High-dimensional instrumental variables regression and confidence sets

Eric Gautier, Alexandre Tsybakov

HIGH-DIMENSIONAL INSTRUMENTAL VARIABLES REGRESSION AND CONFIDENCE SETS

ERIC GAUTIER AND ALEXANDRE B. TSYBAKOV

Abstract. We propose an instrumental variables method for inference in high-dimensional structural equations with endogenous regressors. The number of regressors K can be much larger than the sample size. A key ingredient is sparsity, *i.e.*, the vector of coefficients has many zeros, or approximate sparsity, *i.e.*, it is well approximated by a vector with many zeros. We can have less instruments than regressors and allow for partial identification. Our procedure, called *STIV* (Self Tuning Instrumental Variables) estimator, is realized as a solution of a conic program. The joint confidence sets can be obtained by solving K convex programs. We provide rates of convergence, model selection results and propose three types of joint confidence sets relying each on different assumptions on the parameter space. Under the stronger assumption they are adaptive. The results are uniform over a wide classes of distributions of the data and can have finite sample validity. When the number of instruments is too large or when one only has instruments for an endogenous regressor which are too weak, the confidence sets can have infinite volume with positive probability. This provides a simple one-stage procedure for inference robust to weak instruments which could also be used for low dimensional models. In our *IV* regression setting, the standard tools from the literature on sparsity, such as the restricted eigenvalue assumption are inapplicable. Therefore we develop new sharper sensitivity characteristics, as well as easy to compute data-driven bounds. All results apply to the particular case of the usual high-dimensional regression. We also present extensions to the high-dimensional framework of the two-stage least squares method and method to detect endogenous instruments given a set of exogenous instruments.

Date: This version: August 2014. This is a revision of the 12 May 2011 preprint arXiv:1105.2454.

Keywords: Instrumental variables, sparsity, *STIV* estimator, endogeneity, high-dimensional regression, conic programming, heteroscedasticity, confidence regions, non-Gaussian errors, variable selection, unknown variance, sign consistency.

We thank James Stock and three anonymous referees for comments that greatly improved this paper. We also thank Azeem Shaikh and the seminar participants 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, Zurich as well as participants of SPA, Saint-Flour, ERCIM 2011 and the 2012 CIREQ conference on High Dimensional Problems in Econometrics and 4th French Econometrics Conference for helpful comments. We acknowledge financial support from the grants ANR-13-BSH1-0004 and Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047).

1. Introduction

We consider a structural model of the form

$$y_i = x_i^T \beta + u_i, \quad i = 1, \dots, n, \tag{1.1}$$

where x_i are random vectors of regressors of dimension K and u_i is a zero-mean random error possibly correlated with some or many regressors, called endogenous - as opposed to exogenous - regressors. Endogeneity occurs when a regressor is determined simultaneously with the response variable y_i, when the error term u_i absorbs an unobserved variable which is partially correlated with y_i, in the errors-in-variables model when the measurement error is independent of the underlying variable. This paper provides a computationaly efficient method based on convex optimization for (robust) inference under the sparsity scenario when K is large, possibly much larger than n, and one of the following assumptions is satisfied:

(i) only few coefficients β_k are non-zero (β is *sparse*),

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

We rely on instrumental variables (instruments), *i.e.* random vectors z_i of dimension L which satisfy

$$\forall i = 1, \dots, n, \ \mathbb{E}[z_i u_i] = 0, \tag{1.2}$$

where $\mathbb{E}[\,\cdot\,]$ denotes the expectation. We assume that we have n independent, not necessarily identically distributed, realizations (y_i, x_i^T, z_i^T), $i = 1, \dots, n$ drawn from a probability $\mathbb{P}$. We consider the problem of inference on the set of vectors compatible with (1.1) and (1.2)

$$\mathcal{I}dent \triangleq \left\{\beta : \ \forall i = 1, \dots, n, \ \mathbb{E}[z_i(y_i - x_i^T\beta)] = 0\right\} \ . \tag{1.3}$$

When this affine space is reduced to a point the model is point identified. It is possible to impose some restrictions like known signs, prior upper bounds on the size of the coefficients or, as we study in more details, the number of non-zero coefficients. One typically considers as instrument the regressor which is identically equal to 1, which gives rise to a constant in (1.1), and all regressors which we know are exogenous. For each endogenous regressor, one should find instruments which are exogenous while correlated with the endogenous regressor. Usually such instruments are excluded from the right-hand side of (1.1) and one is in a situation where $L \geq K$. We do not assume this. We allow the last type of instruments that we described to have a direct affect and appear on the right-hand side of (1.1). There, sparsity corresponds to exclusion restrictions which are not known in advance. The large K relative to n problem is a very natural framework in many empirical applications.

Example 1. Economic theory is not explicit enough about which variables belong to the true model. Sala-i-Martin (1997) and Belloni and Chernozhukov (2001b) give examples from development economics where it is unclear which growth determinant should be included in the model. More than 140 growth determinants have been proposed in the litterature and we are typically faced with the situation where n is smaller than K and endogeneity. Searching among 2^{140} (a tredecillion) submodels, for example if one wants to implement BIC, is simply impossible.

Example 2. Rich heterogeneity. When there is a rich heterogeneity one usually wants to control for many variables and possibly interactions, or to carry out a stratified analysis where models are estimated in small population sub-groups (*e.g.*, groups defined by the value of an exogenous discrete variable). In both cases K can be large relative to n.

Example 3. Many nonlinear functions of an endogenous regressor. This occurs when one considers a structural equation of the form

$$y_i = \sum_{k=1}^{K} \alpha_k f_k(x_{\text{end},i}) + u_i$$

where $x_{\text{end},i}$ is a low dimensional vector of endogenous regressor and $(f_k)_{k=1}^K$ are many functions to capture a wide variety of nonlinearities. Exogenous regressors could also be included. Belloni and Chernozhukov (2011b) give an example of a wage equation with many transformations of education. In Engle curves models it is important to include nonlinearities in the total budget (see, *e.g.*, Blundell, Chen and Kristensen (2007)). When one estimates Engle curves using aggregate data, n is usually small. Education in a wage equation and total budget in Engle curves are endogenous.

Example 4. Many exogenous regressors due to nonlinearities. Similarly, one can have to estimate a model of the form

$$y_i = x_{\text{end},i}^T \beta_{\text{end}} + \sum_{k=1}^{K_c} \alpha_k f_k(x_{\text{ex},i}) + u_i$$

where $x_{\text{ex},i}$ and $x_{\text{end},i}$ are respectively vectors of exogenous and endogenous regressors.

Example 5. Many control variables to justify the use of an instrument. Suppose that we are interested in the parameter $\overline{\beta}$ in

$$y_i = \overline{x}_i^T \overline{\beta} + v_i, \tag{1.4}$$

where some of the variables in $\overline{x}_i$ are endogenous and we want to use as an instrument a variable z_i that does not satisfy $\mathbb{E}[z_i v_i] = 0$. Suppose that we also have observations of vectors of controls w_i

such that $\mathbb{E}[v_i|w_i, z_i] = \mathbb{E}[v_i|w_i]$. Then we can rewrite (1.4) as

$$y_i = \overline{x}_i^T \overline{\beta} + f(w_i) + u_i$$

where $f(w_i) = \mathbb{E}[v_i|w_i]$ and $u_i = v_i - \mathbb{E}[v_i|w_i, z_i]$ is such that $\mathbb{E}[z_i u_i] = 0$. If for a sufficiently large and good set of functions $f = \sum_{k=1}^{K_c} \alpha_k f_k$ we get back to our original model.

Statistical inference under the sparsity scenario when the dimension is larger than the sample size is now an active and challenging field. The most studied techniques are the Lasso, the Dantzig selector (see, *e.g.*, Candès and Tao (2007), Bickel, Ritov and Tsybakov (2009); more references can be found in the recent book by Bühlmann and van de Geer (2011), as well as in the lecture notes by Koltchinskii (2011))and aggregation methods (see Dalalyan and Tsybakov (2008), Rigollet and Tsybakov (2011) and the papers cited therein). This literature proposes methods that are computationally feasible in high-dimensional setting. 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 very small. The Dantzig selector is solution of a simple linear program. Some important extensions to model from econometrics have been obtained by Belloni and Chernozhukov (2011a) who study the ℓ_1-penalized quantile regression and by Belloni, Chen, Chernozhukov et al. (2012) who use Lasso type methods to estimate the so-called optimal instruments and obtain optimal confidence for low dimensional structural equations. Caner (2009) studies a Lasso-type GMM estimator. Rosenbaum and Tsybakov (2010) deal with the high-dimensional errors-in-variables problem. The high-dimensional setting in a structural model with endogenous regressors that we are considering here has not yet been analyzed. This paper presents an infererence procedure based on a new estimator that we call the *STIV* estimator (Self Tuning Instrumental Variables estimator).

The *STIV* estimator is an extension of the Dantzig selector of Candès and Tao (2007). It can obviously also be applied in the high-dimensional regression model without endogeneity simply using $z_i = x_i$. The results of this paper extend those on the Dantzig selector (see Candès and Tao (2007), Bickel, Ritov and Tsybakov (2009)) in several ways: By allowing for endogenous regressors when instruments are available, by working under weaker sensitivity assumptions than the restricted eigenvalue assumption of Bickel, Ritov and Tsybakov (2009), which in turns yields tighter bounds, by imposing weak distributional assumptions, by introducing a procedure independent of the noise level and by providing uniform joint confidence sets. The *STIV* estimator is also related to the Square-root Lasso of Belloni, Chernozhukov and Wang (2011). The Square-root Lasso is a method independent of the variance of the errors.The *STIV* estimator adds extra linear constraints coming

from the restrictions (1.2) which allows to deal with endogenous regressors. The implementation of the *STIV* estimator also corresponds to solving a conic program.

The theoretical results of this paper include rates of convergence, variable selection and joint confidence sets for sparse vectors. The first and second classes of confidence sets are based on an the estimation of the set of the non-zero coefficients. Excluding from the high-dimensional parameter space models which are too close to the true model, we obtain an upper estimate $\widehat{J}_2$ on the set of non-zero coefficients. Based on this set, one can obtain a conservative confidence set. Assuming both an upper bound on the number of non-zero coefficients and that the non-zero coefficients are large enough, we obtain a smaller set $\widehat{J}_1$. It corresponds to the true set of non-zero coordinates with probability close to 1. The confidence sets which are based on the set $\widehat{J}_1$ are adaptive in a sense that will be clear in Section 8.3.1. The corresponding confidence sets are obtained by solving $|\widehat{J}|(2K+2)$ simple convex programs where $\widehat{J} = \widehat{J}_2$ or $\widehat{J} = \widehat{J}_1$. The third class of confidence sets requires an upper bound on the number of non-zero coefficients but these coefficients can be arbitrarily close to 0. A base solution to obtain these joint confidence sets requires to solve K convex programs. It is possible to obtain sharper confidence sets by solving $2K$ programs for each coefficient of interest. So, our method is easy and fast to implement in practice. This is an attractive feature even when $K \ll n$. We also present tighter confidence sets that can be obtained in specific situations such as when the number of non-zero coefficients and/or endogenous regressors is small. Similar confidence sets can be obtained for approximately sparse models. For example, for the third type of confidence sets, one uses an upper bound on the size of the best aproximating sparse model.

The core of our analysis is non-asymptotic. We do not make any restriction on the number of regressors nor on the number of instruments relative to the number of potential regressors. We provide a partially identified analysis to allow for the situation where $K > L$. For example, among the large number of regressors, only L are known to be exogenous and used as their own instruments while the number of non-zero coefficients is possibly smaller than L. The results are uniform among wide classes of data generating processes that allow for non-Gaussian errors and sometimes heteroscedasticity. We consider several scenarii on the data generating process under which the confidence sets can have finite sample validity. In the presence of weak instruments the finite sample distribution of the two-stage least squares can be highly non normal and even bimodal (see, *e.g.*, Nelson and Startz (1990a,b)) and inference usually relies on non-standard asymptotics (see, *e.g.*, Stock, Wright and Yogo (2002) and Andrews and Stock (2007) for a review and the references cited therein). We do not need assumptions on the strength of the instruments, and no preliminary test for weak instruments is required. The size

of the confidence sets depends on the best instrument. If all the instruments are individually weak or when the number of instruments is too large the method yields infinite volume confidence sets. We also show that the *STIV* estimator can be used for low dimensional structural equations when there is no uncertainty on the relevance of the regressors and provides confidence sets that are easy to calculate, robust to weak instruments, and do not require a pretest. We present an extension to the high-dimensional framework of the two-stage least squares method where both the structural and reduced form models are high-dimensional. Though in the literature there are no results on optimal joint confidence sets for high-dimensional regression or high-dimensional structural equations, this is a natural procedure to look at. Unlike for low dimensional structural equations, we observe that if one accounts for the uncertainty coming for the estimation of the first stage equation then the two-stage method usually gives larger joint confidence sets than the easy one-stage method. Optimal confidence sets for low dimensional parameters in high-dimensional structural equations will appear in a different paper and uses the *STIV* estimator and the joint confidence sets as a core ingredient. We also present a two-stage method to detect endogenous instruments given a preliminary set of valid instruments. Again, the second stage confidence sets heavily rely on the availibility of joint confidence sets for the first stage. Finally, we conclude with a simulation study. All proofs are given in the appendix.

2. Basic Definitions and Notation

We set $\mathbf{Y} = (y_1, \dots, y_n)^T$, $\mathbf{z}_l = (z_{l1}, \dots, z_{ln})^T$ for $l = 1, \dots, L$, $\mathbf{U} = (u_1, \dots, u_n)^T$, and we denote by $\mathbf{X}$ and $\mathbf{Z}$ the matrices of dimension $n \times K$ and $n \times L$ respectively with rows x_i^T and z_i^T, $i = 1, \dots, n$. The sample mean is denoted by $\mathbb{E}_n[\,\cdot\,]$. We use the notation

$$\mathbb{E}_n[X_k^a U^b] \triangleq \frac{1}{n}\sum_{i=1}^n x_{ki}^a u_i^b, \quad \mathbb{E}_n[Z_l^a U^b] \triangleq \frac{1}{n}\sum_{i=1}^n z_{li}^a u_i^b,$$

where x_{ki} is the kth component of vector x_i, and z_{li} is the lth component of z_i for some $k \in \{1, \dots, K\}$, $l \in \{1, \dots, L\}$, $a \geq 0, b \geq 0$. Similarly, we define the sample mean for vectors; for example, $\mathbb{E}_n[UX]$ is a row vector with components $\mathbb{E}_n[UX_k]$. We also define the corresponding population means:

$$\mathbb{E}[X_k^a U^b] \triangleq \frac{1}{n}\sum_{i=1}^n \mathbb{E}[x_{ki}^a u_i^b], \quad \mathbb{E}[Z_l^a U^b] \triangleq \frac{1}{n}\sum_{i=1}^n \mathbb{E}[z_{li}^a u_i^b]\ .$$

We use the normalization matrices for $\mathbf{D_X}$ and $\mathbf{D_Z}$ to rescale $\mathbf{X}$ and $\mathbf{Z}$. They are diagonal $K \times K$, respectively $L \times L$, matrices. The diagonal entries of $\mathbf{D_X}$ are $(\mathbf{D_X})_{kk} = \mathbb{E}_n[X_k^2]^{-1/2}$ for $k = 1, \dots, K$. The diagonal entries of $\mathbf{D_Z}$ are $(\mathbf{D_Z})_{ll} = \mathbb{E}_n[Z_l^2]^{-1/2}$ for $l = 1, \dots, L$.

For a vector $\beta \in \mathbb{R}^K$, let $J(\beta) = \{k \in \{1, \dots, K\} :\ \beta_k \neq 0\}$ be its support. We denote by $|J|$ the

cardinality of a set $J \subseteq \{1,\dots,K\}$ and by J^c its complement: $J^c = \{1,\dots,K\} \setminus J$. We denote by J_{ex} the subset of $\{1,\dots,K\}$ corresponding to the indices of the regressors that we know to be exogenous. It can be a subset of all the exogenous regressors. The regressors whose index is in J_{ex} are used as their own instruments. The ℓ_p norm of a vector Δ is denoted by $|\Delta|_p$, $1 \leq p \leq \infty$. For $\Delta = (\Delta_1, \dots \Delta_K)^T \in \mathbb{R}^K$ and a set of indices $J \subseteq \{1,\dots,K\}$, we define $\Delta_J \triangleq (\Delta_1 \mathbb{1}_{\{1\in J\}}, \dots, \Delta_K \mathbb{1}_{\{K\in J\}})^T$, where $\mathbb{1}_{\{\cdot\}}$ is the indicator function. For a vector $\beta \in \mathbb{R}^K$, we set $\overrightarrow{\text{sign}(\beta)} \triangleq (\text{sign}(\beta_1), \dots, \text{sign}(\beta_K))$ where

$$\text{sign}(t) \triangleq \begin{cases} 1 & \text{if } t > 0 \\ 0 & \text{if } t = 0 \\ -1 & \text{if } t < 0 \end{cases}$$

For $a \in \mathbb{R}$, we set $a_+ \triangleq \max(0, a)$. We use the convention $\inf \emptyset \triangleq \infty$.

We will sometimes restrict the class of models to sparse models and make inference on the *sparse identifiable parameters*:

$$\mathcal{B}_s = \mathcal{I}dent \bigcap \{\beta : \ |J(\beta)| \leq s\}$$

for some upper bound s in $\{1,\dots,K\}$ on the sparsity. This is the set of vectors of coefficients compatible with (1) the moments restrictions and (2) a prior upper bound on the number of non-zero coefficients. These sets satisfy

$$\forall s \leq s' \leq K, \quad \mathcal{B}_s \subseteq \mathcal{B}_{s'} \subseteq \mathcal{B}_K = \mathcal{I}dent \ .$$

3. The *STIV* Estimator

The sample counterpart of the moment conditions (1.2) can be written in the form

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

This is a system of L equations with K unknown parameters. If $L > K$, it is overdetermined. In general $\text{rank}(\mathbf{Z}^T\mathbf{X}) \leq \min(K, L, n)$, thus when $L \leq K$ or when $n < K$ the matrix does not have full column rank. Furthermore, replacing the population equations (1.2) by (3.1) induces a huge error when L, K or both are larger that n. So, looking for the exact solution of (3.1) in high-dimensional settings makes no sense. However, we can stabilize the problem by restricting our attention to a suitable "small" candidate set of vectors β, for example, to those satisfying the constraint

$$\left|\frac{1}{n}\mathbf{Z}^T(\mathbf{Y} - \mathbf{X}\beta)\right|_\infty \leq \tau, \tag{3.2}$$

where $\tau > 0$ is chosen such that (3.2) holds for β in $\mathcal{I}dent$ with high probability. We can then refine the search of the estimator in this "small" random set of vectors β by minimizing an appropriate

criterion such as, for example, the ℓ_1 norm of β, which leads to a simple optimization problem. It is possible to consider different small sets in (3.2), however the use of the sup-norm makes the inference robust to the presence many weak or irrelevant instruments as explained in Section 6.

In what follows, we use this idea with suitable modifications. First, notice that it makes sense to normalize the matrix $\mathbf{Z}$. This is quite intuitive because, otherwise, the larger the instrumental variable, the more influential it is on the estimation of the vector of coefficients. The constraint (3.2) is modified as follows:

$$\left|\frac{1}{n}\mathbf{D_Z}\mathbf{Z}^T(\mathbf{Y}-\mathbf{X}\beta)\right|_\infty \le \tau. \tag{3.3}$$

Along with the constraint of the form (3.3), we include more constraints to account for the unknown level σ of $\mathbb{E}_n[U^2]$. Specifically, we say that a pair $(\beta,\sigma)\in\mathbb{R}^K\times\mathbb{R}^+$ satisfies the *IV-constraint* if it belongs to the set

$$\widehat{\mathcal{I}} \triangleq \left\{(\beta,\sigma):\ \beta\in\mathbb{R}^K,\ \sigma>0,\ \left|\frac{1}{n}\mathbf{D_Z}\mathbf{Z}^T(\mathbf{Y}-\mathbf{X}\beta)\right|_\infty \le \sigma r,\ \widehat{Q}(\beta)\le\sigma^2\right\} \tag{3.4}$$

for some $r>0$, where the function $\widehat{Q}(\beta)$ is defined as

$$\widehat{Q}(\beta) \triangleq \frac{1}{n}\sum_{i=1}^n (y_i - x_i^T\beta)^2.$$

The choice of r depends on the class of distributions for the data generating process, the number of instruments, the sample size and the confidence level $1-\alpha$. We give the details in the next section. A typical ("reference") behavior is

$$r \sim \sqrt{\frac{\log L}{n}}\,. \tag{3.5}$$

The additional constraint $\widehat{Q}(\beta)\le\sigma^2$ is introduced in (3.4) because $\mathbb{E}[U^2]$ is not identified without instruments. For example, (1.1) allows the variance of u_i to be greater than the variance of y_i. This constraint is crucial to obtain uniform, possibly finite sample, confidence sets under various classes of data generating processes.

Definition 3.1. *We call the* STIV *estimator any solution* $(\widehat{\beta},\widehat{\sigma})$ *of the following minimization problem:*

$$\min_{(\beta,\sigma)\in\widehat{\mathcal{I}}} \left(\left|\mathbf{D}_\mathbf{X}^{-1}\beta\right|_1 + c\sigma\right), \tag{3.6}$$

where c *is a constant in* $(0, r^{-1})$.

The summand $c\sigma$ is included in the criterion to prevent from choosing σ arbitrarily large; indeed, the *IV*-constraint does not prevent from this.

We use $\widehat{\beta}$ as an estimator of β in $\mathcal{I}dent$ and use both $\widehat{\beta}$ and $\widehat{\sigma}$ to construct confidence sets. Finding a solution $(\widehat{\beta}, \widehat{\sigma})$ of the minimization problem (3.6) reduces to the following conic program.

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

$$\min_{(\beta,t,v,w)\in\mathcal{V}} \left(\sum_{k=1}^{K} w_k + c\frac{t}{\sqrt{n}} \right) \tag{3.7}$$

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

$$v = \mathbf{Y} - \mathbf{X}\beta, \qquad -rt\mathbf{1} \leq \frac{1}{\sqrt{n}}\mathbf{D_Z}\mathbf{Z}^T(\mathbf{Y} - \mathbf{X}\beta) \leq rt\mathbf{1},$$
$$-w \leq \mathbf{D_X}^{-1}\beta \leq w, \qquad w \geq \mathbf{0}, \qquad (t, v) \in C.$$

Here $\mathbf{0}$ and $\mathbf{1}$ are vectors of zeros and ones respectively, the inequality between vectors is understood in the componentwise sense, and C is a cone: $C \triangleq \{(t, v) \in \mathbb{R} \times \mathbb{R}^n : \ t \geq |v|_2\}$.

Conic programming is a standard tool in optimization and open source toolboxes are available to implement it (see, *e.g.*, Grant and Boyd (2013)). Computationally, conic programming starts to be difficult when K is of the order of several thousands. In a forthcoming paper, we will show that, under some conditions, is possible to replace the conic program (3.7) by a linear program as it is done in Gautier and Tsybakov (2013) for the usual regression setting without instruments.

Note that the *STIV* estimator is not necessarily unique. Minimizing the ℓ_1 criterion $\left|\mathbf{D_X}^{-1}\beta\right|_1$ is a convex relaxation of minimizing the ℓ_0 norm, *i.e.*, the number of non-zero components of β. This usually ensures that the resulting solution is sparse.

Remark 3.1. *If one knows in advance that some components of β are non-zero, they can be excluded from the ℓ_1 norm in* (3.6)*. The special case where the model is low dimensional and there is no uncertainty on which variable belongs to the model is treated in Section 9.1. This is important because it provides easily computable confidence sets which are robust to weak instruments.*

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 (3.3), which is not present in the conic program for the Square-root Lasso. This is due to the fact that we need to handle the endogeneity.

Our main findings about the *STIV* estimator can be sketched as follows. First, we obtain rates of convergence for $\left|\mathbf{D}_{\mathbf{X}}^{-1}(\widehat{\beta}-\beta)\right|_p$ for $1 \le p \le \infty$ of the order $O(c_{J(\beta)}^{1/p} r)$ for sufficiently sparse vectors. Here, r is essentially as in (3.5). The constant $c_{J(\beta)}$ is given in Table 7. It is of the order of $|J(\beta)|$ without endogeneity or when $0 < c < 1$. When the dimension is large relative to n, one uses $1 \le c < r^{-1}$. The rates then start to be influenced by the number of regressors that are not used as their own instruments. We also analyse the estimation of approximately sparse vectors. Second, we show that based on the *STIV* estimator, we can efficiently construct joint confidence sets for the components of sparse vectors β. We propose two approaches to address this issue. The first, that we call the *sparsity certificate* approach, is applicable when one knows an upper bound on the sparsity s. In the second approach, we use instead an estimator $\widehat{J}$ of the support $J(\beta)$, for example, the support of the *STIV* estimator or a thresholded *STIV* estimator. Both approaches are based on the bounds of Theorem 6.1 that can be stated as follows (here we only display the coordinate-wise bounds):

$$|\widehat{\beta}_k - \beta_k| \le \frac{2\widehat{\sigma} r A_k(|J(\beta)|)}{\mathbb{E}_n[X_k^2]^{1/2}}, \quad k = 1, \dots, K. \tag{3.8}$$

Here, $A_k(t)$ are some explicitly defined coefficients, such that $A_k(t)$ is monotone increasing in t. This motivates the sparsity certificate approach: replace $|J(\beta)|$ in (3.8) by a known upper bound s or display nested confidence sets for various values of s. We provide a base solution where the values $A_k(s)$ can be then computed by solving K simple convex programs. For the second approach, the joint confidence sets are of the form

$$|\widehat{\beta}_k - \beta_k| \le \frac{2\widehat{\sigma} r \bar{A}_k(\widehat{J})}{\mathbb{E}_n[X_k^2]^{1/2}}, \quad k = 1, \dots, K. \tag{3.9}$$

where the explicitly defined constant $\bar{A}_k(\widehat{J})$ is sharper than $A_k(s)$ when $|\widehat{J}| \le s$. In the general case, the computation of $\bar{A}_k(\widehat{J})$ reduces to solving $|\widehat{J}|(2K+2)$ simple convex programs. We also present refinements which are possible in various cases, for example, when there are few endogenous regressors.

