HIGH-DIMENSIONAL INSTRUMENTAL VARIABLES REGRESSION AND CONFIDENCE SETS

ERIC GAUTIER AND ALEXANDRE TSYBAKOV

CREST (ENSAE), 3 avenue Pierre Larousse, 92 245 Malakoff Cedex, France;
eric.gautier@ensae.fr; alexandre.tsybakov@ensae.fr.

Abstract. We propose an instrumental variables method for estimation in linear models with endogenous regressors in the high-dimensional setting where the sample size n can be smaller than the number of possible regressors K, and $L \geq K$ instruments. We allow for heteroscedasticity and we do not need a prior knowledge of variances of the errors. We suggest a new procedure called the *STIV* (Self Tuning Instrumental Variables) estimator, which is realized as a solution of a conic optimization program. The main results of the paper are upper bounds on the estimation error of the vector of coefficients in ℓ_p-norms for $1 \leq p \leq \infty$ that hold with probability close to 1, as well as the corresponding confidence intervals. All results are non-asymptotic. These bounds are meaningful under the assumption that the true structural model is sparse, *i.e.*, the vector of coefficients has few non-zero coordinates (less than the sample size n) or many coefficients are too small to matter. In our *IV* regression setting, the standard tools from the literature on sparsity, such as ther restricted eigenvalue assumption are inapplicable. Therefore, for our analysis we develop a new approach based on data-driven sensitivity characteristics. We show that, under appropriate assumptions, a thresholded *STIV* estimator correctly selects the non-zero coefficients with probability close to 1. The price to pay for not knowing which coefficients are non-zero and which instruments to use is of the order $\sqrt{\log(L)}$ in the rate of convergence. We extend the procedure to deal with high-dimensional problems where some instruments can be non-valid. We obtain confidence intervals for non-validity indicators and we suggest a procedure, which correctly detects the non-valid instruments with probability close to 1.

Date: First version December 2009, this version: August 2011.

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

We thank seminar participants at Bocconi, CREST, Compiègne, Institut Henri Poincaré, Paris 6 and 7, and Toulouse 3 as well as participants of SPA and Saint-Flour 2011 for helpful comments.

1. Introduction

Endogeneity is one of the most important issues in empirical economic research. Consider the linear model

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

where x_i are vectors of explanatory variables of dimension $K \times 1$, u_i is a zero-mean random error possibly correlated with x_i, and β^* is an unknown parameter. We denote by x_{ki}, $k = 1, \dots, K$, the components of x_i. The regressors x_{ki} are called endogenous if they are correlated with u_i and they are called exogenous otherwise. Without loss of generality, we assume that the endogenous variables are $x_{1i}, \dots, x_{k_{\text{end}}i}$ for some $k_{\text{end}} \leq K$. It is well known that endogeneity occurs, for example, when a regressor correlated both with y_i and regressors in the model is unobserved; in the errors-in-variables model when the measurement error is independent of the underlying variable; when a regressor is determined simultaneously with the response variable y_i; in treatment effect models when the individuals can self-select to the treatment (see, *e.g.*, Wooldridge (2002)). The quantities of interest that we would like to estimate are the components β_k^* of β^*. They characterize the partial effect of the variable x_{ki} on the outcome y_i for fixed other variables.

Having access to instrumental variables makes it possible to identify the coefficients β_k^* in such a setting. A random vector z_i of dimension $L \times 1$ with $L \geq K$ will be called a vector of instrumental variables (or instruments) if it satisfies

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

where $\mathbb{E}[\,\cdot\,]$ denotes the expectation. Throughout the paper, we assume that the exogenous variables serve as their own instruments, which means that the components z_{li} of z_i satisfy $z_{li} = x_{l'i}$, where $l' = k_{\text{end}} + l$, $l = 1, \dots, K - k_{\text{end}}$. We consider the problem of inference on the parameter β^* from n independent realizations (y_i, x_i^T, z_i^T), $i = 1, \dots, n$. We allow these observations and the unobserved error terms u_i to be heteroscedastic.

In this paper, we are mainly interested in the high-dimensional setting where the sample size n is small compared to K, and one of the following two assumptions is satisfied:

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

(ii) most of the coefficients β_k^* are too small to matter (β^* is *approximately sparse*).

In this setting, the system of equations (1.2) provides more moment conditions than observations, and β^* cannot be identified by usual instrumental variables methods. Even if $L = K$ and the observations

are identically distributed, the empirical counterpart of the matrix $\mathbb{E}[z_1 x_1^T]$ has rank at most $\min(n, K)$ and is not invertible for $K > n$. To our knowledge, no estimator for this setting is currently available.

Cross-country or cross-states regressions are typical situations where one may want to use high-dimensional procedures. The sample size is usually small and one may want to include many variables. Economic theory is indeed not always explicit about the variables that belong to the true model (see, *e.g.*, Sala-i-Martin (1997) concerning development economics). Cross-country regressions are widely used in macroeconomics, development economics or international finance. One possible application is the estimation of Engle curves using aggregate data where the total expenditure is endogenous and we consider as regressors various transformations of the total expenditure. There are other contexts in economics where high-dimensional methods can be used. For example, it is notably hard to obtain adequate data (legal issues, etc.) in contract economics and the researcher may be interested in studying contracts between governors and public firms or state regulations of private telecommunications company, etc. Even in contexts where sample sizes are relatively large, the full search over the models is exponentially hard in the number of parameters. High-dimensional methods can be extremely useful for this purpose since they provide computationally feasible methods of variable selection. There are indeed many cases where the theory asks for a rich and flexible specification. The list of possible regressors quickly increases when one considers interactions between variables or wants to explore the *IV*-regression in nonparametric setting using linear combinations of elementary functions to approximate the nonparametric function of interest. One may also want to control for many variables when there is a rich heterogeneity or to justify exclusion restrictions and the validity of instruments. Finally, even in cases where the theory is explicit and the selection of variables is not a priori an issue, it becomes important in a stratified analysis when models are estimated in small population sub-groups (for example, in estimating models by cells as defined by the value of an exogenous discrete variable).

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), Belloni and Chernozhukov (2011a); more references can be found in the recent books by Bühlmann and van de Geer (2011), as well as in the lecture notes by Koltchinskii (2011), Belloni and Chernozhukov (2011b)), and the Bayesian-type methods (see Dalalyan and Tsybakov (2008), Rigollet and Tsybakov (2011) and the papers cited therein). In recent years, these techniques became a reference in several areas, such as biostatistcs and imaging. Some first applications are now available in economics. Thus,

Belloni and Chernozhukov (2011a) study the ℓ_1-penalized quantile regression and give an application to cross-country growth analysis. Belloni and Chernozhukov (2010) present various applications of the Lasso to economics including wage regressions, in particular, the selection of instruments in such models. Belloni, Chernozhukov and Hansen (2010) use the Lasso to estimate the optimal instruments with an application to the impact of eminent domain on economic outcomes. Caner (2009) studies a Lasso-type GMM estimator. Rosenbaum and Tsybakov (2010) deal with the high-dimensional errors-in-variables problem where the non-random regressors are observed with error and discuss an application to hedge fund portfolio replication. The high-dimensional setting in a structural model with endogenous regressors that we are considering here has not yet been analyzed. Note that the direct implementation of the Lasso or Dantzig selector fails in the presence of a single endogenous regressor as the zero coefficients in the structural equation (1.1) do not correspond to the zero coefficients in a linear projection type model.

The main message of this paper is that, in model (1.1) containing endogenous regressors, the high-dimensional vector of coefficients can be estimated together with proper confidence intervals using instrumental variables. This is achieved by the *STIV* estimator (Self Tuning Instrumental Variables estimator) that we introduce below. Based on it, we can also perform variable selection. All our results are non-asymptotic and provide meaningful bounds when either (i) or (ii) above holds and $\log(L)$ is small as compared to n. In particular, they can be used in the still troublesome case $K \leq n < L$, *i.e.*, for models with relatively small number of variables and relatively large number of instruments. As exemplified by Angrist and Krugger (1991), under a stronger notion of exogeneity which is based on a zero conditional mean assumption, considering interactions of instruments or functionals of instruments can lead to a large amount of instruments. This is related to the many instruments literature (see, *e.g.,* Andrews and Stock (2007) for a review). Important problems in this context are selection of instruments (see, e.g., Donald and Newey (2001), Hall and Peixe (2003), Okui (2008), Bai and Ng (2009), and Belloni and Chernozhukov (2011b)), estimation of optimal instruments (see, e.g., Amemiya (1974), Chamberlain (1987), Newey (1990), and Belloni, Chen, Chernozhukov et al. (2010)) or various other issues in the many instruments asymptotics (see, e.g., Chao and Swanson (2005), Hansen, Hausman and Newey (2008), Hausman, Newey, Woutersen et al. (2009)). The number of instruments can be much larger than the sample size. Carrasco (2008), building on Carrasco and Florens (2000, 2008), analyzes this setting in the inverse problems framework and proposes a suitable regularization method. The *STIV* estimator also leads to a smoothing procedure which is able to handle this case.

The *STIV* estimator is an extension of the Dantzig selector of Candès and Tao (2007). Like the Square-root Lasso of Belloni, Chernozhukov and Wang (2010), the *STIV* estimator is a pivotal method independent of the variances of the errors, which are allowed to be heteroscedastic. The implementation of the *STIV* estimator corresponds to solving a conic optimization program. The results of this paper extend those on the Dantzig selector (see Candès and Tao (2007), Bickel, Ritov and Tsybakov (2009) and further references in Bühlmann and van de Geer (2011)) 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), by imposing weak distributional assumptions, by introducing a procedure independent of the noise level and by providing finite sample confidence intervals.

We present basic definitions and notation in Section 2 and we introduce the *STIV* estimator in Section 3. In Section 4 we present the sensitivity characteristics, which play a major role in our error bounds and confidence intervals. They provide a generalization of the restricted eigenvalues to non-symmetric and non-square matrices. The main results on the *STIV* estimator are given in Section 5. In Section 6 we consider the setting where some instruments might be non-valid and we wish to detect this. Section 7 discusses some special cases and extensions, in particular, the *STIV* procedure with estimated linear projection type instruments, akin to two-stage least squares. Section 8 considers computational issues and presents a simulation study. All the proofs are given in the appendix. Also in the appendix, we compare our sensitivity analysis to the more standard one based on restricted eigenvalues in the case of symmetric square matrices.

2. Basic definitions and notation

We set $\mathbf{Y} = (y_1, \ldots, y_n)^T$, $\mathbf{U} = (u_1, \ldots, u_n)^T$, 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, $i = 1, \ldots, n$.

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 $k \in \{1, \ldots, K\}$, $l \in \{1, \ldots, L\}$, $a \geq 0, b \geq 0$. 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:

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

and set

$$x_{k*} \triangleq \max_i |x_{ki}|, \quad z_{l*} \triangleq \max_i |z_{li}|$$

for $k = 1, \dots, K$, $l = 1, \dots, L$. We denote by $\mathbf{D_X}$ and $\mathbf{D_Z}$ the diagonal $K \times K$ (respectively, $L \times L$) matrices with diagonal entries x_{k*}^{-1}, $k = 1, \dots, K$ (respectively, z_{l*}^{-1}, $l = 1, \dots, L$).

For a vector $\beta \in \mathbb{R}^K$, let $J(\beta) = \{k \in \{1, \dots, K\} : \ \beta_k \neq 0\}$ be its support, *i.e.*, the set of indices corresponding to its non-zero components β_k. We denote by $|J|$ the cardinality of a set $J \subseteq \{1, \dots, K\}$ and by J^c its complement: $J^c = \{1, \dots, K\} \setminus J$. The subset of indices $\{1, \dots, K\}$ corresponding to endogenous regressors is denoted by J_{end}. 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 consider $\Delta_J \triangleq (\Delta_1 \mathbb{1}_{\{1 \in J\}}, \dots, \Delta_K \mathbb{1}_{\{K \in J\}})^T$, where $\mathbb{1}_{\{.\}}$ 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 $a_+ \triangleq \max(0, a)$, $a_+^{-1} \triangleq (a_+)^{-1}$, and $a/0 \triangleq \infty$ for $a > 0$. We adopt the convention $0/0 \triangleq 0$ and $1/\infty \triangleq 0$.

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 $L \geq K$ equations with K unknown parameters. If $L > K$, it is overdetermined; if $L = K$ the matrix $\mathbf{Z}^T\mathbf{X}$ is not invertible in the high-dimensional case $K > n$, since its rank is at most $\min(n, K)$. Furthermore, replacing the population equations (1.2) by (3.1) induces a huge error when $L, K > n$. So, looking for the exact solution of (3.1) in the high-dimensional setting 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 $\beta = \beta^*$ 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. It is possible to consider different small sets in (3.2), however, the use of the sup-norm makes the inference robust when some (not all) instruments for each endogenous variable are weak.

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. For technical reasons, we choose normalization by the maximal absolute value, *i.e.*, multiplying $\mathbf{Z}$ by $\mathbf{D_Z}$. Then the constraint (3.2) is modified as follows:

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

Along with the constraint of the form (3.3), we include the second constraint to account for the unknown level σ of the "noise" u_i; in particular, if the errors u_i are i.i.d., σ^2 corresponds to their unknown variance. 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

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

for some $r>0$ (to be specified below), and

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

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

$$\min_{(\beta,\sigma)\in\widehat{\mathcal{I}}} \left(\left|\mathbf{D}_\mathbf{X}^{-1}\beta\right|_1 + c\sigma\right), \tag{3.5}$$

where $0<c<1$.

We use $\widehat{\beta}$ as an estimator of β^*. Finding the *STIV* estimator is a conic program; it can be efficiently solved, see Section 8.1. Note that the *STIV* estimator is not necessarily unique. Minimizing the ℓ_1 criterion $\left|\mathbf{D}_\mathbf{X}^{-1}\beta\right|_1$ is a convex relaxation of minimizing the ℓ_0 norm, *i.e.*, the number of non-zero coordinates of β. This usually ensures that the resulting solution is sparse. The term $c\sigma$ is included in the criterion to prevent from choosing σ arbitrarily large; indeed, the *IV*-constraint does not prevent from this. The matrix $\mathbf{D}_\mathbf{X}^{-1}$ arises from re-scaling of $\mathbf{X}$, which is similar to the re-scaling of $\mathbf{Z}$ discussed above. For the particular case where $\mathbf{Z}=\mathbf{X}$, 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 also be related to the Square-root Lasso of Belloni, Chernozhukov and Wang (2010), which solves

the problem of unknown variance in high-dimensional regression with deterministic regressors and i.i.d. errors. The definition of *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 have to handle the endogeneity.

Remark 3.2. *Other normalizations can be used. Instead of matrices* $\mathbf{D_X}$ *and* $\mathbf{D_Z}$ *we can take the diagonal matrices with entries* $\mathbb{E}_n[X_1^2]^{-1/2}, \ldots, \mathbb{E}_n[X_K^2]^{-1/2}$ *and* $\mathbb{E}_n[Z_1^2]^{-1/2}, \ldots, \mathbb{E}_n[Z_L^2]^{-1/2}$ *respectively. Then the proofs become more complicated and we need extra assumptions, in the spirit that for every* $l = 1, \ldots, L$, *and* $i = 1, \ldots, n$, *the variables* u_i^2 *and* z_{li}^2 *are "almost" uncorrelated.*

4. Sensitivity characteristics

The identifiability of β^* relies on the sensitivity characteristics of the problem. 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. More precisely, we will deal with its normalized version

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

In general, Ψ_n is not a square matrix. For $L = K$, it is square matrix but, in the presence of at least one endogenous regressor, Ψ_n is not symmetric. Since the endogenous variables are assumed to be the first variables in the model, the lower right block of matrix Ψ_n is, up to a scaling, the sample correlation matrix of the exogenous variables (when considering x_i as centered). The upper left block accounts for the relation between the endogenous variables and the instruments.

