First Year Report
Sensitivity Analysis for Instrumental Variables

Tobias Freidling
DPMMS, University of Cambridge
taf40@cam.ac.uk

supervised by Qingyuan Zhao

1 Introduction

Instrumental Variables (IV) are one of the most used and researched methods within the area of causal inference and can be traced back to the very formation of the field, see e.g. Wright (1928) and Reiersøl (1945). Initially developed in the econometrics literature, Instrumental Variables became increasingly popular in other disciplines, such as sociology or epidemiology, cf. Hernán and Robins (2006), as well and are still object of current research, both from a statistical and an applied point of view.
The method is concerned with estimating the causal effect of some explanatory variables on a dependent variable, or rather outcome. While there is a multitude of models tailored for this purpose, they rely on the assumption that there is no unmeasured confounding between explanatory and dependent variables. For instance, in a linear regression model, the estimate of the regression coefficients is only unbiased if the error term and the covariates are independent. In many applications, however, such an assumption is dubious. Especially in economics, many involved variables act upon each other simultaneously which gave rise to the notion of endogeneous variables as quantities that are determined jointly within the model. Exogeneous variables, on the other hand, can be seen as predetermined, cf. Davidson and MacKinnon (1993).
The Instrumental Variables method prevents this problem by considering additional data, the so called instrumental variables, that can be used to predict the explanatory variables but is independent of the unmeasured confounder and only acts upon the outcome through the explanatory variables. Theses three assumptions are made more precise in Section 3; Figure 1 provides a sketch of an IV model. In the past three decades, research activity mostly concentrated around

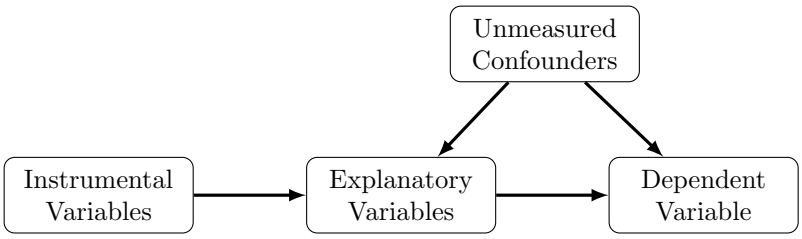

Figure 1: Schematic of Instrumental Variables.

the first condition as it can be tested based on the available data. The plausibility of the second and third assumption, however, cannot be verified statistically and must be argued for by an expert in the respective field. This leaves much room for uncertainty in studies which employ

Instrumental Variables and might incentivise practitioners to ignore doubts on the validity of these assumptions.
For this reason, sensitivity analysis is an important tool to assess the robustness of the findings. The researcher adds a sensitivity parameter δ to the model and specifies a sensitivity set Δ, which represents the degree of violation of the assumptions that she considers plausible. $\Delta = \{0\}$ is interpreted as all instruments being valid. Similar to confidence intervals/regions, we can introduce a sensitivity interval/region as a set that contains the true causal effect of the explanatory variables on the outcome at a prescribed level under the assumption that $\delta \in \Delta$. Previous work considered one-dimensional and point sensitivity sets. Yet, a more general method for constructing sensitivity regions is still a blind spot, to the best of our knowledge. This work proposes such a procedure for linear IV models.
The paper is organised as follows. Section 2 presents a literature review on previously proposed methods and Section 3 formally introduces the linear Instrumental Variables model, summarises important results in IV theory and introduces the sensitivity model. In Section 4 we construct sensitivity regions with the union method, whereas we use the inversion of hypothesis tests to propose three different approaches to sensitivity regions in Section 5. Section 6 addresses the specification of the sensitivity set in more detail and Section 7 summarises tasks for future work.

2 Literature Review

The issue of unverifiable assumptions in IV models caught the attention of practitioners, especially in econometrics, in the last two decades and some approaches based on point sensitivity sets $\Delta = \{\delta_1, \dots, \delta_m\}$ were proposed. DiPrete and Gangl (2004) compare the estimates derived from an IV model for different degrees of deviation from the assumptions and give detailed account of its practical implementation. Conley et al. (2012) construct sensitivity intervals as the union of confidence intervals for different values in Δ. Ashley (2009) formulates his sensitivity model in terms of the covariance matrix of explanatory variables, outcome and instruments and expands this approach to the overidentified case in a later paper, Ashley and Parmeter (2015).
Beyond the primary use of established IV theory, some research on invalid instruments was undertaken in the last decade. Kolesár et al. (2015) consider estimation of the correct causal effect in many instrument models, i.e. the number of exogeneous instruments and regressors grows along the sample size. They derive a consistent estimator assuming that the direct effects of the instruments on the outcome are orthogonal to the direct effects of the instruments on the endogenous regressor. Yet, this condition along with the many instrument setting seems fairly restrictive and does not solve the problem of unverifiable conditions. In recent years, the setting of some valid and some invalid instruments received attention in IV research. For instance, Kang et al. (2020) construct a valid test for the causal effect assuming that all instruments are independent and that at least one instrumental variable is valid. In some applications, such as Mendelian Randomisation, these conditions can be reasonable but in fields like the social sciences or economics this seems overly optimistic.
Closest to our research are the works of Small (2007) and Wang et al. (2018). The former introduce the notion of sensitivity intervals for Instrumental Variables and considers an overidentified model, i.e. more instruments than explanatory variables. Small constructs a joint confidence region for the causal effect and sensitivity parameter via a combination of Sargan's test for δ-values in Δ, cf. Sargan (1958), and the Wald confidence interval for the causal effect. Then the sensitivity parameter is marginalised out yielding a sensitivity region for the causal effect. While instructive from a mathematical perspective, this approach might seem overly complicated to practitioners, heavily relies on overidentification and potentially suffers from loss of power due to the combination of two tests. Wang et al. rely on an extension of the Anderson-Rubin test,

cf. Anderson and Rubin (1949), for one explanatory variable and one instrument. They derive a sensitivity interval based on an F-distribution with noncentrality parameter dependent on Δ. Despite the non-asymptotic validity, this approach is comparatively limited in scope due to the restriction on one-dimensional variables and has the bizarre property that the sensitivity sets $\Delta_1 = \{\delta^*\}$ and $\Delta_2 = [-\delta^*, \delta^*]$, for some δ^*, result in the same sensitivity interval.

3 Linear Instrumental Variables

This section introduces the linear Instrumental Variables model, as commonly used in econometrics, states established results on estimators and confidence intervals that are needed in the following sections and adds a sensitivity parameter to the model.

3.1 Instrumental Variables Model

We consider n independent, identically distributed data points indexed by $i \in \{1, \dots, n\}$. For every observation i, the outcome is denoted by $Y_i \in \mathbb{R}$, the explanatory variables by $D_i \in \mathbb{R}^p$, exogeneous covariates by $X_i \in \mathbb{R}^l$ and the instrumental variables by $Z_i \in \mathbb{R}^k$. We gather all data points in the vector $Y \in \mathbb{R}^n$ and the matrices $D \in \mathbb{R}^{n\times p}$, $X \in \mathbb{R}^{n\times l}$ and $Z \in \mathbb{R}^{n\times k}$. The matrix $W = [Z\colon X] \in \mathbb{R}^{n\times(k+l)}$ is given by concatenating Z and X. We define the projection matrix onto the column space of a matrix $M \in \mathbb{R}^{n\times r}$ by $P_M = M(M^TM)^{-1}M^T$ and the residual projection matrix by $Q_M = \mathrm{Id} - P_M$.
We define the IV model according to Wooldridge (2010).

Definition 3.1 (Single-outcome IV model)**.** The single-outcome Instrumental Variables model with i.i.d. data points is given by

$$\begin{gathered} Y = D\beta + X\psi + \varepsilon_Y, \qquad D = Z\Gamma + X\Psi + \varepsilon_D, \qquad \varepsilon = [\varepsilon_Y\colon \varepsilon_D], \\ \text{where} \quad \mathbb{E}[\varepsilon_i | Z_i, X_i] = 0,\ \mathrm{Var}(\varepsilon_i | Z_i, X_i) = \Sigma \quad \forall i \in \{1, \dots, n\}. \end{gathered} \tag{1}$$

The parameters of the model are $\beta \in \mathbb{R}^p$, $\psi \in \mathbb{R}^l$, $\Gamma \in \mathbb{R}^{p\times k}$, $\Psi \in \mathbb{R}^{p\times l}$ and $\Sigma \in \mathbb{R}^{(p+1)\times(p+1)}$, symmetric and positive definite.

Remark 3.1. We can add an additional column to the matrix X to include the intercept term in the model.

The equations in (1) are to be understood as assignemnts in the sense of Structural Equation Models (SEM), as depicted in Figure 2. It is equally possible to express the IV model in terms of the potential outcome/counterfactual framework; yet, the SEM perspective is more intuitive for our purposes. Intrinsic to the definition of the IV model is the assumption that there are no causal relationships apart from the ones specified. In particular, this means that the instruments Z only affect the outcome through D.

The Instrumental Variables method relies on three assumptions that are needed for the model to be identified and to conduct inference.

Assumptions 3.2.

(A1) $\mathbb{E}[Z^TQ_XD]$ has rank p.

(A2) $\mathbb{E}[W^TW]$ has rank $l + k$.

(A3) $\mathbb{E}[W^T\varepsilon] = 0$.

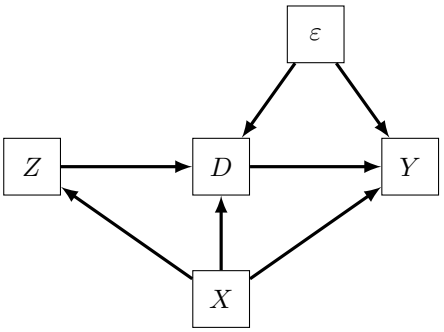

Figure 2: Schematic of the Instrumental Variables model.

Assumption (A1) states that conditional on the exogeneous covariates the instruments are associated with the explanatory variables. In practice, this assumption can be verified. For instance, an F-statistic can be used to test for the association between Z and D where a value of 10 is usually chosen as threshold between strong and weak instruments. Moreover, (A1) implies the rank condition, i.e. $k \geq p$, which demands at least as many instruments as explanatory variables. If $k > p$ holds, one refers to the model as *overidentified* whereas $k = p$ is called the *just identified* case. The second assumption is a mainly technical condition on the second moment of W which is usually fulfilled as the columns of W are almost always linearly independent if the sample size is large enough. Assumption (3) stipulates that both instruments and exogeneous covariates are uncorrelated with the structural error. In general, this assumption is untestable as ε is never observed. However, it can be partially tested via Sargan's test, cf. Sargan (1958), in an overidentified model. Under these three assumptions, the model is indeed identified; see Section 5.2 of Wooldridge (2010) for more details.
Since the main focus lies on inference on the causal effect of the explanatory variables D on the outcome Y, which is parametrised by β, we make use of the following theorem, according to Davidson and MacKinnon (1993), to simplify the model.

Theorem 3.3 (Frisch-Waugh-Lovell)**.** *Let $X_1 \in \mathbb{R}^{n \times k_1}$ and $X_2 \in \mathbb{R}^{n \times k_2}$. Then the ordinary least squares (OLS) estimates for B_2 in the regression models*

$$Y = X_1 B_1 + X_2 B_2 + \varepsilon \quad \text{and} \quad Q_{X_1} Y = Q_{X_1} X_2 B_2 + Q_{X_1} \varepsilon$$

are equal.