4. Sensitivity Characteristics

In the usual linear regression in low dimension, when $\mathbf{Z} = \mathbf{X}$ and the Gram matrix $\mathbf{X}^T\mathbf{X}/n$ is positive definite, the sensitivity is given by the minimal eigenvalue of this matrix. In high-dimensional regression, the theory of the Lasso and the Dantzig selector comes up with a more sophisticated sensitivity analysis; there the Gram matrix cannot be positive definite and the eigenvalue conditions are imposed on its sufficiently small submatrices. This is typically expressed via the restricted isometry property of Candès and Tao (2007) or the more general restricted eigenvalue condition of Bickel,

Ritov and Tsybakov (2009). In our structural model with endogenous regressors, these sensitivity characteristics cannot be used, since instead of a symmetric Gram matrix we have a rectangular matrix $\mathbf{Z}^T\mathbf{X}/n$ involving the instruments. Since we include normalizations, we need to deal with the normalized version of this matrix,

$$\Psi_n \triangleq \frac{1}{n}\mathbf{D_Z}\mathbf{Z}^T\mathbf{X}\mathbf{D_X}.$$

In general, Ψ_n is not a square matrix. For $L = K$, it is a square matrix but, in the presence of at least one endogenous regressor, Ψ_n is not symmetric.

We now introduce some scalar sensitivity characteristics related to the action of the matrix Ψ_n on vectors in the cone

$$C_J \triangleq \left\{\Delta \in \mathbb{R}^K : \ |\Delta_{J^c}|_1 \leq |\Delta_J|_1 + cr|\Delta_{J_{\rm ex}}|_1 + c|\Delta_{J_{\rm ex}^c}|_1\right\} \tag{4.1}$$

where c is the constant in the definition of the *STIV* estimator, J is any subset of $\{1, \ldots, K\}$.

Remark 4.1. *When there are no endogenous regressors the cone can be written as*

$$C_J = \left\{\Delta \in \mathbb{R}^K : \ (1 - cr)|\Delta_{J^c}|_1 \leq (1 + cr)|\Delta_J|_1\right\} \ . \tag{4.2}$$

The use of such cones to define sensitivity characteristics is standard in the literature on Lasso and Dantzig selector (cf. Bickel, Ritov and Tsybakov (2009)).

If the cardinality of J is small, the vectors Δ in the cone C_J have a substantial part of their mass concentrated on a set of small cardinality. This is why C_J in (4.2) is sometimes called the *cone of dominant coordinates*. The set J that will be used later is the set $J(\beta)$, which is small if β is sparse.

Given a subset $J_0 \subseteq \{1, \ldots, K\}$ and $p \in [1, \infty]$, we define the ℓ_p-J_0 *block sensitivity* as

$$\kappa_{p,J_0,J} \triangleq \inf_{\Delta \in C_J:\ |\Delta_{J_0}|_p = 1} |\Psi_n \Delta|_\infty \ . \tag{4.3}$$

By convention, we set $\kappa_{p,\varnothing,J} = \infty$. The sensitivities (4.3) depend both on c and $J_{\rm ex}$ (recall that $J_{\rm ex}$ is the set of potential regressors which are known in advance to be exogenous and used as their own instruments and not necessarily the actual set of exogenous potential regressors) but for brevity we do not make the dependence explicit. Similar but different quantities have been introduced in Ye and Zhang (2010) under the name of cone invertibility factors. They differ from the sensitivities in several respects, in particular, in the definition of the cone C_J and of the matrix Ψ_n. Moreover, unlike the cone invertibility factors, the sensitivities do not involve scaling by $|J(\beta)|^{1/p}$ in the definition. Indeed, by Proposition 4.1 below, the dependence of the sensitivities on $J(\beta)$ is more complex in the presence of endogenous regressors and we do not include any specific scaling for full generality.

Coordinate-wise sensitivities $\kappa^*_{k,J} \triangleq \kappa_{p,\{k\},J}$ correspond to singletons $J_0 = \{k\}$. Since (4.3) is invariant to replacing Δ by $-\Delta$, we also have

$$\kappa^*_{k,J} \triangleq \inf_{\Delta \in C_J:\ \Delta_k = 1} |\Psi_n \Delta|_\infty \ .$$

For the other extreme, $J_0 = \{1, \dots, K\}$ we use the shorthand notation $\kappa_{p,J}$, namely:

$$\kappa_{p,J} \triangleq \inf_{\Delta \in C_J:\ |\Delta|_p = 1} |\Psi_n \Delta|_\infty \ .$$

To explain the role of sensitivity characteristics, let us sketch here some elements of our argument. It will be clear from the proofs that we adjust r in the definition of $\widehat{\mathcal{I}}$ such that for $\Delta = \mathbf{D}_{\mathbf{X}}^{-1}(\widehat{\beta} - \beta)$ where $\beta \in \mathcal{I}dent$, with probability at least $1 - \alpha$ we have:

$$|\Psi_n \Delta|_\infty \le r(2\widehat{\sigma} + r\,|\Delta_{J_{\rm ex}}|_1 + \left|\Delta_{J^c_{\rm ex}}\right|_1), \quad \text{and} \quad \Delta \in C_{J(\beta)}. \tag{4.4}$$

The inequality in (4.4) includes terms of different nature: $|\Psi_n \Delta|_\infty$ on one side, and the ℓ_1-norms on the other. The sensitivities allow one to relate them to each other, since for any $J_0 \subseteq \{1, \dots, K\}$, $1 \le p \le \infty$,

$$|\Delta_{J_0}|_p \le \frac{|\Psi_n \Delta|_\infty}{\kappa_{p,J_0,J(\beta)}}. \tag{4.5}$$

Inequality (4.5) is trivial if $\Delta_{J_0} = 0$ and otherwise immediately follows from

$$\frac{|\Psi_n \Delta|_\infty}{|\Delta_{J_0}|_p} \ge \inf_{\widetilde{\Delta}:\ \widetilde{\Delta} \ne 0,\ \widetilde{\Delta} \in C_{J(\beta)}} \frac{|\Psi_n \widetilde{\Delta}|_\infty}{|\widetilde{\Delta}_{J_0}|_p}.$$

From (4.4) and (4.5) with $p = 1$, $J_0 = J_{\rm ex}$ and $J_0 = J^c_{\rm ex}$ we obtain, with probability at least $1 - \alpha$,

$$|\Psi_n \Delta|_\infty \le r \left(2\widehat{\sigma} + r \frac{|\Psi_n \Delta|_\infty}{\kappa_{1,J_{\rm ex},J(\beta)}} + \frac{|\Psi_n \Delta|_\infty}{\kappa_{1,J^c_{\rm ex},J(\beta)}} \right)$$

and thus

$$|\Psi_n \Delta|_\infty \le 2r\widehat{\sigma} \left(1 - \frac{r^2}{\kappa_{1,J_{\rm ex},J(\beta)}} - \frac{r}{\kappa_{1,J^c_{\rm ex},J(\beta)}} \right)^{-1}_+ . \tag{4.6}$$

Together (4.5) and (4.6) yield for any p in $[1, \infty]$, $J_0 \subseteq \{1, \dots, K\}$, with probability at least $1 - \alpha$,

$$|\Delta_{J_0}|_p \le \frac{2r\widehat{\sigma}}{\kappa_{p,J_0,J(\beta)}} \left(1 - \frac{r^2}{\kappa_{1,J_{\rm ex},J(\beta)}} - \frac{r}{\kappa_{1,J^c_{\rm ex},J(\beta)}} \right)^{-1}_+ \tag{4.7}$$

which yields the desired upper bound on the accuracy of the *STIV* estimator, cf. Theorem 6.1 below.

The coordinate-wise sensitivities can also be written as

$$\kappa^*_{k,J} = \inf_{\Delta \in \widehat{\mathcal{A}}_k} (\mathbf{D}_{\mathbf{X}})_{kk} \max_{l=1,\dots,L} (\mathbf{D}_{\mathbf{Z}})_{ll} \left| \frac{1}{n} \sum_{i=1}^n z_{li} \left(x_{ki} - \sum_{m \ne k} x_{mi} \Delta_m \right) \right| \tag{4.8}$$

for some restricted set $\widehat{\mathcal{A}}_k$ of admissible vectors Δ in $\mathbb{R}^{K-1}$ that is derived from the cone C_J. As an example, when there is only one endogenous regressor which is the regressor with index 1 and it belongs to the true model, the set $\widehat{\mathcal{A}}_1$ is defined as

$$\widehat{\mathcal{A}}_1 \triangleq \left\{\Delta \in \mathbb{R}^{K-1} : \left|\left(\mathbf{D}_{\mathbf{X}}^{-1}\Delta\right)_{J(\beta)^c}\right|_1 \leq (1+c)(\mathbf{D}_{\mathbf{X}})_{11}^{-1} + \left|\left(\mathbf{D}_{\mathbf{X}}^{-1}\Delta\right)_{J(\beta)\setminus\{1\}}\right|_1 + cr\left|\left(\mathbf{D}_{\mathbf{X}}^{-1}\Delta\right)_{\{2,\ldots,K\}}\right|_1\right\} .$$

The coordinate-wise sensitivities are measures of the strength of the instruments. They are also restricted partial empirical correlations. When an exogenous variable $(x_{ki})_{i=1}^n$ serves as its own instrument, unless it is almost colinear to other relevant regressors in the structural model, $\kappa^*_{k,J(\beta)}$ is bounded away from zero. When $(x_{ki})_{i=1}^n$ is endogenous then it is not used as an instruments and $\kappa^*_{k,J(\beta)}$ can be small. Because of the sup-norm, one good instrument is enough to ensure that $\kappa^*_{k,J(\beta)}$ is bounded away from zero. Because of the sup-norm it is small only if all instruments are weak.

The sensitivities also provide sharper results for the analysis of Dantzig selector and of the Lasso in classical high-dimensional regression. We show in Section A.1 that the assumption that the sensitivities $\kappa_{p,J}$ are positive is weaker and more flexible than the restricted eigenvalue (RE) assumption of Bickel, Ritov and Tsybakov (2009). Unlike the RE assumption, it is applicable to non-square non-symmetric matrices.

We explain in Section A.2 how to compute exactly the sensitivities of interest when J is given (in practice, estimated) and $|J|$ is small. The following result is a core ingredient to obtain the lower bounds on the sensitivities in Section 8.1.

Proposition 4.1. *(i) Let $J, \widehat{J}$ be two subsets of $\{1,\ldots,K\}$ such that $J \subseteq \widehat{J}$. Then, for all $J_0 \subseteq \{1,\ldots,K\}$, and all $p \in [1,\infty]$ we have $\kappa_{p,J_0,J} \geq \kappa_{p,J_0,\widehat{J}}$.*

(ii) For all $J_0 \subseteq \{1,\ldots,K\}$ and all $p \in [1,\infty]$ we have $\kappa_{p,J_0,J} \geq \kappa_{p,J}$.

(iii) For all $p \in [1,\infty]$,

$$c_J^{-1/p}\kappa_{\infty,J} \leq \kappa_{p,J} \leq \kappa_{\infty,J} \tag{4.9}$$

and for all $J_0 \subseteq \{1,\ldots,K\}$,

$$|J_0|^{-1/p}\kappa_{\infty,J_0,J} \leq \kappa_{p,J_0,J} \leq \kappa_{\infty,J_0,J} . \tag{4.10}$$

(iv) For all $J_0 \subseteq \{1,\ldots,K\}$ we have $\kappa_{\infty,J_0,J} = \min_{k\in J_0}\kappa^*_{k,J} = \min_{k\in J_0}\min_{\Delta\in C_J,\ \Delta_k=1,\ |\Delta|_\infty\leq 1}|\Psi_n\Delta|_\infty$.

(v) $\kappa_{1,J_{\rm ex},J} \geq \max\left(c^{-1}_{J_{\rm ex},J}\kappa_{\infty,J\cup J^c_{\rm ex},J}, |J_{\rm ex}|^{-1}\kappa_{\infty,J_{\rm ex},J}\right)$.

(vi) $\kappa_{1,J^c_{\rm ex},J} \geq \max\left(\underline{c}^{-1}_{J^c_{\rm ex},J}\kappa_{\infty,J,J}, |J^c_{\rm ex}|^{-1}\kappa_{\infty,J^c_{\rm ex},J}\right)$.

(*vii*) $\kappa_{1,J} \geq \left(\frac{2}{\kappa_{1,J,J}} + \frac{cr}{\kappa_{1,J_{\rm ex},J}} + \frac{c}{\kappa_{1,J^c_{\rm ex},J}} \right)^{-1} \geq c_J^{-1} \kappa_{\infty,J\cup J^c_{\rm ex},J}$.
The constants c_J, $c_{J_{\rm ex},J}$ *and* $\underline{c}_{J^c_{\rm ex},J}$ *are given in Table 7.*

Remark 4.2. *The bound in Proposition 4.1* (vii) *is tighter than* (4.9) *for* $p = 1$ *due to the fact that we replace* $\kappa_{\infty,J}$ *by* $\kappa_{\infty,J\cup J^c_{\rm ex},J}$.

Remark 4.3. *The bounds become simpler when* $J^c_{\rm ex} = \emptyset$. *For example, if* $0 < c < r^{-1}$, *we get*

$$\kappa_{p,J} \geq \left(\frac{2|J|}{1-cr} \right)^{-1/p} \min_{k=1,\dots,K} \kappa^*_{k,J}, \quad \kappa_{1,J} \geq \frac{1-cr}{2} \kappa_{1,J,J} \ .$$

The next proposition gives a a sufficient condition to obtain a lower bound on the ℓ_∞-sensitivity. Proposition 4.1 shows that this is a key element to bound from below all the sensitivities.

Proposition 4.2. *Assume that there exist random variables* η_1 *and* η_2 *such that, on an event* $\mathcal{E}$, $\eta_1 > 0$, $0 < \eta_2 < 1$ *and*

$$(4.11) \qquad \forall k \in \{1,\dots,K\},\ \exists l(k) \in \{1,\dots,L\} : \begin{cases} |(\Psi_n)_{l(k)k}| \geq \eta_1 \,, \\ \max_{k' \neq k} |(\Psi_n)_{l(k)k'}| \leq (1-\eta_2) c_J^{-1} |(\Psi_n)_{l(k)k}| \end{cases} ,$$

then, on $\mathcal{E}$, $\kappa_{\infty,J} \geq \eta_1 \eta_2$.

Assumption (4.11) 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 strong instrument. Indeed, if the regressors and instruments where centered, $|(\Psi_n)_{l(k)k}|$ measures the empirical partial correlation of the $l(k)$th instrument for the kth variable. It should be sufficiently large relative $\max_{k' \neq k} |(\Psi_n)_{l(k)k'}|$.

Remark 4.4. *This paper focuses on sparsity but one can easily incorporate in Algorithm 3.1 constraints of the form* $A\beta \in R$ *where* R *is a rectangle in* $\mathbb{R}^p$ *and* A *is a known* $p \times K$ *matrix, for example known signs or prior upper bounds on the size of the coefficients. This is important for inference in a partially identified setup where the identified set would otherwise be an affine space. One works with the following larger sensitivities*

$$\kappa_{p,J_0,J} \triangleq \inf_{\Delta \in C_J :\ \Delta \neq 0,\ A\mathbf{D}_\mathbf{X}(\mathbf{D}_\mathbf{X}^{-1}\widehat{\beta} - \Delta) \in R} \frac{|\Psi_n \Delta|_\infty}{|\Delta_{J_0}|_p} \ .$$

5. Distributional Assumptions and Choice of r

In this section, we discuss different choices of the parameter r in the definition of $\widehat{\mathcal{I}}$. They rely on a choice of the confidence level $1-\alpha$, and on a choice among five scenarii regarding the distribution $\mathbb{P}(\beta)$ of $z_i u_i$ for $i = 1, \dots, n$ where $u_i = y_i - x_i^T \beta$ and $\beta \in \mathcal{I}dent$. These scenarii are described below. Our main analysis will be carried out on the event

$$\mathcal{G} \triangleq \left\{ \max_{l=1,\dots,L} \frac{|\mathbb{E}_n[Z_l U]|}{\sqrt{\mathbb{E}_n[Z_l^2]\mathbb{E}_n[U^2]}} \leq r \right\},$$

and r is chosen such that the probability of this event is $\geq 1-\alpha$. To achieve this, we first note that $\mathcal{G}$ contains the intersection of two events,

$$\left\{ \max_{l=1,\dots,L} \frac{|\mathbb{E}_n[Z_l U]|}{\sqrt{\mathbb{E}_n[Z_l^2 U^2]}} \leq r_0 \right\} \bigcap \left\{ \max_{l=1,\dots,L} \sqrt{\frac{\mathbb{E}_n[Z_l^2 U^2]}{\mathbb{E}_n[Z_l^2]\mathbb{E}_n[U^2]}} \leq M \right\}$$

where $r_0 > 0$ and $M > 0$ satisfy $r = r_0 M$. The constant M can be chosen as an upper bound of

$$\max_{l=1,\dots,L} \sqrt{\frac{\mathbb{E}_n[Z_l^2 U^2]}{\mathbb{E}_n[Z_l^2]\mathbb{E}_n[U^2]}}$$

on an event of probability close to 1. The simplest choice is to use $M = \max_{\substack{l=1,\dots,L,\\ i=1,\dots,n}} |z_{li}| / \sqrt{\mathbb{E}_n[Z_l^2]^{1/2}}$. Another option that we do not discuss here in detail is to strengthen the distributional assumptions - for example M can be close to 1 if z_i and u_i are independent.

Following this argument, in the first four scenarii below the parameter r is of the form $r = r_0 M$. The constant r_0 is adjusted using a union bound and the fact that $\frac{\mathbb{E}_n[Z_l U]}{\sqrt{\mathbb{E}_n[Z_l^2 U^2]}}$ are self-normlized sums. We use results on moderate deviations of self-normalized sums due to Pinelis (1994), Bertail, Gauthérat and Harari-Kermadec (2009), Efron (1969), and Jing, Shao and Wang (2003). For completeness, they are stated in Section A.4.

Scenario 1. *For every $l = 1, \dots, L$, $z_{li} u_i$ are i.i.d. and symmetric, $\prod_{i=1}^n z_{li} u_i$ is almost surely not equal to 0, and L is such that*

$$L < \frac{9\alpha}{4e^3 \Phi(-\sqrt{n})}$$

where Φ is the CDF of the standard normal distribution. We choose

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

Scenario 2. *For some $\gamma_4 > 0$, for every $l = 1, \dots, L$, $z_{li}u_i$ are i.i.d., $\mathbb{E}[(Z_l U)^4] < \infty$, $\prod_{i=1}^n z_{li}u_i$ is almost surely not equal to 0, and*

$$\max_{l=1,\dots,L} \frac{\mathbb{E}[(Z_l U)^4]}{(\mathbb{E}[(Z_l U)^2])^2} \le \gamma_4, \quad L < \frac{\alpha}{2e+1} \exp\left(\frac{n}{\gamma_4}\right).$$

We choose

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

Scenario 3. *For every $i = 1, \dots, n$ and $l = 1, \dots, L$, $z_{li}u_i$ are symmetric and $\prod_{i=1}^n z_{li}u_i$ is almost surely not equal to 0.*

We choose

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

Scenario 4. *For some $\delta > 0$ and $\gamma_{2+\delta} > 0$, for all n, $l = 1, \dots, L$, $\prod_{i=1}^n z_{li}u_i$ is almost surely not equal to 0, and*

$$\max_{l=1,\dots,L} \frac{\mathbb{E}[|Z_l U|^{2+\delta}]}{(\mathbb{E}[Z_l^2 U^2])^{(2+\delta)/2}} \le \gamma_{2+\delta}, \quad L \le \frac{\alpha}{2\Phi(-n^{\delta/(4+2\delta)}\gamma_{2+\delta}^{-1/(2+\delta)}) \left(1 + A_0 \left(1 + n^{-\delta/(4+2\delta)}\gamma_{2+\delta}^{1/(2+\delta)}\right)^{2+\delta}\right)}$$

where $A_0 > 0$ is the constant in Theorem A.3.

We choose

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

Scenarii 1 and 3 rely on symmetry which is realistic 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 it is reasonable to assume that $n - \gamma_4 \log(L(2e+1)/\alpha) \ge n/2$ one can take $r_0 = 2\sqrt{\log(L(2e+1)/\alpha)/n}$. A two-stage approach is used in Bertail, Gauthérat and Harari-Kermadec (2005) to obtain finite sample confidence sets for inference in quasi-empirical likelihood methods. One starts by choosing r_0 with a rough 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 heteroscedasticity. Scenario 4 relies on an upper bound $\gamma_{2+\delta}$. The proposed choice of r_0 is asymptotic because the moderate deviations result depends on A_0 which is not explicit. Scenarii 1, 2 and 4 require that the number of instruments does not exceed a power of n.

Remark 5.1. *An alternative approach which does not involve a maximum over l is to use $(\mathbf{D_Z})_{ll} = (\max_{i=1,\dots,n} |z_{li}|)^{-1}$ and $(\mathbf{D_X})_{kk} = (\max_{i=1,\dots,n} |x_{ki}|)^{-1}$ as in Gautier and Tsybakov (2011). Using $(\mathbf{D_X})_{kk} = (\max_{i=1,\dots,n} |x_{ki}|)^{-1}$ is useful when $1 \le c < r^{-1}$ but for $0 < c < 1$ one can take $(\mathbf{D_X})_{kk} = \mathbb{E}_n[X_k^2]^{-1/2}$.*

Because relying on a union bound can be conservative when there is correlation between the instruments we present another scenario from Chernozhukov, Chetverikov and Kato (2013).

Scenario 5. *The errors u_i are i.i.d. and independent from the z_i. There exist constants $\underline{c}$, $\overline{C}$, B_n such that: (i) $\forall i = 1,\dots,n,\ l = 1,\dots,L$ we have $|z_{li}| \le B_n$ (a.s.); (ii) $\mathbb{E}[U^4] \le \overline{C}$; (iii) $B_n^4(\log(Ln))^7/n \le \overline{C}n^{-\underline{c}}$.*

Under this scenario, r is not chosen as the product $r_0 M$ but rather is taken as the $1-\alpha$ quantile of $\max_{l=1,\dots,L} \frac{|\mathbb{E}_n[Z_l V]|}{\mathbb{E}_n[Z_l^2]^{1/2}}$ conditional on z_i for $i = 1,\dots,n$ where v_i's are i.i.d. standard normal random variables independent from z_i's. These quantiles can be computed by Monte-Carlo techniques. The corresponding choice of r yields an asymptotic $1-\alpha$ confidence. It follows from Corollary 2.1 in Chernozhukov, Chetverikov and Kato (2013) and the fact that $\mathbb{E}_n[U^2]$ is a consistent estimator of $\mathbb{E}[U^2]$ in view of condition (ii). We present in Section A.8 in the Appendix a two-stage method to obtain confidence sets relaxing the assumption that u_i are homoscedastic. It is similar to Proposition 4.2 in Chernozhukov, Chetverikov and Kato (2013) but we deal with a model with endogenous regressors.

For each $j \in \{1,\dots,5\}$, we denote by $\mathcal{P}_j$ the class of all distributions $\mathbb{P}(\beta)$ satisfying the assumptions of Scenario j. Then, for the event $\mathcal{G}$ we have

$$\inf_{\mathbb{P}\in\mathcal{P}_j} \mathbb{P}(\mathcal{G}) \ge 1-\alpha, \quad j = 1,2,3,$$

$$\underline{\lim}_{n\to\infty,\ \Phi^{-1}\left(\frac{\alpha}{2L}\right)n^{-\frac{\delta}{4+2\delta}}\gamma_{2+\delta}^{\frac{1}{2+\delta}}\to 0} \ \inf_{\mathbb{P}\in\mathcal{P}_4} \mathbb{P}(\mathcal{G}) \ge 1-\alpha$$

$$\underline{\lim}_{n\to\infty:\ B_n^4(\log(Ln))^7/n\le\overline{C}n^{-\underline{c}}} \ \inf_{\mathbb{P}\in\mathcal{P}_5} \mathbb{P}(\mathcal{G}) \ge 1-\alpha.$$

The confidence sets are obtained by working on the event $\mathcal{G}$. This yields sets $\mathcal{C}$ which are functions of the observed data and are uniform confidence sets for identifiable parameters (see Romano and Shaikh (2008)). For example, under Scenarii 1, 2, and 3, we obtain uniform confidence sets for sparse identifiable parameters with finite sample validity such that

$$\inf_{\beta,\mathbb{P}:\ \beta\in\mathcal{B}_s,\ \mathbb{P}(\beta)\in\mathcal{P}_j} \mathbb{P}\left(\beta\in\mathcal{C}\right) \ge 1-\alpha\ , \quad j = 1,2,3,$$

while under the asymptotic Scenario 5,

$$\underline{\lim}_{n\to\infty,\ B_n^4(\log(Ln))^7/n\le\overline{C}n^{-\underline{c}}} \ \inf_{\beta,\mathbb{P}:\ \beta\in\mathcal{B}_s,\ \mathbb{P}(\beta)\in\mathcal{P}_j} \mathbb{P}\left(\beta\in\mathcal{C}\right) \ge 1-\alpha.$$

Since the event $\mathcal{G}$ is independent of the parameter c, confidence sets can be defined as measurable intersections for different values of c. Thus not a single tuning parameter has to be specified. Any feasible intersection contains this set and provides a valid $1-\alpha$ confidence set. For example, we obtain an $1-\alpha$ confidence set by intersecting sets for various values of c on a grid. Uniformity for confidence

sets with asymptotic confidence level garantees that for a given $\epsilon > 0$, there exists n_0 large enough such that for any sample size n larger than n_0, uniformly on the distribution of the observed data, the confidence level is at least $1 - \alpha - \epsilon$. The fact that each proposed confidence set is 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).