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

$$C_J \triangleq \left\{\Delta \in \mathbb{R}^K : \ |\Delta_{J^c}|_1 \leq \frac{1+c}{1-c}|\Delta_J|_1\right\},$$

where $0 < c < 1$ is the constant in the definition of *STIV* estimator, J is a subset of $\{1, \ldots, K\}$, and J^c denotes the complement of J. When 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. We call C_J the *cone of dominant coordinates*. The use of similar cones to define sensitivity characteristics is standard in the literature on the Lasso and the Dantzig selector (see, Bickel, Ritov and Tsybakov (2009)); the particular choice of the constant $\frac{1+c}{1-c}$ will become clear from the proofs. It follows from the definition of C_J that

$$|\Delta|_1 \le \frac{2}{1-c}|\Delta_J|_1 \le \frac{2}{1-c}|J|^{1-1/p}|\Delta_J|_p, \quad \forall\ \Delta \in C_J,\ 1 \le p \le \infty. \tag{4.1}$$

For $p \in [1,\infty]$, we define the ℓ_p *sensitivity* as the following random variable:

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

[EG] These quantities are similar to the cone invertibility factors defined in Ye and Zhang (2010). Given a subset $J_0 \subset \{1,\dots,K\}$, we define the J_0 *block sensitivity* as

$$\kappa^*_{J_0,J} \triangleq \inf_{\Delta \in C_J:\ |\Delta_{J_0}|_1 = 1} |\Psi_n \Delta|_\infty . \tag{4.2}$$

By convention, we set $\kappa^*_{\varnothing,J(\beta^*)} = \infty$. We use the notation $\kappa^*_{k,J}$ for *coordinate-wise sensitivities*, *i.e.*, for block sensitivities when $J_0 = \{k\}$ is a singleton:

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

Note that here we restrict the minimization to vectors Δ with positive kth coordinate, $\Delta_k = 1$, since replacing Δ by $-\Delta$ yields the same value of $|\Psi_n \Delta|_\infty$.

The finite sample bounds that we obtain below (see, *e.g.*, Theorem 5.2) show that the inverse of the sensitivities drive the width of the confidence interval for the true parameter. Thus, it is important to have computable bounds on these characteristics. The following proposition will be useful.

Proposition 4.1. (*i*) *Let $J, \widehat{J}$ be two subsets of $\{1,\dots,K\}$ such that $J \subseteq \widehat{J}$. Then $\kappa_{p,J} \ge \kappa_{p,\widehat{J}}$, and $\kappa^*_{J_0,J} \ge \kappa^*_{J_0,\widehat{J}}$ for all $p \in [1,\infty]$;*

(*ii*) *For all $J_0 \subset \{1,\dots,K\}$, $\kappa^*_{J_0,J} \ge \kappa_{1,J}$.*

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

$$\left(\frac{2|J|}{1-c}\right)^{-1/p} \kappa_{\infty,J} \le \kappa_{p,J} \le \frac{2}{1-c}|J|^{1-1/p}\kappa_{1,J}. \tag{4.3}$$

The proof of Proposition 4.1 is given in Section 9.3.

We can control $\kappa_{p,J(\beta^*)}$ without knowing $J(\beta^*)$ by means of *sparsity certificate*. Assume that we have an upper bound s on the sparsity of β^*, *i.e.*, we know that $|J(\beta^*)| \le s$ for some integer s. Meaningful values of s are small presuming that only few regressors are relevant. In view of (4.1), if

$|J| \le s$, then for any Δ in the cone C_J we have $|\Delta|_1 \le \frac{2s}{1-c}|\Delta|_\infty$. Thus, for all J such that $|J| \le s$, we can bound the coordinate-wise sensitivities as follows:

$$\begin{aligned} \kappa^*_{k,J} &\ge \inf_{\Delta_k=1,\, |\Delta|_1 \le a|\Delta|_\infty} |\Psi_n \Delta|_\infty \\ &\ge \min_{j=1,\dots,K} \left\{ \min_{\Delta_k=1,\ |\Delta|_1 \le a|\Delta_j|} |\Psi_n \Delta|_\infty \right\} \triangleq \kappa^*_k(s), \end{aligned} \tag{4.4}$$

where $a = \frac{2s}{1-c}$. For given s, this bound is data-driven since the minimum in curly brackets can be computed by linear programming (see Section 8.1). Then we can deduce a lower bound on $\kappa_{\infty,J}$ from

$$\kappa_{\infty,J} \ge \min_{k=1,\dots,K} \kappa^*_{k,J}. \tag{4.5}$$

Using (4.3) – (4.5) we get computable lower bounds for all $\kappa_{p,J}$, $p \in [1,\infty]$, which depend only on s and on the data. In particular, for $|J| \le s$,

$$\kappa_{1,J} \ge \frac{1-c}{2s} \min_{k=1,\dots,K} \kappa^*_k(s) \triangleq \kappa_1(s). \tag{4.6}$$

Analogously to (4.4), the sparsity certificate approach yields a bound for block sensitivities:

$$\begin{aligned} \kappa^*_{J_0,J} &\ge \inf_{|\Delta_{J_0}|_1=1,\, |\Delta|_1 \le a|\Delta|_\infty} |\Psi_n \Delta|_\infty \\ &\ge \min_{j=1,\dots,K} \left\{ \min_{|\Delta_{J_0}|_1=1,\ |\Delta|_1 \le a|\Delta_j|} |\Psi_n \Delta|_\infty \right\} \triangleq \kappa^*_{J_0}(s). \end{aligned} \tag{4.7}$$

In Section 8.1 we show that the expression in curly brackets in (4.7) can be computed by solving $2^{|J_0|}$ linear programs. Thus, the values $\kappa^*_{J_0}(s)$ can be readily obtained for sets J_0 of small cardinality.

An alternative to the sparsity certificate approach is to compute $\kappa_{1,J}$ and $\kappa^*_{k,J}$ directly, which is numerically feasible for J of small cardinality. In Section 8.1 we show that obtaining the coordinate-wise sensitivities corresponds to solving $2^{|J|}$ linear programs. Using (4.3) and (4.5) we obtain computable lower bounds for all $\kappa_{p,J}$, $p \in [1,\infty]$. The lower bounds are valid for any given index set J. However, we will need to compute the characteristics for the inaccessible set $J = J(\beta^*)$, where β^* is the true unknown parameter. To circumvent this problem, we can plug in an estimator $\widehat{J}$ of $J(\beta^*)$. For example, we can take $\widehat{J} = J(\widehat{\beta})$. The confidence bounds remain valid whenever $J(\beta^*) \subseteq \widehat{J}$, since then $\kappa_{p,J(\beta^*)} \ge \kappa_{p,\widehat{J}}$ by Proposition 4.1 (i). Theoretical guarantees for the inclusion $J(\beta^*) \subseteq J(\widehat{\beta})$ to hold with probability close to 1 require $|\beta^*_k|$ to be not too small on the support of β^* (see Theorem 5.7 (iv)). On the other hand, one typically observes in simulations that the relevant set $J(\beta^*)$ is either estimated exactly or overestimated by its empirical counterpart $\widehat{J} = J(\widehat{\beta})$, so that the required inclusion is satisfied for such a simple choice of $\widehat{J}$.

We show in Section 9.1 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 and thus allows one to consider the case where several instruments are used for the same endogenous variable.

In the next proposition, we present a simple lower bound on $\kappa_{p,J}$ for general $L \times K$ rectangular matrices Ψ_n. Its proof, as well as other lower bounds on $\kappa_{1,J}$, can be found in Section 9. It is important to note that adding rows to matrix Ψ_n (*i.e.*, adding instruments) increases the sup-norm $|\Psi_n \Delta|_\infty$, and thus potentially increases the sensitivities $\kappa_{p,J(\beta^*)}$. This has a positive effect since the inverse of the sensitivities drive the width of the confidence set for β^*, see Theorem 5.2. Thus, adding instruments potentially improves the confidence set, which is quite intuitive. On the other hand, the price for adding instruments in terms of the rate of convergence is only logarithmic in the number of instruments, as we will see it in the next section.

Proposition 4.2. *Fix $J \subseteq \{1, \ldots, K\}$. Assume that there exist $\eta_1 > 0$ and $0 < \eta_2 < 1$ such that*

$$\forall k \in J,\ \exists l(k) : \begin{cases} |(\Psi_n)_{l(k)k}| \geq \frac{\eta_1}{1-c}\,, \\ \frac{\max_{k' \neq k} |(\Psi_n)_{l(k)k'}|}{|(\Psi_n)_{l(k)k}|} \leq \frac{(1-\eta_2)(1-c)}{2|J|}\,. \end{cases} \tag{4.8}$$

Then

$$\kappa_{p,J} \geq (2|J|)^{-1/p}(1-c)^{-1+1/p}\eta_1\eta_2.$$

The proof of Proposition 4.2 is given in Section 9.3.

Assumption (4.8) is similar in spirit to the coherence condition introduced by Donoho, Elad and Temlyakov (2006) for symmetric matrices, but it is more general because it deals with rectangular matrices. Since the regressors and instruments are random, the values η_1 and η_2 can, in general, be random. Remarkably, for estimation of the coefficients of the endogenous variables, it suffices to have a "good" row of matrix Ψ_n. This means that it is enough to have, among all instruments, one good instrument. The way the instruments are ordered is not important. Good instruments correspond to the rows $l(k)$, for which the value $|(\Psi_n)_{l(k)k}|$ measuring the relevance of the instrument for the kth variable is high. On the other hand, the value $\max_{k' \neq k} |(\Psi_n)_{l(k)k'}|$ accounting for the relation between the instrument and the other variables should be small. An instrument which is well "correlated" with two variables of the model is not satisfactory for this assumption.

5. Main results

We start with introducing some assumptions.

Assumption 5.1. *There exists $\delta > 0$ such that, for all $i = 1, \dots, n$, $l = 1, \dots, L$, the following conditions hold:*

$$\mathbb{E}[|z_{li}u_i|^{2+\delta}] < \infty, \qquad \mathbb{E}[z_{li}u_i] = 0,$$

and neither of $z_{li}u_i$ is almost surely equal to 0.

Define

$$d_{n,\delta} \triangleq \min_{l=1,\dots,L} \frac{\sqrt{\sum_{i=1}^n \mathbb{E}[z_{li}^2 u_i^2]}}{\left(\sum_{i=1}^n \mathbb{E}[|z_{li}u_i|^{2+\delta}]\right)^{1/(2+\delta)}}.$$

If, for any fixed $l \in \{1, \dots, L\}$, the variables $z_{li}u_i$ are i.i.d., then $d_{n,\delta} = n^{\frac{\delta}{4+2\delta}} \min\limits_{l=1,\dots,L} \dfrac{(\mathbb{E}[z_{l1}^2 u_1^2])^{1/2}}{(\mathbb{E}[|z_{l1}u_1|^{2+\delta}])^{1/(2+\delta)}}$.

For $A \geq 1$ set

$$\alpha = 2L\left\{1 - \Phi\big(A\sqrt{2\log(L)}\,\big)\right\} + 2A_0 \frac{\big(1 + A\sqrt{2\log(L)}\,\big)^{1+\delta}}{L^{A^2-1} d_{n,\delta}^{2+\delta}}\,, \tag{5.1}$$

where $A_0 > 0$ is the absolute constant from Theorem 9.4, and $\Phi(\cdot)$ is the standard normal c.d.f.

Theorem 5.2. *Let Assumption 5.1 hold. For $A \geq 1$, define α by (5.1), and set*

$$r = A\sqrt{\frac{2\log(L)}{n}}\,.$$

Assume that $L \leq \exp(d_{n,\delta}^2/(2A^2))$. Then, with probability at least $1 - \alpha$ for any solution $(\widehat{\beta}, \widehat{\sigma})$ of the minimization problem (3.5) we have

$$\left|\mathbf{D_X}^{-1}(\widehat{\beta} - \beta^*)\right|_p \leq \frac{2\widehat{\sigma} r}{\kappa_{p,J(\beta^*)}} \left(1 - \frac{r}{\kappa^*_{J_{\rm end},J(\beta^*)}} - \frac{r^2}{\kappa^*_{J^c_{\rm end},J(\beta^*)}}\right)_+^{-1}, \quad \forall\, p \in [1, \infty], \tag{5.2}$$

and, for all $k = 1, \dots, K$,

$$|\widehat{\beta}_k - \beta^*_k| \leq \frac{2\widehat{\sigma} r}{x_{k*}\ \kappa^*_{k,J(\beta^*)}} \left(1 - \frac{r}{\kappa^*_{J_{\rm end},J(\beta^*)}} - \frac{r^2}{\kappa^*_{J^c_{\rm end},J(\beta^*)}}\right)_+^{-1}. \tag{5.3}$$

Furthermore,

$$\widehat{\sigma} \leq \sqrt{\widehat{Q}(\beta^*)} \left(1 + \frac{r}{c\kappa^*_{J(\beta^*),J(\beta^*)}}\right) \left(1 - \frac{r}{c\kappa^*_{J(\beta^*),J(\beta^*)}}\right)_+^{-1}. \tag{5.4}$$

The proof of Theorem 5.2 is given in Section 9.3.

By convention, $\kappa^*_{\varnothing,J(\beta^*)} = \infty$, so when either J_{end} or J^c_{end} is empty, the term with the corresponding sensitivity disappears from the right hand-side of (5.2) and (5.3). If $J_{\text{end}} = \varnothing$, $L = K$ and $\mathbf{X} = \mathbf{Z}$, the *STIV* estimator yields a pivotal extension of the Dantzig selector of Candès and Tao (2007), in the sense that it allows for the unknown distribution of errors. For this model, which is in the focus of the literature on high-dimensional regression in the recent years, we provide a considerable improvement, since Candès and Tao (2007), Bickel, Ritov and Tsybakov (2009) and the subsequent papers (cf. Bühlmann and van de Geer (2011)) treat the case of i.i.d. errors, which are either Gaussian with known variance or have bounded exponential moment with known parameter. We also improve the Dantzig selector in other aspects by allowing for endogenous regressors, by using weaker sensitivity assumptions than in the previous work, and by providing finite sample confidence intervals.

The bounds (5.2) and (5.3) are meaningful if r is small, *i.e.*, $n \gg \log(L)$. Then under the appropriate conditions on the sensitivities (cf. Remark 5.3), the factor $\tau_1 \triangleq 1 - \frac{r}{\kappa^*_{J_{\text{end}},J(\beta^*)}} - \frac{r^2}{\kappa^*_{J^c_{\text{end}},J(\beta^*)}}$ is close to 1 and the bound on the estimation error in (5.3) 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, which is a modest price for using a large number L of instruments.

Remark 5.3. *Simple sufficient conditions for τ_1 to be close to 1 can be derived from Propositions 4.1 and 4.2. By Proposition 4.1 (ii), we have $\kappa^*_{J_{\text{end}},J(\beta^*)} \geq \kappa_{1,J(\beta^*)}$ and $\kappa^*_{J^c_{\text{end}},J(\beta^*)} \geq \kappa_{1,J(\beta^*)}$. Thus, neglecting the $O(r^2)$ term, we get that τ_1 can be approximately replaced by $1-\frac{r}{\kappa_{1,J(\beta^*)}}$ in (5.2) and (5.3). Therefore, under the premise of Proposition 4.2, for $\tau_1 \approx 1$ it is sufficient to have $|J(\beta^*)| \leq Cr^{-1} = O(\sqrt{n/\log(L)})$ where $C > 0$ is a proper constant. This is quite a reasonable condition on the sparsity $|J(\beta^*)|$ of the true vector β^*. Moreover we get even better conditions if the set of endogenous regressors J_{end} is small. Then the sensitivity $\kappa^*_{J_{\text{end}},J(\beta^*)}$ is large, whereas the small sensitivity of its complement $\kappa^*_{J^c_{\text{end}},J(\beta^*)}$ is compensated by the small value r^2 in the numerator. In the extreme case $J_{\text{end}} = \varnothing$ we have $\frac{r}{\kappa^*_{J_{\text{end}},J(\beta^*)}} = 0$, so that $\tau_1 \leq 1 - \frac{r^2}{\kappa_{1,J(\beta^*)}}$, and it is sufficient to have $|J(\beta^*)| \leq Cr^{-2} = O(n/\log(L))$.*

The assumption $L \leq \exp(d^2_{n,\delta}/(2A^2))$ in Theorem 5.2 is relatively mild. Indeed, in the i.i.d. case, it is equivalent to the condition that $L \leq \exp(Cn^{\delta/(2+\delta)})$ for some $C > 0$.

