. Zhang, T. Cai, and S. Small (2016): "Instrumental Variables Estimation With Some Invalid Instruments and its Application to Mendelian Randomization". *Journal of the American Statistical Association,* 111, 132–144.

[36] Kolesár, M., R. Chetty, J. Fiedman, E. Glaseser, and G. Imbens (2015): "Identification and Inference With Many Invalid Instruments". *Journal of Business & Economic Statistics,* 33, 474–484.

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

[38] Lewbel, A., and K. Pendakur (2009): "Tricks with Hicks: The EASI Demand System". *American Economic Review,* 99, 827–863.

[39] Lounici, K. (2008): "Sup-Norm Convergence Rate and Sign Concentration Property of the Lasso and Dantzig Selector". *Electronic Journal of Statistics,* 2, 90–102.

[40] Nelson, C. R., and R. Startz (1990): "Some Further Results on the Exact Small Sample Properties of the Instrumental Variables Estimator". *Econometrica*, 58, 967–976.

[41] Nevo, A., and A. Rosen (2012): "Identification with Imperfect Instruments". *The Review of Economic and Statistics,* 94, 659–671.

[42] Nickl, R., and S. van de Geer (2013): "Confidence Sets in Sparse Regression". *Annals of Statistics,* 41, 2852–2876.

[43] Romano, J., and M. Wolf (2000): "Finite Sample Nonparametric Inference and Large Sample Efficiency". *Annals of Statistics,* 28, 756–778.

[44] Rose, C. (2016): "Identification of Spillover Effects using Panel Data". Job Market Paper.

[45] Sargan, J. (1958): "The Estimation of Economic Relationships Using Instrumental Variables". *Econometrica,* 26, 393–415.

[46] Sala-i-Martin, X. (1997): "I Just Ran Two Million Regressions". *The American Economic Review,* 87, 178–183.

[47] Staiger, D., and J. Stock (1997): "Instrumental Variables Regression with Weak Instruments". *Econometrica,* 65, 557–586.

[48] van de Geer, S. , P., Bühlmann, Y. Ritov, and R. Dezeure (2014): "On Asymptotically Optimal Confidence Regions and Tests for High-dimensional Models.". *Annals of Statistics*, 42, 1166–1202.

[49] Ye, F., and C.-H. Zhang (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.

[50] Zhang, X., and G. Cheng (2014): "Bootstrapping High-dimensional Time Series". Preprint arXiv:1406.1037v2.

[51] Zhang, C.-H., and S. Zhang (2014): "Confidence Intervals for Low Dimensional Parameters in High-dimensional Linear Models". *Journal of the Royal Statistical Society Series B Statistical Methodology,* 76, 217–242.

[52] Zhu, Y. (2015): "Sparse Linear Models and ℓ_1 Regularized 2SLS with High-Dimensional Endogenous Regressors and Instruments". *Journal of Econometrics,* 202, 196–213.

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

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

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

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

CREST, ENSAE, 5 avenue Henry Le Chatelier, 91764 Palaiseau, France.

E-mail address: alexandre.tsybakov@ensae.fr.

SUPPLEMENTAL APPENDIX FOR “HIGH-DIMENSIONAL INSTRUMENTAL VARIABLES AND CONFIDENCE SETS”

ERIC GAUTIER, CHRISTIERN ROSE, AND ALEXANDRE TSYBAKOV

A.1. **Moderate Deviations for Self-normalized Sums.** Throughout this section $(x_i)_{i=1}^n$ are independent random variables such that, for all i, $\mathbb{E}[x_i] = \mathbf{0}$. The following result is due to Efron (1969).

Theorem A.1. *If $(x_i)_{i=1}^n$ are symmetric, then for all $r > 0$,*

$$\mathbb{P}\left(\frac{\mathbb{E}_n[X]}{\sqrt{\mathbb{E}_n[X^2]}} \geq r\right) \leq \exp\left(-\frac{nr^2}{2}\right).$$

This upper bound is refined in Pinelis (1994) for i.i.d. random variables.

Theorem A.2. *If $(x_i)_{i=1}^n$ are symmetric and identically distributed, then*

$$\forall r \in [0,1),\ \mathbb{P}\left(\frac{\mathbb{E}_n[X]}{\sqrt{\mathbb{E}_n[X^2]}} \geq r\right) \leq \frac{2e^3}{9}\Phi\left(-\sqrt{n}r\right).$$

The following result is Theorem 2.3 in Jing, Shao and Wang (2003).

Theorem A.3. *Assume that $0 < \mathbb{E}[|X|^{2+\delta}] < \infty$ for some $0 < \delta \leq 1$ and set*

$$B_n^2 = n\mathbb{E}[X^2],\ L_{n,\delta} = n\mathbb{E}\left[|X|^{2+\delta}\right],\ d_{n,\delta} = B_n/L_{n,\delta}^{1/(2+\delta)}.$$

Then

$$\forall 0 \leq r \leq \frac{d_{n,\delta}}{\sqrt{n}},\ \mathbb{P}\left(\frac{\mathbb{E}_n[X]}{\sqrt{\mathbb{E}_n[X^2]}} \geq r\right) \leq \Phi(-\sqrt{n}r)\left(1 + A_0\left(\frac{1+\sqrt{n}r}{d_{n,\delta}}\right)^{2+\delta}\right),$$

where $A_0 > 0$ is an absolute constant.

Despite of its interest for large deviations behavior of self-normalized sums, the bound has limited practical use for moderate deviations because A_0 is not an explicit constant.
The following result is a corollary of Theorem 1 in Bertail, Gauthérat, and Harari-Kermadec (2008).

Theorem A.4. *Assume that $(x_i)_{i=1}^n$ are identically distributed and $0 < \mathbb{E}[X^4] < \infty$. Then*

$$\forall r \geq 0,\ \mathbb{P}\left(\frac{|\mathbb{E}_n[X]|}{\sqrt{\mathbb{E}_n[X^2]}} \geq r\right) \leq (2e+1)\exp\left(-\frac{nr^2}{2+\gamma_4 r^2}\right) \tag{A.1}$$

where $\gamma_4 = \frac{\mathbb{E}[X^4]}{\mathbb{E}[X^2]^2}$*, while*

$$\forall r \geq \sqrt{n},\ \mathbb{P}\left(\frac{|\mathbb{E}_n[X]|}{\sqrt{\mathbb{E}_n[X^2]}} \geq r\right) = 0.$$

Proof. Bertail, Gauthérat, and Harari-Kermadec (2008) obtain the upper bound for $r \geq \sqrt{n}$ and that for $0 \leq r < \sqrt{n}$

$$\mathbb{P}\left(\frac{\left|\frac{1}{n}\sum_{i=1}^n x_i\right|}{\sqrt{\frac{1}{n}\sum_{i=1}^n x_i^2}} \geq r\right) \leq \inf_{a>1}\left\{2e\exp\left(-\frac{nr^2}{2(1+a)}\right) + \exp\left(-\frac{n}{2\gamma_4}\left(1-\frac{1}{a}\right)^2\right)\right\}.$$

Because

$$\frac{1}{1+a} = \frac{1}{a}\frac{1}{1+\frac{1}{a}} \geq \frac{1}{a}\left(1-\frac{1}{a}\right)$$

we obtain

$$-\frac{r^2}{1+a} \leq -\frac{r^2}{a}\left(1-\frac{1}{a}\right).$$

This yields (A.1) by choosing a to equate the two exponential terms. □

A.2. **Some Facts From Convex Analysis.** We will use the following property of convex functions that can be found, for example, in Nesterov (2004). Let f be a convex function from $\mathbb{R}^K$ to $\mathbb{R}$. The subdifferential of f at x is defined by

$$\partial f(x) \triangleq \left\{g \in \mathbb{R}^K :\ \forall z \in \mathbb{R}^K,\ f(z) \geq f(x) + g^\top(z-x).\right\}$$

Lemma A.1. *Let* $f(x) = \max_{l=1,\dots,m} f_l(x)$ *where the fonctions* f_l *are convex and defined everywhere. Then* f *is convex and its subdifferential is*

$$\partial f(x) = \mathrm{Conv}\left\{\partial f_l(x) \text{ for } l \text{ such that } f_l(x) = f(x)\right\} \tag{A.2}$$

where Conv *denotes the convex hull.*

A.3. **Proofs.** In this section, we prove the results of the main text.

Proof of the statement for Scenario 5. Let $\beta \in \mathcal{I}dent$. Define the events

$$\mathcal{E}_U \triangleq \left\{\left|(\mathbb{E}_n - \mathbb{E})[U(\beta)^2]\right| \geq \tau\mathbb{E}[U(\beta)^2]\right\};$$

$$\mathcal{E}_Z \triangleq \left\{\left|\mathbf{D}_Z(\mathbb{E}_n - \mathbb{E})\left[ZZ^\top\right]\mathbf{D}_Z\right|_\infty \geq \tau_Z\right\};$$

$$\mathcal{E}'_Z \triangleq \left\{\min_{l\in[L]}\left(\widehat{\mathbf{D}}_{\mathbf{Z}}^{-1}\right)_{ll}(\mathbf{D}_Z)_{ll} \leq \sqrt{1-\tau'_Z} \text{ or } \max_{l\in[L]}\left(\widehat{\mathbf{D}}_{\mathbf{Z}}^{-1}\right)_{ll}(\mathbf{D}_Z)_{ll} \geq \sqrt{1+\tau'_Z}\right\}.$$

Clearly, we have $\mathcal{E}_Z' \subseteq \mathcal{E}_Z$. We obtain, by the Chebychev inequality and (ii)-(iii), $\mathbb{P}\left(\mathcal{E}_U\right) \le m_4/(\tau^2 n)$,

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}_Z\right) &= \mathbb{P}\left(\left|\sum_{i=1}^n \left(\mathbf{D}_Z z_i z_i^\top \mathbf{D}_Z - \mathbb{E}\left[\mathbf{D}_Z Z Z^\top \mathbf{D}_Z\right]\right)\right|_\infty \ge n\tau_Z\right)\\ &\le \frac{1}{n^2\tau_Z^2}\mathbb{E}\left[\left|\sum_{i=1}^n \left(\mathbf{D}_Z z_i z_i^\top \mathbf{D}_Z - \mathbb{E}\left[\mathbf{D}_Z Z Z^\top \mathbf{D}_Z\right]\right)\right|_\infty^2\right]\\ &\le \frac{C_{\mathrm{N}}(L)}{n\tau_Z^2}\mathbb{E}\left[\left|\mathbf{D}_Z Z Z^\top \mathbf{D}_Z - \mathbb{E}\left[\mathbf{D}_Z Z Z^\top \mathbf{D}_Z\right]\right|^2\right] \quad \text{(by the Nemirovski inequality)}\\ &\le \frac{C_{\mathrm{N}}(L^2) M_Z(L)}{n\tau_Z^2},\end{aligned}$$

and similarly

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}_Z'\right) &= \mathbb{P}\left(\max_{l\in[L]}\left|\sum_{i=1}^n \left(\frac{z_{li}^2}{\mathbb{E}\left[Z_l^2\right]} - 1\right)\right| \ge n\tau_Z'\right)\\ &\le \frac{C_{\mathrm{N}}(L) M_Z'(L)}{n\left(\tau_Z'\right)^2}.\end{aligned}$$

Define

$$T \triangleq \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \widehat{\mathbf{D}}_{\mathbf{Z}} z_i \frac{u_i(\beta)}{\sqrt{\widehat{Q}(\beta)}}\right|_\infty, \ T_0 \triangleq \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{D}_Z z_i \frac{u_i(\beta)}{\sigma_{U(\beta)}}\right|_\infty, \ W_0 \triangleq \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{D}_Z z_i e_i\right|_\infty, \ N_0 \triangleq \left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \chi_i\right|_\infty,$$

where χ_i are independent Gaussian vectors of covariance $\mathbb{E}[\mathbf{D}_Z z_i z_i^\top \mathbf{D}_Z]$. Because

$$\mathbb{E}\left[\mathbf{D}_Z z_i z_i^\top \mathbf{D}_Z \frac{u_i(\beta)^2}{\sigma_{U(\beta)^2)}}\right] = \mathbb{E}\left[\mathbf{D}_Z z_i z_i^\top \mathbf{D}_Z \mathbb{E}\left[\left.\frac{u_i(\beta)^2}{\sigma_{U(\beta)^2)}}\right| \mathbf{Z}\right]\right] = \mathbb{E}[\mathbf{D}_Z z_i z_i^\top \mathbf{D}_Z],$$

N_0 is a Gaussian approximation of T_0. Using (vi) and Proposition 2.1 in Chernozhukov, Chetverikov, and Kato (2017), we obtain, for a constant C_2 which can depend only on q_2, for all $t \in \mathbb{R}$,

$$\max\left(\left|\mathbb{P}\left(T_0 \le t\right) - \mathbb{P}\left(N_0 \le t\right)\right|, \left|\mathbb{P}\left(W_0 \le t\right) - \mathbb{P}\left(N_0 \le t\right)\right|\right) \le \rho. \tag{A.3}$$

Note that, by (i), we have $\mathbb{E}\left[\left(\left(\mathbf{D}_Z\right)_{ll} Z_l U(\beta)/\sigma_{U(\beta)}\right)^2\right] = \mathbb{E}\left[\left(\left(\mathbf{D}_Z\right)_{ll} Z_l e_i\right)^2\right] = 1$ for all $l \in [L]$, so condition (M.1) from Chernozhukov, Chetverikov, and Kato (2017) is satisfied. We denote by q_{W_0} the conditional quantile functions of W_0 given $\boldsymbol{Z}$. Lemma 3.1 in Chernozhukov, Chetverikov, and Kato (2013) yields, for all $t \in \mathbb{R}$,

$$\left|\mathbb{P}\left(W_0 \le t\right| \mathbf{Z}) - \mathbb{P}\left(N_0 \le t\right)\right| \le \varphi(\tau_Z) \quad \text{on } \mathcal{E}_Z^c,$$

where φ is the function $x \in (0,1) \to \varphi(x) = C_1 x^{1/3} \max\left(1, \log(L/x)\right)^{2/3}$ and C_1 is a constant (we are in a situation where $c_1 = C_1$ with their notations) and Lemma 3.2 yields, for all $\alpha \in (0,1)$,

$$\mathbb{P}\left(q_{W_0}(\alpha) \le q_{N_0}(\alpha + \varphi(\tau_Z))\right) \ge 1 - \frac{C_{\mathrm{N}}(L^2) M_Z(L)}{n\tau_Z^2},$$
$$\mathbb{P}\left(q_{N_0}(\alpha) \le q_{W_0}(\alpha + \varphi(\tau_Z))\right) \ge 1 - \frac{C_{\mathrm{N}}(L^2) M_Z(L)}{n\tau_Z^2}.$$

If $\mathbb{P}\left(T_0 \le q_{W_0}(\alpha)\right) - \alpha > 0$, we have

$$\left|\mathbb{P}\left(T_0 \le q_{W_0}(\alpha)\right) - \alpha\right| \le \mathbb{P}\left(T_0 \le q_{N_0}(\alpha + \varphi(\tau_Z))\right) - \alpha + \frac{C_{\mathrm{N}}(L^2) M_Z(L)}{n\tau_Z^2},$$

else,

$$\left|\mathbb{P}\left(T_0 \le q_{W_0}(\alpha)\right) - \alpha\right| \le \alpha - \mathbb{P}\left(T_0 \le q_{N_0}(\alpha - \varphi(\tau_Z))\right) + \frac{C_{\mathrm{N}}(L^2) M_Z(L)}{n\tau_Z^2},$$

hence, in both cases,

$$\begin{aligned}\left|\mathbb{P}\left(T_0 \le q_{W_0}(\alpha)\right) - \alpha\right| &\le \mathbb{P}\left(q_{N_0}(\alpha - \varphi(\tau_Z)) \le T_0 \le q_{N_0}(\alpha + \varphi(\tau_Z))\right) + 2\frac{C_{\mathrm{N}}(L^2) M_Z(L)}{n\tau_Z^2} \\ &\le \mathbb{P}\left(q_{N_0}(\alpha - \varphi(\tau_Z)) \le N_0 \le q_{N_0}(\alpha + \varphi(\tau_Z))\right) + 2\frac{C_{\mathrm{N}}(L^2) M_Z(L)}{n\tau_Z^2} + 2\rho \\ &\le 2\varphi(\tau_Z)) + 2\frac{C_{\mathrm{N}}(L^2) M_Z(L)}{n\tau_Z^2} + 2\rho, \qquad \text{(A.4)}\end{aligned}$$

where the second display above uses the first upper bound from (A.32).

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

$$|W - W_0| \le \left(\frac{1}{\sqrt{1 - \tau_Z'}} - 1\right) W_0,$$

hence, by the Markov inequality, law of iterated expectations, and second bound in (A.32),

$$\mathbb{P}\left(\mathbb{P}\left(|W - W_0| > \zeta_1 \middle| \mathbf{Z}\right) > \zeta_2(\zeta_1, \tau_Z')\right) < \frac{1}{\zeta_2(\zeta_1, \tau_Z')} \left(\mathbb{P}\left(N_0 > \zeta_1 \left(\frac{1}{\sqrt{1 - \tau_Z'}} - 1\right)^{-1}\right) + \rho + \frac{C_{\mathrm{N}}(L) M_Z'(L)}{n\left(\tau_Z'\right)^2}\right)$$

(A.5)

$$\mathbb{P}\left(\mathbb{P}\left(|W - W_0| > \zeta_1 \middle| \mathbf{Z}\right) > \zeta_2(\zeta_1, \tau_Z')\right) < \zeta_2(\zeta_1, \tau_Z'),$$

where

$$\zeta_2(\zeta_1, \tau_Z')^2 \triangleq \mathbb{P}\left(N_0 > \zeta_1 \left(\frac{1}{\sqrt{1 - \tau_Z'}} - 1\right)^{-1}\right) + \rho + \frac{C_{\mathrm{N}}(L) M_Z'(L)}{n\left(\tau_Z'\right)^2},$$

C_1 is universal and C_2 is a constant which only depends on q_2. On $\mathcal{E}_Z'^c \cap \mathcal{E}_U^c$, we have

$$|T - T_0| \leq \max\left(\frac{1}{\sqrt{(1-\tau_Z')(1-\tau)}} - 1\right) T_0,$$

hence, by the first bound in (A.32),

(A.6)
$$\mathbb{P}\left(|T - T_0| > \zeta_1\right) \leq \mathbb{P}\left(N_0 > \zeta_1 \left(\frac{1}{\sqrt{(1-\tau_Z')(1-\tau)}} - 1\right)^{-1}\right) + \frac{C_{\mathrm{N}}(L) M_Z'(L)}{n\left(\tau_Z'\right)^2} + \frac{m_4}{\tau^2 n} + \rho = \zeta_2'(\zeta_1).$$

Using Lemma 3.3 in Chernozhukov, Chetverikov, and Kato (2013) and (A.5) for the first display, (A.6) for the second display, and (A.4) for the third display, we get
if $\mathbb{P}\left(T \leq q_W(\alpha)\right) - \alpha > 0$, we have

$$\begin{aligned}
|\mathbb{P}\left(T \leq q_W(\alpha)\right) - \alpha| &\leq \mathbb{P}\left(T \leq q_{W_0}(\alpha + \zeta_2(\zeta_1, \tau_Z')) + \zeta_1\right) - \alpha + \zeta_2(\zeta_1, \tau_Z') \\
&\leq \mathbb{P}\left(T_0 \leq q_{W_0}(\alpha + \zeta_2(\zeta_1, \tau_Z'))\right) - \alpha - \zeta_2(\zeta_1, \tau_Z') + 2\zeta_2(\zeta_1, \tau_Z') + \zeta_2'(\zeta_1) \\
&\leq 2\varphi(\tau_Z)) + 2\frac{C_{\mathrm{N}}(L^2) M_Z(L)}{n\tau_Z^2} + 2\rho + 2\zeta_2(\zeta_1, \tau_Z') + \zeta_2'(\zeta_1),
\end{aligned}$$

else,

$$\begin{aligned}
|\mathbb{P}\left(T \leq q_W(\alpha)\right) - \alpha| &\leq \alpha - \mathbb{P}\left(T \leq q_{W_0}(\alpha - \zeta_2(\zeta_1, \tau_Z')) - \zeta_1\right) + \zeta_2(\zeta_1, \tau_Z') \\
&\leq 2\varphi(\tau_Z)) + 2\frac{C_{\mathrm{N}}(L^2) M_Z(L)}{n\tau_Z^2} + 2\rho + 2\zeta_2(\zeta_1, \tau_Z') + \zeta_2'(\zeta_1),
\end{aligned}$$

hence, in both cases,

$$|\mathbb{P}\left(T \leq q_W(\alpha)\right) - \alpha| \leq 2\varphi(\tau_Z)) + 2\frac{C_{\mathrm{N}}(L^2) M_Z(L)}{n\tau_Z^2} + 2\rho + 2\zeta_2(\zeta_1, \tau_Z') + \zeta_2'(\zeta_1).$$

The result on the deterministic upper bound $\overline{r}$ on r is obtained using Lemma 3.2 and Lemma 3.3 in Chernozhukov, Chetverikov, and Kato (2013). □

Proof of Theorem 4.1. Take $\beta \in \mathcal{I}dent$ and set $\Delta \triangleq \widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\left(\widehat{\beta} - \beta\right)$. Because $\left|\frac{1}{n}\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{Z}^\top(\mathbf{Y} - \mathbf{X}\beta)\right|_\infty = \left|\frac{1}{n}\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{Z}^\top\mathbf{U}(\beta)\right|_\infty$ and $\widehat{Q}(\beta) = \mathbb{E}_n[U(\beta)^2]$, on $\mathcal{G}$, $\beta \in \widehat{\mathcal{I}}\left(r, \sqrt{\widehat{Q}(\beta)}\right)$.
On $\mathcal{G}$, we have:

$$\left|\widehat{\Psi}\Delta\right|_\infty \leq \left|\frac{1}{n}\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{Z}^\top\left(\mathbf{Y} - \mathbf{X}\widehat{\beta}\right)\right|_\infty + \left|\frac{1}{n}\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{Z}^\top(\mathbf{Y} - \mathbf{X}\beta)\right|_\infty \tag{A.7}$$

$$\leq r\left(\widehat{\sigma} + \sqrt{\widehat{Q}(\beta)}\right). \tag{A.8}$$

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

$$\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\widehat{\beta}_P\right|_1+c\widehat{\sigma}\le|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_P|_1+c\sqrt{\widehat{Q}(\beta)}. \tag{A.9}$$

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

$$\begin{aligned}\left|\Delta_{J(\beta)^c\cap P}\right|_1 &= \sum_{k\in J(\beta)^c\cap P}\left|\mathbb{E}_n[X_k^2]^{1/2}\widehat{\beta}_k\right| \\ &\le \sum_{k\in J(\beta)\cap P}\left(\left|\mathbb{E}_n[X_k^2]^{1/2}\beta_k\right|-\left|\mathbb{E}_n[X_k^2]^{1/2}\widehat{\beta}_k\right|\right)+c\left(\sqrt{\widehat{Q}(\beta)}-\widehat{\sigma}\right) \\ &\le \left|\Delta_{J(\beta)\cap P}\right|_1+c\left(\sqrt{\widehat{Q}(\beta)}-\sqrt{\widehat{Q}\left(\widehat{\beta}\right)}\right).\end{aligned} \tag{A.10}$$

The last inequality holds because by construction $\sqrt{\widehat{Q}\left(\widehat{\beta}\right)}\le\widehat{\sigma}$.