Therefore, we can simplify the IV model (1) projecting out the exogeneous covariates and obtain the system

$$\begin{gathered} Y^* = D^* \beta + \varepsilon_Y^*, \qquad D^* = Z^* \Gamma + \varepsilon_D^*, \qquad \varepsilon^* = [\varepsilon_Y^* : \varepsilon_D^*], \\ \text{where} \quad \mathbb{E}[\varepsilon_i^* | Z_i^*] = 0, \ \operatorname{Var}(\varepsilon_i^* | Z_i^*) = \Sigma \quad \forall i \in \{1, \ldots, n\}, \end{gathered} \tag{2}$$

where $Y^* = Q_X Y$, $D^* = Q_X D$, $Z^* = Q_X Z$ and $\varepsilon^* = Q_X \varepsilon$. In order to simplify the notation, we drop the superscript * in the following.

3.2 Estimators and Confidence Intervals

Most of the estimators used for Instrumental Variables models belong to the family of K-class estimators. We define them according to Davidson and MacKinnon (1993).

Definition 3.4 (K-class estimator)**.** Assume a linear IV model and let $K \in \mathbb{R}$. A K-class estimator for the causal effect β is given by

$$\hat{\beta}_K = \left(D^T (\mathrm{Id} - K Q_Z) D\right)^{-1} D^T (\mathrm{Id} - K Q_Z) Y. \tag{3}$$

K-class estimators are mainly popular as they are asymptotically consistent and normally distributed, cf. Amemiya (1985).

Theorem 3.5. *Let $\hat{\beta}_K$ be a K-class estimator. If $K - 1 = o(n^{-1/2})$ as $n \to \infty$, then $\hat{\beta}_K$ is asymptotically consistent and follows the distribution*

$$V^{-1/2}(\hat{\beta}_K - \beta) \xrightarrow{D} \mathcal{N}(0, \mathrm{Id}), \quad \text{as } n \to \infty,$$

where

$$V = \hat{\sigma}^2 \big(D^T(\mathrm{Id} - KQ_Z)D\big)^{-1}, \qquad \hat{\sigma}^2 = \frac{1}{n-l-p}(Y - D\hat{\beta}_K)^T(Y - D\hat{\beta}_K).$$

Setting $K = 0$ corresponds to the usual ordinary least squares (OLS) estimator, whereas $K = 1$ yields the two-stage least squares (TSLS) estimator which takes the form $\hat{\beta}_{TSLS} = (D^T P_Z D)^{-1} D^T P_Z Y$. Another popular choice is the limited information maximum likelihood (LIML) estimator for which K_{LIML} is given as the smallest eigenvalue of

$$(Y^T Q_Z Y)^{-1/2} Y^T Y (Y^T Q_Z Y)^{-1/2}.$$

Although the asymptotic properties of all K-class estimators are the same, the behaviour in finite samples differs. The TSLS estimator is efficient among all IV estimators that are linear in the instrumental variables Z, cf. Theorem 5.3 in Wooldridge (2010), and has as many finite moments as there are overidentification restrictions. However, TSLS tends to be biased towards OLS and is sensitive to weak instruments. The LIML estimator, on the other hand, does not have finite moments but is less prone to bias and weak instruments. To mitigate the shortcomings of LIML, Fuller (1977) proposed an estimator which uses $K = K_{\mathrm{LIML}} - \frac{b}{n-k-p}, b > 0$, which has moments if the sample is large enough. For a more in-depth discussion, we refer to Davidson and MacKinnon (1993).
Based on the asymptotic normality of K-class estimators, we can conduct hypothesis tests for β and construct confidence intervals/regions. For instance, a confidence interval for the j-th entry of β is given by

$$\Big((\hat{\beta}_K)_j - z_{1-\alpha/2}\sqrt{V_{jj}},\, (\hat{\beta}_K)_j + z_{1-\alpha/2}\sqrt{V_{jj}}\Big), \tag{4}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard normal distribution. Alternatively, we could also use the $1-\alpha/2$ quantile of a t-distribution with $n-p-k$ degrees of freedom. If we are interested in a confidence region for a subset of entries of β, we can use Hotelling's T^2 distribution.

3.3 Sensitivity Model

The assessment of the plausibility of Assumption (A1) was studied in many publications in the past. On the other hand, comparatively little research has been undertaken on Assumption (A3) and the implicit assumption that the model contains all causal pathways.
To this end, we expand the linear IV model (2) following Holland (1988). First, we insert a causal path from the instruments to the outcome parametrised by δ_1 and rename Γ

$$\begin{gathered} Y = D\beta + Z\delta_1 + \varepsilon_Y, \qquad D = Z\Gamma_1 + \varepsilon_D, \qquad \varepsilon = [\varepsilon_Y : \varepsilon_D], \\ \text{where} \quad \mathbb{E}[\varepsilon_i | Z_i] = 0,\ \mathrm{Var}(\varepsilon_i | Z_i) = \Sigma \quad \forall i \in \{1, \dots, n\}. \end{gathered}$$

Furthermore, we assume that Z and ε are linearly correlated which is a violation of (A3). Hence, we can express the error terms as $\varepsilon_Y = Z\delta_2 + \varepsilon'_Y$, $\varepsilon_D = Z\Gamma_2 + \varepsilon'_D$, where $(\varepsilon'_Y, \varepsilon'_D) \perp\!\!\!\perp Z$. Inserting these relationships in the model above, we obtain

$$Y = D\beta + Z\delta_1 + Z\delta_2 + \varepsilon'_Y, \qquad D = Z\Gamma_1 + Z\Gamma_2 + \varepsilon'_D, \qquad \varepsilon' = [\varepsilon'_Y : \varepsilon'_D],$$
$$\text{where} \quad \mathbb{E}[\varepsilon'_i | Z_i] = 0, \ \operatorname{Var}(\varepsilon'_i | Z_i) = \Sigma' \quad \forall i \in \{1, \dots, n\}.$$

We can now reduce the number of parameters by defining $\delta = \delta_1 + \delta_2$ and $\Gamma = \Gamma_1 + \Gamma_2$ and dropping the superscript $'$ for notational convenience. This yields the (linear) sensitivity model

$$\begin{aligned} Y = D\beta + Z\delta + \varepsilon_Y, \qquad D = Z\Gamma + \varepsilon_D, \qquad \varepsilon = [\varepsilon_Y : \varepsilon_D], \\ \text{where} \quad \mathbb{E}[\varepsilon_i | Z_i] = 0, \ \operatorname{Var}(\varepsilon_i | Z_i) = \Sigma \quad \forall i \in \{1, \dots, n\}. \end{aligned} \tag{5}$$

Remark 3.2. For a derivation of this sensitivity model from the perspective of potential outcomes we refer to Wang et al. (2018).

The model (5) encapsulates a violation of both the condition (A3) and/or the implicit assumption that the instruments do not directly act upon the outcome and is depicted in Figure 3. Such

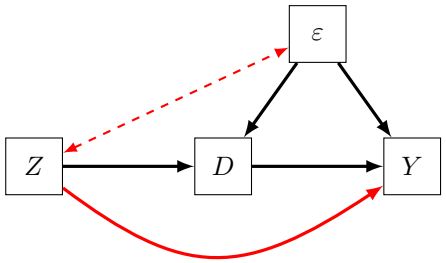

Figure 3: Schematic of the sensitivity model. The red error indicates direct influence of the instruments on the outcome. The dashed red error depicts correlation between the instrumental variables and the error term ε.

a model can be used to investigate the robustness of the findings of the Instrumental Variables method. To this end, a practitioner specifies a bounded *sensitivity set* $\Delta \subset \mathbb{R}^k$ which he deems large enough to cover the potential deviation from the assumptions parametrised by δ. This leaves the causal effect β *partially identified*, i.e. for any fixed δ the causal effect $\beta(\delta)$ and the remaining parameters are identified but, as we only assume that δ is constrained to Δ, only a region can be identified. Hence, we need a version of confidence intervals that handles coverage of the partially identified range of β-values with high probability.

Definition 3.6 (Sensitivity region)**.** Let $\alpha \in (0, 1)$, $\Delta \subset \mathbb{R}^k$ be bounded and introduce the abbreviation $\Pi = (\Gamma, \Sigma)$ for the respective parameters in the linear sensitivity model (5). We denote the distribution of the model by $\mathcal{F}_{\beta,\Pi,\delta}$ and call two sets of parameters (β, Π, δ) and (β', Π', δ') *observationally equivalent* if the corresponding distributions are equal which we denote by $\mathcal{F}_{\beta,\Pi,\delta} \simeq \mathcal{F}_{\beta',\Pi',\delta'}$. Any $1 - \alpha$ sensitivity region S_Δ for the sensitivity set Δ must satisfy

$$\inf_{\mathcal{F}_{\beta,\Pi,\delta} \simeq \mathcal{F}_{\beta_0,\Pi_0,\delta_0}} \mathbb{P}_{\beta_0,\Pi_0,\delta_0}\big(\beta \in S_\Delta\big) \geq 1 - \alpha, \quad \forall\, \beta_0, \Pi_0, \delta_0 \in \Delta. \tag{6}$$

Extending the usual definition of confidence intervals/regions by the concept of observational equivalence naturally arises from partially identified models. In these, several parameter configurations cannot be distinguished from the data and we consequently demand that the sensitivity interval/region covers the considered parameters of all configurations which are observationally equivalent to the truth with high probability.

4 Sensitivity Regions with the Union Method

One approach of constructing sensitivity regions is based on taking unions of confidence intervals corresponding to different values of the sensitivity parameter. This section explores this approach both relying on asymptotic normality and a bootstrap procedure.

4.1 Asymptotic Distribution

Theorem 3.5 provides an asymptotic distribution for K-class estimators which can be used to construct confidence regions/intervals, for example (4). This result can be equally used to build confidence regions for every $\delta \in \Delta$. To this end, we fix $\tilde{\delta} \in \Delta$ and subtract the $Z\tilde{\delta}$ term in the first equation of (5) and treat $Y - Z\tilde{\delta}$ as outcome in Theorem 3.5.
The work of Wang et al. (2018) follows the same idea but uses the Anderson-Rubin test, cf. Anderson and Rubin (1949), to construct confidence intervals as they focus on weak instruments. Moreover, they restrict themselves to one explanatory and one instrumental variable. We, instead, use the asymptotic normality result, assume no detrimental effects of weak instruments and allow for multiple instrumental variables. Furthermore, we only consider one explanatory variable and thus sensitivity intervals, as well.
Suppose that $(I^{(\delta)})_{\delta\in\Delta}$ is a family of $1-\alpha$ confidence intervals for $\beta(\delta), \delta \in \Delta$, we consider the union

$$\bigcup_{\delta\in\Delta} I^{(\delta)} \subset \Big[\inf_{\delta\in\Delta} I^{(\delta)}, \sup_{\delta\in\Delta} I^{(\delta)}\Big].$$

Zhao et al. (2018) prove the intuition that this set is indeed a $1-\alpha$ sensitivity interval.