6. Basic Bounds

In this section, we provide some basic bounds, from which we will deduce in Sections 7 and 8 the rates of convergence and confidence sets for the $STIV$ estimator. In what follows, we write for brevity: "For every β in $\mathcal{I}dent$, on the event $\mathcal{G}$..." instead of: "For every Scenario j with $j = 1, \dots, 5$, and with the corresponding choice of r, and for every distribution $\mathbb{P}$ of the observed data such that β belongs to $\mathcal{I}dent$ and $\mathbb{P}(\beta) \in \mathcal{P}_j$, on the event $\mathcal{G}$...".

6.1. **Upper Bounds for Sparse Vectors.**

Theorem 6.1. *For every β in $\mathcal{I}dent$, on the event $\mathcal{G}$, for any solution $(\widehat{\beta}, \widehat{\sigma})$ of the minimization problem (3.6) we have, for every $J_0 \subseteq \{1, \dots, K\}$, $0 < c < r^{-1}$, $p \geq 1$,*

$$\left| \left(\mathbf{D}_{\mathbf{X}}^{-1} (\widehat{\beta} - \beta) \right)_{J_0} \right|_p \leq \frac{2\widehat{\sigma} r}{\kappa_{p, J_0, J(\beta)}} \left(1 - \frac{r^2}{\kappa_{1, J_{\mathrm{ex}}, J(\beta)}} - \frac{r}{\kappa_{1, J_{\mathrm{ex}}^c, J(\beta)}} \right)_+^{-1} \tag{6.1}$$

and

$$\widehat{\sigma} \leq \sqrt{\widehat{Q}(\beta)} \left(1 + \frac{r}{c\kappa_{1, J(\beta), J(\beta)}} \right) \left(1 - \frac{r}{c\kappa_{1, J(\beta), J(\beta)}} \right)_+^{-1}. \tag{6.2}$$

In the model without endogeneity, we have $\kappa_{1, J_{\mathrm{ex}}^c, J(\beta)} = \infty$ and

$$\sqrt{\frac{1}{n} \sum_{i=1}^n \left(x_i^T (\widehat{\beta} - \beta) \right)^2} \leq \frac{2\widehat{\sigma} r}{\sqrt{\kappa_{1, J(\beta)}}} \left(1 - \frac{r^2}{\kappa_{1, J(\beta)}} \right)_+^{-1}. \tag{6.3}$$

In particular this theorem implies that, on the event $\mathcal{G}$, for every $0 < c < r^{-1}$ and $k = 1, \dots, K$,

$$|\widehat{\beta}_k - \beta_k| \leq \frac{2\widehat{\sigma} r}{\mathbb{E}_n[X_k^2]^{1/2} \, \kappa^*_{k, J(\beta)}} \left(1 - \frac{r^2}{\kappa_{1, J_{\mathrm{ex}}, J(\beta)}} - \frac{r}{\kappa_{1, J_{\mathrm{ex}}^c, J(\beta)}} \right)_+^{-1}. \tag{6.4}$$

When $\tau_1 \triangleq \left(1 - \frac{r^2}{\kappa_{1, J_{\mathrm{ex}}, J(\beta)}} - \frac{r}{\kappa_{1, J_{\mathrm{ex}}^c, J(\beta)}} \right)_+^{-1}$ is close to 1 and the sensitivities are bounded away from zero, the upper bounds in (6.1) and thus (6.4) is of the order $O(r) = O(\sqrt{\log(L)/n})$. Thus, we have an extra $\sqrt{\log(L)}$ factor as compared to the usual root-n rate. It is a modest price for using a large number L of instruments. Also, adding instruments and thus rows to matrix Ψ_n increases the sup-norm $|\Psi_n \Delta|_\infty$, and thus potentially increases the sensitivities $\kappa_{p, J_0, J(\beta)}$ and their computable

data-driven lower bounds that we present in Section 8. This has a positive effect in view of the form of our bounds, cf. , e.g., (6.4). As we will see later, the inverse of the sensitivities drive the rates of convergence and the width of the confidence sets. Thus, adding instruments potentially improves the confidence set. On the other hand, the price for adding instruments in terms of the rate of convergence appears in the constant r and is only logarithmic in the number of instruments.

The upper bounds are infinite on the event $\frac{r^2}{\kappa_{1,J_{\text{ex}},J(\beta)}} + \frac{r}{\kappa_{1,J_{\text{ex}}^c,J(\beta)}} \geq 1$. This occurs either when r is not small enough or when $\kappa_{1,J_{\text{ex}},J(\beta)}$ or $\kappa_{1,J_{\text{ex}}^c,J(\beta)}$ are too small. In scenarii 2-5, r is of the order $\sqrt{\log(L)/n}$, which allows the number of instruments to be of any order smaller than an exponential in n. Proposition 4.1 (v) and (vi) yield that

$$\frac{r^2}{\kappa_{1,J_{\text{ex}},J(\beta)}} + \frac{r}{\kappa_{1,J_{\text{ex}}^c,J(\beta)}} \leq \frac{r\ c_{b,J(\beta)}}{\kappa_{\infty,J(\beta)\cup J_{\text{ex}}^c,J(\beta)}} \tag{6.5}$$

where $c_{b,J} \triangleq r c_{J_{\text{ex}},J} + \underline{c}_{J_{\text{ex}}^c,J}$.

When $1 \leq c < r^{-1}$, one can check that $c_{b,J} \leq (2r\,|J_{\text{ex}} \cap J| + (1+r)\,|J_{\text{ex}}^c \cap J| + (1-r)\,|J_{\text{ex}}^c \cap J^c|)/(1-cr)$. Thus, two conditions $|J_{\text{ex}} \cap J(\beta)| \leq C_1 r^{-2} = O(n/\log(L))$, and $|J_{\text{ex}}^c| \leq C_2 r^{-1}$ with appropriate constants $C_1, C_2 > 0$, are sufficient to ensure that $\frac{r^2}{\kappa_{1,J_{\text{ex}},J(\beta)}} + \frac{r}{\kappa_{1,J_{\text{ex}}^c,J(\beta)}} < 1$. The first condition is implied by a mild and quite standard assumption on the sparsity $|J(\beta)|$, while the second one limits the number of endogenous regressors.

When $0 < c < 1$, one obtains easily the upper bound $c_{b,J} \leq 2|J|(r+1)/(1-c)$. Thus, the condition $|J(\beta)| \leq C_3 r = O(\sqrt{n/\log(L)})$ for some $C_3 > 0$ is sufficient to obtain $\frac{r^2}{\kappa_{1,J_{\text{ex}},J(\beta)}} + \frac{r}{\kappa_{1,J_{\text{ex}}^c,J(\beta)}} < 1$. No restriction on the number of endogenous regressors is needed in this case.

Remark 6.1. *When there is no endogenity,* $c_{b,J} \leq 2r\,|J|\,(1-cr)^{-1}$.

Due to the upper bound in (4.10) and Proposition 4.1 (iv), we have

$$\frac{r|J_{\text{ex}}^c|}{\min_{k\in J_{\text{ex}}^c} \kappa^*_{k,J(\beta)}} \leq \frac{r^2}{\kappa_{1,J_{\text{ex}},J(\beta)}} + \frac{r}{\kappa_{1,J_{\text{ex}}^c,J(\beta)}} .$$

Thus, in the model with only one endogenous regressor, the condition $\min_{k\in J_{\text{ex}}^c} \kappa^*_{k,J(\beta)} \leq r$, implies that $\frac{r^2}{\kappa_{1,J_{\text{ex}},J(\beta)}} + \frac{r}{\kappa_{1,J_{\text{ex}}^c,J(\beta)}} \geq 1$. In other words, if the coordinate-wise sensitivity for the endogenous regressor is too small, we obtain confidence sets of infinite volume. This is in agreement with Dufour (1997) who show that confidence sets of infinite volume cannot be avoided for procedures that are robust to weak instruments.

The case where instruments can have a direct effect on the outcome is studied in Kolesár, Chetty, Friedman, et al. (2011). This can create situations where $L < K$ and identification fails

without further assumptions. The situation can be quite different under sparsity. The set $\mathcal{B}_s$ is the union of the sets of vectors β such that

$$\begin{cases} \beta_{J^c} = 0 \\ \mathbb{E}[Z(Y - X_J^T \beta_J)] = 0 \end{cases}$$

for all subsets J of $\{1, \dots, K\}$ of size s. When $s < L$, the system of equations $\mathbb{E}[ZX_J^T]\beta_J = \mathbb{E}[ZY]$ usually has no solution because it is overdetermined. The condition $s < L$ means that there are exogenous variables that do not have a direct effect on the outcome but one does not know which one in advance. Because the cone constraint restricts the minimization to vectors with s dominant coordinates, the sensitivities on the right-hand side of (6.1) and (6.4) can be different from 0. This is a situation that we study in Section 10.

6.2. **Upper Bounds for Approximately Sparse Vectors.** Define the enlarged cone

$$\widetilde{C}_J \triangleq \left\{\Delta \in \mathbb{R}^K : \ |\Delta_{J^c}|_1 \le 2\left(|\Delta_J|_1 + cr|\Delta_{J_{\rm ex}}|_1 + c|\Delta_{J^c_{\rm ex}}|_1\right)\right\}$$

and define, for $p \in [1, \infty]$ and $J_0 \subseteq \{1, \dots, K\}$

$$\widetilde{\kappa}_{p,J_0,J} \triangleq \inf_{\Delta \in \mathbb{R}^K:\ |\Delta_{J_0}|_p = 1,\ \Delta \in \widetilde{C}_J} |\Psi_n \Delta|_\infty \,.$$

We denote for brevity by $\widetilde{\kappa}_{p,J}$ and $\widetilde{\kappa}^*_{k,J}$ the special cases of $\widetilde{\kappa}_{p,J_0,J}$ when $J_0 = \{1, \dots, K\}$ and $J_0 = \{k\}$ respectively.

Theorem 6.2. *For every β in $\mathcal{I}dent$, on the event $\mathcal{G}$, for any solution $(\widehat{\beta}, \widehat{\sigma})$ of the minimization problem (3.6) we have, for every $J_0 \subseteq \{1, \dots, K\}$, $0 < c < r^{-1}$, $p \ge 1$,*

(6.6)
$$\left|\left(\mathbf{D}_{\mathbf{X}}^{-1}\left(\widehat{\beta} - \beta\right)\right)_{J_0}\right|_p \le \min_{J \subseteq \{1,\dots,K\}} \max\left(\frac{2\widehat{\sigma} r}{\widetilde{\kappa}_{p,J_0,J}}\left(1 - \frac{r^2}{\widetilde{\kappa}_{1,J_{\rm ex},J}} - \frac{r}{\widetilde{\kappa}_{1,J^c_{\rm ex},J}}\right)_+^{-1},\ 6\left|\left(\mathbf{D}_{\mathbf{X}}^{-1}\beta\right)_{J^c}\right|_1\right).$$

In addition, in the model without endogeneity, we have $\widetilde{\kappa}_{1,J^c_{\rm ex},J} = \infty$ and $\sqrt{\frac{1}{n}\sum_{i=1}^n \left(x_i^T(\widehat{\beta} - \beta)\right)^2}$ is less than

$$\min_{J \subseteq \{1,\dots,K\}} \max\left(\frac{2\widehat{\sigma} r}{\sqrt{\widetilde{\kappa}_{1,J(\beta)}}}\left(1 - \frac{r^2}{\widetilde{\kappa}_{1,J(\beta)}}\right)_+^{-1},\ 2\sqrt{3\left|\left(\mathbf{D}_{\mathbf{X}}^{-1}\beta\right)_{J^c}\right|_1\left(r\widehat{\sigma} + \frac{r}{c}\left|\left(\mathbf{D}_{\mathbf{X}}^{-1}\beta\right)_{J^c}\right|_1\right)}\right). \tag{6.7}$$

Applying this theorem with $J_0 = \{k\}$ we obtain that, on the event $\mathcal{G}$, for every $0 < c < r^{-1}$ and $k = 1, \dots, K$,

$$|\widehat{\beta}_k - \beta_k| \le \frac{1}{\mathbb{E}_n[X_k^2]^{1/2}} \min_{J \subset \{1,\dots,K\}} \max\left(\frac{2\widehat{\sigma} r}{\widetilde{\kappa}^*_{k,J}}\left(1 - \frac{r^2}{\widetilde{\kappa}_{1,J_{\rm ex},J}} - \frac{r}{\widetilde{\kappa}_{1,J^c_{\rm ex},J}}\right)_+^{-1},\ 6\left|\left(\mathbf{D}_{\mathbf{X}}^{-1}\beta\right)_{J^c}\right|_1\right). \tag{6.8}$$

Remark 6.2. *To garantee that the right-hand side is small one often assumes that there is $C > 0$, $\alpha > 1$ and sets J such that $|J| \leq s$ and $\left|\left(\mathbf{D}_{\mathbf{X}}^{-1}\beta\right)_{J^c}\right|_1 \leq Cs^{-\alpha+1}$. This holds if there exists a permutation τ of the indices of the coefficients such that, for every k in $\{1,\ldots,K\}$, $|\beta_{\tau(k)}| \leq Ck^{-\alpha}$.*

Inequality (6.6) 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 for all values of the parameters J_0, c and k. For example, for given c, $p = 1$ and $J_0 = \{1,\ldots,K\}$, there exists an optimal set $J = J^*$ such that

$$\left|\mathbf{D}_{\mathbf{X}}^{-1}\left(\widehat{\beta}-\beta\right)\right|_1 \leq \frac{2\widehat{\sigma}r}{\widetilde{\kappa}_{1,J^*}}\left(1-\frac{r^2}{\widetilde{\kappa}_{1,J_{\text{ex}},J^*}}-\frac{r}{\widetilde{\kappa}_{1,J_{\text{ex}}^c,J^*}}\right)_+^{-1}. \tag{6.9}$$

7. Rates of Convergence

In this section, we derive the rates of convergence of the *STIV* estimator. The argument is based on replacing the random right-hand sides of (6.1) and (6.2) by their deterministic upper bounds. To do this, we will need the following assumptions.

Assumption 7.1. *For every $\beta \in \mathcal{B}_s$ and $\gamma_1 \in (0,1)$, there exist $\sigma_* > 0$ and $\kappa_* > 0$ such that with probability at least $1-\gamma_1$,*

$$\max_{l\in I}(\mathbf{D}_{\mathbf{Z}})_{ll}^2\mathbb{E}_n[Z_l^2(Y-X^T\beta)^2] \leq \sigma_*^2, \tag{7.1}$$

$$\kappa_{\infty,J(\beta)} \geq \kappa_*, \tag{7.2}$$

$$\left(1-\frac{r^2}{\kappa_{1,J_{\text{ex}},J(\beta)}}-\frac{r}{\kappa_{1,J_{\text{ex}}^c,J(\beta)}}\right)_+^{-1} \leq \theta_*. \tag{7.3}$$

Due to Proposition 4.1, on the same event,

$$\begin{aligned}
\kappa^*_{k,J(\beta)} &\geq \kappa_*, && \forall k = 1,\ldots,K,\\
\kappa_{p,J(\beta)} &\geq \kappa_* c_{J(\beta)}^{-1/p}, && \forall p \in [1,\infty],\\
\kappa_{1,J(\beta),J(\beta)} &\geq \kappa_*|J(\beta)|^{-1}
\end{aligned}$$

and $\theta_* \leq \left(1-\frac{r\ c_{b,J(\beta)}}{\kappa_*}\right)_+^{-1}$.

Assumption 7.2. *For every $\gamma_2 \in (0,1)$ and $k \in \{1,\ldots,K\}$, there exist constants $v_k > 0$ such that*

$$\mathbb{P}\left(\mathbb{E}_n[X_k^2]^{1/2} \geq v_k,\ \forall\ k \in \{1,\ldots,K\}\right) \geq 1-\gamma_2.$$

Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be the events from Assumptions 7.1 and 7.2 and set $\gamma = \alpha+\gamma_1+\gamma_2$ and

$$\tau_* \triangleq \left(1+\frac{r|J(\beta)|}{c\kappa_*}\right)\left(1-\frac{r|J(\beta)|}{c\kappa_*}\right)_+^{-1}\theta_*$$

where the last term corresponds to (6.5) using κ_* as a lower bound for $\kappa_{\infty,J(\beta)\cup J^c_{\rm ex},J(\beta)}$.

Theorem 7.1. *For every β in $\mathcal{B}_s$, under the assumptions of Theorem 6.1 and Assumption 7.1, the following holds.*

(i) *On the event $\mathcal{G}\cap\mathcal{G}_1$ for any solution $(\widehat\beta,\widehat\sigma)$ of (3.6), we have*

$$\left|\mathbf{D}_{\mathbf{X}}^{-1}\left(\widehat\beta-\beta\right)\right|_p \le \frac{2\sigma_* r c_{J(\beta)}^{1/p}\tau_*}{\kappa_*}, \quad \forall\ p\in[1,\infty], \tag{7.4}$$

and

$$\widehat\sigma \le \sigma_*\left(1+\frac{r|J(\beta)|}{c\kappa_*}\right)\left(1-\frac{r|J(\beta)|}{c\kappa_*}\right)_+^{-1}.$$

In addition, in the model without endogeneity, on the same event,

$$\sqrt{\frac{1}{n}\sum_{i=1}^n\left(x_i^T(\widehat\beta-\beta)\right)^2} \le 2\widehat\sigma r\theta_*\sqrt{\frac{2|J(\beta)|}{(1-cr)\kappa_*}}\ . \tag{7.5}$$

(ii) *Let, in addition, Assumption 7.2 hold. Then on the event $\mathcal{G}\cap\mathcal{G}_1\cap\mathcal{G}_2$, for any solution $\widehat\beta$ of (3.6), we have*

$$|\widehat\beta_k-\beta_k| \le \frac{2\sigma_* r\tau_*}{\kappa_* v_k}, \quad \forall\ k=1,\dots,K. \tag{7.6}$$

(iii) *Let the assumptions of* (ii) *hold, and*

$$|\beta_k| > \frac{2\sigma_* r\tau_*}{\kappa_* v_k} \text{ for all } k\in J(\beta). \tag{7.7}$$

Then, on the event $\mathcal{G}\cap\mathcal{G}_1\cap\mathcal{G}_2$, for any solution $\widehat\beta$ of (3.6), we have

$$J(\beta)\subseteq J(\widehat\beta).$$

For reasonably large sample size ($n\gg\log(L)$), the value r is small, and τ_* is approaching 1 as $r\to 0$. From the discussion that follows Proposition 4.1, we see that when there is no endogeneity ($J^c_{\rm ex}=\emptyset$), the bounds (7.4) and (7.6) are of the order of magnitude $O(r|J(\beta)|^{1/p})$ and $O(r)$ 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 this context, $L=K$ is the dimension of β. Interestingly, under endogeneity we still obtain the $O(r|J(\beta)|^{1/p})$ bound in (7.4) provided that $0<c<1$. However, for larger values of the tuning parameter c, we obtain the rate $O(r(|J(\beta)\cap J_{\rm ex}|+c|J^c_{\rm ex}|)^{1/p})$.

8. Variable Selection and Uniform Joint Confidence Sets

The only unknown ingredient in inequality (6.1) is the set $J(\beta)$ that determines the sensitivities. We propose various strategies to turn these inequalities into valid confidence sets: some relying on an estimator $\widehat{J}$ of $J(\beta)$ and some relying on an upper bound on $|J(\beta)|$ that we call a *sparsity certificate*.

8.1. **Computationally Efficient Lower Bounds on the Sensitivities.** Let us start by stating preliminary lower bounds obtained by minimizing over a set that contains the cone. As a consequence we show that we can bound from below all sensitivities of interest by solving K simple convex programs. They can also be used to obtain sharper bounds which are easy to compute in certain situations.

Proposition 8.1. *When $|J| \leq s$ and $0 < c < r^{-1}$, we have*

$$\kappa_{\infty,J} \geq \kappa_\infty(J,s) \tag{8.1}$$

$$\kappa^*_{k,J} \geq \kappa^*_k(J,s) \tag{8.2}$$

$$\kappa_{p,J_0,J} \geq \kappa_{p,J_0}(J,s) \quad \forall p \in [1,\infty]$$

$$\kappa_{1,J_{\rm ex},J} \geq \kappa_{1,J_{\rm ex}}(J,s) \tag{8.3}$$

$$\kappa_{1,J^c_{\rm ex},J} \geq \underline{\kappa}_{1,J^c_{\rm ex}}(J,s) \tag{8.4}$$

$$\left(1 - \frac{r^2}{\kappa_{1,J_{\rm ex},J}} - \frac{r}{\kappa_{1,J^c_{\rm ex},J}}\right)^{-1}_+ \leq \theta(J,s)$$

with the constants $\kappa_\infty(J,s)$, $\kappa^*_k(J,s)$, $\kappa_{p,J_0}(J,s)$, $\underline{\kappa}_{1,J_{\rm ex}}(J,s)$, $\underline{\kappa}_{1,J^c_{\rm ex}}(J,s)$ and $\theta(J,s)$ of Table 7.

These preliminary bounds can be used to obtain easily computable lower bounds by mimizing over larger sets. We shall use the same notation for the lower bounds of Table 7 and the computable bounds that we now present.

The constants $\kappa_\infty(J,s)$ and $\kappa^*_k(J,s)$ from Table 7 involve the constraint $(1-cr)|\Delta_{J_{\rm ex}}|_1 + (1-c)|\Delta_{J^c_{\rm ex}}|_1 \leq 2s\Delta_j$. For $0 < c < 1$, the function $\Delta \to (1-cr)|\Delta_{J_{\rm ex}}|_1 + (1-c)|\Delta_{J^c_{\rm ex}}|_1$ is convex. For $c > 1$, $\Delta \to (1-c)|\Delta_{J^c_{\rm ex}}|_1$ is not convex but one can handle cases where $J^c_{\rm ex}$ has small cardinality by considering different values of the signs of the coordinates of $\Delta_{J^c_{\rm ex}}$. For example, if only the first regressor is endogenous, the term in curly brackets in the definition of $\kappa^*_k(J,s)$ (cf. Table 7) can be written as

$$\min\left\{ \min_{\substack{\Delta_k=\pm 1,\ \Delta_j\geq 0,\ |\Delta_J|_\infty \leq \Delta_j,\ \Delta_1 \geq 0 \\ (1-cr)|\Delta_{J_{\rm ex}}|_1 + (1-c)\Delta_1 \leq 2s\Delta_j}} |\Psi_n \Delta|_\infty \,,\ \min_{\substack{\Delta_k=\pm 1,\ \Delta_j\leq 0,\ |\Delta_J|_\infty \leq -\Delta_j,\ \Delta_1 \leq 0 \\ (1-cr)|\Delta_{J_{\rm ex}}|_1 - (1-c)\Delta_1 \leq 2s\Delta_j}} |\Psi_n \Delta|_\infty \right\} . \tag{8.5}$$

The constraints $|\Delta_{J_{\rm ex}}|_1 = 1$ in the definition of $\kappa_{1,J_{\rm ex}}(J,s)$, and $|\Delta_{J^c_{\rm ex}}|_1 = 1$ in the definition of $\underline{\kappa}_{1,J^c_{\rm ex}}(J,s)$ can be handled in a similar way when respectively $|J_{\rm ex}|$ or $|J^c_{\rm ex}|$ are not too large. Otherwise, one can minimize on the larger set replacing $|\Delta_{J_{\rm ex}}|_1 = 1$ by the convex constraint $|\Delta_{J_{\rm ex}}|_1 \le 1$, or respectively, $|\Delta_{J^c_{\rm ex}}|_1 = 1$ by $|\Delta_{J^c_{\rm ex}}|_1 \le 1$.

When $1 \le c < r^{-1}$ and $|J^c_{\rm ex}|$ is large we can bound from below these preliminary lower bounds by minimizing over larger sets defined by convex constraints. For example, a lower bound of the term in curly brackets in the definition of $\kappa_\infty(J,s)$ (cf. Table 7) is obtained by solving the following convex program.

Algorithm 8.1. *Find $v > 0$ which achieves the minimum*

$$\min_{\epsilon=\pm 1} \min_{(w,\Delta,v)\in\mathcal{V}_{k,j}} v$$

where $\mathcal{V}_{k,j}$ is the set of (w,Δ,v) with $w \in \mathbb{R}^K$, $\Delta \in \mathbb{R}^K$, $v \in \mathbb{R}$ satisfying:

$$v \ge 0, \quad -v\mathbf{1} \le \Psi_n \Delta \le v\mathbf{1}, \quad \mathbf{0} \le w \le \mathbf{1}, \quad -w\mathbf{1} \le \Delta \le w\mathbf{1},$$
$$\epsilon\Delta_k = 1, \quad -\Delta_j \mathbf{1} \le w_J \le \Delta_j \mathbf{1}, \quad (1-cr)|\Delta_{J_{\rm ex}}|_1 \le 2s\Delta_j + (c-1)\sum_{i\in J^c_{\rm ex}} w_i \ .$$

This leads to an easily computable lower bound for $\kappa_\infty(J,s)$. A similar bound holds for $\kappa^*_k(J,s)$. As a consequence it is possible to obtain lower bounds for all coordinate-wise sensitivities by solving $2K|J|$ convex programs.