Because $\gamma\to\sqrt{\widehat{Q}(\gamma)}$ is convex and

$$w_*\triangleq-\frac{\frac{1}{n}\sum_{i=1}^n x_i(y_i-x_i^\top\beta)}{\sqrt{\frac{1}{n}\sum_{i=1}^n(y_i-x_i^\top\beta)^2}}\mathbb{1}\left\{\frac{1}{n}\sum_{i=1}^n(y_i-x_i^\top\beta)^2\neq0\right\}\in\partial\sqrt{\widehat{Q}(\cdot)}(\beta).$$

we have

$$\begin{aligned}\sqrt{\widehat{Q}(\beta)}-\sqrt{\widehat{Q}\left(\widehat{\beta}\right)} &\le w_*^\top\left(\beta-\widehat{\beta}\right) \\ &= \left(\widehat{\mathbf{D}}_{\mathbf{X}}w_*\right)^\top\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\left(\beta-\widehat{\beta}\right)=-\left(\widehat{\mathbf{D}}_{\mathbf{X}}w_*\right)^\top\Delta.\end{aligned}$$

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

$$\left|\left(\widehat{\mathbf{D}}_{\mathbf{X}}w_*\right)_k\right|\le\frac{|\mathbb{E}_n[X_kU(\beta)]|}{\sqrt{\mathbb{E}_n[X_k^2]\mathbb{E}_n[U(\beta)^2]}}\le1.$$

Finally, we obtain

$$\sqrt{\widehat{Q}(\beta)}-\sqrt{\widehat{Q}\left(\widehat{\beta}\right)}\le r|\Delta_I|_1+|\Delta_{I^c}|_1. \tag{A.11}$$

Combining (A.11) with (A.10), we find that $\Delta\in\widehat{C}_{J(\beta)}$ on $\mathcal{G}$. Using (A.7) and (A.11), we find

$$\begin{aligned}\left|\widehat{\Psi}\Delta\right|_\infty &\le r\left(\widehat{\sigma}+\sqrt{Q\left(\widehat{\beta}\right)}+\sqrt{\widehat{Q}(\beta)}-\sqrt{\widehat{Q}\left(\widehat{\beta}\right)}\right) \\ &\le r\left(2\overline{\sigma}+r|\Delta_I|_1+|\Delta_{I^c}|_1\right).\end{aligned} \tag{A.12}$$

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

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

which implies

$$\left|\widehat{\Psi}\Delta\right|_\infty \leq 2r\overline{\sigma}\left(1 - \frac{r^2}{\widehat{\kappa}_{\sigma,J(\beta)}}\right)_+^{-1}. \tag{A.13}$$

(A.13) and the definition of the sensitivities yield the first upper bound.

To obtain the second bound we use that, by (A.9) and the definition of $\widehat{\kappa}_{1,J(\beta)\cap P,J(\beta)}$,

$$\begin{aligned} c\widehat{\sigma} &\leq |\Delta_{J(\beta)\cap P}|_1 + c\sqrt{\widehat{Q}(\beta)} \\ &\leq \frac{\left|\widehat{\Psi}\Delta\right|_\infty}{\widehat{\kappa}_{1,J(\beta)\cap P,J(\beta)}} + c\sqrt{\widehat{Q}(\beta)} \qquad \text{(A.14)} \end{aligned}$$

Using (A.8) and (A.14) yields the second upper bound. □

Proof of Theorem 4.2. Take $\beta \in \mathcal{I}dent$ and $J \subseteq [K]$. Acting as in (A.10) and assuming that we are on $\mathcal{G}$, we get

$$\begin{aligned} \sum_{k\in J^c\cap P}\left|\mathbb{E}_n[X_k^2]^{1/2}\widehat{\beta}_k\right| + \sum_{k\in J^c\cap P}\left|\mathbb{E}_n[X_k^2]^{1/2}\beta_k\right| &\leq \sum_{k\in J\cap P}\left(\left|\mathbb{E}_n[X_k^2]^{1/2}\beta_k\right| - \left|\mathbb{E}_n[X_k^2]^{1/2}\widehat{\beta}_k\right|\right) \\ &\quad + 2\sum_{k\in J^c\cap P}\left|\mathbb{E}_n[X_k^2]^{1/2}\beta_k\right| + c\left(\sqrt{\widehat{Q}(\beta)} - \sqrt{\widehat{Q}\left(\widehat{\beta}\right)}\right) \\ &\leq |\Delta_{J\cap P}|_1 + 2\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_{J^c\cap P}\right|_1 + cr\left|\Delta_I\right|_1 + c\left|\Delta_{I^c}\right|_1. \end{aligned}$$

This yields

$$|\Delta_{J^c\cap P}|_1 \leq |\Delta_{J\cap P}|_1 + 2\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_{J^c\cap P}\right|_1 + cr\left|\Delta_I\right|_1 + c\left|\Delta_{I^c}\right|_1. \tag{A.15}$$

Let us show the first inequality and consider two cases.

Case 1: $2\left|\left(\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta\right)_{J^c\cap P}\right|_1 \leq |\Delta_{J\cap P}|_1 + cr\left|\Delta_I\right|_1 + c\left|\Delta_{I^c}\right|_1 + |\Delta_{P^c}|_1$, then $\Delta \in \widehat{C}_{\gamma,J}$. From this, using the definition of the sensitivity $\widehat{\gamma}_{q,T,J}$, we get the upper bound corresponding to the first term in the minimum. Also, we have

$$\begin{aligned} \widehat{\sigma} &\leq \frac{1}{c}\left(\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_P\right|_1 - \left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\widehat{\beta}_P\right|_1\right) + \sqrt{\widehat{Q}(\beta)} \\ &\leq \frac{1}{c}\min\left(\left|\Delta_P\right|_1, \left|\Delta_{J\cap P}\right|_1 + \left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_{J^c\cap P}\right|_1\right) + \sqrt{\widehat{Q}(\beta)} \\ &\leq \frac{1}{c}\min\left(\left|\Delta_P\right|_1, \frac{1}{2}\left(3\left|\Delta_{J\cap P}\right|_1 + cr\left|\Delta_I\right|_1 + c\left|\Delta_{I^c}\right|_1 + |\Delta_{P^c}|_1\right)\right) + \sqrt{\widehat{Q}(\beta)} \qquad \text{(A.16)} \end{aligned}$$

$$\leq \frac{\left|\widehat{\Psi}\Delta\right|_\infty}{c\widehat{\gamma}_{Q,J}} + \sqrt{\widehat{Q}(\beta)}, \tag{A.17}$$

which, with (A.8), yields the upper bound corresponding to the second term in the minimum.
Case 2: $2\left|\left(\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta\right)_{J^c\cap P}\right|_1 > |\Delta_{J\cap P}|_1 + cr\left|\Delta_I\right|_1 + c\left|\Delta_{I^c}\right|_1 + |\Delta_{P^c}|_1$, then we have

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

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

Proof of Proposition 5.1. This is a simple consequence of the definition of the sensitivities, the restricted sets $C_{\widehat{J}}$ and $C_{\widehat{J}}^{\gamma}$, and the fact that minimizing on a larger set yields lower bounds on the sensitivities. More specifically, we use that $|\Delta_{J\cap P}|_1 \leq 2\min\left(s, \left|\widehat{J}\right|\right)\left|\Delta_{\widehat{J}\cap P}\right|_\infty$. The last constraint is not convex but the restricted set is a union of sets involving the linear constraint $|\Delta_{J\cap P}|_1 \leq 2\min\left(s, \left|\widehat{J}\right|\right)\Delta_j$, hence the second minimum. One can assume everywhere that $\Delta_j \geq 0$ because the objective function in the sensitivities involves a ℓ_∞-norm so that changing Δ in $-\Delta$ does not change the sensitivities. □

Proof of Proposition 6.1. Define the events:

$$\mathcal{E}_X \triangleq \left\{\min_{k\in[K]}\left(\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\right)_{kk}(\mathbf{D}_X)_{kk} \leq \sqrt{1-\tau_X} \text{ or } \max_{k\in[K]}\left(\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\right)_{kk}(\mathbf{D}_X)_{kk} \geq \sqrt{1+\tau_X}\right\};$$

$$\mathcal{E}_{ZX^\top} \triangleq \left\{\left|\mathbf{D}_Z(\mathbb{E}_n - \mathbb{E})\left[ZX^\top\right]\mathbf{D}_X\right|_\infty \geq r_\Psi\right\}.$$

The event $\mathcal{E}_Z'$ is such that $\mathcal{E}_Z' \subseteq \mathcal{E}_Z$, where $\mathcal{E}_Z$ has been introduced in the appendix for the formal justification of the choice of r from Scenario 5. We use the one above for scenarii 1-4 and the previous one for Scenario 5. Define $\mathcal{G}_\Psi \triangleq \{r \leq \overline{r}\} \cap (\mathcal{E}_Z')^c \cap \mathcal{E}_X^c \cap \mathcal{E}_{ZX^\top}^c \cap \mathcal{E}_U^c$ for scenarii 1-4 and $\mathcal{G}_\Psi \triangleq \{r \leq \overline{r}\} \cap \mathcal{E}_Z^c \cap \mathcal{E}_X^c \cap \mathcal{E}_{ZX^\top}^c \cap \mathcal{E}_U^c$ for Scenario 5. Recall that $\mathbb{P}(r \leq \overline{r}) \geq 1 - \alpha_C(n)$. The probability of the events $\mathcal{E}_U$ and $\mathcal{E}_Z'$ are analyzed in the proof of the statement for Scenario 5. Similarly, we have

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_X\right) &\leq \frac{C_{\mathrm{N}}(K)M_X(K)}{n\tau_X^2};\\
\mathbb{P}\left(\mathcal{E}_{ZX^\top}\right) &= \mathbb{P}\left(\left|\sum_{i=1}^n\left(\mathbf{D}_Z z_i x_i^\top\mathbf{D}_X - \mathbb{E}\left[\mathbf{D}_Z ZX^\top\mathbf{D}_X\right]\right)\right|_\infty \geq nr_\Psi\right)\\
&\leq \frac{1}{n^2r_\Psi^2}\mathbb{E}\left[\left|\sum_{i=1}^n\left(\mathbf{D}_Z z_i x_i^\top\mathbf{D}_X - \mathbb{E}\left[\mathbf{D}_Z ZX^\top\mathbf{D}_X\right]\right)\right|_\infty^2\right]\\
&\leq \frac{C_{\mathrm{N}}(LK)M(L,K)}{nr_\Psi^2}.
\end{aligned}$$

Clearly, on $\mathcal{E}_X^c$, (6.3) holds. Assume now that we work on the event $\mathcal{G}_\Psi$.

Let $J \subseteq [K]$, $l \in \mathcal{L}$, and $\overline{\Delta} \triangleq \mathbf{D}_X^{-1}\widehat{\mathbf{D}}_{\mathbf{X}}\Delta$. Due to (6.3), we have, for all $l \in \mathcal{L}$, in particular ℓ_1-norms of subvectors, $\sqrt{1-\tau_X}\, l\left(\overline{\Delta}\right) \le l(\Delta) \le \sqrt{1+\tau_X}\, l\left(\overline{\Delta}\right)$. This, the fact that $r \le \overline{r}$ and manipulations on the ℓ_1-norm of subvectors used previously, yield $\overline{\Delta} \in C_J$ if $\Delta \in \widehat{C}_J$ and $\overline{\Delta} \in C_{\gamma,J}$ if $\Delta \in \widehat{C}_{\gamma,J}$. Now, because $\mathcal{G}_\Psi \subseteq (\mathcal{E}'_Z)^c \cap \mathcal{E}^c_{ZX^\top}$, we obtain

$$
\begin{aligned}
\left|\widehat{\Psi}\Delta\right|_\infty &\ge \min_{l \in [L]} \left(\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{D}_Z^{-1}\right)_{ll} \left|\mathbf{D}_Z \mathbb{E}_n\left[ZX^\top\right]\mathbf{D}_X\mathbf{D}_X^{-1}\widehat{\mathbf{D}}_{\mathbf{X}}\Delta\right|_\infty \\
&\ge \frac{1}{\sqrt{1+\tau_Z}}\left(\left|\mathbf{D}_Z\mathbb{E}\left[ZX^\top\right]\mathbf{D}_X\overline{\Delta}\right|_\infty - \left|\mathbf{D}_Z(\mathbb{E}_n-\mathbb{E})\left[ZX^\top\right]\mathbf{D}_X\overline{\Delta}\right|_\infty\right) \\
&\ge \frac{1}{\sqrt{1+\tau_Z}}\left(\left|\Psi\overline{\Delta}\right|_\infty - r_\Psi\left|\overline{\Delta}\right|_1\right).
\end{aligned}
$$

Inequalities (6.4) and (6.5) are obtained from the definition of $\kappa_{1,J}$ and $\gamma_{1,J}$ and the fact that, on $\mathcal{G}_\Psi$, $l(\Delta) \le \sqrt{1+\tau_X}\, l\left(\overline{\Delta}\right)$. Finally, we check that $\mathbb{P}(\mathcal{G}\cap\mathcal{G}_\Psi) \ge 1-\alpha_D(n)$. □

Proof of Theorem 6.1. The first inequalities in (i) and (ii) follow from the second bounds in theorems 4.1 and 4.2 and Proposition 6.1.

The second inequality in item (i) is obtained as follows.

Work on the event $\mathcal{G}\cap\mathcal{G}_\Psi$. By (A.14) and (A.8), we have

$$
\begin{aligned}
\left|\widehat{\Psi}\Delta\right|_\infty &\le r\left(\frac{\left|\widehat{\Psi}\Delta\right|_\infty}{c\widehat{\kappa}_{1,J(\beta)\cap P,J(\beta)}} + 2\sqrt{\widehat{Q}(\beta)}\right) \\
&\le 2r\sqrt{\widehat{Q}(\beta)}\left(1-\frac{r}{c\widehat{\kappa}_{1,J(\beta)\cap P,J(\beta)}}\right)_+^{-1} \\
&\le 2\overline{r}\sqrt{\widehat{Q}(\beta)}\left(1-\frac{\overline{r}\sqrt{(1+\tau_Z)(1+\tau_X)}}{c\kappa_{1,J(\beta)\cap P,J(\beta)}}\left(1-\frac{r_\Psi}{\kappa_{1,J}}\right)_+^{-1}\right)_+^{-1} \\
&\le 2\overline{r}\sqrt{1+\tau}\sigma_{U(\beta)}\left(1-\frac{\overline{r}\sqrt{(1+\tau_Z)(1+\tau_X)}}{c\kappa_{1,J(\beta)\cap P,J(\beta)}}\left(1-\frac{r_\Psi}{\kappa_{1,J}}\right)_+^{-1}\right)_+^{-1}
\end{aligned}
$$

This, together with (A.14), yield the last inequality. The middle inequality $\sqrt{\widehat{Q}\left(\widehat{\beta}\right)} \le \widehat{\sigma}$ comes from the definition of the estimator. Finally, the first inequality uses (A.11).

Let us now prove the second inequality in item (ii). With the same arguments as for item (i) using (A.17) instead of (A.14), in case 1, we have

$$
\sigma_{U(\beta)}\left(\sqrt{1-\tau}-\frac{2\overline{r}^2\sqrt{1+\tau}\Theta_\gamma(J)}{\gamma_{\sigma,J}}\right) \le \sqrt{\widehat{Q}\left(\widehat{\beta}\right)} \le \widehat{\sigma} \le \sigma_{U(\beta)}\sqrt{1+\tau}\left(1+\frac{2\overline{r}\Theta_\gamma(J)}{c\gamma_{Q,J}}\right).
$$

In case 2, using the second element in the minimum in (A.16), we get

$$\widehat{\sigma} \le \frac{3}{c}\left|\left(\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta\right)_{J^c\cap P}\right|_1 + \sqrt{\widehat{Q}(\beta)}$$

and, using (A.11),

$$\sqrt{\widehat{Q}\left(\beta\right)} - \frac{2}{c}\left|\left(\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta\right)_{J^c\cap P}\right|_1 \le \sqrt{\widehat{Q}\left(\widehat{\beta}\right)}.$$

Hence the result.

Part (iii) follow from (i) and (ii) with $l(\Delta) = |e_k^\top \Delta|$ and the fact that the assumption on $|\beta_k|$ imply: $\widehat{\beta}_k \neq 0$ for $k \in J(\beta)$ (resp., J_*). □

Proof of Theorem 6.3. Fix s and β in $\mathcal{B}_s$ and work on $\mathcal{G}\cap\mathcal{G}_\Psi$. Using Theorem 6.1 (i), we obtain $\widehat{\omega}_k(s) \le \omega_k(s)$. The following two cases can occur.

First, if $k \in J(\beta)^c$ (so that $\beta_k = 0$) then, using (5.1) for l defined, for all Δ, by $l(\Delta) = \left|e_k^\top \Delta\right|$ we obtain $\left|\widehat{\beta}_k\right| \le \widehat{\omega}_k(s)$, which implies $\widehat{\beta}_k^\omega = 0$.

Second, if $k \in J(\beta)$, then using again (5.1) for the same functional, we get $\left|\left|\widehat{\beta}_k\right| - |\beta_k|\right| \le \left|\widehat{\beta}_k - \beta_k\right| \le \widehat{\omega}_k(s)/\sqrt{(1-\tau_X)\mathbb{E}[X_k^2]} \le \omega_k(s)/\sqrt{(1-\tau_X)\mathbb{E}[X_k^2]}$. Since $|\beta_k| > 2\omega_k(s)/\sqrt{(1-\tau_X)\mathbb{E}[X_k^2]}$ for $k \in J(\beta)$, we obtain $\left|\widehat{\beta}_k\right| > \omega_k(s)/\sqrt{(1-\tau_X)\mathbb{E}[X_k^2]} \ge \widehat{\omega}_k(s)/\sqrt{\mathbb{E}_n[X_k^2]}$, so that $\widehat{\beta}_k^\omega = \widehat{\beta}_k$. □

Proof of Theorem 7.1. Take $\beta \in \mathcal{I}dent$ and work on $\mathcal{G}\cap\mathcal{G}_\Psi$. We have, using the triangle inequality in the second and fourth display, and the definition of $\mathcal{G}$ and the Cauchy-Schwartz inequality in the third display,

$$\begin{aligned}\left|\frac{1}{n}\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{Z}^\top(\mathbf{Y}-\mathbf{X}\beta)\right|_\infty &= \left|\frac{1}{n}\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{Z}^\top\left(\mathbf{W}(\beta)+\mathbf{V}(\beta)\right)\right|_\infty \\ &\le \left|\frac{1}{n}\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{Z}^\top\mathbf{W}(\beta)\right|_\infty + \left|\frac{1}{n}\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{Z}^\top\mathbf{V}(\beta)\right|_\infty \\ &\le r\sqrt{\mathbb{E}_n[W(\beta)^2]} + \sqrt{\mathbb{E}_n[V(\beta)^2]} \\ &\le r\sqrt{\widehat{Q}(\beta)} + (r+1)\sqrt{\mathbb{E}_n[V(\beta)^2]} \\ &\le r\sqrt{\widehat{Q}(\beta)} + (r+1)\rho_E.\end{aligned}$$

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

$$\left|\widehat{\Psi}\Delta\right|_\infty \le r\left(\widehat{\sigma} + \sqrt{\widehat{Q}(\beta)}\right) + 2(r+1)\sqrt{1+\tau}E. \tag{A.18}$$

Moreover, by the inverse triangle inequality, we have

$$\begin{aligned}\sqrt{\widehat{Q}(\beta)} &\geq \sqrt{\mathbb{E}_n[W(\beta)^2]} - \sqrt{\mathbb{E}_n[V(\beta)^2]},\\ &\geq \sqrt{1-\tau}\sigma_W(\beta) - \sqrt{1+\tau}E.\end{aligned}$$

Hence, by convexity, we have

$$\begin{aligned}\sqrt{\widehat{Q}(\beta)} - \sqrt{\widehat{Q}\left(\widehat{\beta}\right)} &\leq \min\left(r + \frac{(r+1)\sqrt{1+\tau}E}{\left(\sqrt{1-\tau}\sigma_W(\beta) - \sqrt{1+\tau}E\right)_+}, 1\right)|\Delta_I|_1 + |\Delta_{I^c}|_1\\ &\leq \min\left(r + (r+1)\left(\sqrt{\frac{1-\tau}{1+\tau}}\frac{\sigma_W(\beta)}{E} - 1\right)_+^{-1}, 1\right)|\Delta_I|_1 + |\Delta_{I^c}|_1\\ &\leq \overline{r}(\beta)\left|\Delta_I\right|_1 + \left|\Delta_{I^c}\right|_1. \qquad \text{(A.19)}\end{aligned}$$

Start by considering the case of a sparse vector β. By definition of $\widehat{\kappa}_{1,J(\beta)\cap P,J(\beta)}$ and (A.18), we have

$$\begin{aligned}\left|\widehat{\Psi}\Delta\right|_\infty &\leq r\left(\frac{\left|\widehat{\Psi}\Delta\right|_\infty}{c\widehat{\kappa}_{1,J(\beta)\cap P,J(\beta)}} + 2\sqrt{\widehat{Q}(\beta)}\right) + 2(r+1)\sqrt{1+\tau}E\\ &\leq 2r\left(\sqrt{\widehat{Q}(\beta)} + (r+1)\sqrt{1+\tau}E\right)\left(1 - \frac{r}{c\widehat{\kappa}_{1,J(\beta)\cap P,J(\beta)}}\right)_+^{-1}\\ &\leq 2\overline{r}\left(\sqrt{\widehat{Q}(\beta)} + (\overline{r}+1)\sqrt{1+\tau}E\right)\left(1 - \frac{\overline{r}\sqrt{(1+\tau_Z)(1+\tau_X)}}{c\kappa_{1,J(\beta)\cap P,J(\beta)}}\left(1 - \frac{r_\Psi}{\kappa_{1,J}}\right)_+^{-1}\right)_+^{-1}\\ &\leq 2\overline{r}\sqrt{1+\tau}\sigma(\beta)\left(1 - \frac{\overline{r}\sqrt{(1+\tau_Z)(1+\tau_X)}}{c\kappa_{1,J(\beta)\cap P,J(\beta)}}\left(1 - \frac{r_\Psi}{\kappa_{1,J}}\right)_+^{-1}\right)_+^{-1} \qquad \text{(A.20)}\end{aligned}$$

The rest is similar to what we have done before. □

Proof of Theorem 7.2. Take β in $\mathcal{I}dent$, set $\Delta_g \triangleq \widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\left(\widehat{\beta}_g - \beta_g\right)$, and work on $\mathcal{G}\cap\mathcal{G}_\Psi$. Most of the arguments are the same as those in the proof of Theorem 7.1 and we do not reproduce them. Rather we stress the main differences. We have, for $g \in [G]$,

$$\left|\frac{1}{n}\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{Z}^\top(\mathbf{Y}_g - \mathbf{X}\beta_g)\right|_\infty \leq r\sqrt{\widehat{Q}(\beta_g)} + (r+1)\rho_{g,E},$$

hence, $\beta \in \widehat{\mathcal{I}}_{SE}\left(r, \sqrt{\widehat{Q}(\beta_1)}, \dots, \sqrt{\widehat{Q}(\beta_G)}\right)$ and

$$\left|\widehat{\Psi}\Delta_g\right|_\infty \leq r\left(\widehat{\sigma}_g + \sqrt{\widehat{Q}(\beta_g)}\right) + 2(r+1)\sqrt{1+\tau}E_g,$$

$$\begin{aligned}\sqrt{\widehat{Q}(\beta_g)} - \sqrt{\widehat{Q}\left(\widehat{\beta}_g\right)} &\le \min\left(r + (r+1)\left(\sqrt{\frac{1-\tau}{1+\tau}}\frac{\sigma_{W_g(\beta_g)}}{E_g} - 1\right)_+^{-1}, 1\right)\left|(\Delta_g)_I\right|_1 + \left|(\Delta_g)_{I^c}\right|_1 \\ &\le \overline{r}(\beta)\left|(\Delta_g)_I\right|_1 + \left|(\Delta_g)_{I^c}\right|_1 . \qquad \text{(A.21)}\end{aligned}$$

Hence we obtain the first inequality of (i). The second is obtained by using the inverse triangle inequality and the definition of $\overline{\kappa}_{\sigma,J(\beta)}$.