Proposition 4.1. *Let $\alpha \in (0,1)$ and assume there exist data-dependent intervals $I^{(\delta)} = [L^{(\delta)}, U^{(\delta)}]$ such that*

$$\liminf_{n\to\infty} \mathbb{P}_{\beta_0,\Pi_0,\delta_0}\big(\beta_0 \in [L^{(\delta_0)}, U^{(\delta_0)}]\big) \geq 1-\alpha, \qquad \forall \beta_0, \Pi_0, \delta_0 \in \Delta.$$

Let $L_\Delta = \inf_{\delta\in\Delta} L^{(\delta)}$, $U_\Delta = \inf_{\delta\in\Delta} U^{(\delta)}$ and suppose the intervals are congruent, i.e. there exists $\alpha' \in [0,\alpha]$ such that for all $\beta_0, \Pi_0, \delta_0 \in \Delta$

$$\limsup_{n\to\infty} \mathbb{P}_{\beta_0,\Pi_0,\delta_0}\big(\beta_0 < L^{(\delta_0)}\big) \leq \alpha' \quad \text{and} \quad \limsup_{n\to\infty} \mathbb{P}_{\beta_0,\Pi_0,\delta_0}\big(\beta_0 > U^{(\delta_0)}\big) \leq \alpha - \alpha'.$$

Then $I_\Delta = [L_\Delta, U_\Delta]$ is an asymptotic $1-\alpha$ sensitivity interval.

Combining this result and Theorem 3.5, we can easily construct a sensitivity interval for β under the model (5).

Corollary 4.2. *Let $\alpha \in (0,1)$, assume the setting of Theorem 3.5 and define the δ-dependent estimates of the causal effect and its variance as*

$$\hat{\beta}_K(\delta) = \frac{D^T(\mathrm{Id} - KQ_Z)(Y - Z\delta)}{D^T(\mathrm{Id} - KQ_Z)D}, \quad V(\delta) = \frac{\|Y - Z\delta - D\hat{\beta}_K(\delta)\|_2^2}{(n-l-p)\, D^T(\mathrm{Id} - KQ_Z)D}.$$

An asymptotic $1-\alpha$ sensitivity interval for β is given by

$$I_\Delta = \left[\inf_{\delta\in\Delta} \hat{\beta}_K(\delta) - z_{1-\alpha/2}\sqrt{V(\delta)},\ \sup_{\delta\in\Delta} \hat{\beta}_K(\delta) + z_{1-\alpha/2}\sqrt{V(\delta)}\right]. \tag{7}$$

Proof. This follows directly from Theorems 3.5 and 4.1 as we take the $\alpha/2$ and $1-\alpha/2$ quantiles of the normal distribution for all confidence intervals, i.e. $\alpha' = \alpha/2$. □

Corollary 4.2 reduces the construction of sensitivity intervals to an optimisation problem that can be solved with standard algorithms if Δ is sufficiently regular.
For the TSLS estimator, K is fixed at 1 and thus independent of the sensitivity parameter. This allows to describe more precisely for which δ-values the upper and lower point of the sensitivity interval (7) are attained.

Lemma 4.3 (Sensitivity Interval for TSLS)**.** *In the setting of Corollary 4.2, the end points of the interval I_Δ are obtained for δ-values on the boundary of Δ, that is*

$$I_\Delta = \left[\inf_{\delta \in \partial\Delta} \hat{\beta}_{TSLS}(\delta) - z_{1-\alpha/2}\sqrt{V(\delta)},\ \sup_{\delta \in \partial\Delta} \hat{\beta}_{TSLS}(\delta) + z_{1-\alpha/2}\sqrt{V(\delta)} \right].$$

Proof. In order to facilitate notation, we introduce the abbreviations

$$c_1 = \frac{D^T P_Z Y}{D^T P_Z D}, \quad d^T = \frac{D^T P_Z Z}{D^T P_Z D}, \quad c_2 = (n-l-p)\, D^T(\mathrm{Id} - KQ_Z)D,$$
$$a = \big(\mathrm{Id} - D(D^T P_Z D)^{-1} D^T P_Z\big)Y, \quad B = \big(\mathrm{Id} - D(D^T P_Z D)^{-1} D^T P_Z\big)Z$$

to shorten the expression of $\hat{\beta}_K(\delta) = c_1 - d^T\delta$ and $V = \|a - B\delta\|_2^2/c_2$. Hence, we can write the upper and lower boundary point of the δ-dependent $1-\alpha$ confidence interval as

$$C(\delta) = c_1 - d^T\delta \pm z_{1-\alpha/2}\ c_2^{-1/2}\|a - B\delta\|_2.$$

In order to prove the lemma, we show that the Hessian matrix of C as a function of δ is positive [negative] semi-definite for the upper [lower] boundary point. Hence, C is convex [concave] and attains its maximum [minimum] on $\partial\Delta$. First, we compute the gradient

$$\nabla C(\delta) = -d \pm z_{1-\alpha/2}\ c_2^{-1/2}\frac{B^T B\delta - B^T a}{\|a - B\delta\|_2},$$

and then the Hessian

$$\begin{aligned}\mathrm{H}(\delta) &= \pm\, z_{1-\alpha/2}\ c_2^{-1/2}\|a - B\delta\|_2^{-3}\Big[\|a - B\delta\|_2^2 B^T B - (B^T B\delta - B^T a)(B^T B\delta - B^T a)^T\Big] \\ &= \pm\, z_{1-\alpha/2}\ c_2^{-1/2}\|a - B\delta\|_2^{-3} B^T\Big[\|a - B\delta\|_2^2\,\mathrm{Id} - (a - B\delta)(a - B\delta)^T\Big]B.\end{aligned}$$

Since $z_{1-\alpha/2}$, $c_2^{-1/2}$ and $\|a - B\delta\|_2^{-3}$ are non-negative scalars, it suffices to show that the term in brackets is positive semi-definite. We abbreviate $v = a - B\delta$, let $w \in \mathbb{R}^n$ be an arbitrary vector and prove positive semi-definiteness as follows

$$w^T(v^T v\,\mathrm{Id} - vv^T)w = v^T v\, w^T w - (w^T v)^2 = \|v\|_2^2\|w\|_2^2 - \langle v, w\rangle^2 \geq 0,$$

where we apply the Cauchy-Schwarz inequality in the last step. This concludes the proof. □

This result provides a more precise description of the δ-values that lead to the upper and lower end point of I_Δ. In settings with moderate or many instruments, however, this insight will not accelerate the optimisation procedure significantly as the dimensions of Δ and its boundary only differ by one.

4.2 Percentile Bootstrap

In the previous section, we relied on asymptotic normality to construct confidence intervals. This naturally raises the question how good the approximation is when the sample size is comparatively small or some instruments are weak. Hence, we also consider the percentile bootstrap method that was introduced by Zhao et al. (2018) to the field of sensitivity analysis. We would like to highlight that we havn't yet investigated theoretical guarantees for its validity but would like to introduce it as another potential method nonetheless.
The given dataset is denoted by $D = (Z_i, D_i, Y_i)_{i=1}^n$ and let $\hat{D}_1, \dots, \hat{D}_B$ be i.i.d. resamples. For any fixed $\delta \in \Delta$, let $\hat{\beta}(\delta)$ be the estimate of the causal effect based on D and $\hat{\hat{\beta}}(\delta) = \{\hat{\hat{\beta}}_b(\delta)\}_{b=1}^B$ be the estimates computed from the resampled data. The percentile bootstrap confidence interval is then given by

$$I^{(\delta)} = [L^{(\delta)}, U^{(\delta)}] = \big[Q_{\alpha/2}(\hat{\hat{\beta}}(\delta)), Q_{1-\alpha/2}(\hat{\hat{\beta}}(\delta))\big],$$

where $Q_\alpha(\hat{\hat{\beta}}(\delta))$ is the α-percentile of the bootstrap distribution. As shown in the previous section, we can obtain a sensitivity interval by maximising [minimising] the upper [lower] boundary point of $I^{(\delta)}$. A key observation of Zhao et al. (2018) is the fact that interchanging optimisation and quantiles leads to a conservative sensitivity interval that can be computed using standard optimisation routines, i.e.

$$\begin{aligned} I_\Delta &= \Big[\inf_{\delta\in\Delta} Q_{\alpha/2}\Big(\hat{\hat{\beta}}(\delta)\Big), \sup_{\delta\in\Delta} Q_{1-\alpha/2}\Big(\hat{\hat{\beta}}(\delta)\Big)\Big] \\ &\subseteq \Big[Q_{\alpha/2}\Big(\Big\{\inf_{\delta\in\Delta} \hat{\hat{\beta}}_b(\delta)\Big\}_{b=1}^B\Big), Q_{1-\alpha/2}\Big(\Big\{\sup_{\delta\in\Delta} \hat{\hat{\beta}}_b(\delta)\Big\}_{b=1}^B\Big)\Big]. \end{aligned}$$

This approach is attractive since it does not require knowledge of the distribution of the involved estimator and promises a potentially superior behaviour for small and moderate sample sizes. Nonetheless, further investigation is needed to establish theoretical guarantess and to assess the method in practice.

5 Sensitivity Regions with Constrained Statistical Inference

This section introduces the duality of hypothesis testing and sensitivity regions and recalls relevant results of the field of constrained statistical inference. Based on these considerations, we formulate three different approaches to likelihood ratio tests (non-adaptive, adaptive and split) and express the IV sensitivity model in this framework.

5.1 Inversion of Tests

The duality of confidence intervals and hypothesis tests is a well-known relationship in the statistical literature, cf. Lehmann and Romano (2005). We demonstrate how it can be also used to construct sensitivity regions.
Suppose we can test H_0: $\beta = \beta^*$ against H_1: $\beta \neq \beta^*$ at a given level α for every value of β^* in presence of the nuisance parameters Π and δ. More precisely, we demand

$$\mathbb{P}_{\beta^*,\Pi,\delta}(D \in A(\beta^*)) \geq 1 - \alpha, \quad \forall\, \Pi, \delta \in \Delta,$$

where D is one sample from the model and $A(\beta^*)$ denotes the acceptance region of the test. If we define the sensitivity region by

$$S_\Delta(d) = \{\beta^* \colon d \in A(\beta^*)\}, \text{ then} \qquad \beta^* \in S_\Delta(d) \Leftrightarrow d \in A(\beta^*)$$

holds and hence

$$\mathbb{P}_{\beta,\Pi,\delta}(\beta \in S_\Delta(D)) \geq 1-\alpha, \quad \forall\, \beta, \Pi, \delta \in \Delta.$$

Finally, dropping D and taking an infimum over the observationally equivalent distributions, we obtain the definition of a sensitivity region

$$\inf_{\mathcal{F}_{\beta,\Pi,\delta} \simeq \mathcal{F}_{\beta_0,\Pi_0,\delta_0}} \mathbb{P}_{\beta,\Pi,\delta}\big(\beta \in S_\Delta\big) = \inf_{\mathcal{F}_{\beta,\Pi,\delta} \simeq \mathcal{F}_{\beta_0,\Pi_0,\delta_0}} \mathbb{P}_{\beta_0,\Pi_0,\delta_0}\big(\beta \in S_\Delta\big) \geq 1-\alpha, \quad \forall\, \beta_0, \Pi_0, \delta_0 \in \Delta.$$

Remark 5.1. There are no requirements on the test procedures that are used as long as the type-I error control is satisfied. Hence, we could, in theory, use different tests for different values of β^*.

The procedure outlined above allows us to reduce the task of constructing sensitivity regions to finding suitable tests.