The set J that matters for our analysis is the set $J(\beta)$. Due to Proposition 4.1 (i), we can get data-driven lower bounds for sensitivities provided we have an estimator $\widehat{J}$ such that of $\widehat{J} \supseteq J(\beta)$. When using the sparsity certificate approach, then lower bounds are obtained by taking $J = \{1,\dots,K\}$ in (8.2), (8.3) and (8.4). In this case, c_J, $c_{J_{\rm ex},J}$, $c_{J^c_{\rm ex},J}$ are replaced by the upper bounds $c(s)$, $c_{J_{\rm ex}}(s)$, $c_{J^c_{\rm ex}}(s)$ that only depend on s, cf. Table 7. We also use an upper bound $c_b(s)$ on $r\max(c^{-1}_{J_{\rm ex},J}, |J_{\rm ex}|^{-1}) + \max(c^{-1}_{J^c_{\rm ex},J}, |J^c_{\rm ex}|^{-1})$. The lower bounds $\kappa_\infty(s)$, $\kappa^*_k(s)$, $\kappa_{p,J_0}(s)$, $\underline{\kappa}_{1,J_{\rm ex}}(s)$, $\underline{\kappa}_{1,J^c_{\rm ex}}(s)$ and $\theta(s)$ on the constants $\kappa_\infty(J,s)$, $\kappa^*_k(J,s)$, $\kappa_{p,J_0}(J,s)$, $\underline{\kappa}_{1,J_{\rm ex}}(J,s)$, $\underline{\kappa}_{1,J^c_{\rm ex}}(J,s)$ and $\theta(J,s)$ respectively are given in Table 7. As explained after Proposition 8.1, these constants can be further bounded from below by easily computable values. For example, if $1 < c < r^{-1}$ and $|J^c_{\rm ex}|$ is large enough, a lower bound for the term in curly brackets in the definition of $\kappa_\infty(s)$ can be obtained by solving the following convex program.

Algorithm 8.2. *Find $v > 0$ which achieves the minimum*

$$\min_{(w,\Delta,v)\in\mathcal{V}_k} v$$

where $\mathcal{V}_k$ is the set of (w, Δ, v) with $w \in \mathbb{R}^K$, $\Delta \in \mathbb{R}^K$, $v \in \mathbb{R}$ satisfying:

$$v \geq 0, \quad -v\mathbf{1} \leq \Psi_n \Delta \leq v\mathbf{1}, \quad \mathbf{0} \leq w \leq \mathbf{1}, \quad -w \leq \Delta \leq w, \quad \Delta_k = 1,$$
$$(1 - cr)|\Delta_{J_{\mathrm{ex}}}|_1 \leq 2s + (c-1) \sum_{i \in J_{\mathrm{ex}}^c} w_i \ .$$

The constant $\kappa_\infty(s)$ can be used as a lower bound for all the coordinate-wise sensitivities. Its computation requires to solve K convex programs. This leads to easily computable joint confidence sets. The direct bounds $\kappa_k^*(s)$ are sharper but their computation requires to solve $2K$ convex programs for each coordinate k. Thus one has to solve $2K|J_0|$ convex programs if one wishes to obtain tight bounds for a subvector β_{J_0}.

8.2. **Selection of Variables.** Theorem 7.1 (iii) provides an upper estimate on the set of non-zero components of β. Exact selection of variables can be performed as well. For this purpose, we use the thresholded *STIV* estimator $\widetilde{\beta}$ whose coordinates are defined by

$$\widetilde{\beta}_k \triangleq \begin{cases} \widehat{\beta}_k & \text{if } |\widehat{\beta}_k| > \omega_k, \\ 0 & \text{otherwise,} \end{cases} \tag{8.6}$$

where $\omega_k > 0$, $k = 1, \ldots, K$, are thresholds specified below. The thresholds depend on s and they can be used if we have a sparsity certificate s. To provide garantees for the thresholding rule, we strengthen Assumption 7.1 as follows.

Assumption 8.1. *There exists positive constants $\kappa_*(s)$ and $\theta_*(s)$ such that, on an event of probability at least $1 - \gamma_1$, we have: $\kappa_\infty(s) \geq \kappa_*(s)$, $\theta(s) \leq \theta_*(s)$, and inequalities* (7.1) *and* (7.2) *hold.*

Denote by $\mathcal{G}_1$ and $\mathcal{G}_2$ the events of Assumptions 8.1 and 7.2 respectively, and set

$$\tau_*(s) \triangleq \left(1 + \frac{r|J(\beta)|}{c\kappa_*}\right)\left(1 - \frac{r|J(\beta)|}{c\kappa_*}\right)_+^{-1} \theta(s)\,.$$

The following theorem shows that, based on thresholding of the *STIV* estimator, we recover the set of non-zero coefficients $J(\beta)$ with probability close to 1. Even more, we achieve the sign consistency, *i.e.,* we recover the vector of signs of the coefficients of β with probability close to 1.

Theorem 8.1. *For every β in $\mathcal{B}_s$, let the assumptions of Theorem 6.1 and Assumptions 8.1 and 7.2 be satisfied. Assume that*

$$|\beta_k| > \frac{4\sigma_* r \tau_*(s)}{\kappa_*(s) v_k} \text{ for all } k \in J(\beta). \tag{8.7}$$

If the thresholds are takes as

$$\omega_k \triangleq \frac{2\widehat{\sigma} r}{\kappa_k^*(s)\mathbb{E}_n[X_k^2]^{1/2}}\theta(s),$$

then, on the event $\mathcal{G} \cap \mathcal{G}_1 \cap \mathcal{G}_2$*, we have* $J(\widetilde{\beta}) = J(\beta)$ *and*

$$\overrightarrow{\operatorname{sign}(\widetilde{\beta})} = \overrightarrow{\operatorname{sign}(\beta)} \ . \tag{8.8}$$

Conditions (7.7) and (8.7) will be referred to as the *beta-min* assumptions.

8.3. **Joint Confidence Sets.** We propose three types of confidence sets under different assumptions.

8.3.1. *Adaptive Confidence Sets Under the Stronger Beta-min Assumption.* The only unknown in the bounds of Theorem 6.1 is the set $J(\beta)$. However, by Theorem 8.1, under the beta-min assumption (8.7) we have $J(\widetilde{\beta}) = J(\beta)$ with probablity close to 1, and thus we can plug in a data-driven $J(\widetilde{\beta})$ instead of $J(\beta)$. This leads to the confidence sets that we refer to as adaptive confidence sets.

Theorem 8.2. *Let* $0 < c < r^{-1}$*, and let the assumptions of Theorem 8.1 hold. Set* $\widehat{J} = J(\widetilde{\beta})$ *where* $\widetilde{\beta}$ *is defined in* (8.6)*. Then, for any* β *in* $\mathcal{B}_s$*, on the event* $\mathcal{G} \cap \mathcal{G}_1 \cap \mathcal{G}_2$*, for any solution* $(\widehat{\beta}, \widehat{\sigma})$ *of the minimization problem (3.6) and any* $J_0 \subseteq \{1, \dots, K\}$*,* $p \geq 1$*, we have*

$$\left| \left(\mathbf{D}_{\mathbf{X}}^{-1}(\widehat{\beta} - \beta) \right)_{J_0} \right|_p \leq \frac{2\widehat{\sigma} r \theta(\widehat{J}, |\widehat{J}|)}{\kappa_{p,J_0}(\widehat{J}, |\widehat{J}|)} . \tag{8.9}$$

In addition, in the model without endogeneity, on the same event,

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(x_i^T(\widehat{\beta} - \beta) \right)^2} \leq \frac{2\widehat{\sigma} r \theta(\widehat{J}, |\widehat{J}|)}{\sqrt{\kappa_1(\widehat{J}, |\widehat{J}|)}} \ . \tag{8.10}$$

Since the set J_0 can be any singleton, we obtain as a corollary that (3.8) holds on the event $\mathcal{G} \cap \mathcal{G}_1 \cap \mathcal{G}_2$; the expression for $A_k(\cdot)$ appearing in (3.8) can be deduced from (8.9). The constant $\gamma_1 + \gamma_2$ in the confidence level of Theorem 8.2 can be much smaller than α for large enough σ_* and small enough κ_* and v_k, $k = 1, \dots, K$.

Remark 8.1. *Posterior to this paper, Nickl and Van de Geer (2013) considered asymptotic joint confidence sets based on the* ℓ_2*-norm in high-dimensional regression with Gaussian errors independent of the regressors. Restricting the parameter space to* s*-sparse vectors and assuming specific distributions of the regressors and non-degeneracy of matrix* $\mathbb{E}[\Psi_n]$ *(in our notation), they propose a confidence set which, for every* $k \leq s$*, has a diameter bounded in probability by* $\sqrt{\frac{\log(K)k}{n}} + n^{-1/4}$ *uniformly over vectors* β *such that* $|J(\beta)| \leq k$*. Its construction is based on the minimal eigenvalue of* $\mathbb{E}[\Psi_n]$ *which is unknown in our paper. They also present necessary and sufficient conditions for the existence of*

confidence sets of diameter bounded in probability by $\sqrt{\frac{\log(K)k}{n}}$ *when* $\sqrt{\frac{\log(K)k}{n}} \gg n^{-1/4}$. *This requires to remove from the parameter space vectors that are too close to k-sparse vectors in the Euclidean distance. The beta-min assumption can be viewed as a sup-norm analogue of this.*

8.3.2. *Non-adaptive Confidence Sets Under the Weaker Beta-min Assumption.* Replacing in Theorem 8.2, Assumption 8.1 by Assumption 7.1, denoting again by $\mathcal{G}_2$ the corresponding event, and replacing the *beta-min* assumption (8.7) from Theorem 8.1 by the weaker assumption (7.7) from Theorem 7.1, we obtain the same result as in Theorem 8.2 with $\widehat{J} = J(\widehat{\beta})$. The confidence sets are no longer adaptive because $J(\widehat{\beta})$ can be larger than $J(\beta)$.

Remark 8.2. *It is possible to obtain an upper bound on* $|J(\widehat{\beta})|$ *for the Lasso (see,* e.g., *Bickel, Ritov and Tsybakov (2009)) using the equalities in the Karush-Kuhn-Tucker condition for the non-zero components of* $\widehat{\beta}$. *This is not possible here because we only have inequalities.*

For the part of the parameter space where the stronger beta-min assumption (8.7) does not hold but the weaker assumption (7.7) holds the thresholded estimator $\widetilde{\beta}$ can miss some of the relevant regressors. In Section A.3 in the Appendix we present error bounds in the case where one uses an estimated set $\widehat{J}$ such that $\widehat{J} = J(\widehat{\beta})$ or $\widehat{J} = J(\widetilde{\beta})$ but the beta-min assumption needed in Section 8.3.1, respectively Section 8.3.2, is not satisfied.

8.3.3. *Non-adaptive Confidence Sets Under a Sparsity Certificate.* The next result presents uniform joint confidence sets with confidence level at least $1 - \alpha$ without beta-min assumptions.

Theorem 8.3. *For any* β *in* $\mathcal{B}_s$, *on the event* $\mathcal{G}$, *for any solution* $(\widehat{\beta}, \widehat{\sigma})$ *of the minimization problem (3.6), and any c in* $(0, r^{-1})$, $J_0 \subseteq \{1, \ldots, K\}$, $p \geq 1$, *we have*

$$\left| \left(\mathbf{D}_{\mathbf{X}}^{-1} (\widehat{\beta} - \beta) \right)_{J_0} \right|_p \leq \frac{2\widehat{\sigma} r \theta(s)}{\kappa_{p,J_0}(s)} . \tag{8.11}$$

In addition, in the model without endogeneity, on the same event,

$$\sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(x_i^T (\widehat{\beta} - \beta) \right)^2} \leq \frac{2\widehat{\sigma} r \theta(s)}{\sqrt{\kappa_1(s)}} . \tag{8.12}$$

Replacing the sensitivities by sensitivities for enlarged cones yields confidence sets for non-sparse models. There, s is an upper bound on the dimension of the best approximating sparse model (see Theorem 6.2). This approach is similar to undersmoothing in nonparametric statistics.

The parameter c appears in the definitions of the *STIV* estimator and of the sensitivities. Choosing smaller c leads to smaller cone $C_{J(\beta)}$ and thus to larger sensitivity. This contributes to

improving the bounds. On the other hand, with smaller c we penalize less for σ in (3.6), which tends to increase the resulting $\widehat{\sigma}$ and thus the bounds. Overall, there might be some optimal c. However, the dependency of the bounds on c does not have a tractable form, which makes the optimization problematic. Importantly, the result of Theorem 8.3 is uniform in $c \in (0, r^{-1})$. Since the procedure is fast to implement, it is possible to vary c on a grid, intersect the obtained sets, and still obtain a valid confidence region of level $1-\alpha$.

The sparsity certificate approach is an alternative to the beta-min assumption. We propose to draw nested sets for increasing values of s in order to obtain honest inference statements for the whole vector of regressors when one is not willing to assume that the non-zero coefficients are large enough and there is uncertainty about the number of non-zero coordinates in the true model. This analysis can be then complemented by the confidence sets of Sections 8.3.1 and 8.3.2. This approach is analyzed numerically in Section 10.4.

9. Further Results on the STIV estimator

9.1. **Low Dimensional Models and Many Weak Instruments.** In this section we suppose that $K < n$ and that we know that all K coefficients are non-zero. We propose the following modification of the $STIV$ estimator which is a simple one stage method that can handle many instruments (L can be much greater than n) and where the strength of the instruments can be arbitrary.

Definition 9.1. *We call the* STIV-R *estimator any solution* $(\widehat{\beta}, \widehat{\sigma})$ *of the minimization problem:*

$$\min_{(\beta,\sigma)\in\widehat{\mathcal{I}}} \sigma. \tag{9.1}$$

To study this estimator, we modify the definitions of the sensitivities by dropping the cone constraints. Accordingly, we drop the index $J(\beta)$ in the notation of the sensitivities. Unlike in the setup of the previous sections (which allows for high dimensionality), these new sensitivities can be directly computed from the data and no lower bounds are required. For example, the coordinate-wise sensitivities without the cone constraint can be written as

$$\kappa_k^* = \inf_{\Delta\in\mathbb{R}^{K-1}} (\mathbf{D_X})_{kk} \max_{l=1,\dots,L} (\mathbf{D_Z})_{ll} \left| \frac{1}{n}\sum_{i=1}^n z_{li} \left(x_{ki} - \sum_{m\neq k} x_{mi}\Delta_m \right) \right| .$$

Obtaining joint confidence sets is much more direct and easy.

Theorem 9.1. *For every β in $\mathcal{I}dent$, on the event $\mathcal{G}$, for any solution $(\widehat{\beta},\widehat{\sigma})$ of the minimization problem (9.1), and any p in $[1,\infty]$, $J_0 \subseteq \{1,\dots,K\}$, we have*

$$\left|\left(\mathbf{D_X}^{-1}(\widehat{\beta}-\beta)\right)_{J_0}\right|_p \leq \frac{2\widehat{\sigma}r}{\kappa_{p,J_0}}\left(1-\frac{r^2}{\kappa_{1,J_{\rm ex}}}-\frac{r}{\kappa_{1,J_{\rm ex}^c}}\right)_+^{-1}, \tag{9.2}$$

$$\widehat{\sigma} \leq \sqrt{\widehat{Q}(\beta)}. \tag{9.3}$$

In the model without endogeneity, we have $\kappa_{1,J_{\rm ex}^c}=\infty$, and

$$\sqrt{\frac{1}{n}\sum_{i=1}^n\left(x_i^T(\widehat{\beta}-\beta)\right)^2} \leq \frac{2\widehat{\sigma}r}{\sqrt{\kappa_1}}\left(1-\frac{r^2}{\kappa_1}\right)_+^{-1}. \tag{9.4}$$

As for high-dimensional structural equations, the factor $\left(1-\frac{r^2}{\kappa_{1,J_{\rm ex}}}-\frac{r}{\kappa_{1,J_{\rm ex}^c}}\right)_+^{-1}$ in the upper bound of (9.2) can lead to infinite volume confidence sets.

9.2. **The *STIV* Estimator with Linear Projection Instrument.** In this section, we consider the case where $L>K$ but we look for a smaller set of instruments of size K. The two-stage least squares is a leading method when the structural equation is low-dimensional. Under the zero conditional mean assumption which is a stronger exogeneity condition than (1.2), optimal instruments provide a semi-parametric efficiency bound (see Amemiya (1974), Chamberlain (1987) and Newey (1990)). In the homoscedastic case, the optimal instruments correspond to the projection of the endogenous variables on the space of variables measurable with respect to all the instruments. These optimal instruments are thus regression functions. Belloni, Chen, Chernozhukov et al. (2012) proposes to use the Lasso or post-Lasso to estimate the optimal instrument and use as a second stage the heteroscedastic robust *IV* estimator to obtain confidence sets for the parameters of the low dimensional structural equation. There are no results yet on optimality of joint confidence sets for high-dimensional regression, nor for high-dimensional structural equations. However, it is a natural question to investigate a version of the classical two-stage least squares for high-dimensional structural equations when $L>K$. In this section, we present some theory for such a two-stage inference. We illustrate it in a simulation study in Section 10.

For simplicity assume that there is only one endogenous regressor $(x_{1i})_{i=1}^n$ in (1.1). We write the first stage reduced form equation as

$$x_{1i} = \sum_{l=1}^L z_{li}\zeta_l + v_i, \quad i=1,\dots,n, \tag{9.5}$$

where $\sum_{l=1}^{L} z_{li}\zeta_l$ is the linear projection instrument, ζ_l are unknown coefficients and

$$\mathbb{E}[z_{li}v_i] = 0 \ . \tag{9.6}$$

The first stage consists in estimating the unknown coefficients ζ_l. If $L \geq K > n$ and if the reduced form model (9.5) 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. It is easy to check, that the *STIV* estimator is, up to the normalization, equivalent to the Square-root Lasso when all the regressors are exogenous. 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

$$\mathcal{G} \triangleq \left\{ \max \left(\max_{l=1,\dots,L} \frac{|\mathbb{E}_n[Z_l V]|}{\sqrt{\mathbb{E}_n[Z_l^2]\mathbb{E}_n[V^2]}}, \max_{k=1,\dots,K} \frac{\left|\mathbb{E}_n[\widetilde{Z}_k U]\right|}{\sqrt{\mathbb{E}_n[\widetilde{Z}_k^2]\mathbb{E}_n[U^2]}} \right) \leq r \right\} \tag{9.7}$$

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).

Remark 9.1. *One could alternatively choose a parameter $p \in (0,1)$ and work on the event*

$$\mathcal{G} \triangleq \left\{ \max_{l=1,\dots,L} \frac{|\mathbb{E}_n[Z_l V]|}{\sqrt{\mathbb{E}_n[Z_l^2]\mathbb{E}_n[V^2]}} \leq r_1 \right\} \bigcap \left\{ \max_{k=1,\dots,K} \frac{\left|\mathbb{E}_n[\widetilde{Z}_k U]\right|}{\sqrt{\mathbb{E}_n[\widetilde{Z}_k^2]\mathbb{E}_n[U^2]}} \leq r_2 \right\}$$

where r_1 and r_2 are such that the probability of each event in the above maximum are respectively $1-p\alpha$ and $1-(1-p)\alpha$. This can be easily achieved under Scenarii 1-4. However, adjusting r based on Scenario 5 allows to properly account for the dependence between the coordinates, indeed many instruments appearing in the two terms in the above maximium are the same.

We can construct the confidence sets for the parameters ζ and β under all three cases discussed in Section 8.3. We present the case where we have a sparsity certificate s_1 for ζ . We obtain, analogously to Theorem 8.3, that for any $c_1 \in (0, r^{-1})$,

$$\left| \mathbf{D}_{\mathbf{Z}}^{-1}(\widehat{\zeta} - \zeta) \right|_1 \leq \frac{2\widehat{\sigma}_1 r \theta(s_1)}{\kappa_1^{(1)}(s_1)} \triangleq C_1(s_1) \tag{9.8}$$

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(z_i^T(\widehat{\zeta}-\zeta)\right)^2} \leq \frac{2\widehat{\sigma}_1 r\theta(s_1)}{\sqrt{\kappa_1^{(1)}(s_1)}} \triangleq C_2(s_1) \tag{9.9}$$

where $\kappa_1^{(1)}(s_1)$ denotes a lower bound on the sensitivity corresponding to the estimation of the high-dimensional reduced form equation.

The second stage makes use of the estimated instrument $(z_i^T\widehat{\zeta})_{i=1}^n$ 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

$$\widehat{\mathcal{I}}^{(2)} \triangleq \left\{(\beta,\sigma):\ \beta\in\mathbb{R}^K,\ \sigma>0,\ \left|\frac{1}{n}\mathbf{D}_{\mathbf{Z}}^{(2)}\left(\mathbf{Z}^{(2)}\right)^T(\mathbf{Y}-\mathbf{X}\beta)\right|_\infty \leq \sigma r,\ \widehat{Q}(\beta)\leq\sigma^2\right\} \tag{9.10}$$

where $\mathbf{D}_{\mathbf{Z}}^{(2)}$ is a $K\times K$ diagonal matrix such that $(\mathbf{D}_{\mathbf{Z}}^{(2)})_{11} = \left(C_1(s_1)+C_2(s_1)+\mathbb{E}_n[(\widehat{\zeta}^TZ)^2]^{1/2}\right)^{-1}$, $(\mathbf{D}_{\mathbf{Z}}^{(2)})_{kk} = (\mathbf{D}_{\mathbf{X}})_{kk}$ for $k=2,\ldots,K$, and the matrix $\mathbf{Z}^{(2)}$ is the stacked matrix of the estimated linear projection instrument $(z_i^T\widehat{\zeta})_{i=1}^n$ 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 Ψ_n, which differs from the original one in that we replace $\mathbf{Z}$ by $\mathbf{Z}^{(2)}$ and $\mathbf{D}_{\mathbf{Z}}$ by $\mathbf{D}_{\mathbf{Z}}^{(2)}$. We assign the upper index (2) to the sensitivities corresponding to this new matrix Ψ_n, for example, $\kappa^{(2)}_{1,J(\beta),J(\beta)}$. Then we have the following theorem.

Theorem 9.2. *For any β in $\mathcal{I}dent$, on the event $\mathcal{G}$, for any solution $(\widehat{\beta},\widehat{\sigma})$ of the minimization problem (3.6) where we replace $\widehat{\mathcal{I}}$ by $\widehat{\mathcal{I}}^{(2)}$ and c by c_2, for any $c_2\in(0,r^{-1})$, p in $[1,\infty]$, and $J_0\subseteq\{1,\ldots,K\}$, we have*

$$\left|\left(\mathbf{D}_{\mathbf{X}}^{-1}(\widehat{\beta}-\beta)\right)_{J_0}\right|_p \leq \frac{2\widehat{\sigma}r}{\kappa^{(2)}_{p,J_0,J(\beta)}}\left(1-\frac{r}{\kappa^{(2)}_{1,J(\beta)}}-\frac{r^2}{\kappa^{(2)}_{1,\{2,\ldots,K\},J(\beta)}}\right)_+^{-1} \tag{9.11}$$

and

$$\widehat{\sigma}\leq\sqrt{\widehat{Q}(\beta)}\left(1+\frac{r}{c\kappa^{(2)}_{1,J(\beta),J(\beta)}}\right)\left(1-\frac{r}{c\kappa^{(2)}_{1,J(\beta),J(\beta)}}\right)_+^{-1}. \tag{9.12}$$

In addition, for any β in $\mathcal{B}_s$, on the event $\mathcal{G}$, for any solution $(\widehat{\beta},\widehat{\sigma})$ of the minimization problem (3.6) where we replace $\widehat{\mathcal{I}}$ by $\widehat{\mathcal{I}}^{(2)}$ and c by c_2, for any c_2 in $(0,r^{-1})$, p in $[1,\infty]$, and $J_0\subseteq\{1,\ldots,K\}$, we have

$$\left|\left(\mathbf{D}_{\mathbf{X}}^{-1}(\widehat{\beta}-\beta)\right)_{J_0}\right|_p \leq \frac{2\widehat{\sigma}r\theta^{(2)}(s_2)}{\kappa^{(2)}_{p,J_0}(s)}. \tag{9.13}$$

Inequality (9.13) yields uniform joint confidence sets for all $k = 1, \dots, K$, c_2 in $(0, r^{-1})$:

$$|\widehat{\beta}_k - \beta_k| \leq \frac{2\widehat{\sigma} r \theta^{(2)}(s_2)}{\mathbb{E}_n[X_k^2]^{1/2}\ \kappa_k^{(2)*}} \tag{9.14}$$

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 7.