Also, by definition of the estimator and the above, we have, for $J_1, \dots, J_G$ in $[K]$,

$$\begin{aligned}|\Delta_{J^c\cap P}|_1 \le{}& |\Delta_{J\cap P}|_1 + 2\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_{J^c\cap P}\right|_1 \\ &+ c\sum_{g\in[G]}\left(\min\left(\overline{r} + (\overline{r}+1)\left(\sqrt{\frac{1-\tau}{1+\tau}}\frac{\sigma_{W_g(\beta_g)}}{E_g} - 1\right)_+^{-1}, 1\right)\left|(\Delta_g)_I\right|_1 + \left|(\Delta_g)_{I^c}\right|_1\right) \\ \le{}& |\Delta_{J\cap P}|_1 + 2\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_{J^c\cap P}\right|_1 + c\left(\overline{r}(\beta)\left|\Delta_{\mathbf{I}}\right|_1 + \left|\Delta_{\mathbf{I}^c}\right|_1\right)\end{aligned}$$

and

$$\begin{aligned}\sum_{g=1}^G\left|\widehat{\Psi}\Delta_g\right|_\infty &\le r\sum_{g=1}^G\left(\widehat{\sigma}_g + \sqrt{\widehat{Q}(\beta_g)}\right) + 2(r+1)\sqrt{1+\tau}\sum_{g=1}^G E_g \\ &\le \frac{r}{c}\left(\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_{\mathbf{P}}\right|_1 - \left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\widehat{\beta}_{\mathbf{P}}\right|_1\right) + 2r\sum_{g=1}^G\sqrt{\widehat{Q}(\beta_g)} + 2(r+1)\sqrt{1+\tau}\sum_{g=1}^G E_g.\end{aligned}$$

The second inequality from (ii) is obtained in a similar manner as in the proof of Theorem 8.1. The last inequality follows from (A.21). □

Complements on the *C-STIV*. The classes $\widetilde{\mathcal{P}}_j$ are defined in a similar manner as in Assumption 6.1 for scenarii 1-4, replacing $\mathbb{P}\left(\left|\mathbf{D}_Z\mathbf{Z}^\top\right|_\infty > B(n,L)\right) \le \alpha_\infty(n)$ by $\mathbb{P}\left(\{\widehat{\rho}_I > \rho_I\} \cup \{\widehat{\rho}_{I^c} > \rho_{I^c}\}\right) \le \alpha_\infty(n)$, where ρ_I depends on $n, \overline{r}_0, L, I$ and ρ_{I^c} on n, L, I^c, and

$$\widehat{\rho}_I = \max_{l\in[L],\ k\in I}\left(\widehat{\mathbf{D}}_{\mathbf{Z}}\right)_{ll}\left(\widehat{\mathbf{D}}_{\mathbf{X}}\right)_{kk}\min\left(\overline{r}_0\max_{i\in[n]}|x_{ki}z_{li}|, \sqrt{\mathbb{E}_n\left[(X_kZ_l)^2\right]}\right)$$

and (S5.ii) by, for $M'_{ZU}(L) > 0$, for all $\left(\beta,\widetilde{\beta}\right), \mathbb{P}$ such that $\left(\beta,\widetilde{\beta}\right) \in \mathcal{I}dent$, where $\mathcal{P} = \overline{\mathcal{P}}_j$, $\mathbb{E}\left[\left|\left(\left(Z_lU(\beta) - \widetilde{\beta}_l\right)^2/\sigma^2_{Z_lU(\beta)-\widetilde{\beta}_l} - 1\right)_{l=1}^L\right|_\infty\right] \le M'_{ZU}(L)$. For simplicity, we still refer to this assumption as Assumption 6.1 and use $\alpha_C(n) = \alpha_\infty(n) + C_{\mathrm{N}}(L)\left(M'_{ZU}(L)/\tau^2 + M'_Z(L)/(\tau'_Z)^2\right)/n$ and $\alpha_D(n) = \alpha_B(n) + \alpha_C(n) + \left(C_{\mathrm{N}}(K)M_X(K)/\tau_X^2 + C_{\mathrm{N}}(LK)M(L,K)/r_\Psi^2\right)/n$.

The restricted sets, for $J \subseteq [K]$ and $\widetilde{J} \subseteq [L]$, are given in Table 17. Denote by $m(\tau_Z,\tau_X) =$

$\sqrt{\min\left(1/(1+\tau_Z),1-\tau_X\right)}$, $M(\tau_Z,\tau_X)=\sqrt{\max\left(1/(1-\tau_Z),1+\tau_X\right)}$, 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 5.1, $\widehat{\Psi}$, $\widehat{C}_J$, $\widehat{C}_J^{\gamma}$, by Ψ, C_J, and C_J^{γ}. Their lower bounds are computed on the sets of Table 17 and for the deterministic bounds we simply replace $\widehat{\rho}_I$ and $\widehat{\rho}_{I^c}$ by ρ_I and ρ_{I^c}. We define similarly $\theta_{\kappa}(s)$ and $\theta_{\gamma}(s)$. The sensitivities, their population counterparts, and lower bounds are now indexed by two sets or two sparsity certificates. Below, we refer to Assumption 6.1 for conciseness, it is indeed the suitable modification based on the elements that we have given.

Proposition A.1. *Under Assumption 6.1, for all* $\left(\beta,\widetilde{\beta}\right),\mathbb{P}$ *such that* $\left(\beta,\widetilde{\beta}\right)\in\mathcal{I}dent$*, on an event* $\mathcal{G}_{\Psi}$ *of probability* $1-\alpha_D(n)$*, we have, for all* $c>0$*,*

$$\begin{aligned}
&F\left(\beta,\widetilde{\beta}\right)\sqrt{1-\tau}\le\widehat{F}\left(\beta,\widetilde{\beta}\right)\le F\left(\beta,\widetilde{\beta}\right)\sqrt{1+\tau}\quad\textit{(see Table 17)};\\
&\forall\left(b,\widetilde{b}\right)\in\mathbb{R}^{K+L},\ l\in\mathcal{L},\ m\left(\tau_Z,\tau_X\right)l\left(\mathbf{D}_X^{-1}b,\mathbf{D}_Z\widetilde{b}\right)\le l\left(\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}b,\widehat{\mathbf{D}}_{\mathbf{Z}}\widetilde{b}\right)\le M(\tau_Z,\tau_X)l\left(\mathbf{D}_X^{-1}b,\mathbf{D}_Z\widetilde{b}\right);\\
&\forall J\subseteq[K],\ \forall\widetilde{J}\subseteq[L],\ l\in\mathcal{L},\ \widehat{\kappa}_{l,J,\widetilde{J}}\ge\frac{\kappa_{l,J,\widetilde{J}}}{\sqrt{1+\tau_Z}m\left(\tau_Z,\tau_X\right)}\left(1-\frac{r_{\Psi}}{\kappa_{1,[K],\emptyset,J,\widetilde{J}}}\right);\\
&\qquad\qquad\qquad\qquad\qquad\qquad\widehat{\gamma}_{l,J,\widetilde{J}}\ge\frac{\gamma_{l,J,\widetilde{J}}}{\sqrt{1+\tau_Z}m\left(\tau_Z,\tau_X\right)}\left(1-\frac{r_{\Psi}}{\gamma_{1,[K],\emptyset,J,\widetilde{J}}}\right).
\end{aligned}$$

The lower bounds in Proposition 5.1 involving the sparsity certificates hold if we remove the hats.

The main elements of the proofs are as follows. Take $\left(\beta,\widetilde{\beta}\right)\in\mathcal{I}dent$. Set $\Delta\triangleq\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\left(\widehat{\beta}-\beta\right)$ and $\widetilde{\Delta}\triangleq\widehat{\mathbf{D}}_{\mathbf{Z}}\left(\widehat{\widetilde{\beta}}-\widetilde{\beta}\right)$. Clearly, on $\mathcal{G}$, $\left(\beta,\widetilde{\beta}\right)$ belongs to $\widehat{\mathcal{I}}_C\left(\overline{r}_0,\widehat{F}\left(\beta,\widetilde{\beta}\right)\right)$. We now work on that event. Following the arguments in the proof of Theorem 4.1, we obtain

$$\begin{aligned}
&\left|\widehat{\Psi}\Delta+\widetilde{\Delta}\right|_{\infty}\le\overline{r}_0\left(\widehat{\sigma}+\widehat{F}\left(\beta,\widetilde{\beta}\right)\right)\\
&\left|\Delta_{J(\beta)^c\cap P}\right|_1+\left|\widetilde{\Delta}_{J\left(\widetilde{\beta}\right)^c}\right|_1\le\left|\Delta_{J(\beta)\cap P}\right|_1+\left|\widetilde{\Delta}_{J\left(\widetilde{\beta}\right)}\right|_1+c\left(\widehat{F}\left(\beta,\widetilde{\beta}\right)-\widehat{F}\left(\widehat{\beta},\widehat{\widetilde{\beta}}\right)\right).
\end{aligned}\tag{A.22}$$

Each function $\gamma\in\mathbb{R}^{K+L}\to\sqrt{\overline{Q}_l(\gamma)}$ is convex and

$$w_{l*}\triangleq-\begin{pmatrix}w_l\\ \widetilde{w}_l\end{pmatrix}\mathbb{1}\left\{\mathbb{E}_n\left[\left(Z_lU(\beta)-\widetilde{\beta}_l\right)^2\right]\ne0\right\}\in\partial\sqrt{\overline{Q}_l}\left(\beta,\widetilde{\beta}\right),$$

TABLE 17. Table of correspondence for the results on the *C-STIV*

STIV	*C-STIV*
$\overline{\sigma},\ r,\overline{r}$	$\left(\widehat{\sigma}+\widehat{F}\left(\widehat{\beta},\widehat{\widetilde{\beta}}\right)\right)/2,\ \overline{r}_0$
$\left\lvert\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_{J^c\cap P}\right\rvert_1$	$\left\lvert\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_{J^c\cap P}\right\rvert_1+\left\lvert\widehat{\mathbf{D}}_{\mathbf{Z}}\widetilde{\beta}_{\widetilde{J}^c}\right\rvert_1$
$\sqrt{1+\tau_X}\left\lvert\mathbf{D}_X^{-1}\beta_{J^c\cap P}\right\rvert_1$	$\sqrt{1+\tau_X}\left\lvert\mathbf{D}_X^{-1}\beta_{J^c\cap P}\right\rvert_1+\frac{1}{\sqrt{1-\tau_Z}}\left\lvert\mathbf{D}_Z\widetilde{\beta}_{\widetilde{J}^c}\right\rvert_1$
$\sqrt{1-\tau_X}\left\lvert\mathbf{D}_X^{-1}\left(\widehat{\beta}-\beta\right)_T\right\rvert_q$	$\sqrt{1-\tau_X}\left\lvert\mathbf{D}_X^{-1}\left(\widehat{\beta}-\beta\right)_T\right\rvert_q+\frac{1}{\sqrt{1+\tau_Z}}\left\lvert\mathbf{D}_Z\left(\widehat{\widetilde{\beta}}-\widetilde{\beta}\right)_{\widetilde{T}}\right\rvert_q$
$\widehat{C}_J$	$\widehat{C}_{J,\widetilde{J}}\triangleq\left\{\begin{array}{l}\left(\Delta,\widetilde{\Delta}\right):\ \left(\widehat{\mathbf{D}}_{\mathbf{X}}\Delta,\widehat{\mathbf{D}}_{\mathbf{Z}}^{-1}\widetilde{\Delta}\right)\in\mathcal{R}_D,\ \Delta_{J^c\cap J\left(\widehat{\beta}\right)^c}=\mathbf{0},\ \widetilde{\Delta}_{\widetilde{J}^c\cap J\left(\widehat{\widetilde{\beta}}\right)^c}=\mathbf{0},\ \lvert\Delta_{J^c\cap P}\rvert_1\\ +\left\lvert\widetilde{\Delta}_{\widetilde{J}^c}\right\rvert_1\le\lvert\Delta_{J\cap P}\rvert_1+\left\lvert\widetilde{\Delta}_{\widetilde{J}}\right\rvert_1+c\left(\widehat{\rho}_I\lvert\Delta_I\rvert_1+\widehat{\rho}_{I^c}\lvert\Delta_{I^c}\rvert_1+\overline{r}_0\left\lvert\widetilde{\Delta}\right\rvert_1\right)\end{array}\right\}$
$\widehat{C}_J^{\gamma}$	$\widehat{C}_{J,\widetilde{J}}^{\gamma}\triangleq\left\{\begin{array}{l}\left(\Delta,\widetilde{\Delta}\right):\ \left(\widehat{\mathbf{D}}_{\mathbf{X}}\Delta,\widehat{\mathbf{D}}_{\mathbf{Z}}^{-1}\widetilde{\Delta}\right)\in\mathcal{R}_D,\lvert\Delta_{J^c\cap P}\rvert_1+\left\lvert\widetilde{\Delta}_{\widetilde{J}^c}\right\rvert_1\\ \le 2\left(\lvert\Delta_{J\cap P}\rvert_1+\left\lvert\widetilde{\Delta}_{\widetilde{J}}\right\rvert_1+c\left(\widehat{\rho}_I\lvert\Delta_I\rvert_1+\widehat{\rho}_{I^c}\lvert\Delta_{I^c}\rvert_1+\overline{r}_0\left\lvert\widetilde{\Delta}\right\rvert_1\right)\right)+\lvert\Delta_{P^c}\rvert_1\end{array}\right\}$
$C_{J,}$	$C_{J,\widetilde{J}}\triangleq\left\{\begin{array}{l}\left(\Delta,\widetilde{\Delta}\right):\ \left(\mathbf{D}_X\Delta,\mathbf{D}_Z^{-1}\widetilde{\Delta}\right)\in\mathcal{R}_D,\left(\sqrt{\frac{1-\tau_X}{1+\tau_X}}-c\rho_I\right)\lvert\Delta_I\rvert_1+\left(\sqrt{\frac{1-\tau_X}{1+\tau_X}}-c\rho_{I^c}\right)\lvert\Delta_{I^c}\rvert_1\\ +(1-\overline{r}_0)\left\lvert\widetilde{\Delta}\right\rvert_1\le 2\lvert\Delta_{J\cap P}\rvert_1+\lvert\Delta_{P^c}\rvert_1+2\left\lvert\widetilde{\Delta}_{\widetilde{J}}\right\rvert_1\end{array}\right\}$
C_J^{γ}	$C_{J,\widetilde{J}}^{\gamma}\triangleq\left\{\begin{array}{l}\left(\Delta,\widetilde{\Delta}\right):\ \left(\mathbf{D}_X\Delta,\mathbf{D}_Z^{-1}\widetilde{\Delta}\right)\in\mathcal{R}_D,\ \left(\sqrt{\frac{1-\tau_X}{1+\tau_X}}-2c\rho_I\right)\lvert\Delta_I\rvert_1+\left(\sqrt{\frac{1-\tau_X}{1+\tau_X}}-2c\rho_{I^c}\right)\lvert\Delta_{I^c}\rvert_1\\ +(1-\overline{r}_0)\left\lvert\widetilde{\Delta}\right\rvert_1\le 3\lvert\Delta_{J\cap P}\rvert_1+2\lvert\Delta_{P^c}\rvert_1+3\left\lvert\widetilde{\Delta}_{\widetilde{J}}\right\rvert_1\end{array}\right\}$
$\widehat{\kappa}_{q,T,J}$	$\widehat{\kappa}_{q,T,\widetilde{T},J,\widetilde{J}}\triangleq\min_{\left(\Delta,\widetilde{\Delta}\right)\in\widehat{C}_{J,\widetilde{J}}:\ \lvert\Delta_T\rvert_q+\left\lvert\widetilde{\Delta}_{\widetilde{T}}\right\rvert_q=1}\left\lvert\widehat{\Psi}\Delta+\widetilde{\Delta}\right\rvert_\infty$
$\widehat{\kappa}_{\sigma,J}$	$\widehat{\kappa}_{\sigma,J,\widetilde{J}}\triangleq\min_{\left(\Delta,\widetilde{\Delta}\right)\in\widehat{C}_{J,\widetilde{J}}:\ \overline{r}_0^{-1}\left(\widehat{\rho}_I\lvert\Delta_I\rvert_1+\widehat{\rho}_{I^c}\lvert\Delta_{I^c}\rvert_1\right)+\left\lvert\widetilde{\Delta}\right\rvert_1=1}\left\lvert\widehat{\Psi}\Delta+\widetilde{\Delta}\right\rvert_\infty$
$\widehat{\gamma}_{Q,J}$	$\widehat{\gamma}_{Q,J,\widetilde{J}}\triangleq\min_{\substack{\left(\Delta,\widetilde{\Delta}\right)\in\widehat{C}_{J,\widetilde{J}}^{\gamma}\\ \min\left(\lvert\Delta_P\rvert_1+\left\lvert\widetilde{\Delta}\right\rvert_1,\frac{1}{2}\left(3\lvert\Delta_{J\cap P}\rvert_1+3\left\lvert\widetilde{\Delta}_{\widetilde{J}}\right\rvert_1+c\left(\widehat{\rho}_I\lvert\Delta_I\rvert_1+\widehat{\rho}_{I^c}\lvert\Delta_{I^c}\rvert_1+\overline{r}_0\left\lvert\widetilde{\Delta}\right\rvert_1\right)+\lvert\Delta_{P^c}\rvert_1\right)\right)=1}}\left\lvert\widehat{\Psi}\Delta+\widetilde{\Delta}\right\rvert_\infty$
$\widehat{B}\left(\widehat{J}\right)$	$\widehat{B}\left(\widehat{J},\widehat{\widetilde{J}}\right)\triangleq\left\{\begin{array}{l}\left(\widehat{\mathbf{D}}_{\mathbf{X}}\Delta,\widehat{\mathbf{D}}_{\mathbf{Z}}^{-1}\widetilde{\Delta}\right)\in\mathcal{R}_D,\ w\ge\mathbf{0},\ -w\le\Delta\le w,\ \widetilde{w}\ge\mathbf{0},\ -\widetilde{w}\le\widetilde{\Delta}\le\widetilde{w}\\ w_{\widehat{J}^c\cap J\left(\widehat{\beta}\right)^c}=\mathbf{0},\ w_{\widehat{\widetilde{J}}^c\cap J\left(\widehat{\widetilde{\beta}}\right)^c}=\mathbf{0},\\ (1-c\widehat{\rho}_I)\left(\sum_{j\in I}w_j\right)+(1-c\widehat{\rho}_{I^c})\left(\sum_{j\in I^c}w_j\right)+(1-c\overline{r}_0)\left(\sum_{l\in\widetilde{P}}\widetilde{w}_l\right)\\ \le 2\left(\sum_{j\in\widehat{J}\cap P}w_j+\sum_{l\in\widehat{\widetilde{J}}}\widetilde{w}_l\right)+\sum_{j\in P^c}w_j\end{array}\right\}$
$\widehat{B}(k)$	$\widehat{B}(k,l)\triangleq\left\{\begin{array}{l}\left(\widehat{\mathbf{D}}_{\mathbf{X}}\Delta,\widehat{\mathbf{D}}_{\mathbf{Z}}^{-1}\widetilde{\Delta}\right)\in\mathcal{R}_D,\ w\ge\mathbf{0},\ -w\le\Delta\le w,\ \widetilde{w}\ge\mathbf{0},\ -\widetilde{w}\le\widetilde{\Delta}\le\widetilde{w}\\ (1-c\widehat{\rho}_I)\left(\sum_{j\in I}w_j\right)+(1-c\widehat{\rho}_{I^c})\left(\sum_{j\in I^c}w_j\right)+(1-c\overline{r}_0)\left(\sum_{l\in\widetilde{P}}\widetilde{w}_l\right)\\ \le 2\left(sw_k+\widetilde{s}\widetilde{w}_l\right)+\sum_{j\in P^c}w_j\end{array}\right\}$
$\widehat{\theta}_\kappa\left(\widehat{J}\right),\widehat{\theta}_\kappa(s)$	$\widehat{\theta}_\kappa\left(\widehat{J},\widehat{\widetilde{J}}\right),\widehat{\theta}_\kappa(s,\widetilde{s})$
$\Theta_\kappa(J)$	$\Theta_\kappa\left(J,\widetilde{J}\right)\triangleq\mu\left(1-\frac{r_\Psi}{\kappa_{1,[K],\emptyset,J,\widetilde{J}}}-\frac{\overline{r}_0\mu}{c\kappa_{1,J\cap P,\widetilde{J},J,\widetilde{J}}}\right)_+^{-1}$
$\Theta_\gamma(J)$	$\Theta_\gamma\left(J,\widetilde{J}\right)\triangleq\mu\left(1-\frac{r_\Psi}{\gamma_{1,[K],\emptyset,J,\widetilde{J}}}-\frac{\overline{r}_0\mu}{c\gamma_{J,\widetilde{J}}^{Q}}\right)_+^{-1}$
$\sigma_{U(\beta)}$	$F\left(\beta,\widetilde{\beta}\right)\triangleq\frac{1}{\sqrt{1-\tau_Z}}\max_{l\in[L]}\left(\mathbf{D}_Z\right)_{ll}\sigma_{Z_lU(\beta)-\widetilde{\beta}_l}$
$\widehat{\beta}^{\omega}$	$\widehat{\beta}_k^{\omega}\triangleq\widehat{\beta}_k\mathbb{1}\left\{\left\lvert\widehat{\beta}_k\right\rvert>\frac{\widehat{\omega}_k(s,\widetilde{s})}{\sqrt{\mathbb{E}_n[X_k^2]}}\right\},\ \widehat{\widetilde{\beta}}_l^{\omega}\triangleq\widehat{\widetilde{\beta}}_l\mathbb{1}\left\{\left\lvert\widehat{\widetilde{\beta}}_l\right\rvert>\widehat{\omega}_l(s,\widetilde{s})\sqrt{\mathbb{E}_n[Z_l^2]}\right\}$
$\widehat{\omega}_k(s)$	$\widehat{\omega}_k(s,\widetilde{s})\triangleq\frac{2\overline{r}_0\overline{\sigma}\widehat{\theta}_\kappa(s,\widetilde{s})}{\widehat{\kappa}^*_{\binom{e_k}{\mathbf{0}}}(s,\widetilde{s})},\ \widehat{\widetilde{\omega}}_l(s,\widetilde{s})\triangleq\frac{2\overline{r}_0\overline{\sigma}\widehat{\theta}_\kappa(s,\widetilde{s})}{\widehat{\kappa}_{\binom{\mathbf{0}}{f_l}}(s,\widetilde{s})}$
$\theta_\kappa(s)$	$\theta_\kappa(s,\widetilde{s})\triangleq\mu\left(1-\frac{r_\Psi}{\kappa_{1,[K],\emptyset}(s,\widetilde{s})}-\frac{\overline{r}_0\mu}{\kappa^{\sigma}(s,\widetilde{s})}\right)_+^{-1}$
$\Theta_\kappa^{\sigma}(s)$	$\Theta_\kappa^{\sigma}(s,\widetilde{s})\triangleq\left(1-\frac{r_\Psi}{\kappa_{1,[K],\emptyset}(s,\widetilde{s})}+\frac{\widetilde{r}_0(s+\widetilde{s})\mu}{c\kappa_\infty(s,\widetilde{s})}\right)\left(1-\frac{r_\Psi}{\kappa_{1,[K],\emptyset}(s,\widetilde{s})}-\frac{\widetilde{r}_0(s+\widetilde{s})\mu}{c\kappa_\infty(s,\widetilde{s})}\right)_+^{-1}$
$\omega_k(s)$	$\omega_k(s,\widetilde{s})\triangleq\frac{\overline{r}_0F\left(\beta,\widetilde{\beta}\right)\mu(\Theta_\kappa^{\sigma}(s,\widetilde{s})+1)\theta_\kappa(s,\widetilde{s})}{\kappa^*_{\binom{e_k}{\mathbf{0}}}(s,\widetilde{s})},\ \widetilde{\omega}_l(s,\widetilde{s})\triangleq\frac{\overline{r}_0F\left(\beta,\widetilde{\beta}\right)\mu(\Theta_\kappa^{\sigma}(s,\widetilde{s})+1)\theta_\kappa(s,\widetilde{s})}{\kappa_{\binom{\mathbf{0}}{f_l}}(s,\widetilde{s})}$