The value $d_{n,\delta}$ depends on the distribution of the errors and in practice it is unknown. However, in the high-dimensional setting when L is large, the term involving $d_{n,\delta}$ in (5.1) is negligible for reasonable values of A (say, $A \geq 2$) and for moderate sample size n. Thus, in practice, we can drop

this term, and choose A large enough to have

$$2L\left\{1-\Phi\big(A\sqrt{2\log(L)}\,\big)\right\}=\alpha, \tag{5.5}$$

where α is a suitable confidence level. This yields

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

Theorem 5.2 holds for arbitrary tuning constant $0<c<1$. This constant appears in the definition of the *STIV* estimator. Choosing a small c increases the sensitivities in the denominators of the bounds in Theorem 5.2 since the cone of dominant coordinates shrinks as c decreases. On the other hand, this yields less penalization for large values of σ and results in higher $\widehat{\sigma}$.

The proof of Theorem 5.2 relies on a bound for moderate deviations for self-normalized sums of random variables, cf. Jing, Shao and Wang (2003). This is a useful tool that was first applied in the context of high-dimensional estimation by Belloni, Chen, Chernozhukov et al. (2010). There, "asymptotically valid penalty loadings" are required but we do not need such an assumption.

The only unknown ingredient of the inequalities (5.2) and (5.3) is the set $J(\beta^*)$ that determines the sensitivities. To turn these inequalities into valid confidence bounds, it suffices to provide data-driven lower estimates on the sensitivities. As discussed in Section 4, there are two ways to do it. The first one is based on the sparsity certificate, *i.e.*, assuming some known upper bound s on $|J(\beta^*)|$; then we get bounds depending only on s and on the data. The second way is to plug in, instead of $J(\beta^*)$, some data-driven upper estimate $\widehat{J}$, *i.e.*, a set satisfying $J(\beta^*)\subseteq\widehat{J}$ with probability close to 1. The next theorem (Theorem 5.7) provides examples of such estimators $\widehat{J}$. In particular, assertion (iv) of Theorem 5.7 guarantees that, under some assumptions, the estimator $\widehat{J}=J(\widehat{\beta})$ has the required property. Moreover, Theorem 5.7 establishes upper bounds on the rate of convergence of $\widehat{\beta}$ in terms of population characteristics. To state the theorem, we need the following additional assumptions. The first one introduces the population "noise level" σ_*.

Assumption 5.4. *There exist constants $\sigma_*>0$ and $0<\gamma_1<1$ such that*

$$\mathbb{P}\left(\mathbb{E}_n[U^2]\le\sigma_*^2\right)\ge 1-\gamma_1.$$

The second assumption concerns the population counterparts of the sensitivities. It is stated in terms of subsets J_0 of $\{1,\dots,K\}$ and constants $p\ge 1$, $k\in\{1,\dots,K\}$ that can differ from case to case and will be specified later.

Assumption 5.5. *There exist constants $c_p > 0$, $c^*_{J_0} > 0$, and $0 < \gamma_2 < 1$ such that, with probability at least $1 - \gamma_2$,*

$$\kappa_{p,J(\beta^*)} \geq c_p |J(\beta^*)|^{-1/p}, \tag{5.6}$$

$$\kappa^*_{J_0,J(\beta^*)} \geq c^*_{J_0}. \tag{5.7}$$

If $J_0 = \{k\}$ is a singleton we write for brevity $c^*_{J_0} = c^*_k$.

The dependence on $|J(\beta^*)|$ of the right hand side of (5.6) is motivated by Proposition 4.2. In (5.7), we do not indicate the dependence of the bounds on $|J(\beta^*)|$ explicitly because it can be different for different sets J_0. For general J_0, combining Proposition 4.1 (ii) and Proposition 4.2 suggests that the value $c^*_{J_0}$ can be bounded from below by a quantity of the order $|J(\beta^*)|^{-1}$. Note, however, that this is a coarse bound valid for any set J_0.

The last assumption defines a population counterpart of x_{k*}.

Assumption 5.6. *There exist constants $v_k > 0$ and $0 < \gamma_3 < 1$ such that*

$$\mathbb{P}\left(x_{k*} \geq v_k,\ \forall\ k \in J(\beta^*)\right) \geq 1 - \gamma_3.$$

We set $\gamma = \alpha + \sum_{j=1}^{3} \gamma_j$, and

$$\tau^* \triangleq \left(1 + \frac{r}{cc^*_{J(\beta^*)}}\right)\left(1 - \frac{r}{cc^*_{J(\beta^*)}}\right)_+^{-1}\left(1 - \frac{r}{c^*_{J_{\text{end}}}} - \frac{r^2}{c^*_{J^c_{\text{end}}}}\right)_+^{-1}.$$

Theorem 5.7. *Under the assumptions of Theorem 5.2 and Assumption 5.4, the following holds.*

(i) *Let part (5.7) of Assumption 5.5 with $J_0 = J(\beta^*)$ be satisfied. Then, with probability at least $1 - \alpha - \gamma_1 - \gamma_2$ for any solution $\widehat{\sigma}$ of (3.5) we have*

$$\widehat{\sigma} \leq \sigma_* \left(1 + \frac{r}{cc^*_{J(\beta^*)}}\right)\left(1 - \frac{r}{cc^*_{J(\beta^*)}}\right)_+^{-1}.$$

(ii) *Fix $p \in [1, \infty]$. Let Assumption 5.5 be satisfied, where (5.7) holds simultaneously for $J_0 = J(\beta^*), J_0 = J_{\text{end}}$ and $J_0 = J^c_{\text{end}}$. Then, with probability at least $1-\alpha-\gamma_1-\gamma_2$, for any solution $\widehat{\beta}$ of (3.5) we have*

$$\left|\mathbf{D_X}^{-1}\left(\widehat{\beta} - \beta^*\right)\right|_p \leq \frac{2\sigma_* r |J(\beta^*)|^{1/p} \tau^*}{c_p}. \tag{5.8}$$

(iii) *Let Assumptions 5.5 and 5.6 be satisfied, where (5.7) holds simultaneously for $J_0 = \{k\}$, $\forall k$, and $J_0 = J(\beta^*), J_0 = J_{\text{end}}, J_0 = J^c_{\text{end}}$. Then with probability at least $1 - \gamma$, for any solution $\widehat{\beta}$*

of (3.5) we have

$$|\widehat{\beta}_k - \beta_k^*| \le \frac{2\sigma_* r\tau^*}{c_k^* v_k}, \qquad k = 1, \dots, K. \tag{5.9}$$

(iv) *Let the assumptions of* (iii) *hold, and* $|\beta_k^*| > \frac{2\sigma_* r\tau^*}{c_k^* v_k}$ *for all* $k \in J(\beta^*)$*. Then, with probability at least* $1-\gamma$*, for any solution* $\widehat{\beta}$ *of (3.5) we have:*

$$J(\beta^*) \subseteq J(\widehat{\beta}).$$

The proof of Theorem 5.7 is given in Section 9.3.

For reasonably large sample size ($n \gg \log(L)$), the value r is small, and τ^* is a constant approaching 1 as $r \to 0$. Thus, the bounds (5.8) and (5.9) 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 and without endogenous variables Candès and Tao (2007), Bickel, Ritov and Tsybakov (2009), Lounici (2008) (see also Bühlmann and van de Geer (2011) for references to more recent work).

From (5.3) and Theorem 5.7 (iv), we obtain the following confidence intervals of level $1-\gamma$ for β_k^*:

$$|\widehat{\beta}_k - \beta_k^*| \le \frac{2\widehat{\sigma} r}{x_{k*}\ \kappa^*_{k,J(\widehat{\beta})}} \left(1 - \frac{r}{\kappa^*_{J_{\text{end}},J(\widehat{\beta})}} - \frac{r^2}{\kappa^*_{J_{\text{end}}^c,J(\widehat{\beta})}}\right)_+^{-1}. \tag{5.10}$$

Theorem 5.7 (iv) provides an upper estimate on the set of non-zero components of β^*. We now consider the problem of the exact selection of variables. For this purpose, we use the thresholded *STIV* estimator whose coordinates are defined by

$$\widetilde{\beta}_k(\omega_k) \triangleq \begin{cases} \widehat{\beta}_k & \text{if } |\widehat{\beta}_k| > \omega_k, \\ 0 & \text{otherwise,} \end{cases} \tag{5.11}$$

where $\widehat{\beta}_k$ are the coordinates of the *STIV* estimator $\widehat{\beta}$, and $\omega_k > 0$, $k = 1, \dots, K$, are thresholds that will be specified below. We will use the sparsity certificate approach, so that the thresholds will depend on the upper bound s on the number of non-zero components of β^*. We will need the following modification of Assumption 5.5.

Assumption 5.8. *Fix an integer* $s \ge 1$*. There exist constants* $c^*_{J(\beta^*)} > 0$, $c^*_{J_0}(s) > 0$, *and* $0 < \gamma_2 < 1$ *such that, with probability at least* $1-\gamma_2$,

$$\kappa^*_{J(\beta^*),J(\beta^*)} \ \ge \ c^*_{J(\beta^*)} \quad \text{and} \quad \kappa^*_{J_0}(s) \ge c^*_{J_0}(s) \tag{5.12}$$

for $J_0 = \{k\}$, $\forall k$, *and* $J_0 = J_{\rm end}, J_0 = J_{\rm end}^c$. *If* $J_0 = \{k\}$ *is a singleton we write for brevity* $c^*_{J_0}(s) = c^*_k(s)$.

Set

$$\tau^*(s) \triangleq \left(1 + \frac{r}{cc^*_{J(\beta^*)}}\right)\left(1 - \frac{r}{cc^*_{J(\beta^*)}}\right)_+^{-1}\left(1 - \frac{r}{c^*_{J_{\rm end}}(s)} - \frac{r^2}{c^*_{J^c_{\rm end}}(s)}\right)_+^{-1}.$$

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

Theorem 5.9. *Let the assumptions of Theorem 5.2 and Assumptions 5.4, 5.6, 5.8 be satisfied. Assume that* $|J(\beta^*)| \le s$, *and* $|\beta^*_k| > \frac{4\sigma_* r\tau^*(s)}{c^*_k(s)v_k}$ *for all* $k \in J(\beta^*)$. *Take the thresholds*

$$\omega_k(s) \triangleq \frac{2\widehat{\sigma}r}{\kappa^*_k(s)x_{k*}}\left(1 - \frac{r}{\kappa^*_{J_{\rm end}}(s)} - \frac{r^2}{\kappa^*_{J^c_{\rm end}}(s)}\right)_+^{-1},$$

and consider the estimator $\widetilde{\beta}$ *with coordinates* $\widetilde{\beta}_k(\omega_k(s)), k = 1, \dots, K$. *Then, with probability at least* $1 - \gamma$, *we have*

$$\overrightarrow{\rm sign}(\widetilde{\beta}) = \overrightarrow{\rm sign}(\beta^*). \tag{5.13}$$

As a consequence, $J(\widetilde{\beta}) = J(\beta^*)$.

The proof of Theorem 5.9 is given in Section 9.3.

Remark 5.10. *Inspection of the proof of Theorem 5.9 shows that the same conclusion as in Theorem 5.9 holds with other definitions of the thresholds. Indeed,* $\kappa^*_k(s), \kappa^*_{J_{\rm end}}(s)$, *and* $\kappa^*_{J^c_{\rm end}}(s)$ *in the definition of* $\omega_k(s)$ *are lower bounds for the sensitivities* $\kappa^*_{k,J(\beta^*)}, \kappa^*_{J_{\rm end},J(\beta^*)}$, *and* $\kappa^*_{J^c_{\rm end},J(\beta^*)}$. *They can be replaced by other* s*-dependent lower bounds on these sensitivities. Then Theorem 5.9 remains valid, with the modifications only in the value of the lower bound on* $|\beta^*_k|$ *required for* $k \in J(\beta^*)$, *and in a slightly different formulation of Assumption 5.8. For example, if there is only one endogenous variable,* $J_{\rm end} = \{1\}$, *we can take the thresholds*

$$\underline{\omega}_k(s) \triangleq \frac{2\widehat{\sigma}r}{\kappa^*_k(s)x_{k*}}\left(1 - \frac{r}{\kappa^*_1(s)} - \frac{r^2}{\kappa_1(s)}\right)_+^{-1}.$$

We now consider the approximately sparse setting. The sparsity assumption is quite natural in empirical economics since usually only a moderate number of covariates is included in the model. However, one might be also interested in the case when the true vector β^* is only approximately sparse. This means that most of the coefficients β_k^* are not exactly zero but too small to matter, whereas the remaining ones are relatively large. This setting received some attention in the statistical literature. For example, the performance of Dantzig selector and MU-selector under such assumptions is studied by Candès and Tao (2007) and Rosenbaum and Tsybakov (2010) respectively. We will derive a similar result for the *STIV* estimator.

Consider the enlarged cone

$$\widetilde{C}_J \triangleq \left\{ \Delta \in \mathbb{R}^K : \ |\Delta_{J^c}|_1 \leq \frac{2+c}{1-c} |\Delta_J|_1 \right\}$$

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

$$\widetilde{\kappa}_{p,J} \triangleq \inf_{\Delta \in \mathbb{R}^K : \ |\Delta|_p = 1, \ \Delta \in \widetilde{C}_J} |\Psi_n \Delta|_\infty \quad \text{and} \quad \widetilde{\kappa}^*_{J_0,J} \triangleq \inf_{\Delta \in \mathbb{R}^K : \ |\Delta_{J_0}|_1 = 1, \ \Delta \in \widetilde{C}_J} |\Psi_n \Delta|_\infty .$$

The following theorem is an analog of the above results for the approximately sparse case.

Theorem 5.11. *Let A, α, and r satisfy the same conditions as in Theorem 5.2. Assume that $L \leq \exp(d_{n,\delta}^2/(2A^2))$ and fix $p \in [1, \infty]$. Let Assumption 5.1 be satisfied. Then with probability at least $1 - \alpha$ for any solution $\widehat{\beta}$ of (3.5) we have*

(5.14)

$$\left| \mathbf{D_X}^{-1} \left(\widehat{\beta} - \beta^* \right) \right|_p \leq \min_{J \subset \{1, \ldots, K\}} \left\{ \max \left(\frac{2\widehat{\sigma} r}{\widetilde{\kappa}_{p,J}} \left(1 - \frac{r}{\widetilde{\kappa}^*_{J_{\text{end}},J}} - \frac{r^2}{\widetilde{\kappa}^*_{J^c_{\text{end}},J}} \right)_+^{-1}, \ \frac{6 \left| \left(\mathbf{D_X}^{-1} \beta^* \right)_{J^c} \right|_1}{1-c} \right) \right\} .$$

We can interpret Theorem 5.11 as the fact that the *STIV* estimator automatically realizes a "bias/variance" trade-off related to a non-linear approximation. Inequality (5.14) means that this estimator performs as well as if the optimal subset J were known.

6. Models with possibly non-valid instruments

In this section, we propose a modification of the *STIV* estimator for the model with possibly non-valid instruments. The main purpose of the suggested method is to construct confidence intervals for non-validity indicators, and to detect non-valid instruments. This question has been addressed in the non high-dimensional case, for example, in Andrews (1999) and Hahn and Hausman (2002) among others; one of the most recent papers is Liao (2010) where one can find more references. The model

can be written in the form:

$$y_i = x_i^T \beta^* + u_i, \tag{6.1}$$

$$\mathbb{E}\,[z_i u_i] = 0, \tag{6.2}$$

$$\mathbb{E}\,[\overline{z}_i u_i] = \theta^*, \tag{6.3}$$

where x_i, z_i, and $\overline{z}_i$ are vectors of dimensions K, L and L_1, respectively. The instruments are decomposed in two parts, z_i and $\overline{z}_i$, where $\overline{z}_i^T = (\overline{z}_{1i}, \dots, \overline{z}_{L_1 i})$ is a vector of possibly non-valid instruments. A component of the unknown vector $\theta^* \in \mathbb{R}^{L_1}$ is equal to 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 L_1 that can be much larger than the sample size.