9.3. **Models with Possibly Non-valid Instruments.** We now study the problem of checking instrument exogeneity when there is overidentification. This 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), Andrews and Lu (2001) and Liao (2013) and the references therein). We propose a two-stage method based on the *STIV* estimator. The main purpose of the suggested method is to construct confidence sets for non-validity indicators, and to detect non-valid (*i.e.*, endogenous) instruments in the high-dimensional framework. The model can be written in the form: for every $i = 1, \dots, n$,

$$y_i = x_i^T \beta + u_i, , \tag{9.15}$$

$$\mathbb{E}\left[z_i u_i\right] = 0, \tag{9.16}$$

$$\mathbb{E}\left[\overline{z}_i u_i\right] = \theta, \tag{9.17}$$

where x_i, z_i, and $\overline{z}_i$ are vectors of dimensions K, L and $\overline{L}$, respectively. We observe realizations of independent random vectors $(y_i, x_i^T, z_i^T, \overline{z}_i^T)$, $i = 1, \dots, n$. For simplicity assume that (9.15)-(9.16) is point identified. The instruments are decomposed in two parts, z_i and $\overline{z}_i$, where $\overline{z}_i^T = (\overline{z}_{1i}, \dots, \overline{z}_{\overline{L}i})$ is a vector of possibly non-valid instruments. A component of the unknown vector $\theta \in \mathbb{R}^{\overline{L}}$ is zero when the corresponding instrument is indeed valid. The component θ_l of θ will be called the *non-validity indicator* of the instrument $\overline{z}_{li}$. Our study covers the models with dimensions K, L and $\overline{L}$ that can be much larger than the sample size. We denote by $\overline{\mathbf{Z}}$ the matrix of dimension $n \times \overline{L}$ with rows $\overline{z}_i^T$, $i = 1, \dots, n$ and by $\mathbf{D}_{\overline{\mathbf{Z}}}$ the normalization matrix such that $(\mathbf{D}_{\overline{\mathbf{Z}}})_{ll} = \mathbb{E}_n[\overline{Z}_l^2]^{-1/2}$ for $l = 1, \dots, \overline{L}$. We set $\overline{\Psi}_n = \frac{1}{n}\mathbf{D}_{\overline{\mathbf{Z}}}\overline{\mathbf{Z}}^T\ \mathbf{X}\mathbf{D}_{\mathbf{X}}$ and $\overline{z}_* = \max_{\substack{l=1,\dots,\overline{L},\\ i=1,\dots,n}} |z_{li}|/\sqrt{\mathbb{E}_n[\overline{Z}_l^2]}$. We assume that we have a pilot estimator $\widehat{\beta}$ and two statistics $\widehat{b}$ and $\widehat{b}_1$ such that, on the event $\mathcal{G}$ of Section 5,

$$\left|\overline{\Psi}_n \mathbf{D}_{\mathbf{X}}^{-1}(\widehat{\beta} - \beta)\right|_\infty \leq \widehat{b}, \quad \left|\mathbf{D}_{\mathbf{X}}^{-1}(\widehat{\beta} - \beta)\right|_1 \leq \widehat{b}_1\ . \tag{9.18}$$

For example, $\widehat{\beta}$ can be the *STIV* estimator based on the vectors of instruments z_i that are known to be valid, with constants c and r. The sensitivity associated to the first bound in (9.18) is defined as

$$\overline{\kappa}_J \triangleq \inf_{\Delta \in C_J:\ \left|\overline{\Psi}_n \Delta\right|_\infty = 1} |\Psi_n \Delta|_\infty \ . \tag{9.19}$$

A lower bound for $\overline{\kappa}_J$ can be obtained using the inclusion $\widehat{J} \supseteq J(\beta)$ for an estimator $\widehat{J}$ or using the sparsity certificate approach and working with the lower bound $\overline{\kappa}(s)$ of Table 7, which reduces to solving $2K\overline{L}$ simple convex programs. This yields the explicit expressions

$$\widehat{b} = \frac{2\widehat{\sigma} r \theta(s)}{\overline{\kappa}(s)}, \quad \widehat{b}_1 = \frac{2\widehat{\sigma} r \theta(s)}{\kappa_1(s)} \ . \tag{9.20}$$

In the case of non-sparse structural equations we use the expressions for the enlarged cone. According Theorem 7.1, if Assumptions 7.1 and 7.2 hold and if $\overline{\kappa}_{J(\beta)} \geq \overline{\kappa}_*$ on the event $\mathcal{G}_1$, then there exist non-random constants b_* and b_{1*} such that on the event $\mathcal{G} \cap \mathcal{G}_1 \cap \mathcal{G}_2$,

$$\widehat{b} \leq \frac{2\sigma_* r \tau_*}{\overline{\kappa}_*} \triangleq b_*, \quad \widehat{b}_1 \leq \frac{2\sigma_* r \tau_* c_{J(\beta)}}{\kappa_*} \triangleq b_{1*} \ . \tag{9.21}$$

We define the *STIV-NV* estimator $(\widehat{\theta}, \widehat{\overline{\sigma}})$ as any solution of the problem

$$\min_{(\theta,\overline{\sigma}) \in \widehat{\mathcal{I}}_1} \left(\left|\mathbf{D}_{\overline{\mathbf{Z}}} \theta\right|_1 + \overline{c}\ \overline{\sigma} \right), \tag{9.22}$$

where $0 < \overline{c} < \overline{r}^{-1}$ and

$$\widehat{\overline{\mathcal{I}}} \triangleq \left\{ (\theta, \overline{\sigma}) :\ \theta \in \mathbb{R}^{\overline{L}},\ \overline{\sigma} > 0,\ \left| \mathbf{D}_{\overline{\mathbf{Z}}} \left(\frac{1}{n} \overline{\mathbf{Z}}^T (\mathbf{Y} - \mathbf{X}\widehat{\beta}) - \theta \right) \right|_\infty \leq \overline{\sigma}\ \overline{r} + \widehat{b},\ F(\theta, \widehat{\beta}) \leq \overline{\sigma} + \widehat{b}_1 \overline{z}_* \right\}$$

for some $\overline{r} > 0$ (to be specified below), where for all $\theta = (\theta_1, \ldots, \theta_{\overline{L}}) \in \mathbb{R}^{\overline{L}}$, $\beta \in \mathbb{R}^K$,

$$F(\theta, \beta) \triangleq \max_{l=1,\ldots,\overline{L}} \sqrt{\widehat{Q}_l(\theta, \beta)}$$

with

$$\widehat{Q}_l(\theta, \beta) \triangleq \frac{\mathbb{E}_n[\left(\overline{Z}_l (Y - X^T\beta) - \theta_l\right)^2]}{\mathbb{E}_n[\overline{Z}_l^2]} \ .$$

It is not hard to see that the optimization problem (9.22) can be re-written as a conic program.

The following theorem provides a basis for constructing confidence sets for the non-validity indicators. Define the random event

$$\mathcal{G}' = \left\{ \max_{l=1,\ldots,\overline{L}} \frac{\left|\mathbb{E}_n\left[\overline{Z}_l U - \theta_l\right]\right|}{\sqrt{\mathbb{E}_n\left[(\overline{Z}_l U - \theta_l)^2\right]}} \leq \overline{r} \right\}$$

where $\overline{r}$ is the analogue of constant r_0 in Scenarios 1-4. For simplicity we do not consider Scenario 5 in this section.

Theorem 9.3. *On the event $\mathcal{G} \cap \mathcal{G}'$, for any estimator $\widehat{\beta}$ satisfying (9.18) on $\mathcal{G}$, any solution $(\widehat{\theta}, \widehat{\overline{\sigma}})$ of the minimization problem (9.22), and any $\overline{c}$ in $(0, \overline{r}^{-1})$, we have*

$$\left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\widehat{\theta}-\theta\right)\right|_\infty \le 2\left(1-\frac{2\overline{r}^2|J(\theta)|}{1-\overline{c}\ \overline{r}}\right)_+^{-1}\left(\widehat{\overline{\sigma}}\ \overline{r}+\widehat{b}+\left(1+\frac{\overline{c}}{1-\overline{c}\ \overline{r}}\right)\widehat{b}_1\overline{z}_*\overline{r}\right) \triangleq V(\widehat{\overline{\sigma}},\widehat{b},\widehat{b}_1,|J(\theta)|) \tag{9.23}$$

$$\left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\widehat{\theta}-\theta\right)\right|_1 \le \frac{2}{1-\overline{c}\ \overline{r}}\left(1-\frac{2\overline{r}^2|J(\theta)|}{1-\overline{c}\ \overline{r}}\right)_+^{-1}\left(2\left(\widehat{\overline{\sigma}}\ \overline{r}+\widehat{b}_1\overline{z}_*\overline{r}+\widehat{b}\right)|J(\theta)|+\overline{c}\widehat{b}_1\overline{z}_*\right). \tag{9.24}$$

This theorem gives a meaningful result when $\overline{r}$ is small, *i.e.*, $n \gg \log(\overline{L})$. In addition, we need small $\widehat{b}$ and $\widehat{b}_1$, which is guaranteed by the results of Section 6 under the condition $n \gg \log(L)$ if the pilot estimator $\widehat{\beta}$ is the *STIV* estimator. Note also that the bounds (9.23) and (9.24) are meaningful if their denominators are positive, which is roughly equivalent to the following bound on the sparsity of θ: $|J(\theta)| = O\left(1/\overline{r}^2\right) = O\left(n/\log(\overline{L})\right)$. Bounds for all ℓ_p norms follow immediately from (9.23) and (9.24) by the standard interpolation argument.

To turn (9.23) and (9.24) into valid confidence bounds, we use a sparsity certificate $\overline{s}$ or replace there $|J(\theta)|$ by $|J(\widetilde{\theta})|$ or $|J(\widehat{\theta})|$, as justified by Theorem 9.4 below. To state the theorem, we need an extra assumption that the random variable $F(\theta,\beta)$ is bounded in probability by a constant $\overline{\sigma}_* > 0$.

Assumption 9.1. *There exist constants $\overline{\sigma}_* > 0$, $\overline{v}_l > 0$ for $l = 1,\dots,\overline{L}$ and $0 < \varepsilon < 1$ such that, with probability at least $1-\varepsilon$,*

$$\max_{l=1,\dots,\overline{L}} \frac{\mathbb{E}_n[\left(\overline{Z}_l U-\theta_l\right)^2]}{\mathbb{E}_n[\overline{Z}_l^2]} \le \overline{\sigma}_*^2, \quad \forall l = 1,\dots,\overline{L},\ \mathbb{E}_n[\overline{Z}_l^2] \le \overline{v}_l\ . \tag{9.25}$$

Denote by $\mathcal{G}_3$ the corresponding event. As in (8.6), we define a thresholded estimator

$$\widetilde{\theta}_l \triangleq \begin{cases} \widehat{\theta}_l & \text{if } |\widehat{\theta}_l| > \omega_l, \\ 0 & \text{otherwise,} \end{cases} \tag{9.26}$$

where $\omega_l > 0$ for $l = 1,\dots,\overline{L}$, are some thresholds.

Theorem 9.4. *Let Assumptions 7.1, 7.2, and 9.1 be satisfied, and let $\overline{\kappa}_{J(\beta)} \ge \overline{\kappa}_*$ on the event $\mathcal{G}_1$. Assume that $|J(\theta)| \le \overline{s}$ and consider the solutions $(\widehat{\beta},\widehat{\sigma})$ and $(\widehat{\theta},\widehat{\overline{\sigma}})$ of the minimization problems (3.6) and (9.22) respectively.*

(i) *On the event $\mathcal{G} \cap \mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}' \cap \mathcal{G}_3$ we have*

$$\widehat{\overline{\sigma}} \le \overline{\sigma}_* + \frac{|J(\theta)|}{\overline{c}} V(\widehat{\overline{\sigma}},\widehat{b},\widehat{b}_1,|J(\theta)|) \le \overline{\sigma}_* + \frac{|J(\theta)|}{\overline{c}} V(\overline{\sigma}_*,b_*,b_{1*},|J(\theta)|), \tag{9.27}$$

$$\left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\widehat{\theta}-\theta\right)\right|_\infty \le \left(1-\frac{2\overline{s}\ \overline{r}}{\overline{c}(1-\overline{c}\ \overline{r})}\right)_+^{-1}\left(1-\frac{2\overline{s}\ \overline{c}\ \overline{r}^2}{\overline{c}(1-\overline{c}\ \overline{r})}\right) V(\overline{\sigma}_*,b_*,b_{1*},\overline{s})\ . \tag{9.28}$$

(ii) *Assume, in addition, that* $|\theta_l| > \overline{v}_l \left(1 - \frac{2\overline{s}\ \overline{r}}{\overline{c}(1-\overline{c}\ \overline{r})}\right)_+^{-1} \left(1 - \frac{2\overline{s}\ \overline{c}\ \overline{r}^2}{\overline{c}(1-\overline{c}\ \overline{r})}\right) V(\overline{\sigma}_*, b_*, b_{1*}, \overline{s})$ *for all* $l \in J(\theta)$. *Then, on the event* $\mathcal{G} \cap \mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}' \cap \mathcal{G}_3$ *we have*

$$J(\theta) \subseteq J(\widehat{\theta}). \tag{9.29}$$

(iii) *Let* $|J(\beta)| \leq s$ *and replace Assumption 7.1 by Assumption 8.1, assuming in addition that* $\overline{\kappa}_{J(\beta)} \geq \overline{\kappa}_*(s)$ *on the event* $\mathcal{G}_1$. *Define*

$$b_*(s) = \frac{2\sigma_* r \tau_*(s)}{\overline{\kappa}_*(s)}, \quad b_{1*}(s) = \frac{2\sigma_* r \tau_*(s) c(s)}{\kappa_*(s)}\,. \tag{9.30}$$

Assume that $|\theta_l| > 2\overline{v}_l \left(1 - \frac{2\overline{s}\ \overline{r}}{\overline{c}(1-\overline{c}\ \overline{r})}\right)_+^{-1} \left(1 - \frac{2\overline{s}\ \overline{c}\ \overline{r}^2}{\overline{c}(1-\overline{c}\ \overline{r})}\right) V(\overline{\sigma}_*, b_*(s), b_{1*}(s), \overline{s})$ *for all* $l \in J(\theta)$. *Let* $\widetilde{\theta}$ *be the thresholded estimator defined in (9.26) where* $\widehat{\theta}$ *is any solution of the minimization problem (9.22), and* $\omega_l = \mathbb{E}_n[\overline{Z}_l^2]^{1/2} \left(1 - \frac{2\overline{s}\ \overline{r}}{\overline{c}(1-\overline{c}\ \overline{r})}\right)_+^{-1} \left(1 - \frac{2\overline{s}\ \overline{c}\ \overline{r}^2}{\overline{c}(1-\overline{c}\ \overline{r})}\right) V(\widehat{\overline{\sigma}}, \widehat{b}, \widehat{b}_1, \overline{s})$. *Then, on the event* $\mathcal{G} \cap \mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}' \cap \mathcal{G}_3$, *we have* $J(\widetilde{\theta}) = J(\theta)$ *and*

$$\overrightarrow{\operatorname{sign}(\widetilde{\theta})} = \overrightarrow{\operatorname{sign}(\theta)}\,. \tag{9.31}$$

Remark 9.2. *We presented here the approach based on working on the two events* $\mathcal{G}$ *and* $\mathcal{G}'$ *and using the union bound. One could alternatively use the approach of Section 9.2.*

10. Simulation Study

In this section, we discuss the performance of the *STIV* estimator on simulated data. We consider the following model:

$$y_i = x_{1i}\beta_1 + \sum_{k=2}^{K} x_{ki}\beta_k + u_i,$$
$$x_{1i} = \sum_{l=1}^{L} z_{li}\zeta_l + v_i,$$

where $(y_i, x_i^T, z_i^T, u_i, v_i)$ are i.i.d., (u_i, v_i) have the normal distribution $\mathcal{N}\left(0, \begin{pmatrix} \sigma_{\text{struct}}^2 & \rho\sigma_{\text{struct}}\sigma_{\text{end}} \\ \rho\sigma_{\text{struct}}\sigma_{\text{end}} & \sigma_{\text{end}}^2 \end{pmatrix}\right)$. We set $\rho = 0.3$, $\sigma_{\text{struct}} = \sigma_{\text{end}} = 1$, $\beta = (1, -2, -0.5, 0.25, -1, 0, \ldots, 0)^T$ and $\alpha = 0.05$. We adjust r everywhere based on Scenario 5 and use the CVX package (see Grant and Boyd (2013)) for the optimization routines. We simulate the estimators both for $K > n$ and $K < n$. For the confidence sets, we will deal only with $K < n$ but we take K so big that the BIC type techniques are not computationally feasible (see, *e.g.*, Andrews and Lu 2001). We use the results of Section A.2 to calculate the lower bounds on the sensitivities when they are based on an estimated support $\widehat{J}$.

10.1. **Estimation when $K > n$, $K \gg L$.** Take $n = 500$, $L = 30$ and $K = 600$. Thus, we are under the high-dimensional scenario. There are much more regressors than variables known to be exogenous and used as instruments. To complete the description of the data generating process we set

$$x_{l'i} = z_{li} \quad \text{for } l' = l + 1 \ \text{ and } \ l \in \{1, \dots, L\}$$

and take $(x_{2i}, \dots, x_{Ki})^T$ to be a vector of independent standard normal random variables truncated to the interval $[-5, 5]$ and independent of (u_i, v_i). We take $\zeta = (0.3, \dots, 0.3)^T$. We choose c very close to the uper bound r^{-1} allowed by the theory of this paper. This corresponds to the smallest shrinkage to zero. The results are summarized in Table 1. In agreement with what is explained at the end of Section 6.1, sparsity provides exclusion restrictions, and for each possible submodel there is overidentification.

TABLE 1. Monte-Carlo study, 1000 replications

	5^{th} percentile	Median	95^{th} percentile		5^{th} percentile	Median	95^{th} percentile
$\widehat{\beta}_1$	1.058	1.131	1.202	$\widehat{\beta}_8$	-1.32 10^{-8}	-4.18 10^{-10}	9.60 10^{-9}
$\widehat{\beta}_2$	-1.939	-1.869	-1.794	β_9	-1.36 10^{-8}	-3.53 10^{-10}	1.07 10^{-8}
$\widehat{\beta}_3$	-0.446	-0.368	-0.289	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\widehat{\beta}_4$	0.026	0.105	0.187	$\widehat{\beta}_{598}$	-1.25 10^{-8}	3.25 10^{-11}	1.21 10^{-8}
$\widehat{\beta}_5$	-0.946	-0.869	-0.796	$\widehat{\beta}_{599}$	-1.28 10^{-8}	2.16 10^{-10}	1.31 10^{-8}
$\widehat{\beta}_6$	-1.39 10^{-8}	-5.84 10^{-10}	8.91 10^{-9}	$\widehat{\beta}_{600}$	-1.13 10^{-8}	-3.21 10^{-12}	1.19 10^{-8}
$\widehat{\beta}_7$	-1.37 10^{-8}	-3.83 10^{-10}	9.95 10^{-9}	$\widehat{\sigma}$	0.947	1.000	1.051

We used: $c = 1/(r * 1.001)$.

In the simulation design, the coefficients of the reduiced form equation ζ as well as β_3 and β_4 are at least 10 times smaller than the detection level (see Theorem 8.1). 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 would expect from Theorem 7.1 that β_3 and β_4 should be impossible to distinguish from 0. The results highlighted in the grey box correspond to the non-zero coefficients of the estimator. All coefficients outside the grey box are at most of the order of 10^{-8}. This is much below the level of accuracy we can expect and can be considered to be 0. The non-zero coefficients of the estimator are the true non-zero coefficients in all cases reported in Table 1.

10.2. **Choice of c for Confidence Sets.** In this simulation study, we did not intersect the sets based on the sparsity certificate approach for different values of c. So, the confidence sets are more conservative than they could have been. We rather started by estimating each model with $c = r^{-1}$

and compared the point estimators obtained for decreasing c. For sufficiently large sample size, the estimators remain almost unchanged when c decreases, until the point where $\widehat{\beta}$ becomes almost zero and $\widehat{\sigma}$ starts to increase. We chose c around that value.

10.3. **Confidence Sets: Less Instruments than Potential Regressors.** Under the simulation design of Section 10.1, the confidence sets are infinite. We now change only two parameters and take $n = 8000$ and $K = 50$. We have still $K \gg L$. The simulations show that the 5 true non-zero coefficients are clearly detected to be non-zero based on the value of the *STIV* estimator. Out of the 45 other estimated coefficients, 9 are of the order of 10^{-2}, all others are of the order of 10^{-3} or below. These larger coefficients correspond to the exogenous variables used as instruments. They also have the larger coordinate-wise sensitivities which are about 0.3. The estimated coefficients for the exogenous regressors that are not used as instruments are at most of the order of 10^{-5} but their sensitivities are as small as 0.07, which is roughly the same as for the endogenous regressors. This leads to wider confidence intervals for them, roughly twice as large as for the endogenous regressors, due to the weights $\mathbf{D_X}$. The confidence sets based on sparsity certificate with $s = 5$ are infinite. However, those based on the estimated support $\widehat{J} = \{1, \ldots, 5\}$ are finite. They are presented in Table 2.

TABLE 2. Less instruments than potential regressors, estimated support

	$\widehat{\beta}_l$	$\widehat{\beta}$	$\widehat{\beta}_u$	κ_k^*		$\widehat{\beta}_l$	$\widehat{\beta}$	$\widehat{\beta}_u$	κ_k^*
β_1	-0.47	0.99	2.44	0.07	β_6	-0.59	0.02	0.62	0.34
β_2	-2.62	-2.01	-1.39	0.33	β_7	-0.6	-0.01	0.58	0.34
β_3	-1.13	-0.48	0.17	0.31	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_4	-0.39	0.27	0.92	0.31	β_{49}	-2.97	0	2.97	0.07
β_5	-1.6	-0.99	-0.38	0.33	β_{50}	-2.95	0	2.95	0.07

Here: $r = 0.035$, $c = 0.1792$ and $\widehat{\sigma} = 1.014$.

10.4. **Confidence Sets: One Strong Instrument and Many Weak Instruments.** We set $K = 150$, $L = 152$, $n = 4000$,

$$x_{l'i} = z_{li} \quad \text{for } l' = l + 1 \ \text{ and } \ l \in \{1, \ldots, K-1\}$$

and we take $(z_{1i}, \ldots, z_{Li})^T$ as a vector of independent standard normal random variables truncated to the interval $[-5, 5]$ and independent of (u_i, v_i). We take $\zeta = (1, a, \ldots, a)^T$ where $a = 0.1$. This model cannot be estimated using the BIC since this would require solving more than 10^{45} least squares problems.

The results are presented in Table 3. We use the notation $\widehat{\beta}_{l,s}$ and $\widehat{\beta}_{u,s}$ for the lower and upper bounds computed using the sparsity certificate for various degrees of sparsity s. The confidence sets are nested. The column "Selection" shows under which sparsity certificate s the variable is selected based on Theorem 8.1. We see from Table 3 that using the thresholding rule for $s = 8$ yields $\widehat{J}_1 \triangleq J(\widetilde{\beta}) = \{1, 2, 3, 5\}$. The choice $s = 8$ is conservative. Indeed we have $|J(\widehat{\beta})| = 5$. Recall that under the premise of Theorem 7.1 we have $J(\beta) \subseteq \widehat{J}_2 \triangleq J(\widehat{\beta}) = \{1, 2, 3, 4, 5\}$. We present in Table 4 the confidence sets obtained using (8.9) with $\widehat{J}_1$ and $\widehat{J}_2$.

TABLE 3. One strong instrument and all others equally weak, sparsity certificate

	$\widehat{\beta}_{l,10}$	$\widehat{\beta}_{l,9}$	$\widehat{\beta}_{l,8}$	$\widehat{\beta}_{l,7}$	$\widehat{\beta}_{l,6}$	$\widehat{\beta}_{l,5}$	$\widehat{\beta}_{l,4}$	$\widehat{\beta}$	Selection	$\widehat{\beta}_{u,4}$	$\widehat{\beta}_{u,5}$	$\widehat{\beta}_{u,6}$	$\widehat{\beta}_{u,7}$	$\widehat{\beta}_{u,8}$	$\widehat{\beta}_{u,9}$	$\widehat{\beta}_{u,10}$
β_1	0.3	0.36	0.42	0.49	0.55	0.62	0.67	0.94	≥ 10	1.2	1.26	1.32	1.39	1.45	1.51	1.57
β_2	-2.46	-2.41	-2.36	-2.32	-2.27	-2.23	-2.19	-1.95	≥ 10	-1.71	-1.67	-1.63	-1.58	-1.54	-1.49	-1.44
β_3	-0.94	-0.89	-0.84	-0.79	-0.75	-0.7	-0.66	-0.43	≤ 8	-0.2	-0.16	-0.11	-0.07	-0.02	0.03	0.08
β_4	-0.35	-0.3	-0.25	-0.2	-0.15	-0.11	-0.06	0.18	< 4	0.43	0.47	0.52	0.56	0.61	0.67	0.72
β_5	-1.51	-1.46	-1.42	-1.37	-1.32	-1.27	-1.23	-0.98	≥ 10	-0.73	-0.68	-0.64	-0.59	-0.54	-0.49	-0.45
β_6	-0.52	-0.47	-0.43	-0.38	-0.34	-0.29	-0.25	0	0	0.25	0.29	0.34	0.38	0.43	0.47	0.52
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_{150}	-0.54	-0.49	-0.43	-0.39	-0.34	-0.29	-0.25	0	0	0.25	0.29	0.34	0.39	0.43	0.49	0.54

Here: $r = 0.057$ and $c = 0.353$ and $\widehat{\sigma} = 1.010$

TABLE 4. One strong instrument and all others equally weak, estimated support

	$\widehat{\beta}_{l,\widehat{J}_2}$	$\widehat{\beta}_{l,\widehat{J}_1}$	$\widehat{\beta}$	$\widehat{\beta}_{u,\widehat{J}_1}$	$\widehat{\beta}_{u,\widehat{J}_2}$
β_1	0.75	0.76	0.94	1.11	1.13
β_2	-2.15	-2.14	-1.95	-1.76	-1.74
β_3	-0.63	-0.62	-0.43	-0.24	-0.23
β_4	-0.02	0	0.18	0.36	0.38
β_5	-1.19	-1.17	-0.98	-0.78	-0.77
β_6	-0.2	-0.19	0	0.19	0.2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_{150}	-0.2	-0.19	0	0.19	0.2

10.5. **Confidence Sets: Sparse Reduced Form.** Consider now the two-stage method from Section 9.2 based on an estimated linear projection instrument, akin to the two-stage least squares. Let $n = 8000$, $K = 70$ and $L = 100$. We take $\zeta_3 = -0.5$, $\zeta_4 = 1$, $\zeta_{98} = -1$, $\zeta_{99} = 1$, $\zeta_{100} = 0.5$, and we set all other coefficients in the reduced form equation equal to zero. Tables 5 and 6 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 $C_\infty(5) = 2\widehat{\sigma}_1 r\theta_1(5)/\kappa_\infty^{(1)}(5) = 0.112$. Using the estimated support $\widehat{J} = J(\widehat{\zeta}) = \{3, 4, 98, 99, 100\}$ yields (with obvious meaning of the notation) $C_1(\widehat{J}) = 0.923$, $C_2(\widehat{J}) = 0.279$. We present in Table 6 the results based on the two approaches for 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 (9.7).