$\mu\triangleq\sqrt{1+\tau_Z}m\left(\tau_Z,\tau_X\right).$

where

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

Hence, for $l \in [L]$, we have, because we work on $\mathcal{G}$, for $k \in I$,

$$\begin{aligned}\left(\widehat{\mathbf{D}}_{\mathbf{X}}\right)_{kk}|(w_l)_k| &\le \frac{\left|\mathbb{E}_n\left[X_kZ_l\left(Z_lU(\beta)-\widetilde{\beta}_l\right)\right]\right|}{\sqrt{\mathbb{E}_n\left[\left(X_kZ_l\left(Z_lU(\beta)-\widetilde{\beta}_l\right)\right)^2\right]}}\sqrt{\frac{\mathbb{E}_n\left[(X_kZ_l)^2\left(Z_lU(\beta)-\widetilde{\beta}_l\right)^2\right]}{\mathbb{E}_n\left[X_k^2\right]\mathbb{E}_n\left[Z_l^2\right]\mathbb{E}_n\left[\left(Z_lU(\beta)-\widetilde{\beta}_l\right)^2\right]}} \\ &\le \overline{r}_0 \max_{i\in[n]}\frac{|x_{ki}z_{li}|}{\sqrt{\mathbb{E}_n\left[X_k^2\right]\mathbb{E}_n\left[Z_l^2\right]}},\end{aligned}$$

but also, by the Cauchy-Schwarz inequality,

$$\left(\widehat{\mathbf{D}}_{\mathbf{X}}\right)_{kk}|(w_l)_k| \le \sqrt{\frac{\mathbb{E}_n\left[(X_kZ_l)^2\right]}{\mathbb{E}_n\left[X_k^2\right]\mathbb{E}_n\left[Z_l^2\right]}},$$

and, for all l' in $[L]$, $\left(\widehat{\mathbf{D}}_{\mathbf{Z}}\right)^{-1}_{l'l'}|(\widetilde{w}_l)_{l'}| \le \overline{r}_0$, also, by the Cauchy-Schwarz inequality, for all $k \in I^c$, $\left(\widehat{\mathbf{D}}_{\mathbf{X}}\right)_{kk}|(w_l)_k| \le \widehat{\rho}_{I^c}$. Taking $w_* = (w^\top, \widetilde{w}^\top)^\top$ as one of the w_{l*} for which $\sqrt{\overline{Q}_l\left(\beta,\widetilde{\beta}\right)} = \widehat{F}\left(\beta,\widetilde{\beta}\right)$ yields an element of $\partial\widehat{F}\left(\beta,\widetilde{\beta}\right)$ by Lemma A.1. By definition of $\partial\widehat{F}\left(\beta,\widetilde{\beta}\right)$, we have

$$\begin{aligned}\widehat{F}\left(\beta,\widetilde{\beta}\right)-\widehat{F}\left(\widehat{\beta},\widehat{\widetilde{\beta}}\right) &\le w_*^\top\begin{pmatrix}\beta-\widehat{\beta}\\ \widetilde{\beta}-\widehat{\widetilde{\beta}}\end{pmatrix} \\ &\le \left|\widehat{\mathbf{D}}_{\mathbf{X}}w_I\right|_\infty|\Delta_I|_1+\left|\widehat{\mathbf{D}}_{\mathbf{X}}w_{I^c}\right|_\infty|\Delta_{I^c}|_1+\left|\widehat{\mathbf{D}}_{\mathbf{Z}}^{-1}\widetilde{w}_I\right|_\infty\left|\widetilde{\Delta}\right|_1 \\ &\le \widehat{\rho}_I|\Delta_I|_1+\widehat{\rho}_{I^c}|\Delta_{I^c}|_1+\overline{r}_0\left|\widetilde{\Delta}\right|_1. \qquad \text{(A.23)}\end{aligned}$$

As a result, we have $\left(\Delta,\widetilde{\Delta}\right)\in\widehat{C}_{J(\beta),J(\widetilde{\beta})}$.

Using (A.22) and (A.23), we find

$$\left|\widehat{\Psi}\Delta+\widetilde{\Delta}\right|_\infty \le \overline{r}_0\left(2\overline{\sigma}+\widehat{\rho}_I|\Delta_I|_1+\widehat{\rho}_{I^c}|\Delta_{I^c}|_1+\overline{r}_0\left|\widetilde{\Delta}\right|_1\right). \qquad \text{(A.24)}$$

Using the definition of the sensitivities, we obtain

$$\left|\widehat{\Psi}\Delta+\widetilde{\Delta}\right|_\infty \le \overline{r}_0\left(2\overline{\sigma}+\overline{r}_0^2\frac{\left|\widehat{\Psi}\Delta\right|_\infty}{\widehat{\kappa}_{\sigma,J(\beta),J(\widetilde{\beta})}}\right)$$

$$\leq 2\overline{r}_0\overline{\sigma}\left(1-\frac{\overline{r}_0^2}{\widehat{\kappa}_{\sigma,J(\beta),J\left(\widetilde{\beta}\right)}}\right)_+^{-1};$$

$$\begin{aligned} c\widehat{\sigma} &\leq |\Delta_{J(\beta)\cap P}|_1+\left|\widetilde{\Delta}_{J\left(\widetilde{\beta}\right)}\right|_1+c\widehat{F}\left(\beta,\widetilde{\beta}\right)\\ &\leq \frac{\left|\widehat{\Psi}\Delta+\widetilde{\Delta}\right|_\infty}{\widehat{\kappa}_{1,J(\beta)\cap P,J\left(\widetilde{\beta}\right),J(\beta),J\left(\widetilde{\beta}\right)}}+c\widehat{F}\left(\beta,\widetilde{\beta}\right). \end{aligned}$$

For a nonsparse vectors, $J\subseteq[K]$, and $\widetilde{J}\subseteq[L]$, we obtain

$$|\Delta_{J^c\cap P}|_1+\left|\widetilde{\Delta}_{\widetilde{J}^c}\right|_1 \leq|\Delta_{J\cap P}|_1+\left|\widetilde{\Delta}_{\widetilde{J}}\right|_1+c\left(\widehat{\rho}_I\left|\Delta_I\right|_1+\widehat{\rho}_{I^c}\left|\Delta_{I^c}\right|_1+\overline{r}_0\left|\widetilde{\Delta}\right|_1\right)+2\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_{J^c\cap P}\right|_1+2\left|\widehat{\mathbf{D}}_{\mathbf{Z}}\widetilde{\beta}_{\widetilde{J}^c}\right|_1.$$

Let us show the first inequality and consider two cases.
First, if $2\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_{J^c\cap P}\right|_1+2\left|\widehat{\mathbf{D}}_{\mathbf{Z}}\widetilde{\beta}_{\widetilde{J}^c}\right|_1 \leq |\Delta_{J\cap P}|_1+\left|\widetilde{\Delta}_{\widetilde{J}}\right|_1+c\left(\widehat{\rho}_I\left|\Delta_I\right|_1+\widehat{\rho}_{I^c}\left|\Delta_{I^c}\right|_1+\overline{r}_0\left|\widetilde{\Delta}\right|_1\right)+|\Delta_{P^c}|_1$, then Δ belongs to $\widehat{C}^{\gamma}_{J,\widetilde{J}}$.
Also, we have

$$\begin{aligned} \widehat{\sigma} &\leq \frac{1}{c}\left(\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_P\right|_1-\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\widehat{\beta}_P\right|_1+\left|\widehat{\mathbf{D}}_{\mathbf{Z}}\widetilde{\beta}_{\widetilde{P}}\right|_1-\left|\widehat{\mathbf{D}}_{\mathbf{Z}}\widehat{\theta}_{\widetilde{P}}\right|_1\right)+\widehat{F}\left(\beta,\widetilde{\beta}\right)\\ &\leq \frac{\left|\widehat{\Psi}\Delta+\widetilde{\Delta}\right|_\infty}{c\widehat{\gamma}_{Q,J,\widetilde{J}}}+\widehat{F}\left(\beta,\widetilde{\beta}\right). \end{aligned}$$

Second, if $2\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta_{J^c\cap P}\right|_1+2\left|\widehat{\mathbf{D}}_{\mathbf{Z}}\widetilde{\beta}_{\widetilde{J}^c}\right|_1 > |\Delta_{J\cap P}|_1+\left|\widetilde{\Delta}_{\widetilde{J}}\right|_1+c\left(\widehat{\rho}_I\left|\Delta_I\right|_1+\widehat{\rho}_{I^c}\left|\Delta_{I^c}\right|_1+\overline{r}_0\left|\widetilde{\Delta}\right|_1\right)+|\Delta_{P^c}|_1$, then we have

$$|\Delta|_1+\left|\widetilde{\Delta}\right|_1=|\Delta_{J^c\cap P}|_1+|\Delta_{J\cap P}|_1+|\Delta_{P^c}|_1+\left|\widetilde{\Delta}_{\widetilde{J}^c}\right|_1+\left|\widetilde{\Delta}_J\right|_1\leq 6\left(\left|\left(\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\beta\right)_{J^c\cap P}\right|_1+\left|\left(\widehat{\mathbf{D}}_{\mathbf{Z}}\widetilde{\beta}\right)_{\widetilde{J}^c}\right|_1\right).$$

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

$$\begin{aligned} \left|\widehat{\Psi}\Delta\right|_\infty &\geq \min_{l\in[L]}\left(\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{D}_Z^{-1}\right)_{ll}\left|\mathbf{D}_Z\mathbb{E}_n\left[ZX^\top\right]\mathbf{D}_X\overline{\Delta}+\overline{\widetilde{\Delta}}\right|_\infty\\ &\geq \frac{1}{\sqrt{1+\tau_Z}}\left(\left|\mathbf{D}_Z\mathbb{E}\left[ZX^\top\right]\mathbf{D}_X\overline{\Delta}+\overline{\widetilde{\Delta}}\right|_\infty-\left|\mathbf{D}_Z(\mathbb{E}_n-\mathbb{E})\left[ZX^\top\right]\mathbf{D}_X\overline{\Delta}\right|_\infty\right)\\ &\geq \frac{1}{\sqrt{1+\tau_Z}}\left(\left|\Psi\overline{\Delta}+\widetilde{\widetilde{\Delta}}\right|_\infty-r_Z\left|\overline{\Delta}\right|_1\right). \end{aligned}$$

The rest is easy. □

Remark A.1. *When we only use the instruments which are known to be exogenous to obtain the* C-STIV*, it simplifies to: for $c > 0$, $\left(\widehat{\beta}, \widehat{\sigma}\right)$ is any solution of*

$$\min_{\beta \in \widehat{\mathcal{I}}_C(\overline{r}_0, \sigma), \sigma \geq 0} \left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1} \beta_P\right|_1 + c\sigma,$$

where

$$\widehat{\mathcal{I}}_C(\overline{r}_0, \sigma) \triangleq \left\{\beta \in \mathcal{R} : \ \left|\widehat{\mathbf{D}}_{\mathbf{Z}} \left(\frac{1}{n}\mathbf{Z}^\top \mathbf{U}(\beta)\right)_{\widetilde{P}^c}\right|_\infty \leq \overline{r}_0 \sigma, \ \widehat{F}(\beta) \leq \sigma\right\};$$

$$\forall \beta \in \mathbb{R}^K, \ \widehat{F}(\beta) \triangleq \max_{l \in \widetilde{P}^c} \sqrt{\overline{Q}_l(\beta)}.$$

The restricted set becomes, for $J \subseteq [K]$,

$$\widehat{C}_J = \left\{\widehat{\mathbf{D}}_{\mathbf{X}} \Delta \in \mathcal{R}_D : \ \Delta_{J^c \cap J\left(\widetilde{\beta}\right)^c} = \mathbf{0}, \ |\Delta_{J^c \cap P}|_1 \leq |\Delta_{J \cap P}|_1 + c\left(\widehat{\rho}_I |\Delta_I|_1 + \widehat{\rho}_{I^c} |\Delta_{I^c}|_1\right)\right\}$$

$$\text{where} \quad \widehat{\rho}_I = \max_{l \in \widetilde{P}^c, \ k \in I} \left(\widehat{\mathbf{D}}_{\mathbf{Z}}\right)_{ll} \left(\widehat{\mathbf{D}}_{\mathbf{X}}\right)_{kk} \min\left(\overline{r}_0 \max_{i \in [n]} |x_{ki} z_{li}|, \sqrt{\mathbb{E}_n\left[(X_k Z_l)^2\right]}\right)$$

$$\widehat{\rho}_{I^c} = \max_{l \in \widetilde{P}^c, \ k \in I^c} \left(\widehat{\mathbf{D}}_{\mathbf{Z}}\right)_{ll} \left(\widehat{\mathbf{D}}_{\mathbf{X}}\right)_{kk} \sqrt{\mathbb{E}_n\left[(X_k Z_l)^2\right]}.$$

Proof of Theorem 7.3. We work on the event on the event $\mathcal{G}_1 \cap \mathcal{G}_2$ in case (1) or $\mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}_\Psi$ in cases (2) or (3). First, we show that $\left(\widetilde{\beta}, F\left(\beta, \widetilde{\beta}\right)\right) \in \widehat{\mathcal{I}}_{NV}$ by the following computations

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

The second constraint is satisfied because, by the triangle inequality and convexity, $\widehat{F}_2\left(\widehat{\beta}, \widetilde{\beta}\right) \leq \widehat{F}_2\left(\beta, \widetilde{\beta}\right) + \widehat{b}^\sigma$. Now, because $\left(\widetilde{\beta}, \widehat{F}_2\left(\beta, \widetilde{\beta}\right)\right) \in \widehat{\mathcal{I}}_{NV}$ and $\left(\widehat{\widetilde{\beta}}, \widehat{\widetilde{\sigma}}\right)$ minimizes (7.8), we have

$$\left|\widetilde{\Delta}_{J(\widetilde{\beta})^c}\right|_1 \leq \left|\widetilde{\Delta}_{J(\widetilde{\beta})}\right|_1 + \widetilde{c}\left(\widehat{F}_2\left(\beta, \widetilde{\beta}\right) - \widehat{\widetilde{\sigma}}\right). \tag{A.25}$$

Using that $\widehat{F}_2\left(\widehat{\beta}, \widetilde{\beta}\right) \leq \widehat{\widetilde{\sigma}} + \widehat{b}^\sigma$ (by definition of the estimator) and the computations from the proofs of the results of Section 7.3, we obtain

$$\widehat{F}_2\left(\beta, \widetilde{\beta}\right) - \widehat{\widetilde{\sigma}} \leq r_2 \left|\widetilde{\Delta}\right|_1 + 2\widehat{b}^\sigma. \tag{A.26}$$

This and (A.25) yield

$$\left|\widetilde{\Delta}_{J(\widetilde{\beta})^c}\right|_1 \leq \left|\widetilde{\Delta}_{J(\widetilde{\beta})}\right|_1 + \widetilde{c} r_2 \left|\widetilde{\Delta}\right|_1 + 2\widetilde{c}\widehat{b}^\sigma$$

and, equivalently,

$$\left|\widetilde{\Delta}_{J(\widetilde{\beta})^c}\right|_1 \le \frac{1+\widetilde{c}r_2}{1-\widetilde{c}r_2}\left|\widetilde{\Delta}_{J(\widetilde{\beta})}\right|_1 + \frac{2\widetilde{c}}{1-\widetilde{c}r_2}\widehat{b}^{\sigma}. \tag{A.27}$$

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

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

This and (A.26) yield

$$\left|\widetilde{\Delta}\right|_{\infty} \le r_2\left(2\widehat{\widetilde{\sigma}}+r_2\left|\widetilde{\Delta}\right|_1+2\widehat{b}^{\sigma}\right)+2\widehat{b}. \tag{A.28}$$

On the other hand, (A.27) implies

$$\begin{aligned}\left|\widetilde{\Delta}\right|_1 &\le \frac{2}{1-\widetilde{c}r_2}\left|\widetilde{\Delta}_{J(\widetilde{\beta})}\right|_1 + \frac{2\widetilde{c}\widehat{b}^{\sigma}}{1-\widetilde{c}r_2} \\ &\le \frac{2\left|J\left(\widetilde{\beta}\right)\right|}{1-\widetilde{c}r_2}\left|\widetilde{\Delta}\right|_{\infty} + \frac{2\widetilde{c}\widehat{b}^{\sigma}}{1-\widetilde{c}r_2}.\end{aligned} \tag{A.29}$$

Inequalities (7.9) and (7.10) follow by simple manipulations of (A.28) and (A.29).

As before, we obtain

$$\widehat{\widetilde{\sigma}} \le \frac{\left|\widetilde{\Delta}_{J(\widetilde{\beta})}\right|_1}{\widetilde{c}} + \sqrt{1+\tau}\widehat{F}_2\left(\beta,\widetilde{\beta}\right) \le \frac{\left|J\left(\widetilde{\beta}\right)\right|\left|\widetilde{\Delta}\right|_{\infty}}{\widetilde{c}} + \sqrt{1+\tau}\widehat{F}_2\left(\beta,\widetilde{\beta}\right), \tag{A.30}$$

which, together with (A.28) and (A.29), yield

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

The rest is easy. □

Proof of Theorem 8.1. Take $(\beta,\Lambda)\in\mathcal{B}_{s,s',s_r}$. Set $\Delta'\triangleq\left(\widehat{\Lambda}-\Lambda\right)\widehat{\mathbf{D}}_{\mathbf{Z}}^{-1}$, and $\overline{\Delta}'\triangleq\Delta'\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{D}_Z^{-1}$. Clearly, on $\mathcal{G}_0'$, Λ belongs to $\widehat{\mathcal{A}}\left(r_0',\widehat{F}\left(\Lambda\right)\right)$. We now work on the event $\mathcal{G}_0'\cap\mathcal{G}_{\Psi}\cap\mathcal{E}_{Z'}^c\cap\mathcal{E}_T^c$.

We start by proving the inequalities from item (i). The arguments in the proof of Theorem 4.1 yield

$$\left|\Delta'\widehat{\Psi}^{\top}\right|_{\infty} \le r_0'\left(\widehat{\nu}+\widehat{F}\left(\Lambda\right)\right) \tag{A.31}$$

$$\left|\Delta'_{J(\Lambda)^c}\right|_1 \le \left|\Delta'_{J(\Lambda)}\right|_1 + \frac{\lambda}{\widehat{\rho}}\left(\widehat{F}\left(\Lambda\right)-\widehat{F}\left(\widehat{\Lambda}\right)\right)$$

and, by those of the proof results of Section 7.3, $\widehat{F}(\Lambda) - \widehat{F}\left(\widehat{\Lambda}\right) \le \widehat{\rho}|\Delta'|_1$.

As a result, we have $\Delta' \in C'_{J(\Lambda)} \subseteq \widehat{C}'_{J(\Lambda)}$ and, using the definition of $\widehat{\kappa}'_{1,J(\Lambda),J(\Lambda)}$ and of the objective function in (8.2) in the first display and (A.31) in the third display,

$$\widehat{\nu} \le \frac{\widehat{\rho}\left|\Delta'\widehat{\Psi}^\top\right|_\infty}{\lambda\widehat{\kappa}'_{1,J(\Lambda),J(\Lambda)}} + \widehat{F}(\Lambda);$$

$$\widehat{\nu} + \widehat{F}(\Lambda) \le 2\widehat{F}(\Lambda)\left(1 - \frac{r'_0\widehat{\rho}}{\lambda\widehat{\kappa}'_{1,J(\Lambda),J(\Lambda)}}\right)_+^{-1};$$

$$\left|\Delta'\widehat{\Psi}^\top\right|_\infty \le 2r'_0\widehat{F}(\Lambda)\left(1 - \frac{r'_0\widehat{\rho}}{\lambda\widehat{\kappa}'_{1,J(\Lambda),J(\Lambda)}}\right)_+^{-1}.$$

To conclude, we use the fact that we work on $\mathcal{G}'_0 \cap \mathcal{G}_\Psi \cap \mathcal{E}^c_{Z'} \cap \mathcal{E}^c_T$.