As above, we assume independence and allow for heteroscedasticity. The difference from the previous sections is only in introducing equation (6.3). In addition to x_i, y_i, z_i, we observe the realizations of mutually independent random vectors $\overline{z}_i$, $i = 1, \dots, n$, with components $\overline{z}_{li}$ satisfying $\mathbb{E}[\overline{z}_{li} u_i] = \theta_l^*$ for all $l = 1, \dots, L_1$, $i = 1, \dots, n$. We denote by $\overline{\mathbf{Z}}$ the matrix of dimension $n \times L_1$ with rows $\overline{z}_i^T$, $i = 1, \dots, n$. Set

$$\overline{z}_* = \max_{l=1,\dots,L_1} \left(\frac{1}{n} \sum_{i=1}^n \overline{z}_{li}^2 \right)^{1/2}.$$

In this section, we assume that we have a pilot estimator $\widehat{\beta}$ and a statistic $\widehat{b}$ such that, with probability close to 1,

$$\left| \mathbf{D_X}^{-1} (\widehat{\beta} - \beta^*) \right|_1 \le \widehat{b}. \tag{6.4}$$

For example, $\widehat{\beta}$ can be the *STIV* estimator based only on the vectors of valid instruments $z_1, \dots, z_n$. In this case, an explicit expression for $\widehat{b}$ can be obtained from (5.2) by replacing there $J(\beta^*)$ by a suitable estimator or upper estimator $\widehat{J}$ (see Theorem 5.7 (iv) and Theorem 5.9).

We define the *STIV-NV* estimator $(\widehat{\theta}, \widehat{\sigma}_1)$ as any solution of the problem

$$\min_{(\theta,\sigma_1) \in \widehat{\mathcal{I}}_1} \big(|\theta|_1 + c\sigma_1 \big), \tag{6.5}$$

where $0 < c < 1$,

$$\widehat{\mathcal{I}}_1 \triangleq \left\{ (\theta, \sigma_1) :\ \theta \in \mathbb{R}^{L_1},\ \sigma_1 > 0,\ \left| \frac{1}{n} \overline{\mathbf{Z}}^T (\mathbf{Y} - \mathbf{X}\widehat{\beta}) - \theta \right|_\infty \le \sigma_1 r_1 + \widehat{b}\overline{z}_*,\ F(\theta, \widehat{\beta}) \le \sigma_1 + \widehat{b}\overline{z}_* \right\}$$

for some $r_1 > 0$ (to be specified below), where for all $\theta = (\theta_1, \dots, \theta_{L_1}) \in \mathbb{R}^{L_1}$, $\beta \in \mathbb{R}^K$,

$$F(\theta, \beta) \triangleq \max_{l=1,\dots,L_1} \sqrt{\widehat{Q}_l(\theta_l, \beta)}$$

with

$$\widehat{Q}_l(\theta_l, \beta) \triangleq \frac{1}{n} \sum_{i=1}^{n} \left(\overline{z}_{li}(y_i - x_i^T \beta) - \theta_l\right)^2 .$$

It is not hard to see that the optimization problem (6.5) can be re-written as a linear program.

Assumption 6.1. *There exists $\delta > 0$ such that, for all $i = 1, \dots, n$, $l = 1, \dots, L_1$, the following conditions hold:*

$$\mathbb{E}[|\overline{z}_{li} u_i|^{2+\delta}] < \infty, \qquad \mathbb{E}[\overline{z}_{li} u_i] = \theta_l^*,$$

where θ_l^* is the lth component of θ^* and neither of $\overline{z}_{li} u_i - \theta_l^*$ is almost surely equal to 0.

Define

$$d_{n,\delta,1} \triangleq \min_{l=1,\dots,L_1} \frac{\sqrt{\sum_{i=1}^n \mathbb{E}[|\overline{z}_{li} u_i - \theta_l^*|^2]}}{\left(\sum_{i=1}^n \mathbb{E}[|\overline{z}_{li} u_i - \theta_l^*|^{2+\delta}]\right)^{1/(2+\delta)}}.$$

For $A \geq 1$ set

$$\alpha_1 = 2L_1 \left\{1 - \Phi\big(A\sqrt{2\log(L_1)}\big)\right\} + 2A_0 \frac{\big(1 + A\sqrt{2\log(L_1)}\big)^{1+\delta}}{L_1^{A^2-1} d_{n,\delta,1}^{2+\delta}}, \tag{6.6}$$

where $A_0 > 0$ is the absolute constant from Theorem 9.4, and $\Phi(\cdot)$ is the standard normal c.d.f.

The following theorem provides a basis for constructing confidence intervals for the non-validity indicators.

Theorem 6.2. *Let Assumption 6.1 hold. For $A \geq 1$, define α_1 by (6.6), and set*

$$r_1 = A\sqrt{\frac{2\log(L_1)}{n}}.$$

Assume that $L_1 \leq \exp(d_{n,\delta,1}^2/(2A^2))$, and that $\widehat{\beta}$ is an estimator satisfying (6.4) with probability at least $1 - \alpha_2$ for some $0 < \alpha_2 < 1$. Then, with probability at least $1 - \alpha_1 - \alpha_2$ for any solution $(\widehat{\theta}, \widehat{\sigma}_1)$ of the minimization problem (6.5) we have

$$|\widehat{\theta} - \theta^*|_\infty \leq \frac{2\Big[\widehat{\sigma}_1 r_1 + (1 + r_1(1-c)^{-1})\widehat{b}\overline{z}_*\Big]}{(1 - 2r_1(1-c)^{-1}|J(\theta^*)|)_+} \triangleq V(\widehat{\sigma}_1, \widehat{b}, |J(\theta^*)|), \tag{6.7}$$

and

$$|\widehat{\theta} - \theta^*|_1 \leq \frac{2\Big[2|J(\theta^*)|\Big(\widehat{\sigma}_1 r_1 + (1 + r_1)\widehat{b}\overline{z}_*\Big) + c\widehat{b}\overline{z}_*]}{(1 - c - 2r_1|J(\theta^*)|)_+}. \tag{6.8}$$

This theorem should be naturally applied when r_1 is small, *i.e.*, $n \gg \log(L_1)$. In addition, we need a small $\widehat{b}$, which is guaranteed by the results of Section 5 under the condition $n \gg \log(L)$ if the pilot estimator $\widehat{\beta}$ is the *STIV* estimator. Note also that the bounds (6.7) and (6.8) are meaningful if their denominators are positive, which is roughly equivalent to the following bound on the sparsity of θ^*: $|J(\theta^*)| = O(1/r_1) = O(\sqrt{n/\log(L_1)})$.

Bounds for all the norms $|\widehat{\theta} - \theta^*|_p$, $\forall\ 1 < p < \infty$, follow immediately from (6.7) and (6.8) by the standard interpolation argument. We note that, in Theorem 6.2, $\widehat{\beta}$ can be any estimator satisfying (6.4), not necessarily the *STIV* estimator.

To turn (6.7) and (6.8) into valid confidence bounds, we can replace there $|J(\theta^*)|$ by $|J(\widehat{\theta})|$, as follows from Theorem 6.4 (ii) below. In addition, Theorem 6.4 establishes the rate of convergence of the *STIV-NV* estimator and justifies the selection of non-valid instruments by thresholding. To state the theorem, we will need an extra assumption that the random variable $F(\theta^*, \beta^*)$ is bounded in probability by a constant $\sigma_{1*} > 0$:

Assumption 6.3. *There exist constants $\sigma_{1*} > 0$ and $0 < \varepsilon < 1$ such that, with probability at least $1 - \varepsilon$,*

$$\max_{l=1,\dots,L_1} \frac{1}{n} \sum_{i=1}^{n} (\overline{z}_{li} u_i - \theta_l^*)^2 \le \sigma_{1*}^2. \tag{6.9}$$

As in (5.11) we define a thresholded estimator

$$\widetilde{\theta}_l \triangleq \begin{cases} \widehat{\theta}_l & \text{if } |\widehat{\theta}_l| > \omega, \\ 0 & \text{otherwise,} \end{cases} \tag{6.10}$$

where $\omega > 0$ is some threshold. For $b_* > 0$, $s_1 > 0$, define

$$\overline{\sigma}_* = \left(1 - \frac{4 r_1 s_1}{c(1 - c - 2 r_1 s_1)_+}\right)_+^{-1} \left[\sigma_{1*} + \frac{2 b_* \overline{z}_* (1 + 2(1 + r_1) s_1 / c)}{(1 - c - 2 r_1 s_1)_+}\right].$$

Theorem 6.4. *Let the assumptions of Theorem 6.2 and Assumption 6.3 be satisfied. Then the following holds.*

(i) *Let $\widehat{\beta}$ be an estimator satisfying*

$$\left| \mathbf{D_X}^{-1} (\widehat{\beta} - \beta^*) \right|_1 \le b_* \tag{6.11}$$

with probability at least $1 - \alpha_2$ for some $0 < \alpha_2 < 1$ and some constant b_*. Assume that $|J(\theta^*)| \le s_1$. Then, with probability at least $1 - \alpha_1 - \alpha_2 - \varepsilon$, for any solution $\widehat{\theta}$ of the

minimization problem (6.5) we have

$$|\widehat{\theta} - \theta^*|_\infty \le V(\overline{\sigma}_*, b_*, s_1). \tag{6.12}$$

(ii) *Let $(\widehat{\beta}, \widehat{\sigma})$ be the STIV estimator, and let the assumptions of all the items of Theorem 5.7 be satisfied (with $p = 1$ in item* (ii)*). Assume that $|J(\theta^*)| \le s_1$, $|J(\beta^*)| \le s$, and $|\theta_l^*| > V(\overline{\sigma}_*, b_*, s_1)$ for all $l \in J(\theta^*)$, where*

$$b_* = \frac{2\sigma_* r s \tau^*(s)}{c_1}\,. \tag{6.13}$$

Then, with probability at least $1 - \alpha_1 - \varepsilon - \gamma$, for any solution $\widehat{\theta}$ of the minimization problem (6.5) we have

$$J(\theta^*) \subseteq J(\widehat{\theta}). \tag{6.14}$$

(iii) *Let the assumptions of item* (ii) *and Assumption 5.8 hold. Assume that $|\theta_l^*| > 2V(\overline{\sigma}_*, b_*, s_1)$ for all $l \in J(\theta^*)$. Let $\tilde{\theta}$ be the thresholded estimator defined in (6.10) where $\widehat{\theta}$ is any solution of the minimization problem (6.5), and the threshold is defined by $\omega = V(\widehat{\sigma}_1, \widehat{b}, s_1)$ with*

$$\widehat{b} = \frac{2\widehat{\sigma} r s}{\kappa_1(s)} \left(1 - \frac{r}{\kappa^*_{J_{\mathrm{end}}}(s)} - \frac{r^2}{\kappa^*_{J^{\mathrm{c}}_{\mathrm{end}}}(s)}\right)_+^{-1}.$$

Then, with probability at least $1 - \alpha_1 - \varepsilon - \gamma$, we have

$$\overrightarrow{\mathrm{sign}(\widetilde{\theta})} = \overrightarrow{\mathrm{sign}(\theta^*)}. \tag{6.15}$$

As a consequence, $J(\widetilde{\theta}) = J(\theta^*)$.

In practice, the parameter s may not be known and it can be replaced by $|J(\widehat{\theta})|$; this is a reasonable upper bound on $|J(\theta^*)|$ as suggested by Theorem 6.4 (ii). It is interesting to analyze the dependence of the rate of convergence in (6.12) on r, r_1, s, and s_1. As discussed above, a meaningful framework is to consider small r,r_1 and the sparsities s,s_1 such that rs, $r_1 s_1$ are comfortably smaller than 1. In this case, the value b_* given in (6.15) is of the order $O(rs)$ and the rate of convergence in (6.12) is of the order $O(r_1) + O(rs)$. We see that the rate does not depend on the sparsity s_1 of θ^* but it does depend on the sparsity s of β^*. It is interesting to explore whether this rate is optimal, *i.e.*, whether it can be improved by estimators different from the *STIV-NV* estimator.

7. Complements

7.1. **Non-pivotal *STIV* estimator.** We consider first a simpler version of the *STIV* estimator which is not pivotal in the sense that it depends on the upper bound σ_* on the "noise level" appearing in Assumption 5.4. The estimator that we consider here is a solution $\widehat{\beta}$ of the following minimization problem:

$$\min_{\beta \in \widehat{\mathcal{I}}_{np}} |\mathbf{D}_{\mathbf{X}}^{-1}\beta|_1, \tag{7.1}$$

where

$$\widehat{\mathcal{I}}_{np} \triangleq \left\{ \beta \in \mathbb{R}^K : \left| \frac{1}{n} \mathbf{D}_{\mathbf{Z}} \mathbf{Z}^T (\mathbf{Y} - \mathbf{X}\beta) \right|_\infty \leq \sigma_* r \right\}.$$

It is not hard to see that (7.1) can be written as a linear program. We have the following bounds on the ℓ_p-errors of this estimator.

Theorem 7.1. *Let Assumptions 5.1 and 5.4 hold. For $A \geq 1$, define α by (5.1), and set*

$$r = A\sqrt{\frac{2\log(L)}{n}}.$$

Assume that $L \leq \exp(d_{n,\delta}^2/(2A^2))$. Then, with probability at least $1 - \alpha - \gamma_1$ for any solution $\widehat{\beta}$ of the minimization problem (7.1) we have

$$\left| \mathbf{D}_{\mathbf{X}}{}^{-1}(\widehat{\beta} - \beta^*) \right|_p \leq \frac{2\sigma_* r}{\kappa_{p,J(\beta^*)}}, \quad \forall\ p \in [1, \infty], \tag{7.2}$$

and, for all $k = 1, \ldots, K$,

$$|\widehat{\beta}_k - \beta_k^*| \leq \frac{2\sigma_* r}{x_{k*}\kappa_{k,J(\beta^*)}^*}. \tag{7.3}$$

Here the sensitivities $\kappa_{p,J(\beta^*)}$ and $\kappa_{k,J(\beta^*)}^*$ are defined on the cone C_J with $c = 0$.

The proof of this result is easily obtained by simplifying the proof of Theorem 5.2.

7.2. ***STIV* estimator with linear projection instruments.** The results of the previous sections show that the *STIV* estimator can handle a very large number of instruments, up to an exponential in the sample size. Moreover, adding instruments always improves the sensitivities. In this section, we consider the case where we look for a smaller set of instruments, namely, of size K. A classical solution with one endogenous regressor in low dimension is the two-stage least squares estimator (see, *e.g.*, Wooldridge (2002)). Under the stronger zero conditional mean assumption, the solution in low dimensions is given by the optimal instruments (see Amemiya (1974), Chamberlain (1987), and Newey (1990)). In the homoscedastic case, it corresponds to the projection of the endogenous variables on

the space of variables measurable with respect to all the instruments. These optimal instruments are expressed in terms of conditional expectations that are not available in practice and should be estimated. When K is large, we are typically facing the curse of dimensionality and extremely large samples would be needed to obtain precise estimates of these ideal instruments. In this setting, Belloni, Chen, Chernozhukov et al. (2010a) propose to use the Lasso. They impose an approximate sparsity assumption, different from the one we considered in this paper, which is quite strong and guarantees that the optimal instruments can be estimated at the root-n (parametric) rate. Then they consider the classical heteroscedastic robust *IV* estimator with these instruments.

We propose to proceed in a different way. As discussed after Proposition 4.2, we can expect to get higher sensitivities and thus to obtain tighter bounds if for each endogenous regressor we use a "good instrument", *i.e.*, the instrument correlated as much as possible with the endogenous variable. Akin to the two-stage least squares, we consider instruments which are the projections of the endogenous variables on the linear span of all the instruments, and do not make the stronger zero conditional mean assumption. Note that, for every $k = 1, \dots, k_{\text{end}}$, we can write the reduced form equations

$$x_{ki} = \sum_{l=1}^{L} z_{li}\zeta_{kl} + v_{ki}, \quad i = 1, \dots, n, \tag{7.4}$$

where ζ_{kl} are unknown coefficients of the linear combination of instruments, and

$$\mathbb{E}[z_{li}v_{ki}] = 0 \tag{7.5}$$

for $i = 1, \dots, n$, $l = 1, \dots, L$. The representation (7.4)–(7.5) holds whenever x_{ki} and z_{li} have finite second moments. We call $\sum_{l=1}^{L} z_{li}\zeta_{kl}$ the linear projection instrument. We now estimate the unknown coefficients ζ_{kl}. If $L \geq K > n$ and if the reduced form model (7.4) has some sparsity, it is natural to use a high-dimensional procedure, such as the Lasso, the Dantzig selector or the Square-root Lasso, to produce estimators $\widehat{\zeta}_{kl}$ of the coefficients. (Since there is no endogeneity in (7.4) we need not apply the *STIV* estimator requiring more computations.) Then we replace the initial L-dimensional vector of instruments by a K-dimensional vector $\widehat{x}_i = (\widehat{x}_{1i}, \dots, \widehat{x}_{Ki})$ whose first k_{end} coordinates are $\sum_{l=1}^{L} \widehat{\zeta}_{kl} z_{li}$ and the remaining coordinates are the exogenous variables. These are estimators of the linear projection instruments that we use on the second stage to estimate β^*. Specifically, on the second stage we apply the *STIV* estimator where we replace the matrices $\mathbf{Z}$, $\mathbf{D}_{\mathbf{Z}}$, and Ψ_n by their estimated counterparts corresponding to new vectors of instruments of size K instead of L (just use $\widehat{x}_i$, instead of z_i). Intuitively this should yield larger sensitivities $\kappa_{p,J}$, $\kappa^*_{k,J}$ and others since the

new instruments are better correlated with the endogenous variables. Also, the $\log(L)$ term in the expression for r and in the rates is reduced to a $\log(K)$ term.