5.2 Constrained Statistical Inference

We briefly introduce the theory of constrained statistical inference relying on Silvapulle and Sen (2005). We present it in a general setting where the parameters of the model are $\theta \in \Theta$; in the following section we apply these results to the linear IV sensitivity model. Since we make use of likelihood ratio statistics, we require additional assumptions. In particular, we need to assume that the family of distributions is known.

5.2.1 Likelihood Ratio Setting

Definition 5.1. Let $X^{(1)}, \ldots, X^{(n)}$ be independently and identically distributed random variables with common probability density function $f(x;\theta), \theta \in \Theta \subseteq \mathbb{R}^p$, where x can be uni- or multivariate. The *log-likelihood* and the entries of the *Fisher information matrix* for one observation are defined by

$$\begin{aligned} \ell_n(\theta) &= \sum_{i=1}^{n} \log f(X^{(i)};\theta), \\ [\mathcal{I}(\theta)]_{i,j} &= \mathrm{E}\left[\left(\frac{\partial}{\partial\theta_i}\log f(X;\theta)\right)\left(\frac{\partial}{\partial\theta_j}\log f(X;\theta)\right)\middle|\theta\right], \quad i,j \in \{1,\ldots,p\}, \end{aligned}$$

respectively. Let $\Theta^* \subset \Theta$. The *maximum likelihood estimator* $\hat{\theta}_{\Theta^*}$ is given by

$$\hat{\theta}_{\Theta^*} = \operatorname*{argmax}_{\theta \in \Theta^*} \ell_n(\theta).$$

We require the following regularity conditions.

Assumptions 5.2. In the setting of Definition 5.1, the following is assumed.

(1) $\hat{\theta}_{\Theta^*}$ is $\sqrt{n}$-consistent whenever the true parameter θ_0 is contained in Θ^*.

(2) Distinct values of θ correspond to distinct distributions.

(3) The first three partial derivatives of $\log f(x;\theta)$ with respect to θ exist almost everywhere.

(4) There exists a $G(y)$ such that $\int G(y)dy < \infty$ and the absolute values of the first three partial derivatives of $\log f(x;\theta)$ with respect to θ are bounded by $G(y)$ in a neighbourhood of θ_0.

(5) The Fisher information matrix $\mathcal{I}(\theta)$ is finite and positive definite.

These conditions are not minimal but facilitate developing the theory without concerning oneself with technical details.

Definition 5.3. Assume the framework stated above and let $\Theta_0 \subseteq \Theta_1 \subseteq \Theta$ be nested models. The *likelihood ratio statistic* λ_n for a sample of size n is defined as

$$\lambda_n = 2\left(\sup_{\theta\in\Theta_1} \ell_n(\theta) - \sup_{\theta\in\Theta_0} \ell_n(\theta)\right) = 2\left(\ell_n(\hat{\theta}_{\Theta_1}) - \ell_n(\hat{\theta}_{\Theta_0})\right).$$

In the well-studied case where Θ_0 is a linear subspace and $\Theta_1 = \Theta$, that is the general alternative, the limiting distribution of the likelihood ratio statistic is known.

Theorem 5.4. *Let $R \in \mathbb{R}^{m\times p}$ and $r = \operatorname{rank}(R)$. Testing $H_0 : R\,\theta = 0$ against $H_1 : \theta \in \Theta$ implies $\lambda_n \xrightarrow{D} \chi^2_r$ as $n \to \infty$ under the null hypothesis, where χ^2_d denotes the Chi-squared distribution with d degrees of freedom.*

If the hypothesis to be tested cannot be represented as a linear space, the limiting distribution of λ_n is more intricate. It involves the local geometry at θ_0, which is expressed in the tangent cone and its regularity.

Definition 5.5. Let $\Theta^* \subseteq \mathbb{R}^p$ and $\theta_0 \in \Theta^*$. The *tangent cone* to Θ^* at θ_0, denoted by $\mathcal{T}(\Theta^*; \theta_0)$, is the set of all vectors w for which a sequence of positive numbers $(t_n)_{n\in\mathbb{N}}$ converging to zero and $(\theta_n)_{n\in\mathbb{N}} \subset \Theta^*$ converging to θ_0 exist such that

$$t_n^{-1}(\theta_n - \theta_0) \to w, \quad \text{as } n \to \infty. \tag{8}$$

Θ^* is *Chernoff-regular* if for all its elements and all such $(t_n)_{n\in\mathbb{N}}$ a corresponding $(\theta_n)_{n\in\mathbb{N}}$ can be found such that (8) holds.

Remark 5.2. The tangent cone $\mathcal{T}(\Theta^*; \theta_0)$ is closed and a cone in the sense that, if $w \in \mathcal{T}(\Theta^*; \theta_0)$ holds, then $\lambda w \in \mathcal{T}(\Theta^*; \theta_0)$ for all $\lambda > 0$.

The importance of the concept of Chernoff-regularity was first discovered by Chernoff (1954) and was later interpreted in the context of different definitions of approximating cones, see e.g. Geyer (1994). While it is not immediately obvious how the upper definition can be verified for a given hypothesis Θ^*, Drton (2009) establishes Chernoff-regularity for a wide class of spaces.

Lemma 5.6. *If $\Theta^* \subseteq \mathbb{R}^p$ is a semi-algebraic set, i.e. a finite union of sets defined by polynomial equations and inequalities, then Θ^* is Chernoff-regular everywhere.*

Furthermore, under the Mangasarian-Fromowitz constraint qualification (MF-CQ) and continuous differentiability we can directly compute $\mathcal{T}(\Theta^*; \theta_0)$ with the following proposition.

Proposition 5.7. *Suppose that $\Theta \subseteq \mathbb{R}^p$ is open and let Θ^* be given by*

$$\Theta^* = \{\theta \in \Theta \colon h_1(\theta) = \ldots = h_l(\theta) = 0, h_{l+1}(\theta) \geq 0, \ldots, h_k(\theta) \geq 0\},$$

where $h_1, \ldots, h_k$ are continuously differentiable. Let $\theta_0 \in \Theta^*$ and let $a_i = (\partial/\partial\theta)h_i(\theta_0)$ for $i = 1, \ldots, k$ and $J(\theta_0) = \{i \colon h_i(\theta_0) = 0, l+1 \leq i \leq k\}$. Assume that the MF-CQ is satisfied at θ_0, i.e. there exists a non-zero $b \in \mathbb{R}^p$ such that $a_1^T b = \ldots = a_l^T b = 0$, $a_1, \ldots, a_l$ are linearly independent and $a_i^T b > 0$ for $i \in J(\theta_0)$. Then $\mathcal{T}(\Theta^*; \theta_0)$ is equal to

$$\left\{\theta \in \mathbb{R}^p \colon a_i^T\theta = 0 \ \ \forall\, i = 1, \ldots, l;\ a_i^T\theta \geq 0 \ \ \forall\, i \in J(\theta_0)\right\}.$$

Hence, an inequality condition only effects the tangent cone if it is *active* for θ_0, that is holds with equality for θ_0.
We consider the general testing problem H_0: $\theta \in \Theta_0$ against H_1: $\theta \in \Theta_1$ for nested models $\Theta_0 \subseteq \Theta_1 \subseteq \Theta$. Let θ_0 be the true parameter and define the norm $\|\cdot\|$ on $\mathbb{R}^p$ as

$$\|x\| = \sqrt{x^T \mathcal{I}(\theta_0) x}, \quad \text{for } x \in \mathbb{R}^p.$$

For $\Theta^* \subseteq \mathbb{R}^p$ and $x \in \mathbb{R}^p$, we use the abbreviation

$$\|x - \Theta^*\| = \inf_{\theta \in \Theta^*} \|x - \theta\|.$$

Moreover, we introduce the Wald statistic which is given by

$$W_n = \min_{\theta \in \Theta_0} n(\hat{\theta} - \theta)^T \mathcal{I}(\hat{\theta})(\hat{\theta} - \theta) - \min_{\theta \in \Theta_1} n(\hat{\theta} - \theta)^T \mathcal{I}(\hat{\theta})(\hat{\theta} - \theta), \tag{9}$$

where $\hat{\theta}$ is the unrestricted MLE.

5.2.2 Non-adaptive Likelihood Ratio Test

Theorem 5.8. *Let $\Theta_0 \subseteq \Theta_1 \subseteq \Theta$ be nested models and assume that Θ is open. If the null hypothesis holds and Θ_0 is Chernoff-regular at θ_0, then*

$$\lambda_n \xrightarrow{D} \|Z - \mathcal{T}(\Theta_0; \theta_0)\|^2 - \|Z - \mathcal{T}(\Theta_1; \theta_0)\|^2, \qquad \text{as } n \to \infty, \tag{10}$$

where $Z \sim \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$. In addition, the likelihood ratio and Wald statistic have the same distribution as $n \to \infty$, i.e. $\lambda_n = W_n + o_p(1)$.

This result gives a precise characterisation of the asymptotic distribution. Nonetheless, it remains dependent on the true parameter θ_0 and thus cannot be characterised explicitely. Under further assumptions on Θ_0 and Θ_1, however, we can describe the distribution in more detail. We turn our attention to testing problems that use the general alternative, i.e. $\Theta_1 = \Theta$, which causes the second term in (10) to vanish. If Θ_0 fulfils additional conditions, Wolak (1989) and Silvapulle and Sen (2005) state the following result.

Theorem 5.9. *Let $h^{(1)}(\theta)$ and $h^{(2)}(\theta)$ be continuously differentiable, vector-valued functions which characterise the null hypothesis in the testing problem*

$$H_0 : \theta \in \Theta_0 = \{\theta \colon h^{(1)}(\theta) \geq 0,\, h^{(2)}(\theta) = 0\} \quad \textit{against} \quad H_1 : \theta \in \Theta \subseteq \mathbb{R}^p.$$

Assume that θ_0 lies on the boundary of Θ_0 and that the MF-CQ is fulfilled. Let a_i denote $(\partial/\partial\theta) h_i^{(1)}(\theta_0)$ and let $\{j_1, \ldots, j_m\}$ denote $\{i \colon h_i^{(1)}(\theta_0) = 0\}$. Set $H^{(1)}(\theta_0) = (a_{j_1}, \ldots, a_{j_m})^T$ and $m = m(\theta_0) = \operatorname{rank}(H^{(1)}(\theta_0))$. If $h^{(2)}$ is not specified, set $r = 0$; otherwise,
$H^{(2)}(\theta_0) = \nabla h^{(2)}(\theta_0)$ and assume that the matrix has full row-rank $r = \operatorname{rank}(H^{(2)}(\theta_0))$. If $r + m \leq p$, then

$$\lambda_n \,|\, \theta = \theta_0 \xrightarrow{D} \sum_{i=0}^{m} w_{m-i}(m, V(\theta_0))\, \chi^2_{r+i}, \quad \text{as } n \to \infty, \tag{11}$$

where

$$\begin{aligned} V(\theta_0) &= H^{(1)}(\theta_0) \mathcal{I}(\theta_0)^{-1} H^{(1)}(\theta_0)^T \\ &\quad - H^{(1)}(\theta_0) \mathcal{I}(\theta_0)^{-1} H^{(2)}(\theta_0)^T (H^{(2)}(\theta_0) \mathcal{I}(\theta_0)^{-1} H^{(2)}(\theta_0)^T)^{-1} H^{(2)}(\theta_0) \mathcal{I}(\theta_0)^{-1} H^{(1)}(\theta_0)^T. \end{aligned}$$

If $h^{(2)}$ is not specified, the second summand above is zero. $\{w_k\}_{k \in \{0,\ldots,m\}}$ are positive weights such that $\sum_{k=0}^m w_k = 1$ and χ^2_d denotes a Chi-squared distribution with d degrees of freedom.