Let us now show the results of item (ii). Take $J \subseteq [O] \times [L]$. We have

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

Let us now distinguish between two cases.

Case 1: $2\left|\Lambda_{J^c}\widehat{\mathbf{D}}_{\mathbf{Z}}^{-1}\right|_1 \le |\Delta'_J|_1$. In that case we have $\overline{\Delta} \in C'_{\gamma,J}$. From this, for each $o \in [O]$, we have

$$\begin{aligned}\sqrt{1-\tau_Z}\left|\left(\widehat{\Lambda} - \Lambda\right)_{\{o\}\times[L]}\mathbf{D}_Z^{-1}\right|_1 &\le 2\frac{r'_0F(\Lambda)\sqrt{1+\tau_T}}{\gamma'_{1,\{o\}\times[L],J}\sqrt{1-\tau_X}}\Theta'_\gamma(J)\\ &\le 2\frac{r'_0F(\Lambda)\sqrt{1+\tau_T}}{\min_{o\in[O]}\gamma'_{1,\{o\}\times[L],J}\sqrt{1-\tau_X}}\Theta'_\gamma(J).\end{aligned}$$

Case 2: $2\left|\Lambda_{J^c}\widehat{\mathbf{D}}_{\mathbf{Z}}^{-1}\right|_1 > |\Delta'_J|_1$. In that case, we have

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

hence

$$|\overline{\Delta}'|_1 \le 2\frac{3+\lambda}{1-\lambda}\sqrt{\frac{1+\tau_Z}{1-\tau_Z}}\left|\Lambda_{J^c}\mathbf{D}_Z^{-1}\right|_1.$$

This allows to conclude. □

Proof of Theorem 8.2. We make the proof in the case $G = 1$. Extension to $G > 1$ is easy.

Take $(\beta, \Lambda) \in \mathcal{B}_{s,s',s_r}$. Set $\Delta \triangleq \widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\left(\widehat{\beta} - \beta\right)$, $\overline{\Delta} \triangleq \mathbf{D}_X^{-1}\left(\widehat{\beta} - \beta\right)$, $\Delta' \triangleq \left(\widehat{\Lambda} - \Lambda\right)\widehat{\mathbf{D}}_{\mathbf{Z}}^{-1}$, $\overline{\Delta}' \triangleq \Delta'\widehat{\mathbf{D}}_{\mathbf{Z}}\mathbf{D}_Z^{-1}$, and $\mathcal{E}_{\Lambda Z} \triangleq \left\{\left|\mathbf{D}_{\Lambda Z}\Lambda(\mathbb{E}_n - \mathbb{E})\left[ZZ^\top\right]\Lambda^\top\mathbf{D}_{\Lambda Z}\right|_\infty \ge \tau_Z\right\}$. Due to Assumption 8.1, we have

$$\mathbb{P}(\mathcal{E}_{\Lambda Z}) = \mathbb{P}\left(\left|\sum_{i=1}^n\left(\mathbf{D}_{\Lambda Z}\Lambda z_iz_i^\top\Lambda^\top\mathbf{D}_{\Lambda Z} - \mathbf{D}_{\Lambda Z}\mathbb{E}\left[\Lambda ZZ^\top\Lambda^\top\right]\mathbf{D}_{\Lambda Z}\right)\right|_\infty \ge n\tau_{\Lambda Z}\right)$$

$$\leq \frac{C_{\mathrm{N}}(OL)M_{\Lambda Z}(O)}{n\tau_{\Lambda Z}^2}.$$

We use the decomposition

$$\sqrt{n}\left(\widehat{\Omega\beta}-\Omega\beta\right)=R+\frac{1}{\sqrt{n}}\widehat{\Lambda}\mathbf{Z}^\top\mathbf{U},$$

where

$$R \triangleq \sqrt{n}\left(\Omega-\frac{1}{n}\widehat{\Lambda}\mathbf{Z}^\top\mathbf{X}\right)\widehat{\mathbf{D}}_{\mathbf{X}}\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\mathbf{D}_X\overline{\Delta}.$$

We have

$$\begin{aligned}|R|_\infty &\leq \sqrt{n}\left|\left(\Omega-\frac{1}{n}\widehat{\Lambda}\mathbf{Z}^\top\mathbf{X}\right)\widehat{\mathbf{D}}_{\mathbf{X}}\right|_\infty\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\mathbf{D}_X\overline{\Delta}\right|_1\\ &\leq \sqrt{n}r_0'\widehat{\nu}\widehat{F}(\Lambda)\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\mathbf{D}_X\overline{\Delta}\right|_1 \quad \text{(by definition of the estimator).}\end{aligned}$$

So, on the event of Theorem 8.1 intersected with $\mathcal{E}_X^c\cap\mathcal{G}$, we have

$$|R|_\infty\leq\sqrt{n}r_0'v_\beta(n)v_\nu(n)\sqrt{1+\tau_X}.$$

Define

$$T_\Lambda=\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\widehat{\mathbf{D}}_{\widehat{\Lambda}\mathbf{Z}}\widehat{\Lambda}z_i\frac{u_i(\beta)}{\sqrt{\widehat{Q}\left(\widehat{\beta}\right)}}\right|_\infty,\ T_{\Lambda 2}=\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\mathbf{D}_{\Lambda Z}\widehat{\Lambda}z_i\frac{u_i(\beta)}{\sqrt{\widehat{Q}\left(\widehat{\beta}\right)}}\right|_\infty,\ T_{\Lambda 1}=\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\mathbf{D}_{\Lambda Z}\widehat{\Lambda}z_i\frac{u_i(\beta)}{\sigma_{U(\beta)}}\right|_\infty,$$

$$T_{\Lambda 0}=\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\mathbf{D}_{\Lambda Z}\Lambda z_i\frac{u_i(\beta)}{\sigma_{U(\beta)}}\right|_\infty,\ N_{\Lambda 0}=\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\chi_i\right|_\infty;$$

$$W_\Lambda=\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\widehat{\mathbf{D}}_{\widehat{\Lambda}\mathbf{Z}}\widehat{\Lambda}z_ie_i\right|_\infty,\ W_{\Lambda 1}=\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\mathbf{D}_{\Lambda Z}\widehat{\Lambda}z_ie_i\right|_\infty,\ W_{\Lambda 0}=\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n\mathbf{D}_{\Lambda Z}\Lambda z_ie_i\right|_\infty;$$

where χ_i are independent Gaussian vectors of covariance $\mathbb{E}[\mathbf{D}_{\Lambda Z}\Lambda z_iz_i^\top\Lambda^\top\mathbf{D}_{\Lambda Z}]$. $N_{\Lambda 0}$ is a Gaussian approximation of $T_{\Lambda 0}$.

By (DGP.1), (DGP.4), and Proposition 3.2 in Chernozhukov, Chetverikov, and Kato (2017), we get

$$\max\left(\left|\mathbb{P}\left(T_{\Lambda 0}\leq t\right)-\mathbb{P}\left(N_{\Lambda 0}\leq t\right)\right|,\left|\mathbb{P}\left(W_{\Lambda 0}\leq t\right)-\mathbb{P}\left(N_{\Lambda 0}\leq t\right)\right|\right)\leq\rho. \tag{A.32}$$

where the constant C_2 in the definition of ρ from Scenario 5 can also depend on s_r. We denote by $q_{N_{\Lambda 0}}$ the quantile function of $N_{\Lambda 0}$ and by $q_{W_{\Lambda 0}}$ the conditional quantile function of $W_{\Lambda 0}$ given $\boldsymbol{Z}$. Lemma 3.1 in Chernozhukov, Chetverikov, and Kato (2013) yields, for all $t\in\mathbb{R}$,

$$\left|\mathbb{P}\left(W_{\Lambda 0}\leq t\right|\mathbf{Z})-\mathbb{P}\left(N_{\Lambda 0}\leq t\right)\right|\leq\varphi_\Lambda(\tau_Z)\quad\text{on }\mathcal{E}_{\Lambda Z}^c.$$

The same analysis as for Scenario 5 yields

$$|\mathbb{P}\left(T_{\Lambda 0} \leq q_{W_{\Lambda 0}}(\alpha)\right) - \alpha| \leq 2\varphi_{\Lambda}(\tau_Z)) + 2\frac{C_{\mathrm{N}}(OL)M_{\Lambda Z}(O)}{n\tau_{\Lambda Z}^2} + 2\rho. \tag{A.33}$$

We now have to prove that, for some sequences ζ_1, $\zeta_2(\zeta_1, \tau_Z')$, and $\zeta_2'(\zeta_1)$ converging to zero,

$$\mathbb{P}\left(\mathbb{P}\left(|W_{\Lambda} - W_{\Lambda 0}| > \zeta_1 | \mathbf{Z}\right) > \zeta_2(\zeta_1, \tau_Z')\right) < \zeta_2(\zeta_1, \tau_Z')$$
$$\mathbb{P}\left(|T_{\Lambda} - T_{\Lambda 0}| > \zeta_1\right) \leq \zeta_2'(\zeta_1).$$

Let us consider the second bound. The first bound can be treated in the same way and the arguments in the analysis of the multiplier bootstrap for Scenario 5.

By (DGP.3), on an event of probability at least $1 - \alpha_{\Lambda}(n) - C_{\mathrm{N}}(O)M_{\Lambda,2}(O)/(n\left(\tau_{\Lambda Z}\right)^2)$, for all positive $\tau_{\Lambda Z}$, we have, for all $o \in [O]$,

$$\begin{aligned}\left|\sqrt{\mathbb{E}_n\left[(\widehat{\Lambda}_{o\cdot}Z)^2\right]} - \sqrt{\mathbb{E}\left[(\Lambda_{o\cdot}Z)^2\right]}\right| &\leq \left|\sqrt{\mathbb{E}_n\left[(\widehat{\Lambda}_{o\cdot}Z)^2\right]} - \sqrt{\mathbb{E}_n\left[(\Lambda_{o\cdot}Z)^2\right]}\right| + \left|\sqrt{\mathbb{E}_n\left[(\Lambda_{o\cdot}Z)^2\right]} - \sqrt{\mathbb{E}\left[(\Lambda_{o\cdot}Z)^2\right]}\right| \\ &\leq \sqrt{\mathbb{E}_n\left[\left(\left(\widehat{\Lambda}_{o\cdot} - \Lambda_{o\cdot}\right)Z\right)^2\right]} + \left(\sqrt{1+\tau_{\Lambda Z}} - 1\right) \triangleq v_{D,1}(n), \\ &\leq v_{\Lambda,2}(n) + \left(\sqrt{1+\tau_{\Lambda Z}} - 1\right) \triangleq v_{D,1}(n),\end{aligned} \tag{A.34}$$

hence

$$\left|\widehat{\mathbf{D}}_{\widehat{\Lambda}\mathbf{Z}} - \mathbf{D}_{\Lambda\mathbf{Z}}\right|_{\infty} \leq v_{D,1}(n)\max_{o\in[O]}\frac{(\mathbf{D}_{\Lambda Z})_{oo}}{(\mathbf{D}_{\Lambda Z})_{oo}^{-1} - v_{D,1}(n)} \triangleq v_{D,2}(n)$$

and

$$|T_{\Lambda} - T_{\Lambda 2}| \leq T_{\Lambda 2}v_{D,2}(n).$$

On the same event, because $|T_{\Lambda 1} - T_{\Lambda 0}| \leq |\mathbf{D}_{\Lambda Z}|_{\infty}\left|\overline{\Delta}'\right|_{\infty,\infty}T_0$, we get

$$|T_{\Lambda 1} - T_{\Lambda 0}| \leq T_0\left|\mathbf{D}_{\Lambda Z}\right|_{\infty}v_{\Lambda}(n).$$

On an event of probability at least $1 - \alpha_{\beta}(n)$, we have $\left|\sqrt{\widehat{Q}\left(\widehat{\beta}\right)} - \sigma_{U(\beta)}\right| \leq v_{\sigma}(n)$, thus

$$\left|\frac{1}{\sqrt{\widehat{Q}\left(\widehat{\beta}\right)}} - \frac{1}{\sigma_{U(\beta)}}\right| \leq \frac{v_{\sigma}(n)}{\sigma_{U(\beta)}(\sigma_{U(\beta)} - v_{\sigma}(n))},$$

hence $|T_{\Lambda 2} - T_{\Lambda 1}| \leq T_{\Lambda 1}v_{\sigma}(n)/(\sigma_{U(\beta)} - v_{\sigma}(n))$.

Thus, on an event of probability at least $1 - \alpha_{\beta}(n) - \alpha_{\Lambda}(n) - C_{\mathrm{N}}(O)M_{\Lambda,2}(O)/(n\left(\tau_{\Lambda Z}'\right)^2)$, we have

$$|T_{\Lambda} - T_{\Lambda 0}| \leq T_{\Lambda 1}\left(\left(1 + \frac{v_{\sigma}(n)}{\sigma_{U(\beta)} - v_{\sigma}(n)}\right)v_{D,2}(n) + \frac{v_{\sigma}(n)}{\sigma_{U(\beta)} - v_{\sigma}(n)}\right) + T_0\left|\mathbf{D}_{\Lambda Z}\right|_{\infty}v_{\Lambda}(n)$$

$$
\begin{aligned}
\leq & T_{\Lambda 0}\left(\left(1+\frac{v_{\sigma}(n)}{\sigma_{U(\beta)}-v_{\sigma}(n)}\right) v_{D,2}(n)+\frac{v_{\sigma}(n)}{\sigma_{U(\beta)}-v_{\sigma}(n)}\right) \\
& + T_0 \left|\mathbf{D}_{\Lambda Z}\right|_{\infty} v_{\Lambda}(n)\left(1+\left(1+\frac{v_{\sigma}(n)}{\sigma_{U(\beta)}-v_{\sigma}(n)}\right) v_{D,2}(n)+\frac{v_{\sigma}(n)}{\sigma_{U(\beta)}-v_{\sigma}(n)}\right).
\end{aligned}
$$

The remaining argument uses the pigeonhole principle and a union bound.

Let us now explain how the model with approximation errors can be dealt with.

We base our analysis on the decomposition

$$
\sqrt{n}\left(\widehat{\Omega\beta}-\Omega\beta-V(\beta)\right)=R+\frac{1}{\sqrt{n}}\widehat{\Lambda}\mathbf{Z}^{\top}\mathbf{V}(\beta)-\sqrt{n}V(\beta)+\frac{1}{\sqrt{n}}\widehat{\Lambda}\mathbf{Z}^{\top}\mathbf{W}(\beta),
$$

where R and the last stochastic term are as before. We have, on the event of Theorem 8.1 intersected with an event of probability at least $1-C_{\mathrm{N}}(O)M_{\Lambda,2}(O)/(n\tau_{\Lambda}^2)-C_{\mathrm{N}}(OL)M_{\Lambda Z}(O)/(n\tau_{\Lambda Z}^2)$

$$
\begin{aligned}
\left|\frac{1}{\sqrt{n}}\widehat{\Lambda}\mathbf{Z}^{\top}\mathbf{V}(\beta)\right|_{\infty} & \leq \left|\widehat{\Lambda}\mathbf{Z}^{\top}\right|_{2,\infty}\rho_E \\
& \leq \left|\widehat{\Lambda}\mathbf{Z}^{\top}\right|_{2,\infty}\frac{v_v(n)}{\sqrt{n}} \\
& \leq v_v(n)\left(\frac{1}{\sqrt{n}}\left|\Lambda\mathbf{Z}^{\top}\right|_{2,\infty}+v_{\Lambda,2}(n)\right) \\
& \leq v_v(n)\left(\left(\sqrt{1+\tau_{\Lambda Z}}-1\right)+\sqrt{1+\tau_{\Lambda}}\left|\left(\sigma_{\Lambda_o\cdot Z}\right)_{o\in[O]}\right|_{\infty}+v_{\Lambda,2}(n)\right)
\end{aligned}
$$

and $|\sqrt{n}V(\beta)|_{\infty}\leq v_v(n)$. □

Proof of Theorem 8.3. We make the proof in the case $G=1$. Extension to $G>1$ is easy. Denote by $\overline{\Delta}_{+}\triangleq\mathbf{D}_X^{-1}\left(\widehat{\beta}_{+}-\beta\right)$ and work on the event $\mathcal{G}_{+}\cap\mathcal{G}_{\Psi+}\cap\mathcal{E}_{X-}^c\cap\mathcal{E}_{Z'-}^c\cap\mathcal{E}_{ZX^{\top}-}^c$ where each are defined like before. We add the indices $+$ and $-$ to the events to make precise the sample we are referring to. The event $\mathcal{E}_{X-}^c\cap\mathcal{E}_{Z'-}^c\cap\mathcal{E}_{ZX^{\top}-}^c$ has probability at least

$$
1-\frac{C_{\mathrm{N}}(K)M_X(K)}{n_{-}\tau_{X-}^2}-\frac{C_{\mathrm{N}}(L)M_Z'(L)}{n_{-}(\tau_{Z-}')^2}-\frac{C_{\mathrm{N}}(KL)M(L,K)}{n_{-}r_{\Psi-}^2}.
$$

We have

$$
\sqrt{n_{+}}\left(\widehat{\Omega\beta}-\Omega\beta\right)=R+\frac{1}{\sqrt{n_{+}}}\widehat{\Lambda}\mathbf{Z}_{+}^{\top}\mathbf{U}_{+},
$$

where

$$
R\triangleq\sqrt{n_{+}}\left(\left(\Omega-\frac{1}{n_{-}}\widehat{\Lambda}\mathbf{Z}_{-}^{\top}\mathbf{X}_{-}\right)+\left(\frac{1}{n_{-}}\widehat{\Lambda}\mathbf{Z}_{-}^{\top}\mathbf{X}_{-}-\frac{1}{n_{+}}\widehat{\Lambda}\mathbf{Z}_{+}^{\top}\mathbf{X}_{+}\right)\right)\mathbf{D}_X\overline{\Delta}_{+}.
$$

We have

$$
|R|_{\infty}\leq\sqrt{n_{+}}\left(\left|\left(\Omega-\frac{1}{n_{-}}\widehat{\Lambda}\mathbf{Z}_{-}^{\top}\mathbf{X}_{-}\right)\widehat{\mathbf{D}}_{\mathbf{X}_{-}}\right|_{\infty}\left|\widehat{\mathbf{D}}_{\mathbf{X}_{-}}^{-1}\mathbf{D}_X\right|_{\infty}+\left|\left(\frac{1}{n_{-}}\widehat{\Lambda}\mathbf{Z}_{-}^{\top}\mathbf{X}_{-}-\frac{1}{n_{+}}\widehat{\Lambda}\mathbf{Z}_{+}^{\top}\mathbf{X}_{+}\right)\mathbf{D}_X\right|_{\infty}\right)\left|\overline{\Delta}_{+}\right|_1.
$$

Denoting by $\widetilde{R}$ the quantity in the middle bracket above, we obtain $|R|_\infty \le \sqrt{n_+} v_\beta(n_+)\widetilde{R}$ and

$$\begin{aligned}\widetilde{R} &\le \left|\left(\Omega - \frac{1}{n_-}\widehat{\Lambda}\mathbf{Z}_-^\top \mathbf{X}_-\right)\widehat{\mathbf{D}}_{\mathbf{X}_-}\right|_\infty \left|\widehat{\mathbf{D}}_{\mathbf{X}_-}^{-1}\mathbf{D}_X\right|_\infty + \left|\widehat{\Lambda}\widehat{\mathbf{D}}_{\mathbf{Z}_-}^{-1}\right|_{\infty,\infty}\left|\widehat{\mathbf{D}}_{\mathbf{Z}_-}\mathbf{D}_Z^{-1}\right|_\infty (r_{\Psi-} + r_{\Psi_+}) \\ &\le \sqrt{1+\tau_{X-}}\left(\left|\left(\Omega - \frac{1}{n_-}\widehat{\Lambda}\mathbf{Z}_-^\top \mathbf{X}_-\right)\widehat{\mathbf{D}}_{\mathbf{X}_-}\right|_\infty + \left|\widehat{\Lambda}\widehat{\mathbf{D}}_{\mathbf{Z}_-}^{-1}\right|_{\infty,\infty}\frac{r_{\Psi-}+r_{\Psi_+}}{\sqrt{(1+\tau_{X-})(1-\tau'_{Z-})}}\right),\end{aligned}$$

and, for $\Lambda_* \in \mathcal{S}_{\lambda_1,\lambda_2}$, using the definition of $\widehat{\Lambda}$ in the third display and denoting by

$$M(\lambda_1) \triangleq \sqrt{1+\tau_{X-}}\max\left(1, \frac{r_{\Psi-}+r_{\Psi_+}}{\lambda_1\sqrt{(1+\tau_{X-})(1-\tau'_{Z-})}}\right),$$

$$\begin{aligned}\text{(A.35)}\quad \widetilde{R} &\le M(\lambda_1)\left(\left|\left(\Omega - \frac{1}{n_-}\widehat{\Lambda}\mathbf{Z}_-^\top \mathbf{X}_-\right)\widehat{\mathbf{D}}_{\mathbf{X}_-}\right|_\infty + \lambda_1\left|\widehat{\Lambda}\widehat{\mathbf{D}}_{\mathbf{Z}_-}^{-1}\right|_{\infty,\infty}\right) \\ &\le M(\lambda_1)\left(\left|\left(\Omega - \frac{1}{n_-}\widehat{\Lambda}\mathbf{Z}_-^\top \mathbf{X}_-\right)\widehat{\mathbf{D}}_{\mathbf{X}_-}\right|_\infty + \lambda_1\left|\widehat{\Lambda}\widehat{\mathbf{D}}_{\mathbf{Z}_-}^{-1}\right|_{\infty,\infty} + \frac{\lambda_2}{\sqrt{n_+}}\left|\widehat{\Lambda}\mathbf{Z}_+^\top\right|_{2,\infty}\right) \\ &\le M(\lambda_1)\left(\left|\left(\Omega - \frac{1}{n_-}\Lambda_*\mathbf{Z}_-^\top \mathbf{X}_-\right)\widehat{\mathbf{D}}_{\mathbf{X}_-}\right|_\infty + \lambda_1\left|\Lambda_*\widehat{\mathbf{D}}_{\mathbf{Z}_-}^{-1}\right|_{\infty,\infty} + \frac{\lambda_2}{\sqrt{n_+}}\left|\Lambda_*\mathbf{Z}_+^\top\right|_{2,\infty}\right) \\ &\le M(\lambda_1)\left(\left|\left(\Omega - \frac{1}{n_-}\Lambda_*\mathbf{Z}_-^\top \mathbf{X}_-\right)\widehat{\mathbf{D}}_{\mathbf{X}_-}\right|_\infty + \lambda_1\left|\Lambda_*\widehat{\mathbf{D}}_{\mathbf{Z}_-}^{-1}\right|_{\infty,\infty} + \frac{\lambda_2}{\sqrt{n_+}}\left|\Lambda_*\mathbf{Z}_+^\top\right|_{2,\infty}\right) \\ &\le M(\lambda_1)\left(\frac{\left|\left(\Omega - \frac{1}{n_-}\Lambda_*\mathbf{Z}_-^\top \mathbf{X}_-\right)\mathbf{D}_X\right|_\infty}{\sqrt{1-\tau_{X-}}} + \lambda_1\sqrt{1+\tau_{Z-}}\left|\Lambda_*\mathbf{D}_Z^{-1}\right|_{\infty,\infty} + \frac{\lambda_2}{\sqrt{n_+}}\left|\Lambda_*\mathbf{Z}_+^\top\right|_{2,\infty}\right).\end{aligned}$$