TABLE 5. Sparse reduced form, sparsity certificate, one stage

	$\widehat{\beta}_{l,10}$	$\widehat{\beta}_{l,9}$	$\widehat{\beta}_{l,8}$	$\widehat{\beta}_{l,7}$	$\widehat{\beta}_{l,6}$	$\widehat{\beta}_{l,5}$	$\widehat{\beta}_{l,4}$	$\widehat{\beta}$	Selection	$\widehat{\beta}_{u,4}$	$\widehat{\beta}_{u,5}$	$\widehat{\beta}_{u,6}$	$\widehat{\beta}_{u,7}$	$\widehat{\beta}_{u,8}$	$\widehat{\beta}_{u,9}$	$\widehat{\beta}_{u,10}$
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 $C_1(s)$, $C_2(s)$, $C_1(\widehat{J})$ and $C_2(\widehat{J})$ 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. Still we find that they are larger than those of the one-stage method. Finding optimal joint confidence sets in high-dimensional linear regression and structural equations is an open problem that requires further study.

TABLE 6. 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 $\widehat{\sigma} = 1.042$:

	$\widehat{\beta}_{l,10}$	$\widehat{\beta}_{l,9}$	$\widehat{\beta}_{l,8}$	$\widehat{\beta}_{l,7}$	$\widehat{\beta}_{l,6}$	$\widehat{\beta}_{l,5}$	$\widehat{\beta}_{l,4}$	$\widehat{\beta}$	Selection	$\widehat{\beta}_{u,4}$	$\widehat{\beta}_{u,5}$	$\widehat{\beta}_{u,6}$	$\widehat{\beta}_{u,7}$	$\widehat{\beta}_{u,8}$	$\widehat{\beta}_{u,9}$	$\widehat{\beta}_{u,10}$
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
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
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 $C_1(\widehat{J})$ and $C_2(\widehat{J})$ with $c = 0.444$ and $\widehat{\sigma} = 1.010$:

	$\widehat{\beta}_{l,10}$	$\widehat{\beta}_{l,9}$	$\widehat{\beta}_{l,8}$	$\widehat{\beta}_{l,7}$	$\widehat{\beta}_{l,6}$	$\widehat{\beta}_{l,5}$	$\widehat{\beta}_{l,4}$	$\widehat{\beta}$	Selection	$\widehat{\beta}_{u,4}$	$\widehat{\beta}_{u,5}$	$\widehat{\beta}_{u,6}$	$\widehat{\beta}_{u,7}$	$\widehat{\beta}_{u,8}$	$\widehat{\beta}_{u,9}$	$\widehat{\beta}_{u,10}$
β_1	0.68	0.69	0.69	0.7	0.71	0.72	0.74	0.99	≥ 10	1.25	1.26	1.28	1.28	1.29	1.3	1.31
β_2	-2.22	-2.22	-2.21	-2.2	-2.19	-2.18	-2.17	-2.01	≥ 10	-1.85	-1.83	-1.82	-1.81	-1.8	-1.8	-1.79
β_3	-0.71	-0.7	-0.7	-0.69	-0.68	-0.66	-0.65	-0.5	≥ 10	-0.34	-0.33	-0.31	-0.3	-0.29	-0.29	-0.28
β_4	-0.11	-0.1	-0.09	-0.08	-0.07	-0.06	-0.04	0.24	≤ 5	0.51	0.53	0.54	0.55	0.56	0.57	0.58
β_5	-1.48	-1.47	-1.45	-1.44	-1.42	-1.4	-1.37	-0.99	≥ 10	-0.61	-0.58	-0.56	-0.54	-0.53	-0.51	-0.5
β_6	-0.19	-0.19	-0.18	-0.17	-0.16	-0.15	-0.13	0.01	0	0.16	0.17	0.18	0.19	0.2	0.21	0.22
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_{70}	-0.2	-0.19	-0.19	-0.18	-0.17	-0.16	-0.15	0	0	0.16	0.17	0.18	0.19	0.19	0.2	0.21

Second stage based on the too optimistic choice $C_1 = 0$ and $C_2 = 0$ with $c = 0.376$ and $\widehat{\sigma} = 1.001$:

	$\widehat{\beta}_{l,10}$	$\widehat{\beta}_{l,9}$	$\widehat{\beta}_{l,8}$	$\widehat{\beta}_{l,7}$	$\widehat{\beta}_{l,6}$	$\widehat{\beta}_{l,5}$	$\widehat{\beta}_{l,4}$	$\widehat{\beta}$	Selection	$\widehat{\beta}_{u,4}$	$\widehat{\beta}_{u,5}$	$\widehat{\beta}_{u,6}$	$\widehat{\beta}_{u,7}$	$\widehat{\beta}_{u,8}$	$\widehat{\beta}_{u,9}$	$\widehat{\beta}_{u,10}$
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
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
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: $r = 0.041$.

TABLE 7. Table of constants

$$
\begin{aligned}
c_J &\triangleq \min\left(\frac{2|J|}{(1-c)_+}, \frac{2|J_{\rm ex}\cap J|+(2+c(1-r))|J_{\rm ex}^c\cap J|+c(1-r)|J_{\rm ex}^c\cap J^c|}{(1-cr)_+}\right)\\
c_{J_{\rm ex},J} &\triangleq \min\left(\frac{2|J|}{(1-c)_+}, \frac{2|J_{\rm ex}\cap J|+(1+c)|J_{\rm ex}^c\cap J|+(c-1)|J_{\rm ex}^c\cap J^c|}{(1-cr)_+}\right)\\
\underline{c}_{J_{\rm ex}^c,J} &\triangleq \frac{2|J_{\rm ex}^c\cap J|+(1+cr)|J_{\rm ex}\cap J|}{(1-c)_+}\\
c(s) &\triangleq \min\left(\frac{2s}{(1-c)_+}, \frac{2s+c(1-r)|J_{\rm ex}^c|}{1-cr}\right)\\
c_{J_{\rm ex}}(s) &\triangleq \min\left(\frac{2s}{(1-c)_+}, \frac{\max(2,1+c)s+(c-1)_+|J_{\rm ex}^c|}{(1-cr)_+}, \frac{2s+2c|J_{\rm ex}^c|}{(1-cr)_+}\right)\\
\underline{c}_{J_{\rm ex}^c}(s) &\triangleq \frac{\min(\max(2,1+cr)s,(1+cr)s+|J_{\rm ex}^c|)}{(1-c)_+}\\
\kappa_\infty(J,s) &\triangleq \min_{\substack{j\in J\\ k=1,\ldots,K}}\left\{\min_{\substack{\Delta_k=\pm1,\ |\Delta_J|_\infty\le\Delta_j,\ |\Delta|_\infty\le1,\\ (1-cr)|\Delta_{J_{\rm ex}}|_1+(1-c)|\Delta_{J_{\rm ex}^c}|_1\le2s\Delta_j}}|\Psi_n\Delta|_\infty\right\}\\
\kappa_k^*(J,s) &\triangleq \min_{j\in J}\left\{\min_{\substack{\Delta_k=\pm1,\ |\Delta_J|_\infty\le\Delta_j\\ (1-cr)|\Delta_{J_{\rm ex}}|_1+(1-c)|\Delta_{J_{\rm ex}^c}|_1\le2s\Delta_j}}|\Psi_n\Delta|_\infty\right\}\\
\kappa_{\infty,J_0}(J,s) &\triangleq \min_{k\in J_0}\kappa_k^*(J,s)\\
\kappa_{p,J_0}(J,s) &\triangleq \max\left(\frac{\kappa_{\infty,J_0}(J,s)}{|J_0|^{1/p}}, \frac{\kappa_\infty(J,s)}{c_J^{1/p}}\right)\\
\kappa_{1,J_{\rm ex}}(J,s) &\triangleq \max\left(\min_{j\in J}\left\{\min_{\substack{|\Delta_J|_\infty\le\Delta_j,\ |\Delta_{J_{\rm ex}}|_1=1,\\ (1-cr)+(1-c)|\Delta_{J_{\rm ex}^c}|_1\le2s\Delta_j}}|\Psi_n\Delta|_\infty\right\}, \frac{\kappa_{\infty,J\cup J_{\rm ex}^c}(J,s)}{c_{J_{\rm ex},J}}, \frac{\kappa_{\infty,J_{\rm ex}}(J,s)}{|J_{\rm ex}|}\right)\\
\underline{\kappa}_{1,J_{\rm ex}^c}(J,s) &\triangleq \max\left(\min_{j\in J}\left\{\min_{\substack{|\Delta_J|_\infty\le\Delta_j,\ |\Delta_{J_{\rm ex}^c}|_1=1,\\ (1-cr)|\Delta_{J_{\rm ex}}|_1+1-c\le2s\Delta_j}}|\Psi_n\Delta|_\infty\right\}, \frac{\kappa_{\infty,J}(J,s)}{c_{J_{\rm ex}^c,J}}, \frac{\kappa_{\infty,J_{\rm ex}^c}(J,s)}{|J_{\rm ex}^c|}\right)\\
\theta(J,s) &\triangleq \left(1-\frac{r^2}{\kappa_{1,J_{\rm ex}}(J,s)}-\frac{r}{\underline{\kappa}_{1,J_{\rm ex}^c}(J,s)}\right)_+^{-1}\\
\kappa_\infty(s) &\triangleq \min_{k=1,\ldots,K}\left\{\min_{\substack{\Delta_k=1,\ |\Delta|_\infty\le1,\\ (1-cr)|\Delta_{J_{\rm ex}}|_1+(1-c)|\Delta_{J_{\rm ex}^c}|_1\le2s}}|\Psi_n\Delta|_\infty\right\}\\
\kappa_k^*(s) &\triangleq \min_{j=1,\ldots,K}\left\{\min_{\substack{\Delta_k=\pm1,\\ (1-cr)|\Delta_{J_{\rm ex}}|_1+(1-c)|\Delta_{J_{\rm ex}^c}|_1\le2s\Delta_j}}|\Psi_n\Delta|_\infty\right\}\\
\kappa_{\infty,J_0}(s) &\triangleq \min_{k\in J_0}\kappa_k^*(s)\\
\kappa_{p,J_0}(s) &\triangleq \max\left(\frac{\kappa_{\infty,J_0}(s)}{|J_0|^{1/p}}, \frac{\kappa_\infty(s)}{c_J^{1/p}}\right)\\
\kappa_{1,J_{\rm ex}}(s) &\triangleq \max\left(\min_{j=1,\ldots,K}\left\{\min_{\substack{|\Delta|_\infty\le\Delta_j,\ |\Delta_{J_{\rm ex}}|_1=1,\\ (1-cr)+(1-c)|\Delta_{J_{\rm ex}^c}|_1\le2s\Delta_j}}|\Psi_n\Delta|_\infty\right\}, \frac{\kappa_\infty(s)}{c_{J_{\rm ex}}(s)}, \frac{\kappa_{\infty,J_{\rm ex}}(s)}{|J_{\rm ex}|}\right)\\
\underline{\kappa}_{1,J_{\rm ex}^c}(s) &\triangleq \max\left(\min_{j=1,\ldots,K}\left\{\min_{\substack{|\Delta|_\infty\le\Delta_j,\ |\Delta_{J_{\rm ex}^c}|_1=1,\\ (1-cr)|\Delta_{J_{\rm ex}}|_1+1-c\le2s\Delta_j}}|\Psi_n\Delta|_\infty\right\}, \frac{\kappa_\infty(s)}{c_{J_{\rm ex}^c}(s)}, \frac{\kappa_{\infty,J_{\rm ex}^c}(s)}{|J_{\rm ex}^c|}\right)\\
\theta(s) &\triangleq \min\left(\left(1-\frac{r^2}{\kappa_{1,J_{\rm ex}}(s)}-\frac{r}{\underline{\kappa}_{1,J_{\rm ex}^c}(s)}\right)_+^{-1}, \left(1-\frac{r\ c_b(s)}{\kappa_\infty(s)}\right)_+^{-1}\right)\\
\overline{\kappa}(s) &\triangleq \min_{\substack{k=1,\ldots,K\\ l=1,\ldots,\overline{L}}}\left\{\min_{\substack{|\Delta|_\infty\le\Delta_k,\ (\overline{\Psi}_n\Delta)_l=\pm1,\ |\overline{\Psi}_n\Delta|_\infty\le1,\\ (1-cr)|\Delta_{J_{\rm ex}}|_1+(1-c)|\Delta_{J_{\rm ex}^c}|_1\le2s\Delta_k}}|\Psi_n\Delta|_\infty\right\}
\end{aligned}
$$

References

[1] Amemiya, T. (1974): "The Non-Linear Two-Stage Least Squares Estimator". *Journal of Econometrics,* 2, 105–110.

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

[3] Andrews, D. W. K., and B. Lu (2001): "Consistent Model and Moment Selection Procedures for GMM Estimation With Application to Dynamic Panel Data Models". *Journal of Econometrics,* 101, 123–164.

[4] Andrews, D. W. K., and J. H. 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.

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

[6] Bai J., and S. Ng (2009): "Selecting Instrumental Variables in a Data Rich Environment". *Journal of Time Series Econometrics,* 1, 105–110.

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

[8] 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.

[9] Belloni, A., and V. Chernozhukov (2013): "Least Squares After Model Selection in High-dimensional Sparse Models". *Bernoulli,* 19, 521–547.

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

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

[12] 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, Berlin.

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

[14] Bertail, P. , E. Gauthérat, and H. Harari-Kermadec (2009): "Exponential Inequalities for Self Normalized Sums". *Electronic Communications in Probability,* 13, 628–640.

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

[16] Blundell, R., X. Chen, and D. Kristensen (2007): "Semi-nonparametric IV Estimation of Shape-invariant Engel Curves". *Econometrica,* 75, 1613–1669.

[17] Bühlmann, P., and S. A. van de Geer (2011): *Statistics for High-Dimensional Data*. Springer, New-York.

[18] Caner, M. (2009): "LASSO Type GMM Estimator". *Econometric Theory,* 25, 1–23.

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

[21] Chernozhukov, V., D. Chetverikov, and K. Kato (2013): "Gaussian Approximations and Multiplier Bootstrap for Maxima of Sums of High-Dimensional Random Vectors". Preprint arXiv:1212.6906v5.

[22] Dalalyan, A., and A. B. Tsybakov (2008): "Aggregation by Exponential Weighting, Sharp PAC-Bayesian Bounds and Sparsity". *Journal of Machine Learning Research,* 72, 39–61.

[23] 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.

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

[25] Efron, B. (1969): "Student's t-test Under Symmetry Conditions". *Journal of American Statistical Society,* 64, 1278–1302.

[26] Gautier, E., and A. Tsybakov (2011): "High-dimensional Instrumental Variables Regression and Confidence Sets". arXiv:1105.2454 preprint.

[27] 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.

[28] Grant, M., and S. Boyd (2013): "CVX: Matlab Software for Disciplined Convex Programming, version 2.0 beta.". http://cvxr.com/cvx.

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

[30] 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.

[31] Kolesár, M., R. Chetty, J. Friedman, E. Glaeser, and G. W. Imbens (2011): "Identification and Inference with Many Invalid Instruments". Preprint.

[32] Koltchinskii, V. (2011): *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems.* Forthcoming in *Lecture Notes in Mathematics*, Springer, Berlin.

[33] Liao, Z. (2013): "Adaptive GMM Shrinkage Estimation with Consistent Moment Selection". *Econometric Theory,* 29, 1–48.

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

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

[36] Nelson, C. R., and Startz, R. (1990b): "The Distribution of the Instrumental Variable Estimator and Its t Ratio When the Instrument Is a Poor One". *Journal of Business,* 63, S125–S140.

[37] Newey, W. K. (1990): "Efficient Instrumental Variables Estimation of Nonlinear Models". *Econometrica,* 58, 809–837.

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

[39] Pinelis, I. (1994): "Probabilistic Problems and Hotelling's t^2 Test Under a Symmetry Condition". *Annals of Statistics,* 22, 357–368.

[40] Polyak, B.T. (1987): *Introduction to Optimization.* Optimization Software.

[41] Rigollet, P., and A. B. Tsybakov (2011): "Exponential Screening and Optimal Rates of Sparse Estimation". *Annals of Statistics,* 35, 731–771.

[42] Romano, J. P., and A. Shaikh (2008): "Inference for Identifiable Parameters in Partially Identified Econometric Models". *Journal of Statistical Planning and Inference,* 138, 2786–2807.

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

[44] Rosenbaum, M., and A. B. Tsybakov (2010): "Sparse Recovery Under Matrix Uncertainty". *The Annals of Statistics,* 38, 2620–2651.

[45] Sargan, J. D. (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] Stock, J. H., J.H. Wright, and M. Yogo (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments". *Journal of Business & Economic Statistics,* 20, 518–529.

[48] 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.

CREST, ENSAE ParisTech, 3 avenue Pierre Larousse, 92 245 Malakoff Cedex, France.

E-mail address: eric.gautier@ensae-paristech.fr, alexandre.tsybakov@ensae-paristech.fr

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

ERIC GAUTIER AND ALEXANDRE TSYBAKOV

A.1. **Lower Bounds on $\kappa_{p,J}$ for Square Matrices Ψ_n.** The following propositions establish lower bounds on $\kappa_{p,J}$ when there are no endogenous regressors and Ψ_n is a square $K \times K$ matrix. Recall that in that case the cone C_J takes the the simple form (4.2). For any $J \subseteq \{1, \dots, K\}$ we define the following restricted eigenvalue (RE) constants

$$\kappa_{\mathrm{RE},J} \triangleq \inf_{\Delta \in \mathbb{R}^K \setminus \{0\}:\ \Delta \in C_J} \frac{|\Delta^T \Psi_n \Delta|}{|\Delta_J|_2^2}, \qquad \kappa'_{\mathrm{RE},J} \triangleq \inf_{\Delta \in \mathbb{R}^K \setminus \{0\}:\ \Delta \in C_J} \frac{|J|\,|\Delta^T \Psi_n \Delta|}{|\Delta_J|_1^2}.$$

Proposition A.1. *For any $J \subseteq \{1, \dots, K\}$ we have*

$$\kappa_{1,J} \geq \frac{1-cr}{2} \kappa_{1,J,J} \geq \frac{(1-cr)^2}{4|J|} \kappa'_{\mathrm{RE},J} \geq \frac{(1-cr)^2}{4|J|} \kappa_{\mathrm{RE},J}.$$

Proof. For Δ such that $|\Delta_{J^c}|_1 \leq \frac{1+cr}{1-cr} |\Delta_J|_1$ we have $|\Delta|_1 \leq \frac{2}{1-cr} |\Delta_J|_1$. Thus, one obtains

$$\frac{|\Delta^T \Psi_n \Delta|}{|\Delta_J|_1^2} \leq \frac{|\Delta|_1 |\Psi_n \Delta|_\infty}{|\Delta_J|_1^2} \leq \frac{2}{1-cr} \frac{|\Psi_n \Delta|_\infty}{|\Delta_J|_1} \leq \frac{4}{(1-cr)^2} \frac{|\Psi_n \Delta|_\infty}{|\Delta|_1}.$$

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 $|\Delta|_1^2 \leq |J| |\Delta_J|_2^2$. □

We now obtain bounds for sensitivities $\kappa_{p,J}$ with $1 < p \leq 2$. For any $s \leq K$, we consider a uniform version of the restricted eigenvalue constant: $\kappa_{\mathrm{RE}}(s) \triangleq \min_{|J| \leq s} \kappa_{\mathrm{RE},J}$.

Proposition A.2. *For any $s \leq K/2$ and $1 < p \leq 2$, we have*

$$\kappa_{p,J} \geq C(p) s^{-1/p} \kappa_{\mathrm{RE}}(2s), \quad \forall\ J:\ |J| \leq s,$$

where $C(p) = 2^{-1/p-1/2}(1-cr) \left(1 + \frac{1+cr}{1-cr} (p-1)^{-1/p}\right)^{-1}$.

Proof. For $\Delta \in R^K$ and a set $J \subseteq \{1, \dots, K\}$, let $J_1 = J_1(\Delta, J)$ be the subset of indices in $\{1, \dots, K\}$ corresponding to the s largest in absolute value components of Δ outside of J. Define $J_+ = J \cup J_1$. If $|J| \leq s$ we have $|J_+| \leq 2s$. It is easy to see that the kth largest absolute value of elements of Δ_{J^c} satisfies $|\Delta_{J^c}|_{(k)} \leq |\Delta_{J^c}|_1 / k$. Thus,

$$|\Delta_{J_+^c}|_p^p = \sum_{j \in J_+^c} |\Delta_j|^p = \sum_{k \geq s+1} |\Delta_{J^c}|_{(k)}^p \leq |\Delta_{J^c}|_1^p \sum_{k \geq s+1} \frac{1}{k^p} \leq \frac{|\Delta_{J^c}|_1^p}{(p-1) s^{p-1}}.$$

For $\Delta \in C_J$, this implies

$$|\Delta_{J_+^c}|_p \le \frac{|\Delta_{J^c}|_1}{(p-1)^{1/p} s^{1-1/p}} \le \frac{c_0 |\Delta_J|_1}{(p-1)^{1/p} s^{1-1/p}} \le \frac{c_0 |\Delta_J|_p}{(p-1)^{1/p}}\,,$$

where $c_0 = \frac{1+cr}{1-cr}$. Therefore, using that $|\Delta_J|_p \le |\Delta_{J_+}|_p$ we get, for $\Delta \in C_J$,

$$|\Delta|_p \le |\Delta_{J_+}|_p + |\Delta_{J_+^c}|_p \le (1 + c_0(p-1)^{-1/p})|\Delta_{J_+}|_p \le (1 + c_0(p-1)^{-1/p})(2s)^{1/p-1/2}|\Delta_{J_+}|_2 \tag{A.1}$$

where the last inequality follows from the bound

$$|\Delta_{J_+}|_p \le |J_+|^{1/p-1/2}|\Delta_{J_+}|_2 \le (2s)^{1/p-1/2}|\Delta_{J_+}|_2.$$

Using (A.1) and the fact that $|\Delta|_1 \le \frac{2}{1-cr}|\Delta_J|_1 \le \frac{2\sqrt{|J|}}{1-cr}|\Delta_J|_2 \le \frac{2\sqrt{s}}{1-cr}|\Delta_J|_2 \le \frac{2\sqrt{s}}{1-cr}|\Delta_{J_+}|_2$ for $\Delta \in C_J$, we get

$$\begin{aligned}
\frac{|\Delta^T \Psi_n \Delta|}{|\Delta_{J_+}|_2^2} &\le \frac{|\Delta|_1 |\Psi_n \Delta|_\infty}{|\Delta_{J_+}|_2^2} \\
&\le \frac{2\sqrt{s}|\Psi_n \Delta|_\infty}{(1-cr)|\Delta_{J_+}|_2} \\
&\le \frac{s^{1/p}|\Psi_n \Delta|_\infty}{C(p)|\Delta|_p}\,.
\end{aligned}$$

Since $|J_+| \le 2s$, this proves the proposition. □

A.2. **Exact Computations of the Sensitivities when $|J|$ is small.** The coordinate-wise sensitivities $\kappa^*_{k,J}$ is obtained by minimizing on the set

$$\left\{\Delta \in \mathbb{R}^K :\ \Delta_k = 1,\ (1-cr)|\Delta_{J^c \cap J_{\rm ex}}|_1 + (1-c)|\Delta_{J^c \cap J_{\rm ex}^c}|_1 \le (1+cr)|\Delta_{J \cap J_{\rm ex}}|_1 + (1+c)|\Delta_{J \cap J_{\rm ex}^c}|_1\right\}$$

and can be computed as follows.

Algorithm A.1. *When $0 < c < 1$ solve*

$$\min_{(\epsilon_j)_{j\in J} \in \{-1,1\}^{|J|}} \min_{(\Delta, v) \in \mathcal{U}_{k,J}} v$$

where $\mathcal{U}_{k,J}$ is the set of (Δ, v) with $\Delta \in \mathbb{R}^K$, $v \in \mathbb{R}$ satisfying:

$$\begin{aligned}
&v \ge 0, \qquad -v\mathbf{1} \le \Psi_n \Delta \le v\mathbf{1}, \qquad \Delta_k = 1, \\
&(1-cr)|\Delta_{J^c \cap J_{\rm ex}}|_1 + (1-c)|\Delta_{J^c \cap J_{\rm ex}^c}|_1 \le (1+cr) \sum_{j \in J \cap J_{\rm ex}} \epsilon_j \Delta_j + (1+c) \sum_{j \in J \cap J_{\rm ex}^c} \epsilon_j \Delta_j\ .
\end{aligned}$$

When $1 \le c < r^{-1}$ solve

$$\min_{(\epsilon_j)_{j \in J \cup (J^c \cap J_{\rm ex}^c)} \in \{-1,1\}^{|J \cup (J^c \cap J_{\rm ex}^c)|}} \min_{(\Delta, v) \in \mathcal{U}_J} v$$

where $\mathcal{U}_J$ *is the set of* (Δ, v) *with* $\Delta \in \mathbb{R}^K$, $v \in \mathbb{R}$ *satisfying:*

$$v \geq 0, \qquad -v\mathbf{1} \leq \Psi_n \Delta \leq v\mathbf{1}, \qquad \Delta_k = 1,$$
$$(1-cr)|\Delta_{J^c \cap J_{\text{ex}}}|_1 \leq (1+cr) \sum_{j \in J \cap J_{\text{ex}}} \epsilon_j \Delta_j + (1+c) \sum_{j \in J \cap J_{\text{ex}}^c} \epsilon_j \Delta_j + (c-1) \sum_{j \in J^c \cap J_{\text{ex}}^c} \epsilon_j \Delta_j \ .$$

When the endogenous regressors are in J, $J^c \cap J_{\text{ex}}^c = \emptyset$ and there are no cases to distinguish. One can calculate in a similar manner $\kappa_{1, J_{\text{ex}}^c, J}$.