The asymptotic distribution (11) is a mixture of χ^2-random variables and is also referred to as Chi-bar-squared distribution and denoted by $\bar{\chi}^2(\Theta_0, V)$. It is possible to derive several properties of its weights $w_i(q, V)$, see further Silvapulle and Sen (2005), and for small q closed form representations are known, e.g. $w_0(1, V) = w_1(1, V) = 0.5$. However, for $q \geq 5$, no closed formulas are available. Hence, we cannot use the additional structure of the asymptotic distribution given by (11) and need to resort to simulating data according to (10) to obtain the desired quantiles.
In general, the weights and thus the limit distribution depend on θ_0 through the matrix $V(\theta_0)$ and the rank $m(\theta_0)$. Intuitively, this issue seems to be resolvable by inserting an asymptotically consistent estimator of θ_0. Yet, the weights and degrees of freedom of the χ^2-terms are functions of the active inequalities and consequently depend on θ_0 in a discontinuous fashion. Hence, the limit distribution can depend on the path on which an asymptotically consistent estimator converges to θ_0. For this reason, we use the p-value

$$\sup_{\theta \in \Theta_0} \sum_{i=0}^{m(\theta)} w_{m(\theta)-i}(m(\theta), V(\theta))\, \mathbb{P}(\chi^2_{r(\theta)+i} \geq t), \tag{12}$$

where all θ-dependencies are made explicit and t is the observed value of the likelihood ratio or Wald statistic. This secures that, regardless of θ_0, the type-I error is not exceeded. The value(s) of θ for which the supremum is attained is called *least favourable null*. It is immediately clear that this value cannot lie in the interior of Θ_0. Otherwise, the tangent cone at θ would be $\mathbb{R}^p$ which induces the limiting statistic (10) to be zero almost surely.

The approach of taking a supremum over the entire null hypothesis set as in (12) can be very conservative if the true parameter value is far away from the least favourable null. Therefore, we also consider a more granular technique that was proposed by Silvapulle (1996) in the context of hypothesis testing with nuisance parameters.
In order to simplify notation, a data sample from the sensitivity IV model is denoted by X, the θ-dependent likelihood ratio or Wald statistic is referred to as T_θ, let $t_\theta(X)$ be its observed value and denote the true parameter value θ_0. Consequently, the p-value is given by

$$p_0(X) = \mathbb{P}_{\theta_0}(T_{\theta_0} \geq t_{\theta_0}(X)|X).$$

Since θ_0, however, is unknown, we might use the p-value (12), i.e.

$$p(X) = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T_\theta \geq t_\theta(X)|X),$$

to control the type-I error at a specified significance level α. In order to mitigate the risk of overly conservative tests, Silvapulle (1996) proposes a two-step procedure: First, a $1 - \alpha_1$ confidence region $C(\alpha_1)$ for θ_0 is constructed, where $0 < \alpha_1 < \alpha$. Second, the p-value is computed as

$$p^*(X) = \alpha_1 + \sup_{\theta \in C(\alpha_1)} \mathbb{P}_\theta(T_\theta \geq t_\theta(X)|X). \tag{13}$$

This method indeed controls the type-I error.

Proposition 5.10. *For a given $\alpha \in (0, 1)$ and $0 < \alpha_1 < \alpha$, rejecting H_0 if $p^*(X) \leq \alpha$ ensures that the type-I error does not exceed α.*

Proof. We estimate the type-I error as follows

$$\begin{aligned}
\mathbb{P}_{\theta_0}(p^*(X) \leq \alpha) &= \mathbb{P}_{\theta_0}\left(p^*(X) \leq \alpha, \theta_0 \in C(\alpha_1)\right) + \mathbb{P}_{\theta_0}(p^*(X) \leq \alpha, \theta_0 \notin C(\alpha_1)) \\
&\leq \mathbb{P}_{\theta_0}\Big(\sup_{\theta \in C(\alpha_1)} \mathbb{P}_\theta(T_\theta \geq t_\theta(X)|X) \leq \alpha - \alpha_1, \theta_0 \in C(\alpha_1)\Big) + \alpha_1 \\
&\leq \mathbb{P}_{\theta_0}\left(\mathbb{P}_{\theta_0}(T_{\theta_0} \geq t_{\theta_0}(X)|X) \leq \alpha - \alpha_1\right) + \alpha_1 \\
&= \mathbb{P}_{\theta_0}(p_0(X) \leq \alpha - \alpha_1) + \alpha_1 \\
&= (\alpha - \alpha_1) + \alpha_1 = \alpha.
\end{aligned}$$

The two inequalities are a direct consequence of the monotonicity of the probability measure and the last equality ensues from the uniform distribution of p-values, as stated in Lehmann and Romano (2005). □

The outlined procedure potentially improves the straightforward approach of using p as p-value because the supremum only considers values of θ that are realistic in view of the data. The envisaged type-I error rate α is divided up into α_1, the error probability for the confidence region of θ_0, and $\alpha_2 = \alpha - \alpha_1$, the maximal value of $\sup_{\theta \in C(\alpha_1)} \mathbb{P}_\theta(T_\theta \geq t_\theta)$ which H_0 is rejected for.

In the important special case where the parameter space of the null hypothesis Θ_0 is shaped by linear equations and inequalities and the Fisher information is independent of the parameter θ we can better characterise the least favourable null value(s). Without loss of generality we assume that the constraints are well-defined and non-redundant.

Proposition 5.11. *Let $\Theta_0 = \{\theta \in \mathbb{R}^p \colon A^{(1)}\theta \geq b^{(1)}, A^{(2)}\theta = b^{(2)}\}$, define the set of active indices for a specific θ by $I_\theta = \{i \in \{1, \dots, p\} \colon (A^{(1)}\theta)_i = b_i^{(1)}\}$ and the set of maximally active θ-values by $M_{\Theta_0} = \{\theta \in \Theta_0 \colon \nexists\, \theta' \text{ s.t. } I_\theta \subsetneq I_{\theta'}\}$. If we use the general alternative and the Fisher information does not depend on θ, then*

$$\operatorname*{argmax}_{\theta_0 \in \Theta_0} \lim_{n \to \infty} \mathbb{P}(T_n \geq t \mid \theta = \theta_0) \subset M_{\Theta_0},$$

where T_n is either the likelihood ratio or Wald statistic and t is its observed value.

Proof. We prove the proposition by contradiction. Therefore, let θ_m be one of the maximisers and assume that there exists a θ^* such that $I_{\theta_m} \subsetneq I_{\theta^*}$. Since we use the general alternative, according to Theorem 5.8

$$\lambda_n|\theta = \theta_0 \xrightarrow{D} \|X - \mathcal{T}(\Theta_0; \theta_0)\|^2, \quad \text{as } n \to \infty,$$

where $X \sim \mathcal{N}(\theta_0, \mathcal{I}^{-1})$ and the norm is induced by the Fisher information matrix. Since the constraints are assumed to be well-defined and non-redundant, the rows of $[(A^{(1)})^T \colon (A^{(2)})^T]^T$ are linearly independent and the Mangasarian-Fromowitz constraint qualification is fulfilled. Therefore, we can apply Proposition 5.7 to derive expressions for the tangent cones $\mathcal{T}(\Theta_0; \theta_m)$ and $\mathcal{T}(\Theta_0; \theta^*)$:

$$\begin{aligned}
\mathcal{T}(\Theta_0; \theta_m) &= \{\theta \colon \mathrm{e}_i^T A^{(1)}\, \theta \geq 0 \;\forall\, i \in I_{\theta_m},\, A^{(2)}\theta = 0\}, \\
\mathcal{T}(\Theta_0; \theta^*) &= \{\theta \colon \mathrm{e}_i^T A^{(1)}\, \theta \geq 0 \;\forall\, i \in I_{\theta^*},\, A^{(2)}\theta = 0\}.
\end{aligned}$$

Hence, we see that $\mathcal{T}(\Theta_0; \theta^*) \subsetneq \mathcal{T}(\Theta_0; \theta_m)$ as $I_{\theta_m} \subsetneq I_{\theta^*}$ and, consequently, obtain

$$\|X - \mathcal{T}(\Theta_0; \theta_m)\|^2 < \|X - \mathcal{T}(\Theta_0; \theta^*)\|^2,$$

which is a contradiction and concludes the proof. □

The set of maximally active θ's does not only include values that achieve the maximum number of active inequalities but also includes values such that no other θ can be active for the same and an additional inequality. More intuitively, M_{Θ_0} contains the vertices of the polytope Θ_0. This result greatly facilitates finding the supremum both in (12) and (13) because we only need to consider a finite number of θ-values.

5.2.3 Adaptive Likelihood Ratio Test

Recently, Al Mohamad et al. (2020) proposed an approach to constrained statistical inference that circumvents the complexity and intractability of Chi-bar-squared distributions. While the method is developed for a rather narrow setting, we expect that extensions are easy to obtain due to the similarity to the already established theory. This, however, is still object of future research. Here we present the already proven results.
Let $X \sim \mathcal{N}(\theta, V)$, where $\theta \in \mathbb{R}^p$ and $V \in \mathbb{R}^{p\times p}$ is a positive definite matrix, and define the norm $\|\cdot\|_V$ by $\|x\|_V^2 = x^T V^{-1} x$. We consider the cone $\Theta_0 = \{\theta \in \mathbb{R}^p \colon A\theta \geq 0\}$ and denote its polar cone $\Theta_0^\circ = \{x \in \mathbb{R}^p \colon x^T V^{-1}\theta \leq 0 \ \forall\, \theta \in \Theta_0\}$. Moreover, there exists a collection of faces of Θ_0°, $\{F_1, \dots, F_K\}$, such that their realtive interiors, $\{\mathrm{ri}(F_1), \dots, \mathrm{ri}(F_K)\}$, form a partition of the polar cone.
The projection of x onto Θ_0 with respect to the norm $\|\cdot\|_V$ is denoted by $\Pi_V(y|\Theta_0)$. If we conduct the hypothesis test H_0: $\theta \in \Theta_0$ against H_1: $\theta \in \Theta$, then it can be shown that the likelihood ratio and Wald statistic are given by

$$W_n = \lambda_n = \|\Pi_V(V|\Theta_0^\circ)\|_V^2 = \sum_{j=1}^{K} \mathbb{1}_{\{\Pi_V(X|\Theta_0^\circ) \in \mathrm{ri}(F_j)\}} \|P_j X\|_V^2,$$

where P_j is the projection matrix onto the linear space spaned by F_j. Based on this finding, the classical, non-adaptive approach proceeds to determine the distribution for every face and then average over the probabilities that the projection lies in the interior of the given faces. Thus, a mixture of chi-squared distributions emerges. Al Mohamad et al. (2020), on the other hand, follow an adaptive approach.
Let $\alpha \in (0, 1)$ and denote the $1-\alpha$ quantile of a Chi-squared distributions with d degrees of freedom by q_d. The adaptive critical value is given by

$$q(\Theta_0, X, \alpha) = \sum_{j=1}^{K} \mathbb{1}_{\{\Pi_V(X|\Theta_0^\circ) \in \mathrm{ri}(F_j)\}} q_{r_j},$$

where $r_j = \mathrm{rank}(P_j)$, and the rejection region is $\{X \colon \|\Pi_V(X|\Theta_0^\circ)\|_V^2 > q(\Theta_0, X, \alpha\}$. Using this critical value indeed controls the type-I error at α.