We do not discuss here a theoretical justification of this method. In Section 8 we show that it works successfully in simulations. Note also that a quick proof can be obtained using the sample splitting argument. Indeed, if the linear projection instruments are obtained from the first subsample, whereas the second subsample independent from the first one is used to estimate β^*, then $\widehat{x}_i$ are valid instruments. Therefore, conditioning on the first subsample, we can apply the theory of Section 5. However, in practice, it seems reasonable to use the whole data set on both steps of the two-stage procedure.

Finally, note that another type of two-stage procedures, not motivated by the endogeneity, is discussed in the literature on sparsity in high-dimensional linear models (see, *e.g.*, Candès and Tao (2007) and Belloni and Chernozhukov (2010)). At the first stage, the support of the true vector is estimated with a high-dimensional procedure, such as the Lasso or Dantzig selector, and at the second stage the OLS is used on the estimated support. Belloni and Chernozhukov (2010) study the theoretical properties of such two stage procedures. An analog of this approach for the setting that we consider here would be a two-stage procedure with the *STIV* estimator at the first stage and some classical *IV* estimator (such as the GMM) at the second stage.

8. Practical implementation

8.1. **Computational aspects.** Finding a solution $(\widehat{\beta}, \widehat{\sigma})$ of the minimization problem (3.5) reduces to the following conic program: find $\beta \in \mathbb{R}^K$ and $t > 0$ ($\sigma = t/\sqrt{n}$), which achieve the minimum

$$\min_{(\beta,t,v,w)\in\mathcal{V}} \left(\sum_{k=1}^{K} w_k + c\sqrt{n}t \right) \tag{8.1}$$

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

$$v = \mathbf{Y} - \mathbf{X}\beta, \qquad -rt\mathbf{1} \le \frac{1}{\sqrt{n}}\mathbf{D}_{\mathbf{Z}}\mathbf{Z}^T(\mathbf{Y} - \mathbf{X}\beta) \le rt\mathbf{1},$$
$$-w \le \mathbf{D}_{\mathbf{X}}^{-1}\beta \le w, \qquad w \ge \mathbf{0}, \qquad (t, v) \in C.$$

Here and below $\mathbf{0}$ 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: $C \triangleq \{(t, v) \in \mathbb{R} \times \mathbb{R}^n : \ t \ge |v|_2\}$. Conic programming is a standard tool in optimization and many open source toolboxes are available to implement it (see, *e.g.*, Sturm (1999)).

The expression in curly brackets in the lower bound (4.4) is equal to the value of the following optimization program:

$$\min_{\epsilon=\pm 1} \min_{(w,\Delta,v)\in\mathcal{V}_{k,j}} v \tag{8.2}$$

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:

$$v \geq 0, \qquad -v\mathbf{1} \leq \Psi_n \Delta \leq v\mathbf{1}, \qquad w \geq \mathbf{0}, \qquad -w_{I^c} \leq \Delta_{I^c} \leq w_{I^c} \quad \text{for } I = \{j, k\},$$
$$w_I = \mathbf{0}, \qquad \Delta_k = 1, \qquad \epsilon\Delta_j \geq 0, \qquad \sum_{i=1}^K w_i + 1 \leq \epsilon(a+g)\Delta_j$$

where g is the constant such that

$$g = \begin{cases} 0 & \text{if } k = j \\ -1 & \text{otherwise.} \end{cases}$$

Note that, here, ϵ is the sign of Δ_j, and (8.2) is the minimum of two terms, each of which is the value of a linear program. Analogously, the expression in curly brackets in (4.7) can be computed by solving $2^{|J_0|}$ linear programs. The reduction is done in the same way as in (8.2) with the only difference that instead of ϵ we introduce a vector $(\epsilon_k)_{k\in J_0}$ of signs of the coordinates Δ_k for indices $k \in J_0$.

The coordinate-wise sensitivities

$$\kappa^*_{k,J} = \inf_{\Delta_k=1,\, |\Delta_{J^c}|_1 \leq \frac{1+c}{1-c}|\Delta_J|_1} |\Psi_n \Delta|_\infty$$

can be efficiently computed for given J when the cardinality $|J|$ is small. Indeed, it is enough to find the minimum of the values of $2^{|J|}$ linear programs:

$$\min_{(\epsilon_j)_{j\in J}\in\{-1,1\}^{|J|}} \min_{(w,\Delta,v)\in\mathcal{U}_{k,J}} v \tag{8.3}$$

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

$$v \geq 0, \qquad -v\mathbf{1} \leq \Psi_n \Delta \leq v\mathbf{1}, \qquad w \geq \mathbf{0}, \qquad -w_{I^c} \leq \Delta_{I^c} \leq w_{I^c} \quad \text{for } I = J \cup \{k\},$$
$$w_I = \mathbf{0}, \qquad \Delta_k = 1, \qquad \epsilon_j\Delta_j \geq 0, \quad \text{for } j \in J, \qquad \sum_{i=1}^K w_i \leq \frac{1+c}{1-c}\sum_{j\in J} \epsilon_j\Delta_j + g.$$

Here $(\epsilon_j)_{j\in J}$ is the vector of signs of the coordinates Δ_j with $j \in J$ and g is the constant defined by

$$g = \begin{cases} 0 & \text{if } k \in J, \\ -1 & \text{otherwise.} \end{cases}$$

8.2. **Simulations.** In this section, we consider the performance of the *STIV* estimator on simulated data. The model is as follows:

$$\begin{aligned} y_i &= \sum_{k=1}^{K} x_{ki}\beta_k^* + u_i, \\ x_{1i} &= \sum_{l=1}^{L-K+1} z_{li}\zeta_l + v_i, \\ x_{l'i} &= z_{li} \quad \text{for } l' = l - L + K \ \text{ and } \ l \in \{L-K+2, \dots, L\}, \end{aligned}$$

where $(y_i, x_i^T, z_i^T, u_i, v_i)$ are i.i.d., (u_i, v_i) have the joint normal distribution

$$\mathcal{N}\left(0, \begin{pmatrix} \sigma_{\text{struct}}^2 & \rho\sigma_{\text{struct}}\sigma_{\text{end}} \\ \rho\sigma_{\text{struct}}\sigma_{\text{end}} & \sigma_{\text{end}}^2 \end{pmatrix}\right),$$

z_i^T is a vector of independent standard normal random variables, and z_i^T is independent of (u_i, v_i). Clearly, in this model $\mathbb{E}[z_i u_i] = 0$. We take $n = 49$, $L = 50$, $K = 25$, $\sigma_{\text{struct}} = \sigma_{\text{end}} = \rho = 0.3$, $\beta^* = (1, 1, 1, 1, 1, 0, \dots, 0)^T$ and $\zeta_l = 0.15$ for $l = 1, \dots, L - K + 1$. We have 50 instruments and only 49 observations, so we are in a framework of application of high-dimensional techniques. We set $c = 0.1$ and take A satisfying (5.5) with $\alpha = 0.05$. The three columns on the left of Table 1 present simulation results for the *STIV* estimator. It is straightforward to see that only the first five variables (the true support of β^*) are eligible to be considered as relevant. This set will be denoted by $\widehat{J}$. The second and third columns in Table 1 present the true coordinate-wise sensitivities $\kappa^*_{k,\widehat{J}}$ as well as their lower bounds $\kappa_k^*(5)$ obtained via the sparsity certificate with $s = 5$. These lower bounds are easy to compute, and we see that they yield reasonable approximations from below of the true sensitivities. The estimate $\widehat{\sigma}$ is 0.247 which is quite close to σ_{struct}. Next, based on (5.3), the fact that $J_{\text{end}} = \{1\}$, and the bounds on the sensitivities in Proposition 4.1 and in (4.4) – (4.7), we have the following formulas for the confidence intervals

$$|\widehat{\beta}_k - \beta_k^*| \le \frac{2\widehat{\sigma} r}{x_{k*}\ \kappa^*_{k,\widehat{J}}} \left(1 - \frac{r}{\kappa^*_{1,\widehat{J}}} - \frac{r^2}{\kappa_{1,\widehat{J}}}\right)_+^{-1}, \tag{8.4}$$

$$|\widehat{\beta}_k - \beta_k^*| \le \frac{2\widehat{\sigma} r}{x_{k*}\ \kappa_k^*(s)} \left(1 - \frac{r}{\kappa_1^*(s)} - \frac{r^2}{\kappa_1(s)}\right)_+^{-1} \quad \text{with } s = 5. \tag{8.5}$$

Here, $\kappa^*_{k,\widehat{J}}$ and $\kappa_k^*(s)$ are computed directly via the programs (8.3) and (8.2) respectively. The value $\kappa_1(s)$ is then obtained from (4.6), and for $\kappa_{1,\widehat{J}}$ we use a lower bound analogous to (4.6):

$$\kappa_{1,\widehat{J}} \ge \frac{1-c}{2|\widehat{J}|} \min_{k=1,\dots,K} \kappa^*_{k,\widehat{J}}.$$

TABLE 1. Results for the *STIV* estimator without and with estimated instruments, $n = 49$

	$\hat{\beta}$ (1)	$\kappa^*_{k,\widehat{J}}$ (1)	$\kappa^*_k(5)$ (1)	$\hat{\beta}$ (2)	$\kappa^*_{k,\widehat{J}}$ (2)	$\kappa^*_k(5)$ (2)
β^*_1	1.03	0.107	0.103	1.03	0.085	0.068
β^*_2	0.98	0.308	0.157	0.98	0.367	0.075
β^*_3	0.96	0.129	0.103	0.96	0.126	0.071
β^*_4	0.95	0.150	0.109	0.95	0.115	0.057
β^*_5	0.90	0.253	0.175	0.90	0.177	0.086
β^*_6	0.00	0.166	0.095	0.00	0.126	0.065
β^*_7	0.00	0.155	0.080	0.00	0.148	0.060
β^*_8	0.00	0.154	0.110	0.00	0.122	0.056
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β^*_{23}	0.02	0.287	0.170	0.02	0.231	0.128
β^*_{24}	0.00	0.243	0.137	0.00	0.195	0.105
β^*_{25}	0.00	0.141	0.109	0.00	0.106	0.067

We use dots because the values that do not appear are similar.

(1): With all the 50 instruments,

(2): With 25 instruments including an estimate of the linear projection instrument.

We get $\kappa^*_{1,\widehat{J}} = 0.0096$ and $\kappa^*_1(5) = 0.0072$. In particular, we have $r/\kappa^*_{1,\widehat{J}} = 4.40 > 1$, so that (8.4) and (8.5) do not provide confidence intervals in this numerical example.

The columns on the right in Table 1 present the results where we use the same data, estimate the linear projection instrument by the Square-root Lasso and then take only K instruments: z_{il}, $l = L - K + 2, \dots, L$, and $\widehat{x}_{i1} = \sum_{l=1}^{L} \widehat{\zeta}_l z_{il}$, where $\widehat{\zeta}_l$ are the Square-root Lasso estimators of ζ_l, $l = 1, \dots, L$. The Square-root Lasso with parameter $c_{\sqrt{\text{Lasso}}} = 1.1$ recommended in Belloni, Chernozhukov and Wang (2010) [1] yields all coefficients equal to zero when keeping only the first three digits. This is disappointing since we get an instrument equal to zero. It should be noted that estimation in this setting is a hard problem since the dimension L is larger than the sample size, the number of non-zero coefficients ζ_l is large ($L - K + 1 = 26$), and their values are relatively small (equal to $0, 15$). To improve the estimation, we adjusted the parameter $c_{\sqrt{\text{Lasso}}}$ empirically, based on the value of the estimates. Ultimately, we have chosen $c_{\sqrt{\text{Lasso}}} = 0.3$. This choice is not covered by the theory of Belloni, Chernozhukov and Wang (2010) because there $c_{\sqrt{\text{Lasso}}}$ should be greater than 1. However, it leads to $\sqrt{\widehat{Q}(\widehat{\beta})} = 0.309$, which is very close to σ_{end}. The corresponding estimates $\widehat{\zeta}_l$ are given in Table

[1] The constant $c_{\sqrt{\text{Lasso}}}$ denoted by c in Belloni, Chernozhukov and Wang (2010) should not be mixed up with $c = c_{STIV}$ in the definition of the *STIV* estimator; $c_{\sqrt{\text{Lasso}}}$ is an equivalent of $\sqrt{n}/c_{STIV}$, up to constants.

TABLE 2. Estimates of the coefficients of the linear projection instrument

$\widehat{\zeta}_1$	$\widehat{\zeta}_2$	$\widehat{\zeta}_3$	$\widehat{\zeta}_4$	$\widehat{\zeta}_6$	$\widehat{\zeta}_8$	$\widehat{\zeta}_9$	$\widehat{\zeta}_{10}$	$\widehat{\zeta}_{14}$	$\widehat{\zeta}_{15}$	$\widehat{\zeta}_{16}$	$\widehat{\zeta}_{17}$	$\widehat{\zeta}_{18}$	$\widehat{\zeta}_{20}$
0.084	0.130	0.190	0.142	0.115	0.083	0.104	0.126	0.176	0.030	0.023	0.157	0.135	0.082
$\widehat{\zeta}_{21}$	$\widehat{\zeta}_{23}$	$\widehat{\zeta}_{24}$	$\widehat{\zeta}_{25}$	$\widehat{\zeta}_{26}$	$\widehat{\zeta}_{27}$	$\widehat{\zeta}_{32}$	$\widehat{\zeta}_{33}$	$\widehat{\zeta}_{34}$	$\widehat{\zeta}_{44}$	$\widehat{\zeta}_{47}$	$\widehat{\zeta}_{49}$	$\widehat{\zeta}_{50}$	
0.100	0.125	0.038	0.025	0.026	-0.058	0.108	0.005	-0.053	-0.006	-0.009	-0.063	0.033	

We only show the non-zero coefficients.

2. We see that they are not very close to the true ζ_l; some of the relevant coefficients are erroneously set to 0 and several superfluous variables are included, sometimes with significant coefficients, such as $\widehat{\zeta}_{32}$. We get $\kappa^*_{1,\widehat{J}} = 0.0076$ and $\kappa^*_1(5) = 0.0040$. Again, $r/\kappa^*_{1,\widehat{J}} > 1$, so that we cannot use (8.4) and (8.5) to get the confidence intervals. Note that this approach based on the estimated linear projection instrument gives sensitivities, which are lower than with the full set of instruments. This is mainly due to the fact that the estimation of the linear projection instrument is quite imprecise. Indeed, we add an instrument $\widehat{x}_{i1}$, which is not so good, and at the same time we drop a large number of other instruments, which may be not so bad. The overall effect on the sensitivities turns out to be negative. Recall that since the sensitivities involve the maximum of the scalar products of the rows of Ψ_n with Δ, the more we have rows (*i.e.*, instruments) the higher is the sensitivity. The same deterioration of the sensitivities occurred in other simulated data sets. In conclusion, the approach based on estimation of the linear projection instrument was not helpful to realize the above confidence intervals in this small sample situation. However, we will see that it achieves the task when the sample size gets large.

Although in this numerical example we were not able to use (8.4) and (8.5) for the confidence intervals, we got evidence that the performance of the *STIV* estimator is quite satisfactory. Table 3 shows a Monte-Carlo study where we keep the same values of the parameters of the model, of the sample size $n = 49$, and of the parameter A defining the set $\widehat{\mathcal{I}}$, simulate 1000 data sets, and compute 1000 estimates. The empirical performance of the *STIV* estimator is extremely good, even for the endogenous variable. The Monte-Carlo estimation of the variability of $\widehat{\beta}_1$ is very similar to that of the exogenous variables. With $c = 0.1$ the estimate $\widehat{\sigma}$ is larger than σ_{struct} in 95% of the simulations. This suggests that there remains some margin to penalize less for the "variance" in (3.5), *i.e.*, to decrease c and thus to obtain higher sensitivities.