Let $\epsilon > 0$ and n_0 such that, for all $n_+ \ge n_0$, $(\sigma - v_2(n_+))/\sigma \ge 1-\epsilon$. If $u_i(\beta)$ is i.i.d. normally distributed and independent of z_i we proceed as follows. For $n_+ \ge n_0$, we have

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{\sqrt{n_+}}\left|\widehat{\mathbf{D}}_{\widehat{\Lambda}\mathbf{Z}_+}\widehat{\Lambda}\mathbf{Z}_+^\top\mathbf{U}_+(\beta)\right|_\infty > \frac{q_{W_+}(1-\alpha)}{\sqrt{n_+}}\sqrt{\widehat{Q}\left(\widehat{\beta}_+\right)}\right) \\ &\le \mathbb{P}\left(\left|\frac{\widehat{\mathbf{D}}_{\widehat{\Lambda}\mathbf{Z}_+}\widehat{\Lambda}\mathbf{Z}_+^\top\mathbf{E}}{\sqrt{n_+}}\right|_\infty > \frac{q_{W_+}(1-\alpha)}{\sqrt{n_+}}\left(1 - \frac{v_\sigma(n_+)}{\sigma}\right)\right) + \alpha_\beta(n_+) \\ &\le \mathbb{E}\left[\mathbb{P}\left(\left|\frac{\widehat{\mathbf{D}}_{\widehat{\Lambda}\mathbf{Z}_+}\widehat{\Lambda}\mathbf{Z}_+^\top\mathbf{E}}{\sqrt{n_+}}\right|_\infty > \frac{q_{W_+}(1-\alpha)}{\sqrt{n_+}}(1-\epsilon)\right)\middle|\mathcal{F}_\infty\right] + \alpha_\beta(n_+) \\ &= 1 - \alpha - \alpha_\beta(n_+).\end{aligned}$$

Otherwise, we obtain under the appropriate scenario, for $n_+ \ge n_0$,

$$\mathbb{P}\left(\frac{1}{n_+}\left|\widehat{\mathbf{D}}_{\widehat{\Lambda}\mathbf{Z}_+}\widehat{\Lambda}\mathbf{Z}_+^\top\mathbf{U}_+(\beta)\right|_\infty > r_+^\Lambda\sqrt{\widehat{Q}\left(\widehat{\beta}_+\right)}\right)$$

$$
\begin{aligned}
&\le \mathbb{P}\left(\frac{1}{n_+}\left|\widehat{\mathbf{D}}_{\widehat{\Lambda}\mathbf{Z}_+}\widehat{\Lambda}\mathbf{Z}_+^\top\mathbf{U}_+\right|_\infty > r_+^\Lambda\left(1-\frac{v_\sigma(n_+)}{\sigma}\right)\right) + \alpha_\beta(n_+)\\
&\le \mathbb{E}\left[\mathbb{P}\left(\frac{1}{n_+}\left|\widehat{\mathbf{D}}_{\widehat{\Lambda}\mathbf{Z}_+}\widehat{\Lambda}\mathbf{Z}_+^\top\mathbf{U}_+\right|_\infty > r_+^\Lambda\left(1-\epsilon\right)\right)\middle|\mathcal{F}_\infty\right] + \alpha_\beta(n_+)\\
&= 1-\alpha-\alpha_\beta(n_+)-\alpha_B(n_+),
\end{aligned}
$$

where $\alpha_B(n_+)$ is nonzero only for Scenario 4.

Let us now explain how the model with approximation errors can be dealt with. We base our analysis on the decomposition

$$\sqrt{n_+}\left(\widehat{\Omega\beta}-\Omega\beta-V(\beta)\right) = R + \frac{1}{\sqrt{n_+}}\widehat{\Lambda}\mathbf{Z}_+^\top\mathbf{V}_+(\beta) - \sqrt{n_+}V(\beta) + \frac{1}{\sqrt{n_+}}\widehat{\Lambda}\mathbf{Z}_+^\top\mathbf{W}_+(\beta),$$

where R and the last stochastic term are as before. We have $|\sqrt{n}V(\beta)|_\infty \le v_v(n)$ and

$$\left|\frac{1}{\sqrt{n}}\widehat{\Lambda}\mathbf{Z}_+^\top\mathbf{V}(\beta)\right|_\infty \le v_v(n)\frac{1}{\sqrt{n_+}}\left|\widehat{\Lambda}\mathbf{Z}_+^\top\right|_{2,\infty}\sqrt{\frac{n_+}{n}}.$$

By (8.4), we have

$$\left|\frac{1}{\sqrt{n}}\widehat{\Lambda}\mathbf{Z}_+^\top\mathbf{V}(\beta)\right|_\infty \le \frac{v_v(n)}{\lambda_2}\sqrt{\frac{n_+}{n}}\left(\left|\left(\Omega-\frac{1}{n_-}\Lambda_*\mathbf{Z}_-^\top\mathbf{X}_-\right)\widehat{\mathbf{D}}_{\mathbf{X}_-}\right|_\infty + \lambda_1\left|\Lambda_*\widehat{\mathbf{D}}_{\mathbf{Z}_-}^{-1}\right|_{\infty,\infty} + \frac{\lambda_2}{\sqrt{n_+}}\left|\Lambda_*\mathbf{Z}_+^\top\right|_{2,\infty}\right),$$

hence the modification of the result without approximation errors follows. □

A.4. **Lower Bounds on $\widehat{\kappa}_{q,J}$ When $\mathbf{Z}=\mathbf{X}$.** The following propositions establish lower bounds on $\widehat{\kappa}_{q,J}$ when there are no endogenous regressors, $P=[K]$, $\mathcal{R}=\mathbb{R}^K$, and $c\in(0,r^{-1})$. Recall that, in that case, $\widehat{C}_J$ is a cone and takes the form (4.2). For all $J\subseteq[K]$, we define the following restricted eigenvalue (RE) constants

$$\widehat{\kappa}_{\mathrm{RE},J} \triangleq \min_{\Delta\in\mathbb{R}^K\setminus\{0\}:\ \Delta\in\widehat{C}_J}\frac{|\Delta^\top\widehat{\Psi}\Delta|}{|\Delta_J|_2^2},\qquad \widehat{\kappa}'_{\mathrm{RE},J} \triangleq \min_{\Delta\in\mathbb{R}^K\setminus\{0\}:\ \Delta\in\widehat{C}_J}\frac{|J||\Delta^\top\widehat{\Psi}\Delta|}{|\Delta_J|_1^2}.$$

Proposition A.2. *For all $J\subseteq[K]$, we have*

$$\widehat{\kappa}_{1,J} \ge \frac{1-cr}{2}\widehat{\kappa}_{1,J,J} \ge \frac{(1-cr)^2}{4|J|}\widehat{\kappa}'_{\mathrm{RE},J} \ge \frac{(1-cr)^2}{4|J|}\widehat{\kappa}_{\mathrm{RE},J}.$$

Proof. For Δ such that $|\Delta_{J^c}|_1 \le \frac{1+cr}{1-cr}|\Delta_J|_1$ we have $|\Delta|_1 \le \frac{2}{1-cr}|\Delta_J|_1$. Thus, one obtains

$$\frac{|\Delta^\top\widehat{\Psi}\Delta|}{|\Delta_J|_1^2} \le \frac{|\Delta|_1\left|\widehat{\Psi}\Delta\right|_\infty}{|\Delta_J|_1^2} \le \frac{2}{1-cr}\frac{\left|\widehat{\Psi}\Delta\right|_\infty}{|\Delta_J|_1} \le \frac{4}{(1-cr)^2}\frac{\left|\widehat{\Psi}\Delta\right|_\infty}{|\Delta|_1}.$$

Taking the infimum over Δ's proves the first two inequalities of the proposition. The second inequality uses the fact that from Hölder's inequality $|\Delta_J|_1^2 \le |J||\Delta_J|_2^2$. □

We now obtain bounds for sensitivities $\widehat{\kappa}_{q,J}$ with $1 < q \le 2$. For all $s \le K$, we consider a uniform version of the restricted eigenvalue constant: $\widehat{\kappa}_{\mathrm{RE}}(s) \triangleq \min_{|J|\le s} \widehat{\kappa}_{\mathrm{RE},J}$.

Proposition A.3. *For all $s \le K/2$ and $1 < q \le 2$, we have*

$$\forall\ J:\ |J| \le s,\ \widehat{\kappa}_{q,J} \ge C(q) s^{-1/q} \widehat{\kappa}_{\mathrm{RE}}(2s),$$

where $C(q) = 2^{-1/q-1/2}(1-cr)\left(1 + \frac{1+cr}{1-cr}\,(q-1)^{-1/q}\right)^{-1}$.

Proof. For $\Delta \in \mathbb{R}^K$ and $J \subseteq [K]$, let $J_1 = J_1(\Delta, J)$ be the subset of indices in $[K]$ corresponding to the s largest in absolute value components of Δ outside of J. Define $J_+ = J \cup J_1$. If $|J| \le s$ we have $|J_+| \le 2s$. It is easy to see that the kth largest absolute value of elements of Δ_{J^c} satisfies $|\Delta_{J^c}|_{(k)} \le |\Delta_{J^c}|_1/k$. Thus,

$$|\Delta_{J_+^c}|_q^q = \sum_{j\in J_+^c} |\Delta_j|^p = \sum_{k\ge s+1} |\Delta_{J^c}|_{(k)}^q \le |\Delta_{J^c}|_1^q \sum_{k\ge s+1} \frac{1}{k^q} \le \frac{|\Delta_{J^c}|_1^q}{(q-1)s^{q-1}}.$$

For $\Delta \in \widehat{C}_J$, this implies

$$|\Delta_{J_+^c}|_q \le \frac{|\Delta_{J^c}|_1}{(q-1)^{1/q}s^{1-1/q}} \le \frac{c_0|\Delta_J|_1}{(q-1)^{1/q}s^{1-1/q}} \le \frac{c_0|\Delta_J|_q}{(q-1)^{1/q}},$$

where $c_0 = \frac{1+cr}{1-cr}$. Therefore, using that $|\Delta_J|_q \le |\Delta_{J_+}|_q$, we get, for $\Delta \in \widehat{C}_J$,

$$|\Delta|_q \le |\Delta_{J_+}|_q + |\Delta_{J_+^c}|_q \le (1 + c_0(q-1)^{-1/q})|\Delta_{J_+}|_q \le (1 + c_0(q-1)^{-1/q})(2s)^{1/q-1/2}|\Delta_{J_+}|_2, \tag{A.36}$$

where the last inequality follows from the bound

$$|\Delta_{J_+}|_q \le |J_+|^{1/q-1/2}|\Delta_{J_+}|_2 \le (2s)^{1/q-1/2}|\Delta_{J_+}|_2.$$

Using (A.36) and $|\Delta|_1 \le \frac{2}{1-cr}|\Delta_J|_1 \le \frac{2\sqrt{|J|}}{1-cr}|\Delta_J|_2 \le \frac{2\sqrt{s}}{1-cr}|\Delta_J|_2 \le \frac{2\sqrt{s}}{1-cr}|\Delta_{J_+}|_2$ for $\Delta \in \widehat{C}_J$, we get

$$\begin{aligned} \frac{|\Delta^\top \widehat{\Psi}\Delta|}{|\Delta_{J_+}|_2^2} &\le \frac{|\Delta|_1 \left|\widehat{\Psi}\Delta\right|_\infty}{|\Delta_{J_+}|_2^2} \\ &\le \frac{2\sqrt{s}\left|\widehat{\Psi}\Delta\right|_\infty}{(1-cr)|\Delta_{J_+}|_2} \\ &\le \frac{s^{1/q}\left|\widehat{\Psi}\Delta\right|_\infty}{C(q)|\Delta|_q}. \end{aligned}$$

Since $|J_+| \le 2s$, this proves the proposition. □

A.5. **Relation Between the Sensitivities in the General Case.** The following result allows to relate the sensitivities for various losses. It is obtained by manipulations of the definition of the sensitivities and requires no assumption. It allows to relate the sensitivities for various losses.

Proposition A.4. *Let P, J be two subsets of $[K]$ and $c > 0$.*

(i) Let $J \subseteq \widehat{J} \subseteq [K]$. Then, for all $l \in \mathcal{L}$, we have $\widehat{\kappa}_{l,J} \geq \widehat{\kappa}_{l,\widehat{J}}$.

(ii) For all $T \subseteq [K]$ and $q \in [1, \infty]$, we have $\widehat{\kappa}_{q,T,J} \geq \widehat{\kappa}_{q,J}$.

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

$$c_\kappa(|J \cap P|)^{-1/q} \widehat{\kappa}_{\infty,J} \leq \widehat{\kappa}_{q,J} \leq \widehat{\kappa}_{\infty,J}, \tag{A.37}$$

$$|\widetilde{J}|^{-1/q} \widehat{\kappa}_{\infty,T,J} \leq \widehat{\kappa}_{q,T,J} \leq \widehat{\kappa}_{\infty,T,J}, \tag{A.38}$$

where, for $s \in [p]$, $c_\kappa(s) \triangleq \min\left((1-cr)_+^{-1}(2s + |P^c| + c(1-r)|I^c|), (1-c)_+^{-1}(2s + |P^c|)\right)$.

(iv) In addition to (A.37), we have, for any partition $\bigcup_{m=1}^M J_m = [K]$,

(A.39)
$$\widehat{\kappa}_{1,J} \geq \max\left(\left(\frac{2}{\widehat{\kappa}_{1,J\cap P,J}} + \frac{1}{\widehat{\kappa}_{1,P^c,J}} + \frac{cr}{\widehat{\kappa}_{\sigma,J}}\right)^{-1}, (1-cr)_+\left(\frac{2}{\widehat{\kappa}_{1,J\cap P,J}} + \frac{1}{\widehat{\kappa}_{1,P^c,J}} + \frac{c(1-r)}{\widehat{\kappa}_{1,I^c,J}}\right)^{-1},\right.$$
$$\left.(1-c)_+\left(\frac{2}{\widehat{\kappa}_{1,J\cap P,J}} + \frac{1}{\widehat{\kappa}_{1,P^c,J}}\right)^{-1}, \left(\sum_{m=1}^M \frac{1}{\widehat{\kappa}_{1,J_m,J}}\right)^{-1}\right).$$

(v) We have

$$\widehat{\kappa}_{\sigma,J} \geq \max\left((1-cr)_+\left(\frac{2}{\widehat{\kappa}_{1,J\cap P,J}} + \frac{1}{\widehat{\kappa}_{1,P^c,J}} + \frac{1-r}{r\widehat{\kappa}_{1,I^c,J}}\right)^{-1}, \left(\frac{1}{\widehat{\kappa}_{1,I,J}} + \frac{1}{r\ \widehat{\kappa}_{1,I^c,J}}\right)^{-1}, r\widehat{\kappa}_{1,J}\right). \tag{A.40}$$

(vi) For all $T \subseteq [K]$, we have

$$\widehat{\kappa}_{\infty,T,J} = \min_{k\in T} \widehat{\kappa}^*_{e_k,J} = \min_{k\in T} \min_{\Delta\in\widehat{C}_J:\ \Delta_k=1,\ |\Delta|_\infty\leq 1} \left|\widehat{\Psi}\Delta\right|_\infty.$$

The above statements hold if we replace the sensitivities based on $\widehat{C}_J$ by those based on $\widehat{C}_{\gamma,J}$, $c_\kappa(s)$, (A.39), and (A.40) by, respectively,

$$c_\gamma(s) \triangleq \min\left(\frac{3s + 2|P^c| + 2c(1-r)|I^c|}{(1-2cr)_+}, \frac{3s + 2|P^c|}{(1-2c)_+}\right);$$

$$\widehat{\gamma}_{1,J} \geq \max\left(\left(\frac{3}{\widehat{\gamma}_{1,J\cap P,J}} + \frac{2}{\widehat{\gamma}_{1,P^c,J}} + \frac{2cr}{\widehat{\gamma}_{\sigma,J}}\right)^{-1}, (1-2cr)_+\left(\frac{3}{\widehat{\gamma}_{1,J\cap P,J}} + \frac{2}{\widehat{\gamma}_{1,P^c,J}} + \frac{2c(1-r)}{\widehat{\gamma}_{1,I^c,J}}\right)^{-1},\right.$$
$$\left.(1-2c)_+\left(\frac{3}{\widehat{\gamma}_{1,J\cap P,J}} + \frac{2}{\widehat{\gamma}_{1,P^c,J}}\right)^{-1}, \left(\sum_{m=1}^M \frac{1}{\widehat{\gamma}_{1,J_m,J}}\right)^{-1}\right);$$

$$\widehat{\gamma}_{\sigma,J} \geq \max\left((1-2cr)_+\left(\frac{3}{\widehat{\gamma}_{1,J\cap P,J}}+\frac{2}{\widehat{\gamma}_{1,P^c,J}}+\frac{1-r}{r}\frac{1}{\widehat{\gamma}_{1,I^c,J}}\right)^{-1}, \left(\frac{1}{\widehat{\gamma}_{1,I,J}}+\frac{1}{r\ \widehat{\gamma}_{1,I^c,J}}\right)^{-1}, r\widehat{\gamma}_{1,J}\right).$$

The bound in (A.37) applies to the case where $q=1$ but sharper and simple bounds can be obtained by using (A.39). For example, using the two middle terms in the maximum and (A.38) yields

$$\widehat{\kappa}_{1,J} \geq \max\left((1-cr)_+\left(\frac{2\,|J\cap P|}{\widehat{\kappa}_{\infty,J\cap P,J}}+\frac{|P^c|}{\widehat{\kappa}_{\infty,P^c,J}}+\frac{c(1-r)\,|I^c|}{\widehat{\kappa}_{\infty,I^c,J}}\right)^{-1}, (1-c)_+\left(\frac{2\,|J\cap P|}{\widehat{\kappa}_{\infty,J\cap P,J}}+\frac{|P^c|}{\widehat{\kappa}_{\infty,P^c,J}}\right)^{-1}\right).$$

Proof of Proposition A.4. We prove the bounds for the sensitivities based on $\widehat{C}_J$, those for the sensitivities based on $\widehat{C}_{\gamma,J}$ are obtained similarly. Parts (i) and (ii) are straightforward. The upper bound in (A.37) 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 $\Delta \neq 0$,

$$\frac{\left|\widehat{\Psi}\Delta\right|_\infty}{|\Delta|_q} \geq \frac{\left|\widehat{\Psi}\Delta\right|_\infty}{|\Delta|_\infty}\left(\frac{|\Delta|_\infty}{|\Delta|_1}\right)^{1/q}. \tag{A.41}$$

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

$$|\Delta_{J^c\cap P}|_1 \leq |\Delta_{J\cap P}|_1 + cr|\Delta|_1 + c(1-r)|\Delta_{I^c}|_1 \tag{A.42}$$

which, by adding $|\Delta_{(J\cap P)\cup P^c}|_1$ on both sides, is equivalent to

$$|\Delta|_1 \leq \frac{1}{(1-cr)_+}\left(2|\Delta_{J\cap P}|_1 + |\Delta_{P^c}|_1 + c(1-r)|\Delta_{I^c}|_1\right). \tag{A.43}$$

From (A.43) and the fact that $\Delta_{J^c\cap J\left(\widehat{\beta}\right)^c} = \mathbf{0}$, we deduce

$$|\Delta|_1 \leq \frac{\left|\Delta_{(J\cup P^c\cup I^c)\cap\left(J\cup J\left(\widehat{\beta}\right)\right)}\right|_\infty}{(1-cr)_+}\left(2|J\cap P| + \left|P^c\cap\left(J\cup J\left(\widehat{\beta}\right)\right)\right| + c(1-r)\left|I^c\cap\left(J\cup J\left(\widehat{\beta}\right)\right)\right|\right).$$

We obtain the first lower bound using the fact that $|\Delta_{J\cup P^c\cup I^c}|_\infty \leq |\Delta|_\infty$ and (A.41).
Let us obtain an alternative lower bound for the case where $c\in(0,1)$. The condition that $\Delta\in\widehat{C}_J$ can also be written as

$$|\Delta_{J^c\cap P}|_1 \leq |\Delta_{J\cap P}|_1 + c(r-1)|\Delta_I|_1 + c|\Delta|_1$$

which implies

$$|\Delta_{J^c\cap P}|_1 \leq |\Delta_{J\cap P}|_1 + c|\Delta|_1$$

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

$$|\Delta|_1 \leq \frac{1}{1-c}\left(2|\Delta_{J\cap P}|_1 + |\Delta_{P^c}|_1\right). \tag{A.44}$$

This yields

$$|\Delta|_1 \le \frac{2|J\cap P| + \left|P^c \cap \left(J \cup J\left(\widehat{\beta}\right)\right)\right|}{(1-c)_+} \left|\Delta_{(J\cup P^c)\cap\left(J\cup J(\widehat{\beta})\right)}\right|_\infty . \tag{A.45}$$

Inequality (A.38) can be proved in a similar manner. The lower bounds follows from the fact that

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

and $|\Delta_{\widetilde{J}}|_1 \le |\widetilde{J}||\Delta_{\widetilde{J}}|_\infty$. While the upper bound holds because $\left|\Delta_{\widetilde{J}}\right|_q \ge \left|\Delta_{\widetilde{J}}\right|_\infty$.
To prove (A.39) it suffices to note that, by definition of the set $\widehat{C}_J$,

$$|\Delta|_1 \le \left(\frac{2}{\widehat{\kappa}_{1,J\cap P,J}} + \frac{1}{\widehat{\kappa}_{1,P^c,J}} + \frac{cr}{\widehat{\kappa}_{\sigma,J}}\right)\left|\widehat{\Psi}\Delta\right|_\infty, \tag{A.46}$$

by (A.43),

$$|\Delta|_1 \le \frac{1}{(1-cr)_+}\left(\frac{2}{\widehat{\kappa}_{1,J\cap P,J}} + \frac{1}{\widehat{\kappa}_{1,P^c,J}} + \frac{c(1-r)}{\widehat{\kappa}_{1,I^c,J}}\right)\left|\widehat{\Psi}\Delta\right|_\infty,$$

and, by (A.44),

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

The last upper bound follows from the fact that

$$|\Delta|_1 = \sum_{m=1}^{M}|\Delta_{J_m}|_1 \le \left(\sum_{m=1}^{M}\frac{1}{\widehat{\kappa}_{1,J_m,J}}\right)\left|\widehat{\Psi}\Delta\right|_\infty.$$

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

$$(1-cr)|\Delta_I|_1 + (1-c)|\Delta_{I^c}|_1 \le 2|\Delta_{J\cap P}|_1 + |\Delta_{P^c}|_1, \tag{A.47}$$

which yields

$$\begin{aligned}|\Delta_I|_1 + \frac{1}{r}|\Delta_{I^c}|_1 &\le \frac{1}{(1-cr)_+}\left(2|\Delta_{J\cap P}|_1 + |\Delta_{P^c}|_1 + \frac{1-r}{r}|\Delta_{I^c}|_1\right)\\ &\le \frac{\left|\widehat{\Psi}\Delta\right|_\infty}{(1-cr)_+}\left(\frac{2}{\widehat{\kappa}_{1,J\cap P,J}} + \frac{1}{\widehat{\kappa}_{1,P^c,J}} + \frac{1-r}{r\widehat{\kappa}_{1,I^c,J}}\right).\end{aligned} \tag{A.48}$$

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

$$\frac{1}{\widehat{\kappa}_{\sigma,J}} = \sup_{\Delta\in\widehat{C}_J:\ \left|\widehat{\Psi}\Delta\right|_\infty = 1}\left(|\Delta_I|_1 + r^{-1}|\Delta_{I^c}|_1\right) \le \sup_{\Delta\in\widehat{C}_J:\ \left|\widehat{\Psi}\Delta\right|_\infty = 1}|\Delta_I|_1 + \frac{1}{r}\sup_{\Delta\in\widehat{C}_J:\ \left|\widehat{\Psi}\Delta\right|_\infty = 1}|\Delta_{I^c}|_1 .$$

The third upper uses that $|\Delta_I|_1 + r^{-1}\left|\Delta_{I^c}\right|_1 \le r^{-1}\left|\Delta\right|_1$.