Theorem 5.12 (Theorem 1, Al Mohamad et al. (2020))**.** *Let $X \sim \mathcal{N}(\theta, V)$ and let Θ_0 be a polyhedral cone. A valid test at level α for H_0: $\theta \in \Theta_0$ against H_1: $\theta \in \Theta$ is given by the rejection region*

$$\{X \colon \|\Pi_V(X|\Theta_0^\circ)\|_V^2 > q(\Theta_0, X, \alpha\}.$$

Moreover,

$$\mathbb{P}_{\theta\in\Theta_0}(\lambda > q(\Theta_0, X, \alpha)) \leq [1 - \mathbb{P}_{\theta\in\Theta_0}(X \in \Theta_0)]\, \alpha.$$

This result is interesting inasmuch as it neither involves often intractable weights nor requires finding the least favourable null. As long as the face of the polyhedral cone, which X is projected onto, and the rank of the projection matrix can be determined, the test can easily be carried out, even in higher dimensions.

Theorem 5.12 requires strong assumptions but we think that it is possible to relax them and still obtain asymptotic type-I error control. For example, Al Mohamad et al. themselves prove that replacing V by a consistent estimate yields asymptotic validity. We might generalise the result to random variables that are asymptotically normal, such as the MLE, and consider Chernoff-regular sets Θ_0 instead of polyhedral cones.

5.2.4 Split Likelihood Ratio Test

Wasserman et al. (2020) introduced in their recent paper a new method for universal inference that is based on the likelihood ratio framework and sample splitting. Let $\{\mathbb{P}_\theta \colon \theta \in \Theta\}$ be a general statistical model and let $X_1, \ldots, X_n$ be an independent and identically distributed data sample. We split the data into two folds $D^{(1)} = \{X_1, \ldots, X_m\}$ and $D = \{X_{m+1}, \ldots, X_n\}$, where $1 \leq m \leq n$, for example $m = \lfloor n \rfloor / 2$. (To our knowledge, a heuristic for a good choice of m has not been proposed yet.) Let the log-likelihoods corresponding to $D^{(1)}$ and $D^{(2)}$ be denoted by $\ell^{(1)}$ and $\ell^{(2)}$, respectively, let $\hat{\theta}^{(1)}$ be the unrestricted MLE based on the first fold and let $\Theta_0 \subset \Theta$ be an arbitrary set.
For the hypothesis $\mathrm{H}_0 \colon \theta \in \Theta_0$ against $\mathrm{H}_1 \colon \theta \in \Theta$, we define the split likelihood ratio statistic as

$$U_n = \ell^{(2)}(\hat{\theta}^{(2)}_{\Theta_0}) - \ell^{(2)}(\hat{\theta}^{(1)})$$

where $\hat{\theta}^{(2)}_{\Theta_0}$ is the MLE based on the data $D^{(2)}$ constrained to Θ_0. We can also define the analogous statistic for swapping the role of $D^{(1)}$ and $D^{(2)}$, i.e.

$$\tilde{U}_n = \ell^{(1)}(\hat{\theta}^{(1)}_{\Theta_0}) - \ell^{(1)}(\hat{\theta}^{(2)}).$$

For a specified significance level α, we reject the null hypothesis if $U_n < \log(\alpha)$ and $(U_n + \tilde{U}_n)/2 < \log(\alpha)$, respectively, which yields type-I error control at the desired level.

Theorem 5.13 (Theorem 3, Wasserman et al. (2020))**.** *Let $\alpha \in (0, 1)$. Then,*

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta\big(U_n < \log(\alpha)\big) \leq \alpha, \qquad \sup_{\theta \in \Theta_0} \mathbb{P}_\theta\big((U_n + \tilde{U}_n)/2 < \log(\alpha)\big) \leq \alpha.$$

This result is remarkable as it does not require knowledge of the finite or asymptotic distribution of the likelihood ratio statistic and instead only relies on optimisation. Hence, as long as we are able to compute the MLE under the general alternative and under the null hypothesis, we obtain a valid test. This makes this approach especially useful for settings with a high-dimensional and/or complicated Θ_0.
However, the outcome of the test is dependent on the split ratio and the data points included in each fold. In general, this method is conservative as knowledge of the model is not used. For this reason, we use the test obtained via the split likelihood ratio statistic as a benchmark for the other methods.

5.3 Application to Linear IV Sensitivity Model

The linear sensitivity model (5) is overparametrised and thus only partially identifiable. For this reason, we also consider the reduced model

$$\begin{aligned} Y = Z\rho + \tilde{\varepsilon}_Y, \qquad D = Z\Gamma + \varepsilon_D, \qquad \varepsilon = [\tilde{\varepsilon}_Y \colon \varepsilon_D], \\ \text{where} \quad \mathbb{E}[\varepsilon_i | Z_i] = 0, \ \mathrm{Var}(\varepsilon_i | Z_i) = \Sigma \quad \forall i \in \{1, \ldots, n\}, \end{aligned} \tag{14}$$

which emerges from inserting the first stage equation for D into the second stage equation for the outcome Y. This model is a vanilla linear regression, hence identifiable and linked to the

original model via the relationship $\rho = \delta + \Gamma\beta$. A similar relationship also holds for the covariance matrix. However, we suppress this issue in the notation as no restrictions arise from this and Σ can be replaced by a consistent estimate as explained below.
Moreover, we make the assumption that the error terms follow a Gaussian distribution. On the one hand, this, of course, is a strong restriction, on the other hand, many practitioners, particularly econometricians, are willing to make this assumption. Moreover, we believe that the steps of our proposal can be extended to error distributions that satisfy the Assumptions 5.2 as our approach is based on the theory developed in Section 5.2.

We introduce some notation and state the log-likelihood. Denote the space of covariance matrices $\mathcal{M} = \{\Sigma \in \mathbb{R}^{(p+1)\times(p+1)} \colon \Sigma = \Sigma^T, \Sigma \text{ positive definite}\}$ and let $\widehat{\Sigma}$ be the empirical version of Σ. The log-likelihood is given by

$$\ell(\Gamma, \rho, \Sigma) = -\frac{n}{2}(p+1)\log(2\pi) - \frac{n}{2}\log\det(\Sigma) - \frac{1}{2}\sum_{i=1}^{n}\Big([Y_i \colon D_i] - Z_i[\rho \colon \Gamma]\Big)\Sigma^{-1}\Big([Y_i \colon D_i] - Z_i[\rho \colon \Gamma]\Big)^T,$$

where the observed values $Y_i, D_i, Z_i, i \in \{1, \dots, n\}$, are scalars and row vectors, respectively. To abbreviate notation, we set $\theta = (\rho, \operatorname{vec}(\Gamma)) \in \mathbb{R}^{k+kp}$. Hence, for any two sets $\Theta_0 \subset \Theta_1 \subset \mathbb{R}^{k+kp}$ the likelihood ratio statistic for the corresponding hypothesis test is given by

$$\lambda_n = 2\left(\sup_{\theta\in\Theta_1, \Sigma\in\mathcal{M}} \ell(\theta, \Sigma) - \sup_{\theta\in\Theta_0, \Sigma\in\mathcal{M}} \ell(\theta, \Sigma)\right).$$

We also consider the Wald statistic

$$\begin{aligned} W_n = &\inf_{\theta\in\Theta_0, \Sigma\in\mathcal{M}} n\Big(\big(\hat{\theta}, \operatorname{vec}(\widehat{\Sigma})\big) - \big(\theta, \operatorname{vec}(\Sigma)\big)\Big)^T \mathcal{I}(\hat{\theta}, \widehat{\Sigma})\Big(\big(\hat{\theta}, \operatorname{vec}(\widehat{\Sigma})\big) - \big(\theta, \operatorname{vec}(\Sigma)\big)\Big) \\ &- \inf_{\theta\in\Theta_1, \Sigma\in\mathcal{M}} n\Big(\big(\hat{\theta}, \operatorname{vec}(\widehat{\Sigma})\big) - \big(\theta, \operatorname{vec}(\Sigma)\big)\Big)^T \mathcal{I}(\hat{\theta}, \widehat{\Sigma})\Big(\big(\hat{\theta}, \operatorname{vec}(\widehat{\Sigma})\big) - \big(\theta, \operatorname{vec}(\Sigma)\big)\Big), \end{aligned}$$

where $\mathcal{I}(\hat{\theta}, \widehat{\Sigma})$ is the Fisher information. It exhibits the following block structure

$$\mathcal{I}(\hat{\theta}, \widehat{\Sigma}) = \begin{pmatrix} \widehat{\Sigma}^{-1} \otimes Z^T Z & 0 \\ 0 & \mathcal{I}(\widehat{\Sigma}) \end{pmatrix},$$

where $\otimes$ denotes the Kronecker product. The infimum in the Wald statistic is taken over $\mathcal{M}$ in both the first and second term. Due to the structure of the Fisher information the θ- and Σ-related terms can be separated and the latter consequently cancel. Therefore, we obtain the simplified formula

$$W_n = \inf_{\theta\in\Theta_0} n(\hat{\theta} - \theta)^T \widehat{\Sigma}^{-1} \otimes Z^T Z(\hat{\theta} - \theta) - \inf_{\theta\in\Theta_1} n(\hat{\theta} - \theta)^T \widehat{\Sigma}^{-1} \otimes Z^T Z(\hat{\theta} - \theta).$$

This shows that the Wald statistic for testing the nested models Θ_0 and Θ_1 does not involve the parameter space of the covariance matrix. As the likelihood ratio and the Wald statistic have the same asymptotic distribution according to Theorem 5.8, this finding also holds true for λ_n as $n \to \infty$.