Next, we study the empirical behavior of the non-pivotal *STIV* estimator. We consider the same model and the same values of all the parameters, and we choose $\sigma_* = 2 \cdot 0.233$ where 0.233 is the median of $\widehat{\sigma}$ from Table 3. Indeed $\mathbb{P}\left(\mathbb{E}_n[U^2] \leq \sigma_*^2\right)$ should be close to 1 (see Assumption 5.4). The results are given in Table 4. The non-pivotal procedure seems to better estimate as zeros the

TABLE 3. Monte-Carlo study, 1000 replications

	5^{th} percentile	Median	95^{th} percentile		5^{th} percentile	Median	95^{th} percentile
β_1^*	0.872	0.986	1.093	β_8^*	-0.057	0.000	0.055
β_2^*	0.877	0.970	1.048	β_9^*	-0.052	0.000	0.059
β_3^*	0.879	0.970	1.049	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_4^*	0.886	0.971	1.051	β_{23}^*	-0.051	0.000	0.051
β_5^*	0.877	0.968	1.049	β_{24}^*	-0.057	0.000	0.051
β_6^*	-0.048	0.000	0.055	β_{25}^*	-0.053	0.000	0.049
β_7^*	-0.059	0.000	0.063	$\hat{\sigma}$	0.181	0.233	0.291

TABLE 4. Monte-Carlo study of the non-pivotal estimator, 1000 replications

	5^{th} percentile	Median	95^{th} percentile		5^{th} percentile	Median	95^{th} percentile
β_1^*	0.714	0.914	1.110	β_8^*	-0.003	0.000	0.016
β_2^*	0.803	0.909	1.010	β_9^*	0.000	0.000	0.024
β_3^*	0.789	0.904	1.019	β_{10}^*	0.000	0.000	0.018
β_4^*	0.793	0.904	1.023	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_5^*	0.796	0.907	1.017	β_{23}^*	0.000	0.000	0.021
β_6^*	0.000	0.000	0.020	β_{24}^*	0.000	0.000	0.016
β_7^*	0.000	0.000	0.021	β_{25}^*	0.000	0.000	0.005

zero coefficients. This is because we minimize the ℓ_1 norm of the coefficients without an additional $c\sigma$ term. On the other hand, the non-zero coefficients are better estimated using the pivotal estimator. The non-pivotal procedure yields some shrinkage to zero (especially for large σ_*). Using the pivotal procedure in the first place allows us to have a good initial guess of σ_*.

Let us now increase n to see whether we can obtain interval estimates and take advantage of thresholding for variable selection. We consider the same model as above and the same values of the parameters of the method but we replace $n = 49$ by $n = 8000$. Then we are no longer in a situation where we must use specific high-dimensional techniques. However, it is still a challenging task to select among 25 candidate variables, one of them being endogenous. Indeed, classical selection procedures like the BIC would require to solve 2^{25} least squares problems. Our methods are much less numerically intensive. They are based on linear and conic programming, and their computational cost is polynomial in the dimension. We study both the setting with all the 50 instruments and the setting where we estimate the linear projection instrument.

TABLE 5. Confidence intervals and selection of variables, $n = 8000$

	$\hat{\beta}_{l,SC}$	$\hat{\beta}_{l,\widehat{J}}$	$\hat{\beta}$	$\hat{\beta}_{u,\widehat{J}}$	$\hat{\beta}_{u,SC}$	$\kappa^*_{k,\widehat{J}}$	$\kappa^*_k(5)$	$\omega_{k,\widehat{J}}$	$\omega_{k,SC}$
β^*_1	0.131	0.135	1.048	1.960	1.965	0.134	0.134	0.912	0.917
β^*_2	0.795	0.804	0.995	1.185	1.195	0.897	0.855	0.191	0.200
β^*_3	0.824	0.829	1.004	1.179	1.185	0.796	0.775	0.175	0.180
β^*_4	0.817	0.822	0.998	1.173	1.178	0.858	0.833	0.175	0.181
β^*_5	0.833	0.834	1.001	1.168	1.168	0.793	0.790	0.167	0.168
β^*_6	-0.163	-0.160	0.003	0.166	0.169	0.807	0.791	0.163	0.166
β^*_7	-0.173	-0.168	0.002	0.172	0.177	0.846	0.823	0.170	0.175
β^*_8	-0.173	-0.170	0.001	0.173	0.175	0.789	0.779	0.172	0.174
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β^*_{23}	-0.190	-0.188	0.003	0.194	0.197	0.802	0.793	0.191	0.193
β^*_{24}	-0.171	-0.166	0.001	0.168	0.172	0.842	0.821	0.167	0.172
β^*_{25}	-0.172	-0.169	-0.005	0.158	0.162	0.828	0.811	0.163	0.167

Consider first the case where we use all the instruments. Set for brevity

$$\widehat{w} \triangleq \left(1 - \frac{r}{\kappa^*_{1,\widehat{J}}} - \frac{r^2}{\kappa_{1,\widehat{J}}}\right)_+^{-1}, \quad w(5) \triangleq \left(1 - \frac{r}{\kappa^*_1(5)} - \frac{r^2}{\kappa_1(5)}\right)_+^{-1}.$$

These are the quantities appearing in (8.4) and (8.5). As above, we take $\widehat{J}$ equal to the set of the first five coordinates; $w(5)$ corresponds to the sparsity certificate approach with $s = 5$. Computing the exact coordinate-wise sensitivities we obtain the bound $\widehat{w} \leq 1.6277$. The sparsity certificate approach with $s = 5$ yields $w(5) \leq 1.6306$. We obtain $\widehat{\sigma} = 0.2970$ and the estimates in Table 5. The values $\hat{\beta}_{l,\widehat{J}}$ and $\hat{\beta}_{u,\widehat{J}}$ are the lower and upper confidence limits respectively obtained from (8.4); $\hat{\beta}_{l,SC}$ and $\hat{\beta}_{u,SC}$ are the lower and upper confidence limits obtained from (8.5) (sparsity certificate approach with $s = 5$). The thresholds $\omega_{k,\widehat{J}}$ and $\omega_k(5)$ are computed from the formulas

$$\omega_{k,\widehat{J}} = \frac{2 \cdot 1.6277 \hat{\sigma} r}{x_{k*} \kappa^*_{k,\widehat{J}}}, \qquad \omega_k(5) = \frac{2 \cdot 1.6306 \hat{\sigma} r}{x_{k*} \kappa^*_k(5)}.$$

Table 5 shows that in this example thresholding works well: The true support of β^* is recovered exactly by selecting the variables, for which the estimated coefficient is larger than the threshold. Note that the threshold for the endogenous variable is very close to the estimate of the first coefficient $\widehat{\beta}_1$ since the confidence intervals are wider for the endogenous variable.

We now consider the case where we use only 25 instruments; the 24 exogenous variables serve as their own instruments and the Square-root Lasso estimator of the linear projection instrument is

TABLE 6. Estimates of the coefficients in the linear projection instrument

$\widehat{\zeta}_1$	$\widehat{\zeta}_2$	$\widehat{\zeta}_3$	$\widehat{\zeta}_4$	$\widehat{\zeta}_5$	$\widehat{\zeta}_6$	$\widehat{\zeta}_7$	$\widehat{\zeta}_8$	$\widehat{\zeta}_9$	$\widehat{\zeta}_{10}$	$\widehat{\zeta}_{11}$	$\widehat{\zeta}_{12}$	$\widehat{\zeta}_{13}$	$\widehat{\zeta}_{14}$
0.142	0.145	0.134	0.136	0.137	0.135	0.139	0.139	0.134	0.140	0.146	0.140	0.134	0.136
$\widehat{\zeta}_{15}$	$\widehat{\zeta}_{16}$	$\widehat{\zeta}_{17}$	$\widehat{\zeta}_{18}$	$\widehat{\zeta}_{19}$	$\widehat{\zeta}_{20}$	$\widehat{\zeta}_{21}$	$\widehat{\zeta}_{22}$	$\widehat{\zeta}_{23}$	$\widehat{\zeta}_{24}$	$\widehat{\zeta}_{25}$	$\widehat{\zeta}_{26}$		
0.137	0.138	0.141	0.128	0.142	0.137	0.133	0.135	0.135	0.142	0.137	0.138		

We only show the non-zero coefficients (keeping only three digits).

TABLE 7. Confidence intervals and selection of variables, $n = 8000$

	$\hat{\beta}_{l,SC}$	$\hat{\beta}_{l,\widehat{J}}$	$\hat{\beta}$	$\hat{\beta}_{u,\widehat{J}}$	$\hat{\beta}_{u,SC}$	$\kappa^*_{k,\widehat{J}}$	$\kappa^*_k(5)$	$\omega_{k,\widehat{J}}$	$\omega_{k,SC}$
β^*_1	0.901	0.909	1.048	1.187	1.194	0.556	0.531	0.139	0.146
β^*_2	0.872	0.883	0.995	1.106	1.118	0.968	0.882	0.111	0.123
β^*_3	0.896	0.905	1.004	1.103	1.112	0.888	0.821	0.099	0.108
β^*_4	0.885	0.893	0.998	1.102	1.110	0.907	0.848	0.105	0.113
β^*_5	0.899	0.902	1.001	1.100	1.103	0.843	0.823	0.099	0.102
β^*_6	-0.098	-0.092	0.003	0.099	0.104	0.868	0.822	0.095	0.101
β^*_7	-0.103	-0.098	0.002	0.102	0.107	0.907	0.869	0.100	0.105
β^*_8	-0.099	-0.095	0.001	0.098	0.102	0.886	0.853	0.096	0.101
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β^*_{23}	-0.115	-0.109	0.003	0.115	0.121	0.862	0.825	0.112	0.118
β^*_{24}	-0.104	-0.099	0.001	0.101	0.106	0.888	0.848	0.100	0.105
β^*_{25}	-0.109	-0.104	-0.005	0.093	0.098	0.870	0.830	0.098	0.103

used for the endogenous variable. This time, we apply the Square-root Lasso with the recommended choice $c_{\sqrt{\text{Lasso}}} = 1.1$. We get $\sqrt{\widehat{Q}(\widehat{\beta})} = 0.3012$. The estimates of $\widehat{\zeta}_l$ are given in Table 6. Next, we use (8.4) and (8.5) to obtain the confidence intervals. Computing the exact coordinate-wise sensitivities we get the bound $\widehat{w} \leq 1.0941$. The sparsity certificate approach with $s = 5$ yields $w(5) \leq 1.0990$. We also get $\widehat{\sigma} = 0.2970$. The thresholds $\omega_{k,\widehat{J}}$ and $\omega_k(5)$ are obtained from the formulas

$$\omega_{k,\widehat{J}} = \frac{2 \cdot 1.0941 \hat{\sigma} r}{x_{k*} \kappa^*_{k,\widehat{J}}}, \quad \omega_k(5) = \frac{2 \cdot 1.0990 \hat{\sigma} r}{x_{k*} \kappa^*_k(5)}.$$

The results are summarized in Table 7. Note that the confidence intervals and the thresholds are sharper than in the approach including all the instruments. The particularly good news is that the confidence interval for the coefficient of the endogenous variable becomes much tighter.

In conclusion, when the sample size is large, the coordinate-wise sensitivities based on the sparsity certificate work remarkably well for estimation, confidence intervals, and variable selection.

We also get a significant improvement from using the two-stage procedure with estimated linear projection instrument.

9. Appendix

9.1. **Lower bounds on $\kappa_{p,J}$ for square matrices Ψ_n.** The following propositions establish lower bounds on $\kappa_{p,J}$ when Ψ_n is a square $K \times K$ matrix. 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 9.1. *For any $J \subseteq \{1,\dots,K\}$ we have*

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

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

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

This proves the first inequality of the proposition. The second inequality is obvious. □

Proposition 9.2. *Let $J \subseteq \{1,\dots,K\}$ be such that*

$$\inf_{\Delta\in\mathbb{R}^K\setminus\{0\}:\ \Delta\in C_J} \frac{|\mathbf{X}\mathbf{D}_\mathbf{X}\Delta|_2}{\sqrt{n}|\Delta_J|_2} \geq \widetilde{\kappa} \tag{9.1}$$

for some $\widetilde{\kappa} > 0$, and let there exist $0 < \delta < 1$ such that

$$\left|\frac{1}{n}\left(\mathbf{X}\mathbf{D}_\mathbf{X} - \mathbf{Z}\mathbf{D}_\mathbf{Z}^*\right)^T\mathbf{X}\mathbf{D}_\mathbf{X}\right|_\infty \leq \frac{\delta(1-c)^2\widetilde{\kappa}^2}{4|J|}. \tag{9.2}$$

Then

$$\kappa_{1,J} \geq \frac{(1-\delta)(1-c)^2\widetilde{\kappa}^2}{4|J|}.$$

Proof. We have

$$\begin{aligned}|\Psi_n\Delta|_\infty|\Delta|_1 &\geq |\Delta^T\Psi_n\Delta| \\ &\geq \left|\Delta^T\frac{1}{n}\mathbf{D}_\mathbf{X}\mathbf{X}^T\mathbf{X}\mathbf{D}_\mathbf{X}\Delta\right| - \left|\Delta^T\frac{1}{n}(\mathbf{X}\mathbf{D}_\mathbf{X} - \mathbf{Z}\mathbf{D}_\mathbf{Z})^T\mathbf{X}\mathbf{D}_\mathbf{X}\Delta\right|\end{aligned}$$

where

$$\begin{aligned}\left|\Delta^T\frac{1}{n}\left(\mathbf{X}\mathbf{D}_\mathbf{X} - \mathbf{Z}\mathbf{D}_\mathbf{Z}\right)^T\mathbf{X}\mathbf{D}_\mathbf{X}\Delta\right| &\leq \left|\frac{1}{n}\left(\mathbf{X}\mathbf{D}_\mathbf{X} - \mathbf{Z}\mathbf{D}_\mathbf{Z}\right)^T\mathbf{X}\mathbf{D}_\mathbf{X}\right|_\infty |\Delta|_1^2 \\ &\leq \frac{\alpha(1-c)^2\widetilde{\kappa}^2}{4|J|}|\Delta|_1^2.\end{aligned}$$

Combining these inequalities and using that $|\Delta|_1^2 \le \frac{4}{(1-c)^2}|J||\Delta_J|_2^2$ for all $\Delta \in C_J$ (cf. proof of Proposition 9.2) we get the result. □

Note that (9.1) is the restricted eigenvalue condition of Bickel, Ritov and Tsybakov (2009) for the Gram matrix of X-variables, up to the normalization by $\mathbf{D_X}$. Relation (9.2) accounts for the closeness between the instruments and the original set of variables suspected to be endogenous.

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

Proposition 9.3. *For any $s \le K/2$ and $1 < p \le 2$, we have*

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

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

Proof. For $\Delta \in R^K$ and a set $J \subset \{1,\dots,K\}$, 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 $|J| \le s$ we have $|J_+| \le 2s$. It is easy to see that the kth largest absolute value of elements of Δ_{J^c} satisfies $|\Delta_{J^c}|_{(k)} \le |\Delta_{J^c}|_1/k$. Thus,

$$|\Delta_{J_+^c}|_p^p \le |\Delta_{J^c}|_1^p \sum_{k \ge s+1} \frac{1}{k^p} \le \frac{|\Delta_{J^c}|_1^p}{(p-1)s^{p-1}}.$$

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 $c_0 = \frac{1+c}{1-c}$. Therefore, for $\Delta \in C_J$,

$$|\Delta|_p \le (1 + c_0(p-1)^{-1/p})|\Delta_{J_+}|_p \le (1 + c_0(p-1)^{-1/p})(2s)^{1/p-1/2}|\Delta_{J_+}|_2. \tag{9.3}$$

Using (9.3) and the fact that $|\Delta|_1 \le \frac{2}{1-c}|\Delta_J|_1 \le \frac{2\sqrt{s}}{1-c}|\Delta_J|_2$ for $\Delta \in C_J$, we get

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

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

The lower bounds in Propositions 9.1 and 9.3 require to control from below $|\Delta^T \Psi_n \Delta|$ (where Ψ_n is a non-symmetric possibly non-positive definite matrix) by a quadratic form with many zero eigenvalues for vectors in a cone of dominant coordinates. This is potentially a strong restriction on the instruments that we can use. In other words, the sensitivity characteristics $\kappa_{p,J}$ can be much larger than the above bounds. The propositions of this section imply that, even in the case of symmetric matrices, these characteristics are more general and potentially lead to better results than the restricted eigenvalues $\kappa_{\mathrm{RE}}(\cdot)$ appearing in the usual RE condition of Bickel, Ritov and Tsybakov (2009).