If we replace $\widehat{C}_J$ by $\widehat{C}_{\gamma,J}$, the second and third inequalities hold for the corresponding sensitivities in the same way. Else, we use that the set $\widehat{C}_{\gamma,J}$ can be written as

$$(1-2cr)|\Delta_I|_1 + (1-2c)|\Delta_{I^c}|_1 \le 3|\Delta_{J\cap P}|_1 + 2|\Delta_{P^c}|_1.$$

Let us now prove (vi). Because for all k in $\widetilde{J}$, $|\Delta_{\widetilde{J}}|_\infty \ge |\Delta_k|$, one obtains that for all k in $\widetilde{J}$,

$$\widehat{\kappa}_{\infty,T,J} = \min_{\Delta\in\widehat{C}_J} \frac{\left|\widehat{\Psi}\Delta\right|_\infty}{|\Delta_{\widetilde{J}}|_\infty} \le \min_{\Delta\in\widehat{C}_J} \frac{\left|\widehat{\Psi}\Delta\right|_\infty}{|\Delta_k|} = \widehat{\kappa}^*_{e_k,J}.$$

Thus

$$\widehat{\kappa}_{\infty,T,J} \le \min_{k\in\widetilde{J}} \widehat{\kappa}^*_{e_k,J}.$$

But one also has

$$\text{(A.49)} \qquad \widehat{\kappa}_{\infty,T,J} = \min_{k\in\widetilde{J}} \min_{\Delta\in\widehat{C}_J:\ |\Delta_k|=|\Delta_{\widetilde{J}}|_\infty=1} \left|\widehat{\Psi}\Delta\right|_\infty \ge \min_{k\in\widetilde{J}} \min_{\Delta\in\widehat{C}_J:\ |\Delta_k|=1} \left|\widehat{\Psi}\Delta\right|_\infty .$$

□

The next proposition gives a sufficient condition to obtain a lower bound on the ℓ_∞ sensitivity which, by Proposition A.4, is a key element to bound from below all the sensitivities.

Proposition A.5. *If there exist random variables η_1 and η_2 such that, on an event $\mathcal{E}$, $\eta_1 > 0$, $\eta_2 \in (0,1)$, and*

$$\text{(A.50)} \qquad \forall k\in[K],\ \exists l(k)\in[L]:\ \left|\left(\widehat{\Psi}\right)_{l(k)k}\right| \ge \eta_1,\ \max_{k'\neq k}\left|\left(\widehat{\Psi}\right)_{l(k)k'}\right| \le \frac{1-\eta_2}{c_\kappa(|J\cap P|)}\left|\left(\widehat{\Psi}\right)_{l(k)k}\right|,$$

then, on $\mathcal{E}$, $\widehat{\kappa}_{\infty,J} \ge \eta_1\eta_2$. The same holds for the sensitivity based on $\widehat{C}_{\gamma,J}$ replacing c_κ by c_γ.

Assumption (A.50) is similar to the coherence condition in Donoho, Elad, and Temlyakov (2006) for symmetric matrices, but it is more general because it deals with rectangular matrices. It means that there exists one sufficiently "good" instrument. Indeed, if the regressors and instruments are centered, $|(\widehat{\Psi})_{l(k)k}|$ measures the empirical correlation between the $l(k)$th instrument for the kth regressor. It should be sufficiently large relative $\max_{k'\neq k}\left|\left(\widehat{\Psi}\right)_{l(k)k'}\right|$.

Proof of Proposition A.5. Take $k\in[K]$ and $l\in[L]$, we have

$$\left|\left(\widehat{\Psi}\Delta\right)_l - \left(\widehat{\Psi}\right)_{lk}\Delta_k\right| \le |\Delta|_1 \max_{k'\neq k}\left|\left(\widehat{\Psi}\right)_{lk'}\right|,$$

which yields

$$\left|\left(\widehat{\Psi}\right)_{lk}\right| |\Delta_k| \le |\Delta|_1 \max_{k'\neq k}\left|\left(\widehat{\Psi}\right)_{lk'}\right| + \left|\left(\widehat{\Psi}\Delta\right)_l\right|.$$

The two inequalities of the assumption yield

$$\left|\left(\widehat{\Psi}\right)_{l(k)k}\right| |\Delta_k| \le |\Delta|_1 \frac{1-\eta_2}{c_\kappa(|J\cap P|)} \left|\left(\widehat{\Psi}\right)_{l(k)k}\right| + \frac{1}{\eta_1}\left|\left(\widehat{\Psi}\Delta\right)_{l(k)}\right| \left|\left(\widehat{\Psi}\right)_{l(k)k}\right|.$$

This inequality, together with the fact that $\left|\left(\widehat{\Psi}\Delta\right)_{l(k)}\right| \le \left|\widehat{\Psi}\Delta\right|_\infty$ and the upper bounds from the proof of the upper bound (A.37) of Proposition A.4, yield

$$|\Delta_k| \le (1-\eta_2)|\Delta|_\infty + \frac{\left|\widehat{\Psi}\Delta\right|_\infty}{\eta_1}$$

and thus

$$\eta_1\eta_2|\Delta|_\infty \le \left|\widehat{\Psi}\Delta\right|_\infty .$$

One concludes using the definition of the ℓ_∞-sensitivity. □

A.6. **Sharp Computations of the Sensitivities when $|J \cup P^c \cup I^c|$ is small and $\mathcal{R}_D$ is convex.** Let us consider, for example, the sensitivity $\widehat{\kappa}^*_{e_k,J}$. It can be computed exactly as follows.

Algorithm A.1. *When $c \in (0,1]$ solve*

$$\min_{(\epsilon_j)_{j\in J\cup P^c}\in\{-1,1\}^{|J\cup P^c|}} \ \min_{(\Delta,v)\in\mathcal{U}_{k,J\cup P^c}} v$$

where $\mathcal{U}_{k,J\cup P^c}$ is the set of (Δ,v) with $\widehat{\mathbf{D}}_{\mathbf{X}}\Delta \in \mathcal{R}_D$ and $v\in\mathbb{R}$ satisfying:

$$\begin{gathered} v\ge 0, \qquad -v\mathbf{1} \le \widehat{\Psi}\Delta \le v\mathbf{1}, \qquad \Delta_k = 1, \qquad \Delta_{J^c\cap J\left(\widehat{\beta}\right)^c} = \mathbf{0}, \\ (1-cr)|\Delta_I|_1 + (1-c)|\Delta_{I^c}|_1 \le 2\textstyle\sum_{j\in J\cap P}\epsilon_j\Delta_j + \sum_{j\in P^c}\epsilon_j\Delta_j. \end{gathered}$$

When $c\in(1,r^{-1})$ solve

$$\min_{(\epsilon_j)_{j\in J\cup P^c\cup I^c}\in\{-1,1\}^{|J\cup P^c\cup I^c}} \ \min_{(\Delta,v)\in\mathcal{U}_{k,J\cup P^c\cup I^c}} v$$

where $\mathcal{U}_{k,J\cup P^c\cup I^c}$ is the set of (Δ,v) with $\widehat{\mathbf{D}}_{\mathbf{X}}\Delta\in\mathcal{R}_D$ and $v\in\mathbb{R}$ satisfying:

$$\begin{aligned} & v\ge 0, \qquad -v\mathbf{1}\le\widehat{\Psi}\Delta\le v\mathbf{1}, \qquad \Delta_k = 1, \qquad \Delta_{J^c\cap J\left(\widehat{\beta}\right)^c} = \mathbf{0}, \\ & (1-cr)|\Delta_I|_1 \le 2\sum_{j\in J\cap P_\perp}\epsilon_j\Delta_j + (1+c)\sum_{j\in J\cap P_{\not\perp}}\epsilon_j\Delta_j + \sum_{j\in P^c\cap I}\epsilon_j\Delta_j + c\sum_{j\in P^c\cap I^c}\epsilon_j\Delta_j + (c-1)\sum_{j\in J^c\cap P_{\not\perp}}\epsilon_j\Delta_j. \end{aligned}$$

For the constants $\widehat{\kappa}_{1,J}$ and $\widehat{\kappa}_{\sigma,J}$ one can compute sharp lower bounds as follows. Using (A.43), we obtain

$$\begin{aligned} |\Delta|_1 \le \frac{1}{(1-cr)_+}\big(& 2|\Delta_{J\cap P_\perp}|_1 + (2+c(1-r))|\Delta_{J\cap P_{\not\perp}}|_1 + |\Delta_{P^c\cap I}|_1 \\ & + (1+c(1-r))|\Delta_{P^c\cap I^c}|_1 + c(1-r)|\Delta_{J^c\cap P_{\not\perp}}|_1\big), \end{aligned}$$

hence,

$$\widehat{\kappa}_{1,J} \geq (1-cr)_+ \min_{\substack{\Delta\in\widehat{C}_J \\ 2|\Delta_{J\cap P_\perp}|_1+(2+c(1-r))|\Delta_{J\cap P_{\not\perp}}|_1+|\Delta_{P^c\cap I}|_1 \\ +(1+c(1-r))|\Delta_{P^c\cap I^c}|_1+c(1-r)|\Delta_{J^c\cap P_{\not\perp}}|_1=1}} \left|\widehat{\Psi}\Delta\right|_\infty \triangleq \underline{\widehat{\kappa}}_{1,J}$$

and one can compute $\underline{\widehat{\kappa}}_{1,J}$ as follows.

Algorithm A.2. *When $c\in(0,r^{-1})$, solve*

$$\min_{(\epsilon_j)_{J\cup P^c\cup I^c}\in\{-1,1\}^{|J\cup P^c\cup I^c|}} \ \min_{(\Delta,v)\in\mathcal{U}_{1,J\cup P^c\cup I^c}\in\{-1,1\}^{|J\cup P^c\cup I^c|}} v$$

where $\mathcal{U}_{1,J\cup P^c\cup I^c}$ is the set of (Δ,v) with $\widehat{\mathbf{D}}_{\mathbf{X}}\Delta\in\mathcal{R}_D$ and $v\in\mathbb{R}$ satisfying:

$$\begin{aligned}
& v\geq 0, \qquad -v\mathbf{1}\leq\widehat{\Psi}\Delta\leq v\mathbf{1}, \qquad \Delta_{J^c\cap J(\widehat{\beta})^c}=\mathbf{0}, \\
& 2\sum_{j\in J\cap P_\perp}\epsilon_j\Delta_j+(2+c(1-r))\sum_{j\in J\cap P_{\not\perp}}\epsilon_j\Delta_j+\sum_{j\in P^c\cap I}\epsilon_j\Delta_j+(1+c(1-r))\sum_{j\in P^c\cap I^c}\epsilon_j\Delta_j \\
& +c(1-r)\sum_{j\in J^c\cap P_{\not\perp}}\epsilon_j\Delta_j=1-cr, \\
& (1-cr)|\Delta_I|_1\leq 2\sum_{j\in J\cap P_\perp}\epsilon_j\Delta_j+(1+c)\sum_{j\in J\cap P_{\not\perp}}\epsilon_j\Delta_j+\sum_{j\in P^c\cap I}\epsilon_j\Delta_j+c\sum_{j\in P^c\cap I^c}\epsilon_j\Delta_j \\
& \qquad +(c-1)\sum_{j\in J^c\cap P_{\not\perp}}\epsilon_j\Delta_j.
\end{aligned}$$

When we use the restricted set based on $\widehat{C}_{\gamma,J}$ we have

$$\underline{\widehat{\gamma}}_{1,J} \triangleq (1-2cr)_+ \min_{\substack{\Delta\in\widehat{C}_{\gamma,J} \\ 3|\Delta_{J\cap P_\perp}|_1+(3+2c(1-r))|\Delta_{J\cap P_{\not\perp}}|_1+2|\Delta_{P^c\cap I}|_1 \\ +(2+2c(1-r))|\Delta_{P^c\cap I^c}|_1+c(1-r)|\Delta_{J^c\cap P_{\not\perp}}|_1=1}} \left|\widehat{\Psi}\Delta\right|_\infty .$$

Using (A.48), we obtain

$$|\Delta_I|_1+\frac{1}{r}|\Delta_{I^c}|_1\leq\frac{1}{(1-cr)_+}\left(2|\Delta_{J\cap P_\perp}|_1+\frac{1+r}{r}|\Delta_{J\cap P_{\not\perp}}|_1+|\Delta_{P^c\cap I}|_1+\frac{1}{r}|\Delta_{P^c\cap I^c}|_1+\frac{1-r}{r}|\Delta_{J^c\cap P_{\not\perp}}|_1\right),$$

hence,

$$\text{(A.51)} \qquad \widehat{\kappa}_{\sigma,J} \geq (1-cr)_+ \min_{\substack{\Delta\in\widehat{C}_J \\ 2|\Delta_{J\cap P_\perp}|_1+\frac{1+r}{r}|\Delta_{J\cap P_{\not\perp}}|_1+|\Delta_{P^c\cap I}|_1+\frac{1}{r}|\Delta_{P^c\cap I^c}|_1+\frac{1-r}{r}|\Delta_{J^c\cap P_{\not\perp}}|_1=1}} \left|\widehat{\Psi}\Delta\right|_\infty \triangleq \underline{\widehat{\kappa}}_{\sigma,J}$$

and one can compute $\underline{\widehat{\kappa}}_{\sigma,J}$ as follows.

Algorithm A.3. *When $c \in (0, r^{-1})$, solve*

$$\min_{(\epsilon_j)_{J\cup P^c\cup I^c}\in\{-1,1\}^{|J\cup P^c\cup I^c|}} \ \min_{(\Delta,v)\in\mathcal{U}_{\sigma,J\cup P^c\cup I^c}\in\{-1,1\}^{|J\cup P^c\cup I^c|}} v$$

where $\mathcal{U}_{\sigma,J\cup P^c\cup I^c}$ is the set of (Δ, v) with $\widehat{\mathbf{D}}_{\mathbf{X}}\Delta \in \mathcal{R}_D$ and $v \in \mathbb{R}$ satisfying:

$$\begin{aligned}
& v \ge 0, \qquad -v\mathbf{1} \le \widehat{\Psi}\Delta \le v\mathbf{1}, \qquad \Delta_{J^c\cap J(\widehat{\beta})^c} = \mathbf{0},\\
& 2\sum_{j\in J\cap P_\perp}\epsilon_j\Delta_j + \left(1+\frac{1}{r}\right)\sum_{j\in J\cap P_{\not\perp}}\epsilon_j\Delta_j + \sum_{j\in P^c\cap I}\epsilon_j\Delta_j + \frac{1}{r}\sum_{j\in P^c\cap I^c}\epsilon_j\Delta_j + \left(\frac{1}{r}-1\right)\sum_{j\in J^c\cap P_{\not\perp}}\epsilon_j\Delta_j = 1-cr,\\
& (1-cr)|\Delta_I|_1 \le 2\sum_{j\in J\cap P_\perp}\epsilon_j\Delta_j + (1+c)\sum_{j\in J\cap P_{\not\perp}}\epsilon_j\Delta_j + \sum_{j\in P^c\cap I}\epsilon_j\Delta_j + c\sum_{j\in P^c\cap I^c}\epsilon_j\Delta_j + (c-1)\sum_{j\in J^c\cap P_{\not\perp}}\epsilon_j\Delta_j.
\end{aligned}$$

When we use the restricted set based on $\widehat{C}_{\gamma,J}$ we have

$$\underline{\widehat{\gamma}}_J^{\sigma} \triangleq (1-2cr)_+ \min_{\Delta\in\widehat{C}_{\gamma,J}:\ 3|\Delta_{J\cap P_\perp}|_1+\left(2+\frac{1}{r}\right)|\Delta_{J\cap P_{\not\perp}}|_1+2|\Delta_{P^c\cap I}|_1+\left(1+\frac{1}{r}\right)|\Delta_{P^c\cap I^c}|_1+\left(\frac{1}{r}-1\right)|\Delta_{J^c\cap P_{\not\perp}}|_1=1} \left|\widehat{\Psi}\Delta\right|_\infty.$$

A.7. **Bounds on the Population Sensitivities in Benchmark Cases.** We now present lower bounds on the population sensitivities based on C_J, when $P = [K]$ and $\mathcal{R} = \mathbb{R}^K$, in benchmark cases.

Example O1. Besides possibly a constant, regressors and instruments are mean zero and of variance 1, regressors are uncorrelated, for the regressors of index in $N \subseteq I^c$ there could be no instrument while for those of index in $N^c \cap I^c$ there are l_k instruments with correlation ρ_{jk} for $j = 1, \dots, l_k$ with them and 0 with the other regressors. Assuming that $\rho_k \triangleq \max_{j=1,\dots,l_k}|\rho_{jk}| > 0$ for $k \in I^c \bigcap N^c$ and $\rho_k \triangleq 1$ for $k \in I$, if $J \cap N = \emptyset$, we have

$$\begin{aligned}
&\kappa_{1,J,J} \ge \left(\sum_{k\in J}\frac{1}{\rho_k}\right)^{-1}; \quad \kappa^*_{e_k,J} \ge \rho_k \quad \forall k \in N^c;\\
&\kappa^*_{e_k,J} \ge \left(\left(\sqrt{\frac{1-\tau_X}{1+\tau_X}} - c\right)^{-1}\left(2 - \sqrt{\frac{1-\tau_X}{1+\tau_X}} + c\right)\sum_{k\in J}\frac{1}{\rho_k}\right)^{-1} \quad \forall k \in N;\\
&\kappa_{1,J} \ge \max\left(\left(\sqrt{\frac{1-\tau_X}{1+\tau_X}} - c\right)\left(2\sum_{k\in J}\frac{1}{\rho_k}\right)^{-1},\right.\\
&\qquad\qquad \left.\left(\sqrt{\frac{1-\tau_X}{1+\tau_X}} - c\overline{r} - c(1-\overline{r})\mathbb{1}\{I^c\cap N \neq \emptyset\}\right)\left(2\sum_{k\in J}\frac{1}{\rho_k} + c(1-\overline{r})\sum_{k\in I^c\cap N^c}\frac{1}{\rho_k}\right)\right);\\
&\kappa_{\sigma,J} \ge \max\left(\overline{r}\left(\sqrt{\frac{1-\tau_X}{1+\tau_X}} - c\right)\left(2\sum_{k\in J}\frac{1}{\rho_k}\right)^{-1}\left(\sqrt{\frac{1-\tau_X}{1+\tau_X}} - c\overline{r}\right)\left(2\sum_{k\in J}\frac{1}{\rho_k} + \frac{1-\overline{r}}{\overline{r}}\sum_{k\in I^c}\frac{1}{\rho_k}\right)^{-1}\mathbb{1}\{N = \emptyset\}\right).
\end{aligned}$$

In this example, only the instruments with maximum correlation with the endogenous regressors play a role in the lower bounds. It is sometimes possible to consider situations where $L < K$ when the regressors for which there are no instruments have a coefficient which is zero.

Proof. Note that

$$|\Psi\Delta|_\infty = \max_{k\in N^c} \rho_k \left|\Delta_k\right| . \tag{A.52}$$

Because $J \cap N = \emptyset$, $|\Delta_J|_1 \le \left(\sum_{k\in J}\rho_k^{-1}\right)|\Psi\Delta|_\infty$, which yields the lower bound on $\kappa_{1,J,J}$.

By (A.52), we have, if $k \in N^c$, $\rho_k\left|\Delta_k\right| \le |\Psi\Delta|_\infty$. Using that, if $c \in (0, \sqrt{(1-\tau_X)/(1+\tau_X)})$, on $\mathcal{G}_\Psi$,

$$\left(\sqrt{\frac{1-\tau_X}{1+\tau_X}} - c\right)|\Delta|_1 \le 2|\Delta_J|_1, \tag{A.53}$$

hence,

$$|\Delta_{J^c}|_1 \le \left(\sqrt{\frac{1-\tau_X}{1+\tau_X}} - c\right)^{-1}\left(2 - \sqrt{\frac{1-\tau_X}{1+\tau_X}} + c\right)|\Delta_J|_1 ,$$

and that, because $N \subseteq J^c$, we have, for $k \in N$, $|\Delta_k| \le |\Delta_{J^c}|_1$ yields the lower bounds for the sensitivities $\kappa^*_{e_k,J}$.