Following the idea of inversion of tests as described in Section 5.1, we need to construct a family of hypothesis tests for different values β_0 of the parameter β. Our proposal is based on the observation that the condition $\delta \in \Delta$ poses restrictions on the parameters θ of the reduced model through the relationship $\rho = \delta + \Gamma\beta$. Moreover, if we conduct the test $\mathrm{H}_0 \colon \beta = \beta_0$ against

H_1: $\beta \neq \beta_0$, the set of θ-values associated with the null hypothesis is further restricted. More precisely, the general parameter space, the space associated with the alternative and the null hypothesis are given by

$$\begin{aligned}\Theta &:= \{(\Gamma, \rho) \in \mathbb{R}^{k\times p} \times \mathbb{R}^k\},\\ \Theta_1 &:= \{(\Gamma, \rho) \in \mathbb{R}^{k\times p} \times \mathbb{R}^k \mid \exists\, \delta \in \Delta\, \exists\, \beta \in \mathbb{R}^p \colon \rho = \delta + \Gamma\beta\},\\ \Theta_0 &:= \{(\Gamma, \rho) \in \mathbb{R}^{k\times p} \times \mathbb{R}^k \mid \exists\, \delta \in \Delta \colon \rho = \delta + \Gamma\beta_0\},\end{aligned} \tag{15}$$

where we identified θ with (ρ, Γ). We can express Θ_1 and Θ_0 in terms of projections instead of existential quantifiers. To this end, we introduce the linear space spanned by the columns of Γ as $\operatorname{span}(\Gamma)$ and its orthocomplement as $\operatorname{span}(\Gamma)^\perp$, hence $\mathbb{R}^k = \operatorname{span}(\Gamma) \oplus \operatorname{span}(\Gamma)^\perp$. We decompose ρ as follows

$$\rho = P_\Gamma \rho + Q_\Gamma \rho = P_\Gamma(\Gamma\beta + \delta) + Q_\Gamma(\Gamma\beta + \delta) = \Gamma\beta + P_\Gamma\delta + Q_\Gamma\delta.$$

For the parameter space of the alternative hypothesis, the $P_\Gamma\rho$ component is unrestricted as β can take any value in $\mathbb{R}^p$; the $Q_\Gamma\rho$ component, however, can only take values in $Q_\Gamma\Delta = \{Q_\Gamma\,\delta \colon \delta \in \Delta\}$. This condition also holds for Θ_0 but additionally $P_\Gamma\rho - \Gamma\beta_0 \in P_\Gamma\Delta = \{P_\Gamma\delta \colon \delta \in \Delta\}$ must be fulfilled. Thus, we can express the parameter spaces as

$$\begin{aligned}\Theta_1 &:= \{(\Gamma, \rho) \in \mathbb{R}^{k\times p} \times \mathbb{R}^k \mid Q_\Gamma\rho \in Q_\Gamma\Delta\}\\ \Theta_0 &:= \{(\Gamma, \rho) \in \mathbb{R}^{k\times p} \times \mathbb{R}^k \mid Q_\Gamma\rho \in Q_\Gamma\Delta, P_\Gamma\rho - \Gamma\beta_0 \in P_\Gamma\Delta\}.\end{aligned}$$

This highlights the additional restriction under the null hypothesis.

In view of the parameter spaces (15), we can conduct three different hypothesis tests.
First test: Θ_1 vs. Θ According to Theorem 5.8 the asymptotic distribution of the likelihood ratio statistic for this test is given by $\|\hat\theta - \mathcal{T}(\Theta_1; \theta_0)\|^2$, where $\hat\theta$ is the unrestricted MLE. We assess whether the restriction on the parameters $\theta = (\rho, \Gamma)$, which stem from the definition of the sensitivity model, are sensible compared to the general alternative. In this sense, it is similar to Sargan's test but more general as the latter was not developed for sensitivity models. Therefore, we suspect that the power against many alternatives is small. Moreover, this test can only be applied in an overidentified setting as $\Theta_1 = \Theta$ holds for $k = p$.
Second test: Θ_0 vs. Θ_1 The asymptotic distribution of the likelihood ratio statistic is $\|\hat\theta - \mathcal{T}(\Theta_0; \theta_0)\|^2 - \|\hat\theta - \mathcal{T}(\Theta_1; \theta_0)\|^2$. This test tacitly presumes that the sensitivity model is correctly specified. Therefore, we must intially conduct the first test and only proceed when the null hypothesis, that the sensitivity model is reasonable, was not rejected. Its acceptance equals the existence of combinations of parameter values that are close to the truth. The second test then assesses more precisely which values of β correspond to those. In order to make the notation more acurate, we add a subscript to the null hypothesis parameter space to indicate the value of β_0, i.e. Θ_{0,β_0}. We note that the sensitivity interval based on the second test (assuming the first test did not reject) is never empty as $\Theta_1 = \cup_{\beta_0=-\infty}^{\infty}\Theta_{0,\beta_0}$. If the sensitivity model is well-specified, the $\|\hat\theta - \mathcal{T}(\Theta_1; \theta_0)\|^2$ term of the asymptotic distribution is small which renders the second test similar to the third.
Third test: Θ_0 vs. Θ The asymptotic distribution of the likelihood ratio statistic is given by $\|\mathcal{T}(\Theta_0; \theta_0)\|^2$. This test directly assesses if the model with parameter combinations corresponding to β_0 is reasonable. In this respect, we test both the sensitivity model and the parameter assumption, whereas the first test only checks the former and the second test only checks the latter. As a consequence, we may get an empty sensitivity region. This either indicates that the sensitivity set Δ is not large enough or that the model itself should be doubted.

We argue that the third approach is more suitable than the second one to construct a family of hypothesis tests and thus a sensitivity region. Since the second test demands conducting a prior test, which we suspect to have little power against many alternatives, we need to control for multiple testing. Moreover, the characterisation of the distribution of $\|\hat{\theta} - \mathcal{T}(\Theta_0; \theta_0)\|^2 - \|\hat{\theta} - \mathcal{T}(\Theta_1; \theta_0)\|^2$ is more intricate than using the general alternative in the third test, cf. Theorem 5.9 and Theorem 5.12. From a practical perspective the second approach bears risks, as well. Practitioners might be tempted to skip testing the plausibility of the sensitivity model via the first test and directly proceed to the second test. Thus, they always obtain a sensitivity region even if the model is ill-specified. For these reasons, we advocate using the general alternative, that is the third test.

In order to use the two-step procedure for the non-adaptive test and compute a potentially superior p-value (13), we need to construct confidence regions for the parameters $\theta = (\rho, \Gamma)$. If the assumptions of Proposition 5.11 are, however, fulfilled, it is not necessary to explicitly state the confidence region. Instead, it suffices to only test which of the finitely many candidates for the least favourable null are contained in the confidence region.
In order to facilitate notation, we introduce $B = [\rho \colon \Gamma] \in \mathbb{R}^{k \times (1+p)}$ and $\tilde{Y} = [Y \colon D] \in \mathbb{R}^{n \times (1+p)}$ for the respective variables and parameters of the reduced form model (14). We assume that $n - k \geq p + 1$ and define

$$\hat{B} = Z(Z^T Z)^{-1} \tilde{Y}, \qquad \tilde{\Sigma} = \tilde{Y}^T Q_Z \tilde{Y}, \qquad A = \hat{B}^T Z^T Z \hat{B}.$$

The literature on multivariate regression models with multiple outcomes, also referred to as multivariate analysis of variance (MANOVA), provides four established methods to test H_0: $B = 0$ against H_1: $B \neq 0$. The hypothesis $B = B_0$ can be tested in the same way by simply replacing $\tilde{Y}$ with $\tilde{Y} - ZB_0$. We refer to the eigenvalues of $A\tilde{\Sigma}^{-1}$ as $\lambda_1, \ldots, \lambda_{p+1}$ and list commonly used statistics, relying on Eaton (1983) and references therein

$$\begin{aligned}
&\text{Wilks' Lambda:} && \Lambda_{\text{Wilks}} = \prod_{j=1}^{p+1} \frac{1}{1+\lambda_j} = \det(\tilde{\Sigma}) / \det(\tilde{\Sigma} + A), \\
&\text{Hotelling-Lawley trace:} && \Lambda_{\text{HL}} = \sum_{j=1}^{p+1} \lambda_j = \operatorname{tr}(A\tilde{\Sigma}^{-1}), \\
&\text{Pillai trace:} && \Lambda_{\text{Pillai}} = \prod_{j=1}^{p+1} \frac{\lambda_j}{1+\lambda_j} = \operatorname{tr}(A(\tilde{\Sigma} + A)^{-1}), \\
&\text{Roy's greatest root:} && \Lambda_{\text{Roy}} = \max_{j=1,\ldots,p+1} \lambda_j.
\end{aligned}$$

In general, the distribution of these statistics cannot be derived but approximations are known. For instance, the R function `summary.manova` fits an F-distribution and computes the p-value based on this approximation. As long as we only need to test a moderate number of null hypotheses, this approach is feasible. However, the construction of confidence regions can only be achieved by testing different values of B on a potentially high-dimensional grid as, except for Roy's greatest root, none of the tests can be inverted.

6 Sensitivity set

This section investigates classes of sensitivity sets that are interesting for practitioners aiming to facilitate their specification for users of our method. This section is still subject to ongoing

research.

6.1 Ellipsoid

We assume that the sensitivity set is given by $\Delta = \{\delta \in \mathbb{R}^k \colon (\delta - v)^T A(\delta - v) \leq 1\}$, where $v \in \mathbb{R}^k$ and $A \in \mathbb{R}^{k \times k}$ is a positive definite matrix. This is the general formula of a geometric body bounded by an ellipsoid which is centred at v. An important special case occurs when $v = 0$ and A is a diagonal matrix: The ellipsoid is centred at the origin, its principal axes coincide with the coordinate axes and the largest extension in any dimension $j \in \{1, \ldots, k\}$ is given by $\pm 1/\sqrt{a_{jj}}$.
The parameter space associated with the null hypothesis is given by $\Theta_0 = \{(\Gamma, \rho) \in \mathbb{R}^{k \times p} \times \mathbb{R}^k \colon (\rho - \Gamma\beta_0)^T A(\rho - \Gamma\beta_0) \leq 1\}$. Since the constraint is not linear, Al Mohamad et al.'s method is infeasible; however the classical, non-adaptive approach is remarkably easy as there is only one inequality involved. Hence, according to Theorem 5.9, the limit distribution of the Wald and likelihood ratio statistic is given by a mixture of two χ^2's where the weights are $w_0(1, V(\theta_0)) = w_1(1, V(\theta_0)) = 0.5$. Since these are the same for all $\theta_0 \in \Theta_0$, we can avoid the usual difficulty of finding the least favourable null and directly obtain $0.5\chi^2_0 + 0.5\chi^2_1$ as limit distribution.
Moreover, the two-step procedure only makes sense in settings where the asymptotic distribution depends on the true parameter value. Therefore, we do not apply it for sensitivity sets described by ellipsoids.
We tacitly assumed that the instruments are continuous random variables. While categorical IV's are possible, see further Section 6.2.3, it is not immediately clear how ellipsoidal constraints can be formulated.
An ellipsoidal sensitivity set is by definition symmetric. For instance, in the case of a diagonal matrix A and centre at the origin, every component of the vector δ can violate the IV assumptions in the positive direction as much as in the negative direction. We can relax this property of ellipsoids by defining the sensitivity region in a piecewise fashion via multiple ellipsoids, see for example Figure 4. This does not alter the theoretical result on the limit distribution, if the surface of the sensitivity set remains continuously differentiable.

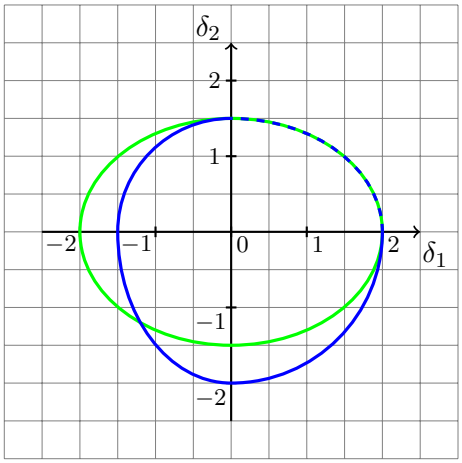

Figure 4: Ellipsoid (green), Piecewise sensitivity set defined by different ellipsoids in each quadrant (blue).

6.2 Polytope

Polytopes are a natural class of geometric bodies to define sensitivity sets, both from a theoretical and practical point of view. Unlike ellipsoids, however, the limit distributions that are obtained

from Theorem 5.9 typically involve higher degrees of freedom.