Algorithm A.2. *When* $0 < c < r^{-1}$, *solve*

$$\min_{(\epsilon_j)_{J \cup (J^c \cap J_{\text{ex}}^c)} \in \{-1,1\}^{|J \cup (J^c \cap J_{\text{ex}}^c)|}} \ \min_{(\Delta, v) \in \mathcal{U}_{J_{\text{ex}}^c, J}} v$$

where $\mathcal{U}_{J_{\text{ex}}^c, J}$ *is the set of* (Δ, v) *with* $\Delta \in \mathbb{R}^K$, $v \in \mathbb{R}$ *satisfying:*

$$v \geq 0, \qquad -v\mathbf{1} \leq \Psi_n \Delta \leq v\mathbf{1}, \qquad \sum_{j \in J_{\text{ex}}^c} \epsilon_j \Delta_j = 1,$$
$$(1-cr)|\Delta_{J^c \cap J_{\text{ex}}}|_1 \leq (1+cr) \sum_{j \in J \cap J_{\text{ex}}} \epsilon_j \Delta_j + \sum_{j \in J \cap J_{\text{ex}}^c} \epsilon_j \Delta_j - \sum_{j \in J^c \cap J_{\text{ex}}^c} \epsilon_j \Delta_j + c \ .$$

A.3. **Error Bounds When** $J(\beta) \subsetneq \widehat{J}$**.** Belloni and Chernozhukov (2011a,2013) give bounds on the error made by a post-model selection procedure when we do not have $J(\beta) \subseteq \widehat{J}$, where $\widehat{J}$ is obtained by a selection procedure. Belloni and Chernozhukov (2011a) considers the high-dimensional quantile regression model and the case of the ℓ_2 loss. Belloni and Chernozhukov (2013) considers the high-dimensional linear model and prediction loss. Theorem 6.2 gives the following bound for the *STIV* estimator, for arbitrary estimated support $\widehat{J}$: For every β in $\mathcal{I}dent$, on the event $\mathcal{G}$, for any solution $(\widehat{\beta}, \widehat{\sigma})$ of the minimization problem (3.6) we have, for every $J_0 \subseteq \{1, \ldots, K\}$, $0 < c < r^{-1}$, $p \geq 1$,

$$\left| \left(\mathbf{D_X}^{-1} \left(\widehat{\beta} - \beta \right) \right)_{J_0} \right|_p \leq \max \left(\frac{2\widehat{\sigma} r \theta(\widehat{J}, |\widehat{J}|)}{\widetilde{\kappa}_{p, J_0}(\widehat{J}, |\widehat{J}|)}, 6 |(\mathbf{D}_{\mathbf{X}}^{-1} \beta)_{\widehat{J}^c}|_1 \right) . \tag{A.2}$$

If one takes as $\widehat{J} = J(\widehat{\beta})$, due to Theorem 7.1 (iii), under the assumptions of (ii), the non-zero coordinates that could be missed in $J(\widehat{\beta})$ are smaller in absolute value than $\frac{2\sigma_* r \tau_*}{\kappa_* v_k}$ on $\mathcal{G} \cap \mathcal{G}_1 \cap \mathcal{G}_2$. This yields, in the special case of the ℓ_1 norm and for fixed $I \subseteq \{1, \ldots, L\}$ and $0 < c < r^{-1}$:
For every β in $\mathcal{I}dent$, on the event $\mathcal{G} \cap \mathcal{G}_1 \cap \mathcal{G}_2$, for any solution $(\widehat{\beta}, \widehat{\sigma})$ of the minimization problem (3.6) and for $\widehat{J} = J(\widehat{\beta})$, we have

$$\left| \mathbf{D_X}^{-1} \left(\widehat{\beta} - \beta \right) \right|_1 \leq \max \left(\frac{2\widehat{\sigma} r \theta(\widehat{J}, |\widehat{J}|)}{\widetilde{\kappa}_{p, J_0}(\widehat{J}, |\widehat{J}|)}, \ \frac{12 \sigma_* r}{\kappa_*} \tau_* \left| J(\beta) \setminus \widehat{J} \right| \right) . \tag{A.3}$$

A similar inequality can be obtained using $\widehat{J} = J(\widetilde{\beta})$ and makes the assumptions of Theorem 8.1.

These error bounds are are not confidence sets due to the presence of the term $\frac{12\sigma_* r}{\kappa_*}\tau_* \left|J(\beta) \setminus \widehat{J}\right|$ which depends on the unknown. Unlike equations (A.2) and (A.3), we obtain valid confidence sets in Section 8.3.1 and Section 8.3.2 because, there, we assume sufficient beta-min assumptions.

A.4. **Moderate Deviations for Self-normalized Sums.** Throughout this section $x_1, \dots, x_n$ are independent random variables such that, for every i, $\mathbb{E}[X] = 0$. The following result is due to Efron (1969).

Theorem A.1. *If x_i for $i = 1, \dots, n$ are symmetric, then for every r positive,*

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

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

Theorem A.2. *If x_i for $i = 1, \dots, n$ are symmetric and identically distributed, then for every r in $[0, 1)$,*

$$\mathbb{P}\left(\frac{\left|\frac{1}{n}\sum_{i=1}^n x_i\right|}{\sqrt{\frac{1}{n}\sum_{i=1}^n x_i^2}} \geq r\right) \leq \frac{4e^3}{9}\Phi(-\sqrt{n}r).$$

The following result is from 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*

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

Then

$$\forall 0 \leq r \leq \frac{d_{n,\delta}}{\sqrt{n}},\ \mathbb{P}\left(\frac{\left|\frac{1}{n}\sum_{i=1}^n x_i\right|}{\sqrt{\frac{1}{n}\sum_{i=1}^n x_i^2}} \geq r\right) \leq 2\Phi(-\sqrt{n}r)\left(1 + A_0\left(\frac{1+\sqrt{n}r}{d_{n,\delta}}\right)^{2+\delta}\right)$$

where $A_0 > 0$ is an absolute constant.

Despite of its great interest to understand the large deviations behavior of self-normalized sums, the bound has limited practical use because A_0 is not an explicit constant.
The following result is a corollary of Theorem 1 in Bertail, Gauthérat and Harari-Kermadec (2009).

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

$$\forall r \geq 0,\ \mathbb{P}\left(\frac{\left|\frac{1}{n}\sum_{i=1}^n x_i\right|}{\sqrt{\frac{1}{n}\sum_{i=1}^n x_i^2}} \geq r\right) \leq (2e+1)\exp\left(-\frac{nr^2}{2+\gamma_4 r^2}\right) \tag{A.4}$$

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

$$\forall r \geq \sqrt{n},\ \mathbb{P}\left(\frac{\left|\frac{1}{n}\sum_{i=1}^n x_i\right|}{\sqrt{\frac{1}{n}\sum_{i=1}^n x_i^2}} \geq r\right) = 0.$$

Proof. Bertail, Gauthérat and Harari-Kermadec (2009) 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.4) by choosing a to equate the two exponential terms. □

A.5. **Some Facts From Convex Analysis.** We will use the following properties of convex functions that can be found, for example, in Polyak (1987), Section 5.1.4. Let f be a convex function on $\mathbb{R}^K$. Denote by ∂f its subdifferential, *i.e.*, the set of all $a \in \mathbb{R}^K$ such that $f(x+y) - f(x) \geq \langle a, y\rangle,\ \forall\, x, y \in \mathbb{R}^K$, where $\langle\cdot,\cdot\rangle$ is the standard inner product in $\mathbb{R}^K$.

Lemma A.1. *Let* $f(x) = \max_{l=1,\ldots,m} f_l(x)$ *where the finctions* f_l *are convex. Then* f *is convex and its subdifferential is contained in the convex hull of the union of the subdifferentials* ∂f_l*:*

$$\partial f \subseteq \mathrm{Conv}\left(\bigcup_{l=1}^m \partial f_l\right). \tag{A.5}$$

A.6. **Proofs. Proof of Proposition 4.1.** Parts (i) and (ii) are straightforward. We now prove (4.9) with the constant

$$c_J = \min\left(\frac{2|J|}{(1-c)_+}, \frac{2\,|J_{\mathrm{ex}} \cap J| + (2 + c(1-r))\,|J_{\mathrm{ex}}^c \cap J| + c(1-r)\,|J_{\mathrm{ex}}^c \cap J^c|}{(1-cr)_+}\right).$$

The upper bound in (4.9) follows from the fact that $|\Delta|_p \geq |\Delta|_\infty$. We obtain the lower bound as follows. Because $|\Delta|_p \leq |\Delta|_1^{1/p}|\Delta|_\infty^{1-1/p}$, we get that, for $\Delta \neq 0$,

$$\frac{|\Psi_n\Delta|_\infty}{|\Delta|_p} \geq \frac{|\Psi_n\Delta|_\infty}{|\Delta|_\infty}\left(\frac{|\Delta|_\infty}{|\Delta|_1}\right)^{1/p}. \tag{A.6}$$

Furthermore, for $\Delta \in C_J$, by the definition of the cone, we have

$$|\Delta_{J^c}|_1 \leq |\Delta_J|_1 + cr|\Delta_{J_{\mathrm{ex}}\cap J}|_1 + cr|\Delta_{J_{\mathrm{ex}}\cap J^c}|_1 + c|\Delta_{J_{\mathrm{ex}}^c\cap J}|_1 + c|\Delta_{J_{\mathrm{ex}}^c\cap J^c}|_1 \tag{A.7}$$

which implies that

$$(1-cr)|\Delta_{J_{\rm ex}\cap J^c}|_1 \le (1+cr)|\Delta_{J_{\rm ex}\cap J}|_1 + (1+c)|\Delta_{J_{\rm ex}^c\cap J}|_1 + (c-1)|\Delta_{J_{\rm ex}^c\cap J^c}|_1$$

and thus

$$|\Delta_{J_{\rm ex}\cap J^c}|_1 \le \frac{1+cr}{1-cr}|\Delta_{J_{\rm ex}\cap J}|_1 + \frac{1+c}{1-cr}|\Delta_{J_{\rm ex}^c\cap J}|_1 + \frac{c-1}{1-cr}|\Delta_{J_{\rm ex}^c\cap J^c}|_1 \ . \tag{A.8}$$

Adding $|\Delta_J|_1$ on both sides of (A.7) and injecting (A.8) into (A.7) yields that

$$\begin{aligned}
|\Delta|_1 &\le 2|\Delta_J|_1 + cr|\Delta_{J_{\rm ex}\cap J}|_1 + cr|\Delta_{J_{\rm ex}\cap J^c}|_1 + c|\Delta_{J_{\rm ex}^c\cap J}|_1 + c|\Delta_{J_{\rm ex}^c\cap J^c}|_1 \\
&\le \left(2+cr+cr\frac{1+cr}{1-cr}\right)|\Delta_{J_{\rm ex}\cap J}|_1 + \left(2+c+cr\frac{1+c}{1-cr}\right)|\Delta_{J_{\rm ex}^c\cap J}|_1 + \left(c+cr\frac{c-1}{1-cr}\right)|\Delta_{J_{\rm ex}^c\cap J^c}|_1 \\
&\le \frac{2}{1-cr}\,|\Delta_{J_{\rm ex}\cap J}|_1 + \frac{2+c(1-r)}{1-cr}|\Delta_{J_{\rm ex}^c\cap J}|_1 + \frac{c(1-r)}{1-cr}|\Delta_{J_{\rm ex}^c\cap J^c}|_1 \\
&\le \left(\frac{2}{1-cr}\,|J_{\rm ex}\cap J| + \frac{2+c(1-r)}{1-cr}|J_{\rm ex}^c\cap J| + \frac{c(1-r)}{1-cr}|J_{\rm ex}^c\cap J^c|\right)|\Delta_{(J_{\rm ex}^c\cap J^c)\cup J}|_\infty
\end{aligned}$$

We obtain the first lower bound using the fact that $|\Delta_{(J_{\rm ex}^c\cap J^c)\cup J}|_\infty \le |\Delta|_\infty$ and (A.6). This lower bound holds for any $0 < c < r^{-1}$. Let us obtain an alternative lower bound for the case where $0 < c < 1$. We can deduce from (A.7) that

$$|\Delta_{J^c}|_1 \le \frac{1+c}{1-c}|\Delta_J|_1$$

and thus

$$|\Delta|_1 = |\Delta_{J^c}|_1 + |\Delta_J|_1 \le \frac{2}{1-c}|\Delta_J|_1 \le \frac{2|J|}{1-c}|\Delta_J|_\infty \le \frac{2|J|}{1-c}|\Delta|_\infty \ .$$

Inequality (4.10) can be proved in a similar manner. The lower bounds follows from the fact that

$$\frac{|\Psi_n\Delta|_\infty}{|\Delta_{J_0}|_p} \ge \frac{|\Psi_n\Delta|_\infty}{|\Delta_{J_0}|_\infty}\left(\frac{|\Delta_{J_0}|_\infty}{|\Delta_{J_0}|_1}\right)^{1/p}$$

and $|\Delta_{J_0}|_1 \le |J_0||\Delta_{J_0}|_\infty$. While the upper bound holds because $|\Delta_{J_0}|_p \ge |\Delta_{J_0}|_\infty$.

Let us now prove (iv). Because for every k in J_0, $|\Delta_{J_0}|_\infty \ge |\Delta_k|$, one obtains that for every k in J_0,

$$\kappa_{\infty,J_0,J} = \inf_{\Delta\in C_J}\frac{|\Psi_n\Delta|_\infty}{|\Delta_{J_0}|_\infty} \le \inf_{\Delta\in C_J}\frac{|\Psi_n\Delta|_\infty}{|\Delta_k|} = \kappa^*_{k,J}.$$

Thus

$$\kappa_{\infty,J_0,J} \le \min_{k\in J_0}\kappa^*_{k,J}.$$

But one also has

$$\kappa_{\infty,J_0,J} = \min_{k\in J_0}\inf_{\Delta\in C_J:\ |\Delta_k|=|\Delta_{J_0}|_\infty=1}|\Psi_n\Delta|_\infty \ge \min_{k\in J_0}\inf_{\Delta\in C_J:\ |\Delta_k|=1}|\Psi_n\Delta|_\infty. \tag{A.9}$$

The bounds (v) and (vi) with constants with

$$c_{J_{\rm ex},J} = \min\left(\frac{2|J|}{(1-c)_+}, \frac{2\,|J_{\rm ex}\cap J| + (1+c)\,|J_{\rm ex}^c\cap J| + (c-1)\,|J_{\rm ex}^c\cap J^c|}{(1-cr)_+}\right)\ ,$$

and

$$c_{J_{\rm ex}^c,J} = \frac{2\,|J_{\rm ex}^c\cap J| + (1+cr)\,|J_{\rm ex}\cap J|}{(1-c)_+}$$

are obtained by rewritting the cone condition as

$$(1-cr)|\Delta_{J_{\rm ex}\cap J^c}|_1 + (1-c)|\Delta_{J_{\rm ex}^c\cap J^c}|_1 \le (1+cr)|\Delta_{J_{\rm ex}\cap J}|_1 + (1+c)|\Delta_{J_{\rm ex}^c\cap J}|_1$$

and using $|\Delta_{J_{\rm ex}}|_1 = |\Delta_{J_{\rm ex}\cap J^c}|_1 + |\Delta_{J_{\rm ex}\cap J}|_1$ and $|\Delta_{J_{\rm ex}^c}|_1 = |\Delta_{J_{\rm ex}^c\cap J^c}|_1 + |\Delta_{J_{\rm ex}^c\cap J}|_1$.
To prove (vii) it suffices to note that, because $|\Delta|_1 \le 2|\Delta_J|_1 + cr\,|\Delta_{J_{\rm ex}}|_1 + c\left|\Delta_{J_{\rm ex}^c}\right|_1$,

$$|\Delta|_1 \le \left(\frac{2}{\kappa_{1,J,J}} + \frac{cr}{\kappa_{1,J_{\rm ex},J}} + \frac{c}{\kappa_{1,J_{\rm ex}^c,J}}\right)|\Psi_n\Delta|_\infty\ .$$

This implies that

$$\kappa_{1,J} \ge \left(\frac{2}{\kappa_{1,J,J}} + \frac{cr}{\kappa_{1,J_{\rm ex},J}} + \frac{c}{\kappa_{1,J_{\rm ex}^c,J}}\right)^{-1.}$$

and we conclude using (4.10), (v) and (vi). □

Proof of Proposition 4.2. Take $1 \le k \le K$ and $1 \le l \le L$,

$$|(\Psi_n\Delta)_l - (\Psi_n)_{lk}\Delta_k| \le |\Delta|_1 \max_{k'\ne k}|(\Psi_n)_{lk'}|,$$

which yields

$$|(\Psi_n)_{lk}|\,|\Delta_k| \le |\Delta|_1 \max_{k'\ne k}|(\Psi_n)_{lk'}| + |(\Psi_n\Delta)_l|\,.$$

The two inequalities of the assumption yield

$$\left|(\Psi_n)_{l(k)k}\right|\,|\Delta_k| \le |\Delta|_1\frac{1-\eta_2}{c_J}|(\Psi_n)_{l(k)k}| + \frac{1}{\eta_1}\left|(\Psi_n\Delta)_{l(k)}\right|\,\left|(\Psi_n)_{l(k)k}\right|\,.$$

This inequality, together with the fact that $\left|(\Psi_n\Delta)_{l(k)}\right| \le |\Psi_n\Delta|_\infty$ and the upper bounds from the proof of the upper bound (4.9) of Proposition 4.1, yield

$$|\Delta_k| \le (1-\eta_2)|\Delta|_\infty + \frac{|\Psi_n\Delta|_\infty}{\eta_1}$$

and thus

$$\eta_2\eta_1|\Delta|_\infty \le |\Psi_n\Delta|_\infty\ .$$

One concludes using the definition of the ℓ_∞-sensitivity. □

Proof of Theorem 6.1. Consider here a fixed $\beta \in \mathcal{I}dent$ and denote by $u_i = y_i - x_i^T\beta$. Because $\left|\frac{1}{n}\mathbf{D_Z}\mathbf{Z}^T(\mathbf{Y} - \mathbf{X}\beta)\right|_\infty = \left|\frac{1}{n}\mathbf{D_Z}\mathbf{Z}^T\mathbf{U}\right|_\infty$ and $\widehat{Q}(\beta) = \mathbb{E}_n[U^2]$, on the event $\mathcal{G}$, $\left(\beta, \sqrt{\widehat{Q}(\beta)}\right)$ belongs to $\widehat{\mathcal{I}}$.

Set $\Delta \triangleq \mathbf{D}_\mathbf{X}^{-1}(\widehat{\beta} - \beta)$. On the event $\mathcal{G}$, we have:

$$|\Psi_n\Delta|_\infty \le \left|\frac{1}{n}\mathbf{D_Z}\mathbf{Z}^T(\mathbf{Y} - \mathbf{X}\widehat{\beta})\right|_\infty + \left|\frac{1}{n}\mathbf{D_Z}\mathbf{Z}^T(\mathbf{Y} - \mathbf{X}\beta)\right|_\infty \tag{A.10}$$

$$\le r\left(\widehat{\sigma} + \sqrt{\widehat{Q}(\beta)}\right). \tag{A.11}$$

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

$$\left|\mathbf{D}_\mathbf{X}^{-1}\widehat{\beta}\right|_1 + c\widehat{\sigma} \le |\mathbf{D}_\mathbf{X}^{-1}\beta|_1 + c\sqrt{\widehat{Q}(\beta)}. \tag{A.12}$$

This implies, again on the event $\mathcal{G}$,

$$\begin{aligned}\left|\Delta_{J(\beta)^c}\right|_1 &= \sum_{k\in J(\beta)^c}\left|\mathbb{E}_n[X_k^2]^{1/2}\widehat{\beta}_k\right| \\ &\le \sum_{k\in J(\beta)}\left(\left|\mathbb{E}_n[X_k^2]^{1/2}\beta_k\right| - \left|\mathbb{E}_n[X_k^2]^{1/2}\widehat{\beta}_k\right|\right) + c\left(\sqrt{\widehat{Q}(\beta)} - \sqrt{\widehat{Q}(\widehat{\beta})}.\right).\end{aligned} \tag{A.13}$$

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

For γ such that $\widehat{Q}(\gamma) \neq 0$, $\gamma \to \sqrt{\widehat{Q}(\gamma)}$ is differentiable and its gradient $\nabla\sqrt{\widehat{Q}}(\gamma)$ is a vector with components

$$\left(\nabla\sqrt{\widehat{Q}}(\gamma)\right)_k = -\frac{\frac{1}{n}\sum_{i=1}^n x_{ki}(y_i - x_i^T\gamma)}{\sqrt{\frac{1}{n}\sum_{i=1}^n(y_i - x_i^T\gamma)^2}}, \quad k = 1, \dots, K.$$

In each scenario we only allowed the denominator to be 0 at β with 0 probability. Therefore the subgradient of $\sqrt{\widehat{Q}}$ at β is a gradient and we have

$$\nabla\sqrt{\widehat{Q}}(\beta) = \frac{\frac{1}{n}\mathbf{X}^T\mathbf{U}}{\sqrt{\widehat{Q}(\beta)}}$$

that we denote for brevity w. This and the fact that the function $\gamma \to \sqrt{\widehat{Q}(\gamma)}$ is convex imply that

$$\begin{aligned}\sqrt{\widehat{Q}(\beta)} - \sqrt{\widehat{Q}(\widehat{\beta})} &\le \langle w, \beta - \widehat{\beta}\rangle \\ &= \langle \mathbf{D}_\mathbf{X}w, \mathbf{D}_\mathbf{X}^{-1}(\beta - \widehat{\beta})\rangle = -\langle \mathbf{D}_\mathbf{X}w, \Delta\rangle\ .\end{aligned}$$

Now for any row of index k in the set J_{ex}, $|(\mathbf{D}_{\mathbf{X}}w)_k| \leq r$ on the event $\mathcal{G}$. This is because these regressors serve as their own instrument and, on the event $\mathcal{G}$, $\left(\beta, \sqrt{\widehat{Q}(\beta)}\right)$ belongs to $\widehat{\mathcal{I}}$. On the other hand, for any row of index k in the set J_{ex}^c, the Cauchy-Schwarz inequality yields that

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

Finally we obtain

$$\sqrt{\widehat{Q}(\widehat{\beta})} - \sqrt{\widehat{Q}(\beta)} \leq r|\Delta_{J_{\text{ex}}}|_1 + |\Delta_{J_{\text{ex}}^c}|_1 \ . \tag{A.14}$$

Combining this inequality with (A.13) we find that $\Delta \in C_{J(\beta)}$ on the event $\mathcal{G}$. Using (A.10) and arguing as in (A.13) we find

$$\begin{aligned} |\Psi_n\Delta|_\infty &\leq r\left(2\widehat{\sigma} + \sqrt{\widehat{Q}(\beta)} - \widehat{\sigma}\right) \qquad (A.15) \\ &\leq r\left(2\widehat{\sigma} + \sqrt{\widehat{Q}(\beta)} - \sqrt{\widehat{Q}(\widehat{\beta})}\right) \\ &\leq r\left(2\widehat{\sigma} + r|\Delta_{J_{\text{ex}}}|_1 + |\Delta_{J_{\text{ex}}^c}|_1\right) \ . \qquad (A.16) \end{aligned}$$

Using the definition of the sensitivities we get that, on the event $\mathcal{G}$,

$$|\Psi_n\Delta|_\infty \leq r\left(2\widehat{\sigma} + r\frac{|\Psi_n\Delta|_\infty}{\kappa_{1,J_{\text{ex}},J(\beta)}} + \frac{|\Psi_n\Delta|_\infty}{\kappa_{1,J_{\text{ex}}^c,J(\beta)}}\right),$$

which implies

$$|\Psi_n\Delta|_\infty \leq 2\widehat{\sigma}r\left(1 - \frac{r^2}{\kappa_{1,J_{\text{ex}},J(\beta)}} - \frac{r}{\kappa_{1,J_{\text{ex}}^c,J(\beta)}}\right)_+^{-1}. \tag{A.17}$$

This inequality and the definition of the sensitivities yield (6.1).

In the case without endogeneity where $L = K$ and $\boldsymbol{Z} = \boldsymbol{X}$,

$$\frac{1}{n}\sum_{i=1}^n (x_i^T\mathbf{D}_{\mathbf{X}}\Delta)^2 \leq |\Delta|_1 \, |\Psi_n\Delta|_\infty \, ,$$

(6.1) and (A.17) yield (6.3).