9.2. **Moderate deviations for self-normalized sums.** We use of the following result from Jing, Shao and Wang (2003), formula (2.11).

Theorem 9.4. *Let $X_1, \dots, X_n$ be independent random variables such that, for every i, $\mathbb{E}[X_i] = 0$ and $0 < \mathbb{E}[|X_i|^{2+\delta}] < \infty$ for some $0 < \delta \leq 1$. Set*

$$S_n = \sum_{i=1}^n X_i, \quad B_n^2 = \sum_{i=1}^n \mathbb{E}[X_i^2], \quad V_n^2 = \sum_{i=1}^n X_i^2, \quad L_{n,\delta} = \sum_{i=1}^n \mathbb{E}\left[|X_i|^{2+\delta}\right], \ d_{n,\delta} = B_n / L_{n,\delta}^{1/(2+\delta)}.$$

Then

$$\forall 0 \leq x \leq d_{n,\delta}, \ |\mathbb{P}(S_n/V_n \geq x) - (1 - \Phi(x))| \leq A_0 (1+x)^{1+\delta} e^{-x^2/2} / d_{n,\delta}^{2+\delta}$$

where $A_0 > 0$ is an absolute constant.

9.3. **Proofs. Proof of Proposition 4.1.** Parts (i) and (ii) of the proposition are straightforward. The upper bound in (4.3) follows immediately from (4.1). Next, obviously, $|\Delta|_p \leq |\Delta|_1^{1/p} |\Delta|_\infty^{1-1/p}$ and we get that, for $\Delta \neq 0$,

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

Furthermore, (4.1) implies $|\Delta|_1 \leq \frac{2}{1-c} |J| |\Delta|_\infty$ for $\Delta \in C_J$. Combining this with the above inequality we obtain the lower bound in (4.3). □

Proof of Proposition 4.2. For all $1 \leq k \leq K$ and $1 \leq l \leq L$,

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

which yields

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

The two inequalities of the assumption yield

$$\left|(\Psi_n)_{l(k)k}\right| |\Delta_k| \leq |\Delta|_1 \frac{(1-\eta_2)(1-c)}{2|J|} |(\Psi_n)_{l(k)k}| + \frac{1-c}{\eta_1} \left|(\Psi_n \Delta)_{l(k)}\right| \left|(\Psi_n)_{l(k)k}\right|.$$

Now, using that $\left|(\Psi_n\Delta)_{l(k)}\right| \le |\Psi_n\Delta|_\infty$ we obtain

$$|\Delta_j| \le |\Delta|_1 \frac{(1-\eta_2)(1-c)}{2|J|} + \frac{1-c}{\eta_1}\,|\Psi_n\Delta|_\infty \tag{9.4}$$

Summing the inequalities over j in J, yields

$$|\Delta_J|_1 \le \frac{(1-\eta_2)(1-c)}{2}|\Delta|_1 + \frac{|J|(1-c)}{\eta_1}\,|\Psi_n\Delta|_\infty\,.$$

This and the first inequality in (4.1) imply that we can take

$$\kappa_{1,J} = \frac{\eta_1\eta_2}{2|J|}. \tag{9.5}$$

Next, from (9.4) and (9.5) we deduce

$$\begin{aligned}|\Delta_j| &\le \left(\frac{1-\eta_2}{\eta_1\eta_2} + \frac{1}{\eta_1}\right)(1-c)\,|\Psi_n\Delta|_\infty \\ &\le \frac{1-c}{\eta_1\eta_2}\,|\Psi_n\Delta|_\infty\,,\end{aligned}$$

which implies

$$\kappa_{\infty,J} \ge \frac{\eta_1\eta_2}{1-c}.$$

This and the lower bound in (4.3) yield the result. □

Proof of Theorem 5.2. Define the event

$$\mathcal{G} = \left\{\left|\frac{1}{n}\mathbf{D_Z}\mathbf{Z}^T\mathbf{U}\right|_\infty \le \sqrt{\widehat{Q}(\beta^*)}r\right\}.$$

Since $\widehat{Q}(\beta^*) = \mathbb{E}_n[U^2]$, the union bound yields

$$\begin{aligned}\mathbb{P}(\mathcal{G}^c) &\le \sum_{l=1}^L \mathbb{P}\left(\frac{1}{n}\left|\frac{\sum_{i=1}^n z_{li}u_i}{z_{l*}\sqrt{\mathbb{E}_n[U^2]}}\right| \ge r\right) \\ &\le \sum_{l=1}^L \mathbb{P}\left(\left|\frac{\sum_{i=1}^n z_{li}u_i}{\sqrt{\sum_{i=1}^n (z_{li}u_i)^2}}\right| \ge A\sqrt{2\log(L)}\right).\end{aligned} \tag{9.6}$$

By Theorem 9.4, for all $l = 1, \dots, L$,

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^n z_{li}u_i}{\sqrt{\sum_{i=1}^n (z_{li}u_i)^2}}\right| \ge A\sqrt{2\log(L)}\right) \le 2\left(1 - \Phi(A\sqrt{2\log(L)})\right) + 2A_0\frac{(1+A\sqrt{2\log(L)})^{1+\delta}}{L^{A^2}d_{n,\delta}^{2+\delta}}. \tag{9.7}$$

Thus, the event $\mathcal{G}$ holds with probability at least $1-\alpha$, by the definition of α in (5.1).

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

$$|\Psi_n\Delta|_\infty \le \left|\frac{1}{n}\mathbf{D_Z}\mathbf{Z}^T(\mathbf{Y} - \mathbf{X}\widehat{\beta})\right|_\infty + \left|\frac{1}{n}\mathbf{D_Z}\mathbf{Z}^T(\mathbf{Y} - \mathbf{X}\beta^*)\right|_\infty \tag{9.8}$$

$$\leq r\widehat{\sigma} + \left|\frac{1}{n}\mathbf{D_Z}\mathbf{Z}^T\mathbf{U}\right|_\infty$$
$$\leq r\left(\widehat{\sigma} + \sqrt{\widehat{Q}(\beta^*)}\right).$$

Notice that, on the event $\mathcal{G}$, the pair $\left(\beta^*, \sqrt{\widehat{Q}(\beta^*)}\right)$ belongs to the set $\widehat{\mathcal{I}}$. On the other hand, $(\widehat{\beta}, \widehat{\sigma})$ minimizes the criterion $\left|\mathbf{D_X}^{-1}\beta\right|_1 + c\sigma$ on the same set $\widehat{\mathcal{I}}$. Thus, on the event $\mathcal{G}$,

$$\left|\mathbf{D_X}^{-1}\widehat{\beta}\right|_1 + c\widehat{\sigma} \leq |\mathbf{D_X}^{-1}\beta^*|_1 + c\sqrt{\widehat{Q}(\beta^*)}. \tag{9.9}$$

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

$$\begin{aligned}
\left|\Delta_{J(\beta^*)^c}\right|_1 &= \sum_{k\in J(\beta^*)^c}\left|x_{k*}\widehat{\beta}_k\right| \\
&\leq \sum_{k\in J(\beta^*)}\left(|x_{k*}\beta_k^*| - \left|x_{k*}\widehat{\beta}_k\right|\right) + c\left(\sqrt{\widehat{Q}(\beta^*)} - \sqrt{\widehat{Q}(\widehat{\beta})}\right) \\
&\leq \left|\Delta_{J(\beta^*)}\right|_1 + c\left(\sqrt{\widehat{Q}(\beta^*)} - \sqrt{\widehat{Q}(\widehat{\beta})}\right) \\
&\leq \left|\Delta_{J(\beta^*)}\right|_1 + c\left|\frac{\mathbb{E}_n[UX^T]\mathbf{D_X}\Delta}{\sqrt{\mathbb{E}_n[U^2]}}\right| \quad \left(\text{by convexity of } \beta \mapsto \sqrt{\widehat{Q}(\beta)}\right) \\
&\leq \left|\Delta_{J(\beta^*)}\right|_1 + c\left|\frac{\mathbb{E}_n[UX^T]\mathbf{D_X}}{\sqrt{\mathbb{E}_n[U^2]}}\right|_\infty |\Delta|_1 \\
&\leq \left|\Delta_{J(\beta^*)}\right|_1 + c|\Delta|_1 \quad \text{(by the Cauchy-Schwarz inequality).}
\end{aligned} \tag{9.10}$$

Note that (9.10) can be re-written as a cone condition:

$$\left|\Delta_{J(\beta^*)^c}\right|_1 \leq \frac{1+c}{1-c}\left|\Delta_{J(\beta^*)}\right|_1. \tag{9.11}$$

Thus, $\Delta \in C_{J(\beta^*)}$ on the event $\mathcal{G}$. Using (9.8) and arguing as in (9.10) we find

(9.12)

$$\begin{aligned}
|\Psi_n\Delta|_\infty &\leq r\left(2\widehat{\sigma} + \sqrt{\widehat{Q}(\beta^*)} - \widehat{\sigma}\right) \\
&\leq r\left(2\widehat{\sigma} + \sqrt{\widehat{Q}(\beta^*)} - \sqrt{\widehat{Q}(\widehat{\beta})}\right) \quad (\text{since } \sqrt{\widehat{Q}(\widehat{\beta})} \leq \widehat{\sigma}) \\
&\leq r\left(2\widehat{\sigma} + \left|\frac{\mathbb{E}_n[UX^T]\mathbf{D_X}\Delta}{\sqrt{\mathbb{E}_n[U^2]}}\right|\right)
\end{aligned}$$

$$\leq r\left(2\widehat{\sigma}+\max_{j\in J_{\rm end}}\left|\frac{\mathbb{E}_n[UX_j]}{\sqrt{\mathbb{E}_n[X_j^2U^2]}}\right||\Delta_{J_{\rm end}}|_1+\max_{j\in J_{\rm end}^c}\left|\frac{\mathbb{E}_n[UX_j]}{\sqrt{\mathbb{E}_n[X_j^2U^2]}}\right||\Delta_{J_{\rm end}^c}|_1\right)$$

$$\leq r\left(2\widehat{\sigma}+|\Delta_{J_{\rm end}}|_1+\max_{j\in J_{\rm end}^c}\left|\frac{\mathbb{E}_n[UX_j]}{\sqrt{\mathbb{E}_n[X_j^2U^2]}}\right||\Delta_{J_{\rm end}^c}|_1\right) \quad \text{(by the Cauchy-Schwarz inequality).}$$

Since $L\geq K$ and $x_{ji}=z_{j'i}$ where $j'=j-k_{\rm end}$, $j\in J_{\rm end}^c=\{k_{\rm end}+1,\dots,K\}$ (the exogenous variables serve as their own instruments), from (9.7) we obtain that, on the event $\mathcal{G}$,

$$\max_{j\in J_{\rm end}^c}\left|\frac{\mathbb{E}_n[UX_j]}{\sqrt{\mathbb{E}_n[X_j^2U^2]}}\right|\leq r.$$

Combining this with (9.12) and using the definition of the block sensitivity $\kappa_{J_0,J(\beta^*)}$ with $J_0=J_{\rm end}$, $J_0=J_{\rm end}^c$, we get that, on the event $\mathcal{G}$,

$$|\Psi_n\Delta|_\infty\leq r\left(2\widehat{\sigma}+\sqrt{\widehat{Q}(\beta^*)}-\widehat{\sigma}\right)$$
$$\leq r\left(2\widehat{\sigma}+\frac{|\Psi_n\Delta|_\infty}{\kappa^*_{J_{\rm end},J(\beta^*)}}+r\frac{|\Psi_n\Delta|_\infty}{\kappa^*_{J_{\rm end}^c,J(\beta^*)}}\right), \tag{9.13}$$

which implies

$$|\Psi_n\Delta|_\infty\leq 2\widehat{\sigma}r\left(1-\frac{r}{\kappa^*_{J_{\rm end},J(\beta^*)}}-\frac{r^2}{\kappa^*_{J_{\rm end}^c,J(\beta^*)}}\right)_+^{-1}. \tag{9.14}$$

This inequality and the definition of the sensitivities yield (5.2) and (5.3).

To prove (5.4), it suffices to note that, by (9.9) and by the definition of $\kappa^*_{J(\beta^*),J(\beta^*)}$,

$$c\widehat{\sigma}\leq|\Delta_{J(\beta^*)}|_1+c\sqrt{\widehat{Q}(\beta^*)}$$
$$\leq \frac{|\Psi_n\Delta|_\infty}{\kappa^*_{J(\beta^*),J(\beta^*)}}+c\sqrt{\widehat{Q}(\beta^*)},$$

and to combine this inequality with (9.8). □

Proof of Theorem 5.7. Part (i) of the theorem is a consequence of (5.4) and Assumptions 5.4 and 5.5. Parts (ii) and (iii) follow immediately from (5.2), (5.3), and Assumptions 5.4 and 5.5. Part (iv) is straightforward in view of (5.9). □

Proof of Theorem 5.9. Let $\mathcal{G}_j$ be the events of probabilities at least $1-\gamma_j$ respectively appearing in Assumptions 5.4, 5.6, 5.8. Assume that all these events hold, as well as the event $\mathcal{G}$. Then

$$\omega_k(s) \leq \frac{2\sigma_* r}{c_k^*(s) v_k} \left(1 + \frac{r}{c c^*_{J(\beta^*)}}\right) \left(1 - \frac{r}{c c^*_{J(\beta^*)}}\right)_+^{-1} \left(1 - \frac{r}{c^*_{J_{\mathrm{end}}}(s)} - \frac{r^2}{c^*_{J^c_{\mathrm{end}}}(s)}\right)_+^{-1} \triangleq \omega_k^*.$$

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 (5.3) and Assumptions 5.4 and 5.8, we obtain $|\widehat{\beta}_k| \leq \omega_k$, which implies $\widetilde{\beta}_k = 0$. Second, if $k \in J(\beta^*)$, then using again (5.3) we get $||\beta_k^*| - |\widehat{\beta}_k|| \leq |\beta_k^* - \widehat{\beta}_k| \leq \omega_k \leq \omega_k^*$. Since $|\beta_k^*| > 2\omega_k^*$ for $k \in J(\beta^*)$, we obtain that $|\widehat{\beta}_k| > \omega_k$, 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 5.11. Fix an arbitrary subset J of $\{1, \ldots, K\}$. Acting as in (9.10) with J instead of $J(\beta^*)$, we get:

$$\begin{aligned}\sum_{k \in J^c} \left|x_{k*}\widehat{\beta}_k\right| + \sum_{k \in J^c} |x_{k*}\beta_k^*| &\leq \sum_{k \in J} \left(|x_{k*}\beta_k^*| - \left|x_{k*}\widehat{\beta}_k\right|\right) + 2\sum_{k \in J^c} |x_{k*}\beta_k^*| \\ &\quad + c\left(\sqrt{\widehat{Q}(\beta^*)} - \sqrt{\widehat{Q}(\widehat{\beta})}\right) \\ &\leq |\Delta_J|_1 + 2\left|\left(\mathbf{D}_{\mathbf{X}}{}^{-1}\beta^*\right)_{J^c}\right|_1 + c|\Delta|_1.\end{aligned}$$

This yields

$$|\Delta_{J^c}|_1 \leq |\Delta_J|_1 + 2\left|\left(\mathbf{D}_{\mathbf{X}}{}^{-1}\beta^*\right)_{J^c}\right|_1 + c|\Delta|_1. \tag{9.15}$$

Assume now that we are on the event $\mathcal{G}$. Consider the two possible cases. First, if $2\left|\left(\mathbf{D}_{\mathbf{X}}{}^{-1}\beta^*\right)_{J^c}\right|_1 \leq |\Delta_J|_1$, then $\Delta \in \widetilde{C}_J$ and, in particular, (9.13) holds with the sensitivities $\widetilde{\kappa}_{\bullet,J}$ instead of $\kappa_{\bullet,J(\beta^*)}$. From this, using the definition of the sensitivity $\widetilde{\kappa}_{p,J}$, we get that $|\Delta|_p$ is bounded from above by the first term of the maximum in (5.14). Second, if $2\left|\left(\mathbf{D}_{\mathbf{X}}{}^{-1}\beta^*\right)_{J^c}\right|_1 > |\Delta_J|_1$, then for any $p \in [1, \infty]$ we have a simple bound

$$|\Delta|_p \leq |\Delta|_1 = |\Delta_{J^c}|_1 + |\Delta_J|_1 \leq \frac{6}{1-c}\left|\left(\mathbf{D}_{\mathbf{X}}{}^{-1}\beta^*\right)_{J^c}\right|_1.$$

In conclusion, $|\Delta|_p$ is smaller than the maximum of the two bounds. □

Proof of Theorem 6.2. Throughout the proof, we assume that we are on the event of probability at least $1-\alpha_2$ where (6.4) holds. It follows easily from (6.4) that

$$\left|\frac{1}{n}\overline{\mathbf{Z}}^T\mathbf{X}(\widehat{\beta} - \beta^*)\right|_\infty \leq \widehat{b}\overline{z}_*. \tag{9.16}$$

Next, an argument similar to (9.6) and Theorem 9.4 yield that, with probability at least $1-\alpha_1$,

$$\left|\frac{1}{n}\overline{\mathbf{Z}}^T\mathbf{U}-\theta^*\right|_\infty \le r_1 \max_{l=1,\dots,L_1}\sqrt{\frac{1}{n}\sum_{i=1}^n(\overline{z}_{li}u_i-\theta^*_l)^2} = r_1F(\theta^*,\beta^*). \tag{9.17}$$

In what follows, we assume that we are on the event of probability at least $1-\alpha_1-\alpha_2$ where both (9.16) and (9.17) are satisfied.