We now prove the lower bound on $\kappa_{1,J}$ and work on $\mathcal{G}_\Psi$. When $c \in (0,1]$, using (A.53), we get

$$\left(\sqrt{\frac{1-\tau_X}{1+\tau_X}} - c\right)|\Delta|_1 = 2\sum_{k\in J}\frac{1}{\rho_k}|\rho_k\Delta_k| = |\Psi\Delta|_\infty 2\sum_{k\in J}\frac{1}{\rho_k},$$

while, when $c > 1$, we get

$$\left(\sqrt{\frac{1-\tau_X}{1+\tau_X}} - c\overline{r}\right)|\Delta|_1 \le \left(|\Psi\Delta|_\infty\left(2\sum_{k\in J}\frac{1}{\rho_k} + c(1-\overline{r})\sum_{k\in I^c\cap N^c}\frac{1}{\rho_k}\right) + c(1-\overline{r})|\Delta_{I^c\cap N}|_1\right),$$

which yields the result. Now, by simple manipulations, if $\Delta \in C_J$, then we have

$$\begin{aligned}\left(\sqrt{\frac{1-\tau_X}{1+\tau_X}} - c\overline{r}\right)\left(|\Delta_I|_1 + \frac{1}{\overline{r}}|\Delta_{I^c}|_1\right) &\le 2|\Delta_J|_1 + \frac{1-\overline{r}}{\overline{r}}\sqrt{\frac{1-\tau_X}{1+\tau_X}}|\Delta_{I^c}|_1, \\ &\le 2\sum_{k\in J}\frac{1}{\rho_k} + \frac{1-\overline{r}}{\overline{r}}\sqrt{\frac{1-\tau_X}{1+\tau_X}}\sum_{k\in I^c}\frac{1}{\rho_k},\end{aligned}$$

supplementing this with the previous lower bound using (A.40) yields the bound for $\kappa_{\sigma,J}$. □

Example O2. Let $\sigma > 0$ and $(H_k)_{k=0}^{K-1}$ be the first Hermite polynomials. Assume $(\widetilde{z}_i, v_i)_{i=1}^n$ are i.i.d. of distribution $\mathcal{N}(0, I_2)$, $\widetilde{x}_i = \zeta\widetilde{z}_i + \sigma v_i$, $x_i = \left(H_k\left(\widetilde{x}_i/\sqrt{\zeta^2+\sigma^2}\right)\right)_{k=0}^{K-1}$, and $z_i = \left(H_k\left(\widetilde{z}_i\right)\right)_{k=0}^{K-1}$. This is a particular case of Example O1 with $\rho_k = \left(\zeta/\sqrt{\zeta^2+\sigma^2}\right)^{k-1}$ and $N = \emptyset$.

Proof. Hermite polynomials are orthonormal in $L^2(\mu) \triangleq \{f : \int_{\mathbb{R}} f^2(x)e^{-\frac{x^2}{2}}dx < \infty\}$ equipped with $(f,g)_{L^2(\mu)} \triangleq (\int_{\mathbb{R}} f(x)g(x)e^{-\frac{x^2}{2}}dx)/\sqrt{2\pi}$ defined for $f,g \in L^2(\mu)$, hence $\mathbf{D}_Z = \mathbf{D}_X = I_K$. Basic properties of these polynomials yield that

$$\begin{aligned}\Psi_{lk} &= \mathbb{E}\left[H_{l-1}(\widetilde{z}_i)H_{k-1}(\widetilde{x}_i/\sqrt{\zeta^2+\sigma^2})\right]\\ &= \mathbb{E}\left[H_{l-1}(\widetilde{z}_i)\mathbb{E}\left[H_{k-1}(\widetilde{x}_i/\sqrt{\zeta^2+\sigma^2})|\widetilde{z}_i\right]\right]\\ &= \mathbb{E}\left[H_{l-1}(\widetilde{z}_i)\left(\frac{\zeta}{\sqrt{\zeta^2+\sigma^2}}\right)^{k-1}H_{k-1}(\widetilde{z}_i)\right]\\ &= \left(\frac{\zeta}{\sqrt{\zeta^2+\sigma^2}}\right)^{k-1}\mathbb{1}\{l=k\}.\end{aligned}$$

□

Example O3. Let $I^c = \{1\}$, x_{1i} be endogenous, $J = \{1,2\}$. Assume that we do not use excluded variables to instrument x_{1i}, $L = K-1$, $\rho \in \mathbb{R}^L$, $\Psi = (\rho\ I_L)$. We have, among others,

$$\kappa_{1,J} \geq \max_{S\subseteq[L]}\left(\frac{1}{2}\left(\sqrt{\frac{1-\tau_X}{1+\tau_X}}-c\right)-\frac{1+|\rho_1|}{|\rho_S|_1}\right)\left(\frac{(1+|\rho_1|)|S|}{|\rho_S|_1}+1\right)^{-1}.$$

Proof. We have $|\Psi\Delta|_\infty = \max\left(|\rho_1\Delta_1+\Delta_2|,\dots,|\rho_L\Delta_1+\Delta_{L+1}|\right)$, hence, for all $l \in [L]$, $|\rho_l||\Delta_1| \leq |\Psi\Delta|_\infty + |\Delta_{l+1}|$, which by summing up these inequalities yield

$$\begin{aligned}|\rho_S|_1|\Delta_1| &\leq |S|\,|\Psi\Delta|_\infty + \sum_{l\in S}|\Delta_{l+1}|\\ &\leq |S|\,|\Psi\Delta|_\infty + |\Delta|_1.\end{aligned}$$

Now, because $J = \{1,2\}$, we have, using $\Delta \in C_J$ and the triangle inequality,

$$\begin{aligned}\frac{1}{2}\left(\sqrt{\frac{1-\tau_X}{1+\tau_X}}-c\right)|\Delta|_1 &\leq |\Delta_1|+|\Delta_2|\\ &\leq (1+|\rho_1|)|\Delta_1| + |\rho_1\Delta_1+\Delta_2|\\ &\leq \frac{1+|\rho_1|}{|\rho_S|_1}|\rho_S|_1\,|\Delta_1| + |\Psi\Delta|_\infty\\ &\leq \frac{1+|\rho_1|}{|\rho_S|_1}\left(|S||\Psi\Delta|_\infty + |\Delta|_1\right) + |\Psi\Delta|_\infty,\end{aligned}$$

hence the result follows from

$$\left(\frac{1}{2}\left(\sqrt{\frac{1-\tau_X}{1+\tau_X}}-c\right)-\frac{1+|\rho_1|}{|\rho_S|_1}\right)|\Delta|_1 \leq \left(\frac{(1+|\rho_1|)|S|}{|\rho_S|_1}+1\right)|\Psi\Delta|_\infty.$$

□

A.8. **The *STIV* Estimator with Linear Projection Instrument.** The 2SLS is a leading method when the structural equation is low-dimensional. Proceeding in two stages is problematic when L is of the order of n, when, in case (1.2) is maintained, the linear projection of the endogenous regressor on the instruments is not sparse or approximately sparse (in the presence of many weak instruments) or $z \to \mathbb{E}[x_i|z_i = z]$ is not smooth. This is even more problematic when there are many endogenous variables and/or L is larger than n. In these cases, the first stage might not even be consistent and relying on the plug-in principle for inference is not possible.

In this section, we analyze the case where $L > K$, K and L can be much larger n, and the linear projections and (1.1) are sparse. We illustrate the method in a simulation study in Section 9 where we find that: even for very sparse models, the baseline one-stage method with all the instruments has smaller confidence sets than the two-stage method akin to 2SLS.

For simplicity, we take $P = [K]$ and $\mathcal{R} = \mathbb{R}^K$ and assume that there is only one endogenous regressor $(x_{1i})_{i=1}^n$ in (1.1). We write reduced form (or first stage) equation as

$$x_{1i} = \sum_{l=1}^{L} z_{li}\zeta_l + v_i, \quad i \in [n], \tag{A.54}$$

where $\sum_{l=1}^{L} z_{li}\zeta_l$ is the linear projection instrument, ζ_l are unknown coefficients and $\mathbb{E}[z_{li}v_i] = 0$.

The first stage consists in estimating the unknown coefficients ζ_l. If $L \geq K > n$ and if the reduced form model (A.54) is sparse or approximately sparse, it is natural to use a high-dimensional procedure, such as the Lasso, the Dantzig selector or the Square-root Lasso to find estimators $\widehat{\zeta}_l$ of the coefficients. Denote by $(\widehat{\zeta}, \widehat{\sigma}_1)$ the *STIV* estimator with parameter $c = c_1 \in (0, r^{-1})$ for the reduced form equation model. Our analysis is now carried out on the event

$$\mathcal{G} \triangleq \left\{ \max\left(\max_{l \in [L]} \frac{|\mathbb{E}_n[Z_l V]|}{\sqrt{\mathbb{E}_n[Z_l^2]\mathbb{E}_n[V^2]}}, \max_{k \in [K]} \frac{\left|\mathbb{E}_n[\widetilde{Z}_k U(\beta)]\right|}{\sqrt{\mathbb{E}_n[\widetilde{Z}_k^2]\mathbb{E}_n[U(\beta)^2]}} \right) \leq r \right\} \tag{A.55}$$

where $\widetilde{Z}_k$ are the exogenous regressors in the structural equation, the linear projection instrument V stands for a generic variable corresponding to the v_i's from the reduced form equation, and r is adjusted so that $\mathbb{P}(\mathcal{G}) \geq 1 - \alpha$. Since there is no access to the theoretical linear projection instrument, we adjust r as usual, excluding the linear projection instrument from the maximum, and setting $\alpha = 0.5(L + K - 1)/(L + K)$. This is the usual union bound scaling (see, *e.g.*, Scenarii 1-4).

We can construct the confidence sets for the parameters ζ and β under all three cases discussed in Section 6.5. We present the case where we have a sparsity certificate s_1 for ζ. We obtain, analogously to Theorem 5.1, that for all $c_1 \in (0, r^{-1})$ and using the notation $\widehat{\kappa}_1^1(s_1)$ and $\widehat{\theta}_\kappa(s_1)$ to make precise that these quantities are related to the estimation of the high-dimensional reduced form equation

$$\left|\widehat{\mathbf{D}}_{\mathbf{Z}}^{-1}\left(\widehat{\zeta}-\zeta\right)\right|_1 \le \frac{2\widehat{\sigma}_1 r\widehat{\theta}_{\kappa}(s_1)}{\widehat{\kappa}_1^1(s_1)} \triangleq C_1(s_1); \tag{A.56}$$

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(z_i^\top\left(\widehat{\zeta}-\zeta\right)\right)^2} \le \frac{2\widehat{\sigma}_1 r\widehat{\theta}_{\kappa}(s_1)}{\sqrt{\widehat{\kappa}_1^1(s_1)}} \triangleq C_2(s_1). \tag{A.57}$$

The second stage makes use of the estimated instrument $(z_i^\top\widehat{\zeta})_{i=1}^n$ to obtain confidence sets for the vector of coefficients in the structural equation. We use a modified *STIV* estimator which differs from the original one in that we replace $\widehat{\mathcal{I}}$ by the enlarged set

$$\widehat{\mathcal{I}}^2(r,\sigma) \triangleq \left\{(\beta,\sigma):\ \beta\in\mathbb{R}^K,\ \sigma>0,\ \left|\frac{1}{n}\widehat{\mathbf{D}}_{\mathbf{Z}}^2\left(\mathbf{Z}^2\right)^\top(\mathbf{Y}-\mathbf{X}\beta)\right|_\infty \le \sigma r,\ \widehat{Q}(\beta)\le\sigma^2\right\}, \tag{A.58}$$

where $\widehat{\mathbf{D}}_{\mathbf{Z}}^2$ is a $K\times K$ diagonal matrix such that $\left(\widehat{\mathbf{D}}_{\mathbf{Z}}^2\right)_{11} = \left(C_1(s_1)+C_2(s_1)+\mathbb{E}_n\left[\left(\widehat{\zeta}^\top Z\right)^2\right]^{1/2}\right)^{-1}$, $\left(\widehat{\mathbf{D}}_{\mathbf{Z}}^2\right)_{kk} = (\widehat{\mathbf{D}}_{\mathbf{X}})_{kk}$ for $k=2,\dots,K$, and $\mathbf{Z}^2$ is the stacked matrix of the estimated linear projection instrument $\left(z_i^\top\widehat{\zeta}\right)_{i=1}^n$ and the exogenous regressors. We enlarge the *IV*-constraint set to account for the estimation error in the linear projection instrument. We now define a new $\widehat{\Psi}$, which differs from the original one in that we replace $\mathbf{Z}$ by $\mathbf{Z}^2$ and $\widehat{\mathbf{D}}_{\mathbf{Z}}$ by $\widehat{\mathbf{D}}_{\mathbf{Z}}^2$. We assign the upper index (2) to the sensitivities corresponding to this new matrix $\widehat{\Psi}$, for example, $\widehat{\kappa}^2_{1,J(\beta),J(\beta)}$.

Theorem A.5. *For all $\beta\in\mathcal{B}_s$, on $\mathcal{G}$, for all solution $(\widehat{\beta},\widehat{\sigma})$ of the minimization problem (3.7) where we replace $\widehat{\mathcal{I}}$ by $\widehat{\mathcal{I}}^2$ and c by c_2, for all c_2 in $(0,r^{-1})$, $q\in[1,\infty]$, and $T\subseteq[K]$, we have*

$$\left|\widehat{\mathbf{D}}_{\mathbf{X}}^{-1}\left(\widehat{\beta}-\beta\right)_T\right|_q \le \frac{2\widehat{\sigma}r\widehat{\theta}^2(s_2)}{\widehat{\kappa}^2_{q,T}(s)}. \tag{A.59}$$

Proof of Theorem A.5. Take $\beta\in\mathcal{B}_s$, we have on $\mathcal{G}$, where $\mathcal{G}$ is defined in Section 3.5 adding the extra instrument $\zeta^\top z_i$ for $i\in[n]$.

$$\begin{aligned}
\frac{1}{n}|\widehat{\zeta}^\top\mathbf{Z}^\top\mathbf{U}| &\le |\widehat{\mathbf{D}}_{\mathbf{Z}}^{-1}(\widehat{\zeta}-\zeta)|_1\sqrt{\widehat{Q}(\beta)}r + \frac{1}{n}|(\zeta^T\mathbf{Z})^\top\mathbf{U}|\\
&\le \left(C_1(r,s_1)+\mathbb{E}_n[(\zeta^\top Z)^2]^{1/2}\right)\sqrt{\widehat{Q}(\beta)}r\\
&\le \left(C_1(r,s_1)+C_2(r,s_1)+\mathbb{E}_n[(\widehat{\zeta}^\top Z)^2]^{1/2}\right)\sqrt{\widehat{Q}(\beta)}r.
\end{aligned}$$

The rest of the proof is the same as for Theorem 4.1. Equation (A.59) is a consequence of Theorem 5.1 calculating the value of $c_b(s)$ when $I^c=\{1\}$. □

Inequality (A.59) yields uniform joint confidence sets for all $k \in [K]$, c_2 in $(0, r^{-1})$:

$$\left|\widehat{\beta}_k - \beta_k\right| \le \frac{2\widehat{\sigma} r \theta^2(s_2)}{\mathbb{E}_n[X_k^2]^{1/2}\widehat{\kappa}_k^{(2)*}} \tag{A.60}$$

with finite sample validity under Scenarii 1-3. One can also obtain adaptive confidence sets under a beta-min assumption, using the plug-in strategy where we replace $J(\beta)$ by an estimate $\widehat{J}$, as well as rates of convergence and model selection results similar to those of Section 6.2.

We illustrate numerically this method. We take $n = 8000$, $K = 70$, $L = 100$ and $I^c = \{1\}$. We take $\zeta_3 = -0.5$, $\zeta_4 = 1$, $\zeta_{98} = -1$, $\zeta_{99} = 1$, $\zeta_{100} = 0.5$. All other entries of ζ are equal to zero and the remaining part of the data generating process is that of Section 9. Tables 18 and 19 present the simulation results for the one-stage and the two-stage *STIV* estimator with estimated linear projection instrument. Lower bounds on the sensitivities based on the sparsity certificate for $s = 5$ yield: $C_1(5) = 1.125$, $C_2(5) = 0.308$, and $C_\infty(5) = 2\widehat{\sigma}_1 r \theta_1(5)/\widehat{\kappa}_\infty^1(5) = 0.112$. We present in Table 19 the results based on the the first stage. We set $\alpha = 0.05(L + K - 1)/(L + K) = 0.0471$. The value r for the two-stage approach is computed with this α by excluding the linear projection instrument from the maximum in (A.55).

TABLE 18. Sparse reduced form, sparsity certificate, one stage

	$\beta_{l,10}$	$\beta_{l,9}$	$\beta_{l,8}$	$\beta_{l,7}$	$\beta_{l,6}$	$\beta_{l,5}$	$\beta_{l,4}$	$\widehat{\beta}$	Selection	$\beta_{u,4}$	$\beta_{u,5}$	$\beta_{u,6}$	$\beta_{u,7}$	$\beta_{u,8}$	$\beta_{u,9}$	$\beta_{u,10}$
β_1	0.83	0.83	0.84	0.84	0.84	0.85	0.86	0.97	≥ 10	1.08	1.08	1.09	1.09	1.1	1.1	1.1
β_2	-2.12	-2.11	-2.11	-2.11	-2.1	-2.09	-2.09	-1.97	≥ 10	-1.85	-1.84	-1.83	-1.83	-1.82	-1.82	-1.82
β_3	-0.61	-0.6	-0.6	-0.6	-0.59	-0.58	-0.58	-0.46	≥ 10	-0.33	-0.33	-0.32	-0.32	-0.31	-0.31	-0.3
β_4	-0.02	-0.01	-0.01	0	0	0.01	0.02	0.19	≤ 6	0.35	0.36	0.37	0.38	0.38	0.39	0.39
β_5	-1.19	-1.19	-1.18	-1.18	-1.17	-1.16	-1.14	-0.93	≥ 10	-0.72	-0.71	-0.69	-0.68	-0.68	-0.67	-0.67
β_6	-0.14	-0.14	-0.14	-0.13	-0.13	-0.12	-0.11	0	0	0.11	0.12	0.13	0.13	0.14	0.14	0.14
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_{70}	-0.14	-0.14	-0.14	-0.13	-0.13	-0.12	-0.12	0	0	0.12	0.12	0.13	0.13	0.14	0.14	0.14

Here: $r = 0.0388$, $c = 0.298$ and $\widehat{\sigma} = 1.014$.

The two-stage method gives wider confidence sets than the one-stage method. For brevity, we do not display the sensitivities. Noteworthy, the two-stage method yields smaller sensitivities for all regressors. Since the constants $C_1(s)$, $C_2(s)$ can be too large, we construct the confidence sets for the overly optimistic case where $C_1 = C_2 = 0$. These sets are obviously not valid because they ignore the estimation error from the first stage. We find that they are larger than those of the one-stage method.

Table 19. Sparse reduced form, sparsity certificate, two stage

first stage:

$\widehat{\zeta}_1$	$\widehat{\zeta}_2$	$\widehat{\zeta}_3$	$\widehat{\zeta}_4$	$\widehat{\zeta}_5$	$\widehat{\zeta}_{97}$	$\widehat{\zeta}_{98}$	$\widehat{\zeta}_{99}$	$\widehat{\zeta}_{100}$	$\widehat{\sigma}_1$	c_1
0	0	-0.48	0.97	0	0	-0.98	1.00	0.47	1.02	0.188

Second stage based on $C_1(5)$ and $C_2(5)$, with $c = 0.479$ and $\widehat{\sigma} = 1.042$:

	$\beta_{l,10}$	$\beta_{l,9}$	$\beta_{l,8}$	$\beta_{l,7}$	$\beta_{l,6}$	$\beta_{l,5}$	$\beta_{l,4}$	$\widehat{\beta}$	Selection	$\beta_{u,4}$	$\beta_{u,5}$	$\beta_{u,6}$	$\beta_{u,7}$	$\beta_{u,8}$	$\beta_{u,9}$	$\beta_{u,10}$
β_1	0.56	0.57	0.58	0.59	0.6	0.61	0.63	0.91	≥ 10	1.2	1.21	1.22	1.23	1.24	1.25	1.26
β_2	-2.2	-2.19	-2.18	-2.17	-2.16	-2.15	-2.13	-1.96	≥ 10	-1.8	-1.78	-1.77	-1.76	-1.75	-1.74	-1.73
β_3	-0.68	-0.67	-0.66	-0.66	-0.64	-0.63	-0.62	-0.45	≥ 10	-0.29	-0.27	-0.26	-0.25	-0.24	-0.23	-0.22
β_4	-0.22	-0.21	-0.2	-0.19	-0.18	-0.16	-0.14	0.15	≤ 5	0.45	0.47	0.49	0.5	0.51	0.52	0.53
β_5	-1.41	-1.39	-1.38	-1.36	-1.34	-1.32	-1.29	-0.87	≥ 10	-0.46	-0.43	-0.41	-0.39	-0.37	-0.35	-0.34
β_6	-0.22	-0.21	-0.2	-0.19	-0.18	-0.17	-0.16	0	0	0.16	0.17	0.18	0.19	0.2	0.21	0.22
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
β_{70}	-0.22	-0.21	-0.2	-0.2	-0.19	-0.18	-0.16	0	0	0.16	0.18	0.19	0.2	0.2	0.21	0.22

Second stage based on the too optimistic choice $C_1 = 0$ and $C_2 = 0$ with $c = 0.376$ and $\widehat{\sigma} = 1.001$:

	$\beta_{l,10}$	$\beta_{l,9}$	$\beta_{l,8}$	$\beta_{l,7}$	$\beta_{l,6}$	$\beta_{l,5}$	$\beta_{l,4}$	$\widehat{\beta}$	Selection	$\beta_{u,4}$	$\beta_{u,5}$	$\beta_{u,6}$	$\beta_{u,7}$	$\beta_{u,8}$	$\beta_{u,9}$	$\beta_{u,10}$
β_1	0.79	0.79	0.79	0.8	0.81	0.81	0.82	1	≥ 10	1.17	1.18	1.18	1.19	1.2	1.2	1.2
β_2	-2.18	-2.18	-2.17	-2.17	-2.16	-2.15	-2.14	-2.01	≥ 10	-1.87	-1.86	-1.85	-1.84	-1.84	-1.83	-1.83
β_3	-0.68	-0.67	-0.67	-0.66	-0.65	-0.64	-0.63	-0.5	≥ 10	-0.36	-0.35	-0.34	-0.33	-0.32	-0.32	-0.31
β_4	-0.03	-0.03	-0.02	-0.01	0	0.01	0.02	0.24	≤ 5	0.45	0.46	0.47	0.48	0.49	0.5	0.5
β_5	-1.35	-1.35	-1.34	-1.33	-1.31	-1.3	-1.28	-0.99	≥ 10	-0.71	-0.69	-0.67	-0.66	-0.65	-0.64	-0.64
β_6	-0.16	-0.16	-0.15	-0.14	-0.13	-0.13	-0.12	0.01	0	0.14	0.15	0.16	0.17	0.18	0.18	0.19
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
β_{70}	-0.17	-0.17	-0.16	-0.16	-0.15	-0.14	-0.13	0	0	0.13	0.14	0.15	0.16	0.17	0.17	0.17

Everywhere: $r = 0.041$.

References

[1] Bertail, P. , E. Gauthérat, and H. Harari-Kermadec (2008): "Exponential Inequalities for Self Normalized Sums". *Electronic Communications in Probability,* 13, 628–640.

[2] Donoho, D. L., M. Elad, and V. N. Temlyakov (2006): "Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise". *IEEE Transactions on Information Theory,* 52, 6–18.

[3] Efron, B. (1969): "Student's t-test Under Symmetry Conditions". *Journal of American Statistical Society,* 64, 1278–1302.

[4] Jing, B.-Y., Q. M. Shao, and Q. Wang (2003): "Self-Normalized Cramér-Type Large Deviations for Independent Random Variables". *Annals of Probability,* 31, 2167–2215.

[5] Nesterov, Y. (2004): *Introductory Lectures on Convex Optimization.* Springer.

[6] Pinelis, I. (1994): "Probabilistic Problems and Hotelling's t^2 Test Under a Symmetry Condition". *Annals of Statistics,* 22, 357–368.