To prove (6.2), it suffices to note that, by (A.12) and by the definition of $\kappa_{1,J(\beta),J(\beta)}$,

$$\begin{aligned} c\widehat{\sigma} &\leq |\Delta_{J(\beta)}|_1 + c\sqrt{\widehat{Q}(\beta)} \\ &\leq \frac{|\Psi_n\Delta|_\infty}{\kappa_{1,J(\beta),J(\beta)}} + c\sqrt{\widehat{Q}(\beta)}, \end{aligned}$$

and to combine this inequality with (A.10). □

Proof of Theorem 6.2. Take β in $\mathcal{I}dent$. Fix an arbitrary subset J of $\{1,\dots,K\}$. Acting as in (A.13) with J instead of $J(\beta)$, we get:

$$\sum_{k\in J^c}\left|\mathbb{E}_n[X_k^2]^{1/2}\widehat{\beta}_k\right|+\sum_{k\in J^c}\left|\mathbb{E}_n[X_k^2]^{1/2}\beta_k\right|\le\sum_{k\in J}\left(\left|\mathbb{E}_n[X_k^2]^{1/2}\beta_k\right|-\left|\mathbb{E}_n[X_k^2]^{1/2}\widehat{\beta}_k\right|\right)+2\sum_{k\in J^c}\left|\mathbb{E}_n[X_k^2]^{1/2}\beta_k\right|$$
$$+c\left(\sqrt{\widehat{Q}(\beta)}-\sqrt{\widehat{Q}(\widehat{\beta})}\right)$$
$$\le|\Delta_J|_1+2\left|\left(\mathbf{D}_{\mathbf{X}}^{-1}\beta\right)_{J^c}\right|_1+cr\,|\Delta_{J_{\mathrm{ex}}}|_1+c\left|\Delta_{J_{\mathrm{ex}}^c}\right|_1\ .$$

This yields

$$|\Delta_{J^c}|_1\le|\Delta_J|_1+2\left|\left(\mathbf{D}_{\mathbf{X}}^{-1}\beta\right)_{J^c}\right|_1+cr\,|\Delta_{J_{\mathrm{ex}}}|_1+c\left|\Delta_{J_{\mathrm{ex}}^c}\right|_1. \tag{A.18}$$

Assume now that we are on the event $\mathcal{G}$. Consider the two possible cases. First, if $2\left|\left(\mathbf{D}_{\mathbf{X}}^{-1}\beta\right)_{J^c}\right|_1\le|\Delta_J|_1+cr\,|\Delta_{J_{\mathrm{ex}}}|_1+c\left|\Delta_{J_{\mathrm{ex}}^c}\right|_1$, then $\Delta\in\widetilde{C}_J$. From this, using the definition of the sensitivity $\widetilde{\kappa}_{p,J_0,J}$, we get that $|\Delta_{J_0}|_p$ is bounded from above by the first term of the maximum in (6.6). Second, if $2\left|\left(\mathbf{D}_{\mathbf{X}}^{-1}\beta\right)_{J^c}\right|_1>|\Delta_J|_1+cr\,|\Delta_{J_{\mathrm{ex}}}|_1+c\left|\Delta_{J_{\mathrm{ex}}^c}\right|_1$, then for any $p\in[1,\infty]$ we have a simple bound

$$|\Delta_{J_0}|_p\le|\Delta|_1=|\Delta_{J^c}|_1+|\Delta_J|_1\le6\left|\left(\mathbf{D}_{\mathbf{X}}^{-1}\beta\right)_{J^c}\right|_1.$$

In conclusion, $|\Delta_{J_0}|_p$ is smaller than the maximum of the two bounds. The bound for the prediction loss under exogeneity combines this inequality with (A.16) and (A.17) for enlarged cones, distingishing the two cases. □

Proof of Theorem 7.1. Part (i) of the theorem follows from (6.1) and (6.2) and Assumptions 7.1 together with (6.5). Part (ii) follows immediately from (6.4), and Assumptions 7.1. To prove part (iii), note that (7.6) and the assumption on $|\beta_k|$ imply: $\widehat{\beta}_k\neq0$ for $k\in J(\beta)$. □

Proof of Proposition 8.1. Any Δ in the cone C_J is such that

$$|\Delta_{J^c}|_1\le|\Delta_J|_1+cr|\Delta_{J_{\mathrm{ex}}}|_1+c|\Delta_{J_{\mathrm{ex}}^c}|_1\ .$$

Adding $|\Delta_J|_1$ to both sides yields

$$|\Delta|_1\le2|\Delta_J|_1+cr|\Delta_{J_{\mathrm{ex}}}|_1+c|\Delta_{J_{\mathrm{ex}}^c}|_1$$

or equivalently

$$(1-cr)|\Delta_{J_{\mathrm{ex}}}|_1+(1-c)|\Delta_{J_{\mathrm{ex}}^c}|_1\le2|\Delta_J|_1\ .$$

Because if $|J|\le s$, one gets

$$|\Delta|_1\le2s|\Delta_J|_\infty+cr|\Delta_{J_{\mathrm{ex}}}|_1+c|\Delta_{J_{\mathrm{ex}}^c}|_1\ . \tag{A.19}$$

This set of vectors Δ contains the cone and the lower bounds are obtained by minimizing on this larger set. One can assume everywhere that $\Delta_j \geq 0$ because the objective function in the sensitivities involves a sup-norm so that changing Δ in $-\Delta$ does not change the sensitivities.

Note that we have included the constraint $|\Delta|_\infty \leq 1$ in (8.1) because of (A.9). The rest follows from Proposition 4.1. □

Proof of Theorem 8.1. Fix β in $\mathcal{B}_s$. Let $\mathcal{G}_j$ be the events of probabilities at least $1-\gamma_j$ respectively appearing in Assumptions 7.2 and 8.1. Assume that all these events hold, as well as the event $\mathcal{G}$. Then, using Theorem 7.1 (i),

$$\omega_k(s) \leq \frac{2\sigma_* r}{\kappa_*(s) v_k} \left(1 + \frac{r|J(\beta)|}{c\kappa_*}\right) \left(1 - \frac{r|J(\beta)|}{c\kappa_*}\right)_+^{-1} \theta(s) \triangleq \omega_k^*.$$

By assumption, $|\beta_k| > 2\omega_k^*$ for $k \in J(\beta)$. Note that the following two cases can occur. First, if $k \in J(\beta)^c$ (so that $\beta_k = 0$) then, using (6.4) and Assumption 8.1, we obtain $|\widehat{\beta}_k| \leq \omega_k^*$, which implies $\widetilde{\beta}_k = 0$. Second, if $k \in J(\beta)$, then using again (6.4) we get $||\beta_k| - |\widehat{\beta}_k|| \leq |\beta_k - \widehat{\beta}_k| \leq \omega_k(s) \leq \omega_k^*$. Since $|\beta_k| > 2\omega_k^*$ for $k \in J(\beta)$, we obtain that $|\widehat{\beta}_k| > \omega_k^*$ thus $|\widehat{\beta}_k| > \omega_k(s)$, so that $\widetilde{\beta}_k = \widehat{\beta}_k$ and the signs of β_k and $\widehat{\beta}_k$ coincide. This yields the result. □

Proof of Theorem 9.1. The only difference with the proof of Theorem 6.1 is that, because we do not have the ℓ_1 norm in the objective function (9.1), we drop the discussion leading to the cone constraint. □

Proof of Theorem 9.2. Take β in $\mathcal{B}_s$, we have on $\mathcal{G}$, where $\mathcal{G}$ is defined in Section 5 adding the extra instrument $\zeta^T z_i$ for $i = 1, \ldots, n$.

$$\begin{aligned}
\frac{1}{n}|\widehat{\zeta}^T \mathbf{Z}^T \mathbf{U}| &\leq |\mathbf{D}_{\mathbf{Z}}^{-1}(\widehat{\zeta} - \zeta)|_1 \sqrt{\widehat{Q}(\beta) r} + \frac{1}{n}|(\zeta^T \mathbf{Z})^T \mathbf{U}| \\
&\leq \left(C_1(r, s_1) + \mathbb{E}_n[(\zeta^T Z)^2]^{1/2}\right) \sqrt{\widehat{Q}(\beta) r} \\
&\leq \left(C_1(r, s_1) + C_2(r, s_1) + \mathbb{E}_n[(\widehat{\zeta}^T Z)^2]^{1/2}\right) \sqrt{\widehat{Q}(\beta) r} \ .
\end{aligned}$$

The rest of the proof is the same as for Theorem 6.1. Equation (9.13) is a consequence of Theorem 8.3 calculating the value of $c_b(s)$ when $J_{\text{ex}}^c = \{1\}$. □

Proof of Theorem 9.3. Throughout the proof, we assume that we are on the event $\mathcal{G} \cap \mathcal{G}'$ where $\mathcal{G}$ is the event where (9.18) holds and $\mathcal{G}'$ is the event such that

$$\max_{l=1,\ldots,\overline{L}} \frac{\left|\mathbb{E}_n\left[\overline{Z}_l U - \theta_l\right]\right|}{\sqrt{\mathbb{E}_n\left[(\overline{Z}_l U - \theta_l)^2\right]}} \leq \overline{r}.$$

One has that on the event $\mathcal{G}'$

$$\left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\frac{1}{n}\overline{\mathbf{Z}}^T\mathbf{U}-\theta\right)\right|_\infty \le \overline{r}\max_{l=1,\dots,\overline{L}}\sqrt{\frac{\mathbb{E}_n[(\overline{Z}_l U-\theta_l)^2]}{\mathbb{E}_n[\overline{Z}_l^2]}} = \overline{r}F(\theta,\beta)\ . \tag{A.20}$$

We now use the properties of $F(\theta,\beta)$ stated in the next lemma that we prove in Section A.7.

Lemma A.2. *We have*

$$F(\theta,\beta)-F(\widehat{\theta},\beta) \le \overline{r}\left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\widehat{\theta}-\theta\right)\right|_1, \tag{A.21}$$

$$|F(\theta,\widehat{\beta})-F(\theta,\beta)| \le \overline{z}_*\left|\mathbf{D}_{\mathbf{X}}{}^{-1}(\widehat{\beta}-\beta^*)\right|_1 \le \widehat{b}_1\overline{z}_* \tag{A.22}$$

$$F(\widehat{\theta},\beta)-F(\widehat{\theta},\widehat{\beta}) \le \overline{z}_*\left|\mathbf{D}_{\mathbf{X}}{}^{-1}(\widehat{\beta}-\beta^*)\right|_1 \le \widehat{b}_1\overline{z}_*. \tag{A.23}$$

We proceed now to the proof of Theorem 9.3. First, we show that the pair $(\theta,\overline{\sigma})=(\theta,F(\theta,\beta))$ belongs to the set $\widehat{\overline{\mathcal{I}}}$. Indeed, from (A.20) we get

$$\begin{aligned}\left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\frac{1}{n}\overline{\mathbf{Z}}^T(\mathbf{Y}-\mathbf{X}\widehat{\beta})-\theta\right)\right|_\infty &\le \left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\frac{1}{n}\overline{\mathbf{Z}}^T\mathbf{U}-\theta\right)\right|_\infty + \left|\frac{1}{n}\mathbf{D}_{\overline{\mathbf{Z}}}\overline{\mathbf{Z}}^T\mathbf{X}(\widehat{\beta}-\beta)\right|_\infty \\ &\le \overline{r}F(\theta,\beta)+\widehat{b}\ .\end{aligned}$$

Thus, the pair $(\theta,\overline{\sigma})=(\theta,F(\theta,\beta))$ satisfies the first constraint in the definition of $\widehat{\overline{\mathcal{I}}}$. It satisfies the second constraint as well, since $F(\theta,\widehat{\beta})\le F(\theta,\beta)+\widehat{b}_1\overline{z}_*$ by (A.22).

Now, as $(\theta,F(\theta,\beta))\in\widehat{\overline{\mathcal{I}}}$ and $(\widehat{\theta},\widehat{\overline{\sigma}})$ minimizes $|\theta|_1+\overline{c}\,\overline{\sigma}$ over $\widehat{\overline{\mathcal{I}}}$, we have

$$\left|\mathbf{D}_{\overline{\mathbf{Z}}}\widehat{\theta}\right|_1+\overline{c}\,\widehat{\overline{\sigma}} \le \left|\mathbf{D}_{\overline{\mathbf{Z}}}\theta\right|_1+\overline{c}\,F(\theta,\beta), \tag{A.24}$$

which implies

$$|\overline{\Delta}_{J(\theta)^c}|_1 \le |\overline{\Delta}_{J(\theta)}|_1+\overline{c}(F(\theta,\beta)-\widehat{\overline{\sigma}}), \tag{A.25}$$

where $\overline{\Delta}=\mathbf{D}_{\overline{\mathbf{Z}}}\left(\widehat{\theta}-\theta\right)$. Using the fact that $F(\widehat{\theta},\widehat{\beta})\le\widehat{\overline{\sigma}}+\widehat{b}\overline{z}_*$ (by the definition of the estimator), (A.21), and (A.22) we obtain

$$\begin{aligned}F(\theta,\beta)-\widehat{\overline{\sigma}} &\le F(\theta,\beta)-F(\widehat{\theta},\widehat{\beta})+\widehat{b}_1\overline{z}_* \\ &= (F(\theta,\beta)-F(\widehat{\theta},\beta))+(F(\widehat{\theta},\beta)-F(\widehat{\theta},\widehat{\beta}))+\widehat{b}_1\overline{z}_* \\ &\le \overline{r}\left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\widehat{\theta}-\theta\right)\right|_1+2\widehat{b}_1\overline{z}_*.\end{aligned} \tag{A.26}$$

This inequality and (A.25) yield

$$|\overline{\Delta}_{J(\theta)^c}|_1 \le |\overline{\Delta}_{J(\theta)}|_1+\overline{c}\,\overline{r}\left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\widehat{\theta}-\theta\right)\right|_1+2\overline{c}\widehat{b}_1\overline{z}_*,$$

or equivalently,

$$|\overline{\Delta}_{J(\theta)^c}|_1 \leq \frac{1+\overline{c}\,\overline{r}}{1-\overline{c}\,\overline{r}}\,|\overline{\Delta}_{J(\theta)}|_1 + \frac{2\overline{c}}{1-\overline{c}\,\overline{r}}\widehat{b}_1\overline{z}_*. \tag{A.27}$$

Next, using (A.20) and the second constraint in the definition of $(\widehat{\theta},\widehat{\overline{\sigma}})$, we find

$$\begin{aligned}\left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\widehat{\theta}-\theta\right)\right|_\infty &\leq \left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\frac{1}{n}\overline{\mathbf{Z}}^T(\mathbf{Y}-\mathbf{X}\widehat{\beta})-\widehat{\theta}\right)\right|_\infty \\ &\quad + \left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\frac{1}{n}\overline{\mathbf{Z}}^T\mathbf{U}-\theta\right)\right|_\infty + \left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\frac{1}{n}\overline{\mathbf{Z}}^T\mathbf{X}(\widehat{\beta}-\beta)\right)\right|_\infty \\ &\leq \overline{r}(\widehat{\overline{\sigma}}+F(\theta,\beta))+2\widehat{b}.\end{aligned}$$

This and (A.26) yield

$$\left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\widehat{\theta}-\theta\right)\right|_\infty \leq \overline{r}\left(2\widehat{\overline{\sigma}}+\overline{r}\left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\widehat{\theta}-\theta\right)\right|_1\right)+2\overline{r}\widehat{b}_1\overline{z}_*+2\widehat{b}. \tag{A.28}$$

On the other hand, (A.27) implies

$$\begin{aligned}\left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\widehat{\theta}-\theta\right)\right|_1 &= |\overline{\Delta}|_1 = |\overline{\Delta}_{J(\theta)}|_1 + |\overline{\Delta}_{J(\theta)^c}|_1 \\ &\leq \frac{2}{1-\overline{c}\,\overline{r}}\,|\overline{\Delta}_{J(\theta)}|_1 + \frac{2\overline{c}}{1-\overline{c}\,\overline{r}}\widehat{b}_1\overline{z}_* \\ &\leq \frac{2|J(\theta)|}{1-\overline{c}\,\overline{r}}\left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\widehat{\theta}-\theta\right)\right|_\infty + \frac{2\overline{c}}{1-\overline{c}\,\overline{r}}\widehat{b}_1\overline{z}_*.\end{aligned} \tag{A.29}$$

Inequalities (9.23) and (9.24) follow from solving (A.28) and (A.29) with respect to $\left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\widehat{\theta}-\theta\right)\right|_\infty$ and $\left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\widehat{\theta}-\theta\right)\right|_1$ respectively. □

Proof of Theorem 9.4. We assume everywhere that we are on the event $\mathcal{G}\cap\mathcal{G}_1\cap\mathcal{G}_2\cap\mathcal{G}'\cap\mathcal{G}_3$ where (A.20), (9.25), and (9.21) are simultaneously satisfied.

We first prove part (i). From (A.24) and the fact that (9.25) can be written as $F(\theta,\beta)\leq\overline{\sigma}_*$ we obtain

$$\widehat{\overline{\sigma}} \leq \frac{|\overline{\Delta}_{J(\theta)}|_1}{\overline{c}} + \overline{\sigma}_* \leq \frac{|J(\theta)|\left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\widehat{\theta}-\theta\right)\right|_\infty}{\overline{c}} + \overline{\sigma}_*. \tag{A.30}$$

Item (i) and (ii) now follow easily using the fact that V is increasing in all of its arguments.

To prove part (iii), note that the thresholds ω_l satisfies

$$\begin{aligned}\omega_l &\triangleq \mathbb{E}_n[\overline{Z}_l^2]^{1/2}\left(1-\frac{2\overline{s}\,\overline{r}}{\overline{c}(1-\overline{c}\,\overline{r})}\right)_+^{-1}\left(1-\frac{2\overline{s}\,\overline{c}\,\overline{r}^2}{\overline{c}(1-\overline{c}\,\overline{r})}\right)V(\widehat{\overline{\sigma}},\widehat{b},\widehat{b}_1,J(\widehat{\theta})) \\ &\leq \overline{v}_l\left(1-\frac{2\overline{s}\,\overline{r}}{\overline{c}(1-\overline{c}\,\overline{r})}\right)_+^{-1}\left(1-\frac{2\overline{s}\,\overline{c}\,\overline{r}^2}{\overline{c}(1-\overline{c}\,\overline{r})}\right)V(\overline{\sigma}_*,b_*,b_{1*},\overline{s}) \triangleq \omega_l^*\end{aligned}$$

on the event. On the other hand, (9.23) guarantees that $|\widehat{\theta}_l - \theta^*_l| \le \omega_l$ and, by assumption, $|\theta^*_l| > 2\omega^*_l > 2\omega_l$ for all $l \in J(\theta^*)$. In addition, by (6.1) and (9.23) for all $l \in J(\theta^*)^c$ we have $|\theta^*_l| < \omega_l$, which implies $\widetilde{\theta}_l = 0$. We finish the proof in the same way as the proof of Theorem 7.1. □

A.7. **Proof of Lemma A.2.** We set $f_l(\theta) \triangleq \sqrt{\widehat{Q}_l(\theta,\beta)}$, and $f(\theta) \triangleq \max_{l=1,\dots,\overline{L}} f_l(\theta) \equiv F(\theta,\beta)$. Each function f_l is convex and

$$(\nabla f_l(\theta))_l = -\frac{\mathbb{E}_n\left[\overline{Z}_l(Y - X^T\beta) - \theta_l\right]}{\sqrt{\mathbb{E}_n[\overline{Z}_l^2]\mathbb{E}_n\left[\left(\overline{Z}_l(Y - X^T\beta) - \theta_l\right)^2\right]}}$$

is such that $|(\nabla f_l(\theta))_l| \le \frac{\overline{r}}{\mathbb{E}_n[\overline{Z}_l^2]^{1/2}}$ while $(\nabla f_l(\theta))_m = 0$ for $m \ne l$. This implies, in view of Lemma A.1, that $\partial f(\theta) \subseteq \left\{w \in \mathbb{R}^{\overline{L}} : \left|\mathbf{D}_{\overline{\mathbf{Z}}}^{-1} w\right|_\infty \le \overline{r}\right\}$. Thus

$$f(\theta) - f(\widehat{\theta}) \le \langle w, \theta^* - \widehat{\theta}\rangle = \left\langle \mathbf{D}_{\overline{\mathbf{Z}}}^{-1} w, \mathbf{D}_{\overline{\mathbf{Z}}}\left(\theta^* - \widehat{\theta}\right)\right\rangle \le \overline{r}\left|\mathbf{D}_{\overline{\mathbf{Z}}}\left(\widehat{\theta} - \theta^*\right)\right|_1, \quad \forall\, w \in \partial f(\theta^*),$$

where $\langle\cdot,\cdot\rangle$ denotes the standard inner product in $\mathbb{R}^{\overline{L}}$. Thus (A.21) follows. The proof of (A.22) and (A.23) are based on similar arguments. Let us prove for example (A.22). Instead of f_l, we now introduce the functions g_l defined by $g_l(\beta) \triangleq \sqrt{\widehat{Q}_l(\theta,\beta)}$, and set $g(\beta) \triangleq \max_{l=1,\dots,\overline{L}} g_l(\beta) \equiv F(\theta,\beta)$. Each function g_l is convex and their gradient $\nabla g_l(\beta)$ is

$$\nabla g_l(\beta) = -\frac{\mathbb{E}_n\left[\overline{Z}_l X\left(\overline{Z}_l(Y - X^T\beta) - \theta_l\right)\right]}{\sqrt{\mathbb{E}_n[\overline{Z}_l^2]\mathbb{E}_n\left[\left(\overline{Z}_l(Y - X^T\beta) - \theta_l\right)^2\right]}}\,.$$

Using the Cauchy-Schwarz inequality for all $k = 1,\dots,K$, we get

$$|(\nabla g_l(\beta))_k| \le \overline{z}_*\sqrt{\mathbb{E}_n[X_k^2]} = \overline{z}_*(\mathbf{D}_{\mathbf{X}})^{-1}_{kk}.$$

Because the functions g_l are convex, Lemma A.1 yields that the subdifferential of the function $g(\cdot) = \max_{l=1,\dots,\overline{L}} g_l(\cdot)$ is included in the same hyperrectangle: $\partial g(\beta) \subseteq \{w \in \mathbb{R}^K : |w_k| \le \overline{z}_*(\mathbf{D}_{\mathbf{X}})^{-1}_{kk}\}$ for all $\beta \in \mathbb{R}^K$. Hence, for any $w \in \partial g(\beta)$ we have $|\mathbf{D}_{\mathbf{X}} w|_\infty \le \overline{z}_*$. This and the definition of subdifferential imply that, for any $\beta, \beta' \in \mathbb{R}^K$ and any $w \in \partial g(\beta)$,

$$g(\beta) - g(\beta') \le \langle w, \beta - \beta'\rangle \le |\mathbf{D}_{\mathbf{X}} w|_\infty \left|\mathbf{D}_{\mathbf{X}}{}^{-1}(\beta - \beta')\right|_1 \le \overline{z}_* \left|\mathbf{D}_{\mathbf{X}}{}^{-1}(\beta - \beta')\right|_1,$$

which immediately implies (A.22). □

A.8. **Extending Scenario 5 to Heteroscedastic Errors.** For the extension to heteroscedastic errors we make the following assumption.

The errors u_i are independent from the z_i's, (z_i, u_i) are independent, there exists constants $\underline{c}$, $\overline{C}$, B_n such that: (i) $\forall i = 1, \ldots, n,\ l = 1, \ldots, L\ |z_{li}| \leq B_n$ a.s.; (ii) $\mathbb{E}[U^4] \leq \overline{C}$; (iii) $B_n^4(\log(Ln))^7/n \leq \overline{C}n^{-\underline{c}}$.

We make use of a first stage *STIV* estimator $(\widehat{\beta}_1, \widehat{\sigma}_1)$ obtained for some constant c_1 and r_1 which is associated to a confidence level $1 - \alpha_1$ under a scenario 4, the associated confidence set with sparsity certificate s and the following statistic

$$W \triangleq \max_{l=1,\ldots,L} \left| \mathbb{E}_n \left[\frac{Z_l(Y - X^T\widehat{\beta}_1)V}{\mathbb{E}_n[Z_l^2]^{1/2}} \right] \right|$$

where v_i are independent standard normal random variables independent of z_i, y_i and x_i. For a confidence level $1 - \alpha$ one adjusts r, α_1, r_1 and η so that

$$\mathbb{P}\left(W \geq r - \eta | y_i, x_i, z_i,\ i = 1, \ldots, n\right) + \mathbb{P}\left(\frac{2\widehat{\sigma}_1 r_1 \theta(s)}{\kappa_1(s)} \left| \mathbf{D}_{\mathbf{Z}} \mathbb{E}_n[ZX^T V] \mathbf{D}_{\mathbf{X}} \right|_\infty \geq \eta \middle| y_i, x_i, z_i,\ i = 1, \ldots, n \right)$$

$$\leq \alpha - \alpha_1$$

where both probabilities can be approximated by Monte-Carlo. Then one uses as a second stage the non-pivotal *STIV* estimator from Gautier and Tsybakov (2011) with $\sigma = 1$.

Proof of the validity of the procedure. Define

$$T_0 \triangleq \max_{l=1,\ldots,L} \left| \mathbb{E}_n \left[\frac{Z_l U}{\mathbb{E}_n[Z_l^2]^{1/2}} \right] \right|$$

$$W_0 \triangleq \max_{l=1,\ldots,L} \left| \mathbb{E}_n \left[\frac{Z_l U V}{\mathbb{E}_n[Z_l^2]^{1/2}} \right] \right| .$$

We need to prove that

$$\overline{\lim}_{n\to\infty,\ B_n^4(\log(Ln))^7/n \leq \overline{C}n^{-\underline{c}}} \mathbb{P}(T_0 \geq r) \leq \alpha .$$

By Corollary 2.1 of Chernozhukov, Chetverikov and Kato (2013), one obtains that for some positive constants c_2 and C_2 and $B_n^4(\log(Ln))^7/n \leq \overline{C}n^{-\underline{c}}$

$$\sup_{t\in\mathbb{R}} |\mathbb{P}\left(T_0 \leq t | z_{li}, i = 1, \ldots, n,\ l = 1, \ldots, L\right) - \mathbb{P}\left(W_0 \leq t | z_{li}, i = 1, \ldots, n,\ l = 1, \ldots, L\right)| \leq C_2 n^{-c_2} .$$

So one has to prove that

$$\overline{\lim}_{n\to\infty,\ B_n^4(\log(Ln))^7/n \leq \overline{C}n^{-\underline{c}}} \mathbb{P}(W_0 \geq r) \leq \alpha .$$

One can conclude using the pigeonhole principle and the fact that

$$W_0 - W \leq \max_{l=1,\ldots,L} \left| \mathbb{E}_n \left[\frac{Z_l X^T(\beta - \widehat{\beta})V}{\mathbb{E}_n[Z_l^2]^{1/2}} \right] \right| \leq \left| \mathbf{D}_{\mathbf{X}}^{-1}(\beta - \widehat{\beta}) \right|_1 \left| \mathbf{D}_{\mathbf{Z}} \mathbb{E}_n[ZX^T V] \mathbf{D}_{\mathbf{X}} \right|_\infty \qquad \square$$

CREST, ENSAE ParisTech, 3 avenue Pierre Larousse, 92 245 Malakoff Cedex, France.

E-mail address: eric.gautier@ensae-paristech.fr, alexandre.tsybakov@ensae-paristech.fr