We will use the properties of $F(\theta,\beta)$ stated in the next lemma that we prove in Section 9.4.

Lemma 9.5. *We have*

$$F(\theta^*,\widehat{\beta})-F(\widehat{\theta},\widehat{\beta}) \le |\widehat{\theta}-\theta^*|_1, \tag{9.18}$$

$$|F(\theta^*,\widehat{\beta})-F(\theta^*,\beta^*)| \le \overline{z}_*\left|\mathbf{D_X}^{-1}(\widehat{\beta}-\beta^*)\right|_1 \le \widehat{b}\overline{z}_*. \tag{9.19}$$

We proceed now to the proof of Theorem 6.2. First, we show that the pair $(\theta,\sigma_1)=(\theta^*,F(\theta^*,\beta^*))$ belongs to the set $\widehat{\mathcal{I}}_1$. Indeed, from (9.16) and (9.17) we get

$$\begin{aligned}\left|\frac{1}{n}\overline{\mathbf{Z}}^T(\mathbf{Y}-\mathbf{X}\widehat{\beta})-\theta^*\right|_\infty &\le \left|\frac{1}{n}\overline{\mathbf{Z}}^T\mathbf{U}-\theta^*\right|_\infty+\left|\frac{1}{n}\overline{\mathbf{Z}}^T\mathbf{X}(\widehat{\beta}-\beta^*)\right|_\infty\\ &\le r_1F(\theta^*,\beta^*)+\widehat{b}\overline{z}_*.\end{aligned}$$

Thus, the pair $(\theta,\sigma_1)=(\theta^*,F(\theta^*,\beta^*))$ satisfies the first constraint in the definition of $\widehat{\mathcal{I}}_1$. It satisfies the second constraint as well, since $F(\theta^*,\widehat{\beta})\le F(\theta^*,\beta^*)+\widehat{b}\overline{z}_*$ by (9.19).

Now, as $(\theta^*,F(\theta^*,\beta^*))\in\widehat{\mathcal{I}}_1$ and $(\widehat{\theta},\widehat{\sigma}_1)$ minimizes $|\theta|_1+c\sigma_1$ over $\widehat{\mathcal{I}}_1$, we have

$$|\widehat{\theta}|_1+c\widehat{\sigma}_1\le|\theta^*|_1+cF(\theta^*,\beta^*), \tag{9.20}$$

which implies

$$|\overline{\Delta}_{J(\theta^*)^c}|_1\le|\overline{\Delta}_{J(\theta^*)}|_1+c(F(\theta^*,\beta^*)-\widehat{\sigma}_1), \tag{9.21}$$

where $\overline{\Delta}=\widehat{\theta}-\theta^*$. Using the fact that $F(\widehat{\theta},\widehat{\beta})\le\widehat{\sigma}_1+\widehat{b}\overline{z}_*$, (9.18), and (9.19) we obtain

$$\begin{aligned}F(\theta^*,\beta^*)-\widehat{\sigma}_1 &\le F(\theta^*,\beta^*)-F(\widehat{\theta},\widehat{\beta})+\widehat{b}\overline{z}_*\\ &\le |\widehat{\theta}-\theta^*|_1+2\widehat{b}\overline{z}_*.\end{aligned} \tag{9.22}$$

This inequality and (9.21) yield

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

or equivalently,

$$|\overline{\Delta}_{J(\theta^*)^c}|_1 \le \frac{1+c}{1-c}\,|\overline{\Delta}_{J(\theta^*)}|_1 + \frac{2c}{1-c}\widehat{b}\overline{z}_*. \tag{9.23}$$

Next, using (9.16), (9.17) and the second constraint in the definition of $(\widehat{\theta}, \widehat{\sigma}_1)$, we find

$$\begin{aligned} |\widehat{\theta} - \theta^*|_\infty &\le \left|\frac{1}{n}\overline{\mathbf{Z}}^T(\mathbf{Y} - \mathbf{X}\widehat{\beta}) - \widehat{\theta}\right|_\infty \\ &\quad + \left|\frac{1}{n}\overline{\mathbf{Z}}^T\mathbf{U} - \theta^*\right|_\infty + \left|\frac{1}{n}\overline{\mathbf{Z}}^T\mathbf{X}(\widehat{\beta} - \beta^*)\right|_\infty \\ &\le r_1(\widehat{\sigma}_1 + F(\theta^*, \beta^*)) + 2\widehat{b}\overline{z}_*. \end{aligned}$$

This and (9.22) yield

$$|\widehat{\theta} - \theta^*|_\infty \le r_1(2\widehat{\sigma}_1 + |\widehat{\theta} - \theta^*|_1) + 2(1+r_1)\widehat{b}\overline{z}_*. \tag{9.24}$$

On the other hand, (9.23) implies

$$\begin{aligned} |\widehat{\theta} - \theta^*|_1 &\le \frac{2}{1-c}\,|\overline{\Delta}_{J(\theta^*)}|_1 + \frac{2c}{1-c}\widehat{b}\overline{z}_* \\ &\le \frac{2|J(\theta^*)|}{1-c}\,|\widehat{\theta} - \theta^*|_\infty + \frac{2c}{1-c}\widehat{b}\overline{z}_*. \end{aligned} \tag{9.25}$$

Inequalities (6.7) and (6.8) follow from solving (9.24) and (9.25) with respect to $|\widehat{\theta} - \theta^*|_\infty$ and $|\widehat{\theta} - \theta^*|_1$ respectively. □

Proof of Theorem 6.4. We first prove part (i). We will assume that we are on the event of probability at least $1 - \alpha_1 - \alpha_2 - \varepsilon$ where (9.17), (6.9), and (6.11) are simultaneously satisfied. From (9.20) and the fact that (6.9) can be written as $F(\theta^*, \beta^*) \le \sigma_{1*}$ we obtain

$$\widehat{\sigma}_1 \le |\widehat{\theta} - \theta^*|_1/c + \sigma_{1*}. \tag{9.26}$$

Note also that the argument in the proof of Theorem 6.2 and the results of that theorem remain obviously valid with $\widehat{b}$ replaced by b_*. Thus, we can use (6.8) with $\widehat{b}$ replaced by b_*, and combining it with (9.26) we obtain

$$\widehat{\sigma}_1 \le \overline{\sigma}_*. \tag{9.27}$$

This and (6.7) yield (6.12).

We now prove part (ii) of the theorem. In the rest of the proof, we assume that we are on the event $\mathcal{G}'$ of probability at least $1 - \alpha_1 - \varepsilon - \gamma$ where (9.17), (6.9), and the events $\mathcal{G}$, $\mathcal{G}_j$ defined in the proofs of Theorems 5.2, 5.7 are simultaneously satisfied. Then item (ii) of Theorem 5.7 with $p = 1$ implies (6.11) with b_* defined in (6.13). This and (6.12) easily give part (ii) of the theorem.

To prove part (iii), note that, by Theorem 5.7 (i) and Assumption 5.8,

$$\widehat{b} = \frac{2\widehat{\sigma} r s}{\kappa_1(s)} \left(1 - \frac{r}{\kappa^*_{J_{\rm end}}(s)} - \frac{r^2}{\kappa^*_{J^c_{\rm end}}(s)} \right)^{-1}_{+} \le b_* \tag{9.28}$$

for b_* defined in (6.13). This and (9.27) imply that the threshold ω satisfies $\omega \triangleq V(\widehat{\sigma}_1, \widehat{b}, J(\widehat{\theta})) \le V(\overline{\sigma}_*, b_*, s_1) \triangleq \omega^*$ on the event $\mathcal{G}'$. On the other hand, (6.7) guarantees that $|\widehat{\theta}_l - \theta^*_l| \le \omega$ and, by assumption, $|\theta^*_l| > 2\omega^*$ for all $l \in J(\theta^*)$. In addition, by (5.2) and (6.7) for all $l \in J(\theta^*)^c$ we have $|\theta^*_l| < \omega$, which implies $\widetilde{\theta}_l = 0$. We finish the proof in the same way as the proof of Theorem 5.7. □

9.4. **Proof of Lemma 9.5.** Set $f_l(\theta_l) \triangleq \sqrt{\widehat{Q}_l(\theta_l, \widehat{\beta})}$, and $f(\theta) \triangleq \max_{l=1,\dots,L_1} f_l(\theta_l) \equiv F(\theta, \widehat{\beta})$. The mappings $\theta \mapsto f_l(\theta_l)$ are convex, so that by the Dubovitsky-Milutin theorem (see, *e.g.*, Alekseev, Tikhomirov and Fomin (1987), Chapter 2), the subdifferential of their maximum f is contained in the convex hull of the union of the subdifferentials of the f_l:

$$\partial f \subseteq \mathrm{Conv} \left(\bigcup_{l=1}^{L_1} \partial f_l \right). \tag{9.29}$$

Since, obviously, $\partial f_l(\theta_l) \subseteq [-1, 1]$, we find that $\partial f(\theta) \subseteq \{w \in \mathbb{R}^{L_1} : |w|_\infty \le 1\}$ for all $\theta \in \mathbb{R}^{L_1}$. Using this property and the convexity of f, we get

$$f(\theta^*) - f(\widehat{\theta}) \le \langle w, \theta^* - \widehat{\theta} \rangle \le |\widehat{\theta} - \theta^*|_1, \quad \forall\, w \in \partial f(\theta^*),$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product in $\mathbb{R}^{L_1}$. This yields (9.18). The proof of (9.19) is based on similar arguments. Instead of f_l, we now introduce the functions g_l defined by $g_l(\beta) \triangleq \sqrt{\widehat{Q}_l(\theta^*_l, \beta)}$, and set $g(\beta) \triangleq \max_{l=1,\dots,L_1} g_l(\beta) \equiv F(\theta^*, \beta)$. Next, notice that the subdifferential of g_l satisfies $\partial g_l(\beta) \subseteq \{w \in \mathbb{R}^K : |w_k| \le a_{lk},\ k = 1, \dots, K\}$ for all $\beta \in \mathbb{R}^K$, $l = 1, \dots, L_1$, where

$$a_{lk} = \frac{\left| \frac{1}{n} \sum_{i=1}^n \overline{z}_{li} x_{ki} \left(\overline{z}_{li} (y_i - x_i^T \beta) - \theta^*_l \right) \right|}{\sqrt{\frac{1}{n} \sum_{i=1}^n \left(\overline{z}_{li} (y_i - x_i^T \beta) - \theta^*_l \right)^2}}.$$

Consequently, by the Cauchy-Schwarz inequality, $\mathbf{D_X} \partial g_l(\beta) \subseteq \{w \in \mathbb{R}^K : |w|_\infty \le \overline{z}_*\}$ for all $\beta \in \mathbb{R}^K$, $l = 1, \dots, L_1$. This and (9.29) with g, g_l instead of f, f_l imply $\mathbf{D_X} \partial g(\beta) \subseteq \{w \in \mathbb{R}^K : |w|_\infty \le \overline{z}_*\}$ for all $\beta \in \mathbb{R}^K$. Using this property and the convexity of g, we get

$$g(\beta) - g(\beta') \le \langle w, (\beta - \beta') \rangle \le |\mathbf{D_X} w|_\infty \left| \mathbf{D_X}^{-1} (\beta - \beta') \right|_1 \le \overline{z}_* \left| \mathbf{D_X}^{-1} (\beta - \beta') \right|_1, \quad \forall\, w \in \partial g(\beta),$$

for any $\beta, \beta' \in \mathbb{R}^K$. This proves (9.19). □

References

[1] Alekseev, V.M, V. M. Tikhomirov, and S. V. Fomin (1987): *Optimal Control*. Consultants Bureau, New York.

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

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

[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] Angrist, J. D., and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?". *Quarterly Journal of Economics,* 106, 979-1014.

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

[7] Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2010): "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain". Preprint: arXiv:1010.4345.

[8] Belloni, A., and V. Chernozhukov (2010): "Post L1-Penalized Estimators in High-Dimensional Linear Regression models". Preprint: arXiv:1001.0188v2.

[9] Belloni, A., V. Chernozhukov, and L. Wang (2010): "Square-Root Lasso: Pivotal Recovery of Sparse Signals Via Conic Programming". Preprint: arXiv:1009.5689.

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

[11] 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.

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

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

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

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

[16] Carrasco, M., and J. P. Florens (2000): "Generalization of GMM to a Continuum of Moment Conditions". *Econometric Theory,* 16, 797–834.

[17] Carrasco, M., and J.-P. Florens (2008): "On the Asymptotic Efficiency of GMM". Working Paper.

[18] Carrasco, M. (2008): "A Regularization Approach to the Many Instruments Problem". Working Paper.

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

[20] Chao, J. C., and N. R. Swanson (2005): "Consistent Estimation with a Large Number of Weak Instruments". *Econometrica,* 73, 1673-1692.

[21] 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.

[22] Donald, S. G., and W. K. Newey (2001): "Choosing the Number of Instruments". *Econometrica,* 69, 1161–1191.

[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] Hahn, J., and J. Hausman (2002): "A New Specification Test for the Validity of Instrumental Variables". *Econometrica,* 70, 163-189.

[25] Hall, A. R., and F. P. M. Peixe (2003): "A Consistent Method for the Selection of Relevant Instruments". *Econometric Reviews,* 22, 269-287.

[26] Hansen, C., J. Hausman, and W. K. Newey (2008): "Estimation with Many Instrumental Variables". *Journal of Business and Economic Statistics*, 26, 398–422.

[27] Hausman, J., W. K. Newey, T. Woutersen, J. Chao, and N. Swanson (2009): "Instrumental Variable Estimation with Heteroskedasticity and Many Instruments". Working Paper.

[28] Jing, B.-Y., Q. M. Shao, and Q. Wang (2003): "Self-Normalized Cramér-Type Large Deviations for Independent Random Variables". *The Annals of Probability,* 31, 2167–2215.

[29] Koltchinskii, V. (2009): "The Dantzig Selector and Sparsity Oracle Inequalities". *Bernoulli*, 15, 799–828.

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

[31] Liao, Z. (2010): "Adaptive GMM Shrinkage Estimation with Consistent Moment Selection ". Working Paper.

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

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

[34] Okui, R. (2008): "Instrumental Variable Estimation in the Presence of Many Moment Conditions". *Journal of Econometrics,* forthcoming.

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

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

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

[38] Surm, J. F. (1999): "Using SeDuMi 1.02, a Matlab Toolbox for Optimization Over Symmetric Cones". *Optimization Methods and Software,* 11, 625-653.

[39] Wooldridge, J. M. (2002): *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge.

[40] 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.