6.2.1 Representation of Polytopes

We summarise relevant results from the theory of polyhedra and in particular polytopes using Fukuda (2020) as reference.

Definition 6.1. Let $A \in \mathbb{R}^{m \times d}, b \in \mathbb{R}^m$. Then, the set

$$P = \{x \in \mathbb{R}^d \colon Ax \leq b\}$$

is called *convex polyhedron*. A bounded convex polyhedron is called *convex polytope*.

The subsequent considerations extend to convex polyhedra; however, we concentrate on convex polytopes as we assume the sensitivity set to be bounded. Moreover, we drop the specifier 'convex' in the following.
Polytopes can be either defined as an intersection of hyperplanes (H-representation) or as convex combination of its vertices (V-representation). We make these notions precise and establish their equivalence.

Theorem 6.2 (Minkowski-Weyl)**.** *For $w_1, \dots, w_q \in \mathbb{R}^d$, define the convex hull of the set $S = \{w_1, \dots, w_q\}$ as*

$$\operatorname{conv}(S) = \left\{ \sum_{j=1}^{q} \lambda_j w_j \colon \sum_{j}^{q} \lambda_j = 1, \lambda_j \geq 0 \, \forall \, j \in \{1, \dots, q\} \right\}.$$

Let $P \subset \mathbb{R}^d$. Then the following statements are equivalent:

(a) P is a polytope, i.e. there exist $A \in \mathbb{R}^{m \times d}, b \in \mathbb{R}^m$ such that $P = \{x \in Ax \leq b\}$ and P is bounded.

(b) P is bounded and finitely generated, i.e. there exist vectors $v_1, \dots, v_s \in \mathbb{R}^d$ such that $P = \operatorname{conv}(\{v_1, \dots, v_s\})$.

The H-representation corresponds to statement (a) as every inequality describes a hyperplane. The V-representation, on the other hand, equals statement (b) where the vectors $v_1, \dots, v_s$ are the vertices of the polytope.
While we know of the theoretical equivalence of both representations, there is no closed-form solution to obtain one from the other for a general polytope. The conversion from (a) to (b) is known as vertex enumeration problem, whereas the conversion in the opposite direction is called the facet enumeration problem. Yet, for some special classes of polytopes the conversion in both directions can be expressed explicitely, as we see in the next section.

6.2.2 Continuous Instruments

The theory of constrained statistical inference can be applied to all kinds of polytopes. However, practitioners might be particularly interested in k-cubes and k-cross polytopes, where k is the dimension of the instrument vector. For every single continuous instrument Z_j the user can stipulate a range of values $[\delta_l^{(j)}, \delta_u^{(j)}]$ which she expects to contain the true value. We refer to

this specification as range definition. The corresponding k-cube and k-cross polytope are then given as

$$\text{Cube}(k) = \text{conv}\left(\bigtimes_{j=1}^{k}\{\delta_l^{(j)}, \delta_u^{(j)}\}\right) = \bigtimes_{j=1}^{k}[\delta_l^{(j)}, \delta_u^{(j)}] = \left\{\delta \in \mathbb{R}^k \colon \begin{pmatrix} \text{Id} \\ -\text{Id} \end{pmatrix} \delta \leq \begin{pmatrix} \delta_u \\ -\delta_l \end{pmatrix}\right\},$$

$$\text{Cross}(k) = \text{conv}\left(\{\delta_u^{(j)} \mathrm{e}_j, \delta_l^{(j)} \mathrm{e}_j : j \in \{1, \dots, k\}\right) = \left\{\delta \in \mathbb{R}^k \colon a^T \delta \leq 1 \, \forall \, a \in \bigtimes_{j=1}^{k}\{1/\delta_l^{(j)}, 1/\delta_u^{(j)}\}\right\}.$$

Remark 6.1. Unlike our more general definition, the k-cube and k-cross polytope are usually referred to the geometric bodies that ensue from setting $\delta_l^{(j)} = -1$ and $\delta_u^{(j)} = 1$ for all $j \in \{1, \dots, k\}$. In this case, they correspond to the k-dimensional unit balls of the L^∞- and L^1-norm, respectively.

As we can see from the upper definitions, the V-representation of the cube requires 2^k vertices and the H-representation of the cross polytope requires 2^k hyperplanes. Therefore, the range definition is convenient for practitioners as only two values have to be specified for each instrument.

Moreover, cubes and cross polytopes defined in this way can be easily interpreted. A k-cube is specified componentwise, that is constraints for different instruments do not interact. Thus, we allow that maximal deviations from the usual IV assumptions can occur simultaneously for all instruments, e.g. $\delta_j = \delta_u^{(j)} \, \forall \, j \in \{1, \dots, k\}$. The k-cross polytope, on the other hand, can be viewed as linear interpolation between worst-case deviations for individual instruments: A maximal deviation from the IV assumptions in one instrument can only occur when all other instrumental variables are valid. Thus, we assume a restriction on the overall deviation from the IV assumption. Ellipsoids, which correspond to the unit ball of the L^2-norm for $A = \text{Id}, v = 0$, are a compromise between cubes and cross polytopes. We refer to Figure 5 for an example with two instruments.

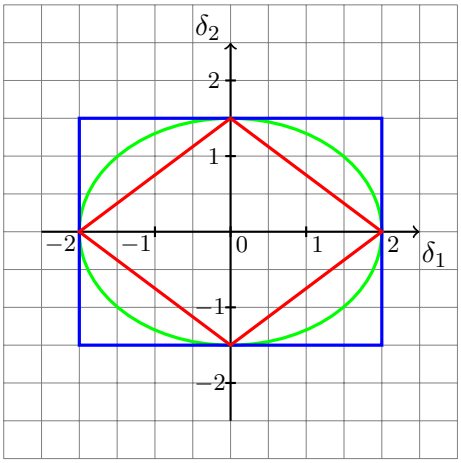

Figure 5: Sensitivity sets for two continuous instruments corresponding to the ranges $[\delta_l^{(1)}, \delta_u^{(1)}] = [2, -2]$ and $[\delta_l^{(2)}, \delta_u^{(2)}] = [1.5, -1.5]$: 2-cube (blue), 2-cross polytope (red), ellipsoid (green).

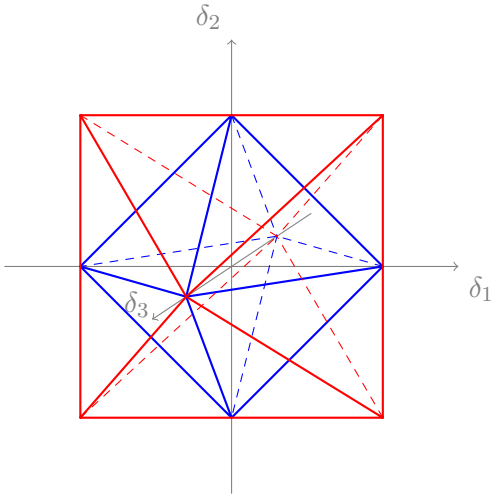

Figure 6: Two sensitivity sets corresponding to the ranges $[\delta_l^{(j)}, \delta_u^{(j)}] = [-2, 2] \, \forall \, j \in \{1, 2, 3\}$: 3-cross polytope (blue), combination of 2-cube, for δ_1 and δ_2, and cross-polytope for δ_3 (red).

In some applications, it can be desirable to specify a sensitivity set that combines the properties of cubes and cross polytopes. For instance, a researcher may want to allow for the simultaneous

occurence of worst-case deviations for the first and second instrument but put a constraint on the overall deviation of these and the third instrument, cf. Figure 6. For this kind of polytopes we have not yet derived the H- and V-representation that ensues from the range defintion. We suspect that the point of view of bi-pyramids, cf. Grünbaum (2003), can be helpful to this end. Besides, while we can provide an interpretation of different shapes of the sensitivity set, we still need to find heuristics on how to choose the ranges in order to give guidance on the application in real-world datasets to practitioners.

6.2.3 Categorical Instruments

Categorical covariates in regression models are usually expressed via one-hot enconding, cf. Fahrmeir et al. (2013). For instance, assuming that an instrument $\tilde{Z}$ has l categories $c_1, \dots, c_l$, we encode the data by adding $l-1$ dummy instruments $\mathbb{1}_{\{\tilde{Z}=c_1\}}, \dots, \mathbb{1}_{\{\tilde{Z}=c_{l-1}\}}$. Here the last category c_l is called reference category and is not added in order to avoid linear dependence, or rather the value of $\mathbb{1}_{\{\tilde{Z}=c_l\}}$ can be inferred from the other dummies.
Unlike continuous instruments, $l-1$ entries of the δ-vector correspond to the same categorical instrument. This requires additional considerations on a good definition of a sensitivity set. In particular, we aim to achieve a charactisation that is invariant to the choice of reference category and resembles that the $(l-1)$-dimensional set describes a single instrument, i.e. the influence of different dummy variables on the outcome cannot differ too much. For $\eta > 0$, we propose the sensitivity set

$$\Delta_\eta = \left\{\delta \in \mathbb{R}^{l-1} \colon |\delta_j| \leq \eta, |\delta_j - \delta_{j'}| \leq \eta \ \forall\, j, j' \in \{1, \dots, l-1\}\right\}. \tag{16}$$

Proposition 6.3. *The sensitivity set Δ_η as defined in (16) is invariant to the choice of reference category.*

Proof. Work in progress. □

In the case of an instrument with three or four levels/categories, we can depict the sensitivity set Δ_η, see Figures 7 and 8.

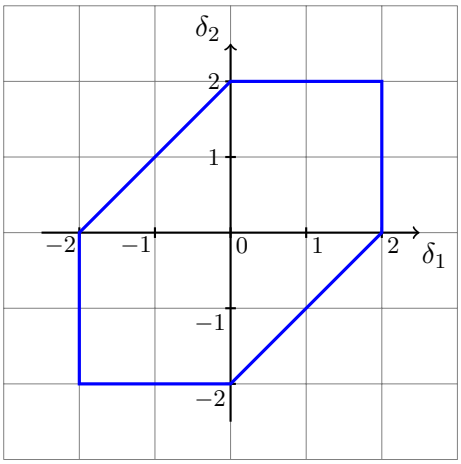

Figure 7: Sensitivity set Δ_2 for a 3-level categorical instrument.

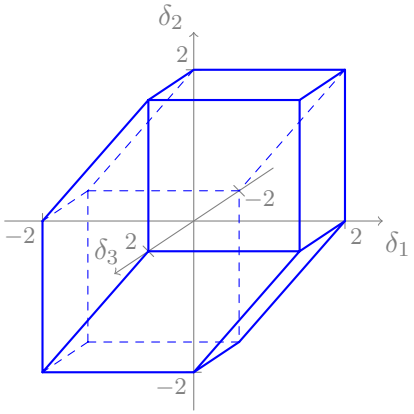

Figure 8: Sensitivity set Δ_2 for a 4-level categorical instrument.

These plots strongly hint that Δ_η can be expressed as Minkowski sum.

Proposition 6.4. *For arbitrary $P, Q \subset \mathbb{R}^m$, the Minkowski sum of P and Q is defined as $P + Q = \{p + q \colon p \in P,\, q \in Q\}$. The sensitivity region Δ_η can be expressed as Minkowski sum*

$$\Delta_\eta = \mathop{+}_{j=1}^{l-1} \big[0, \eta\, \mathrm{e}_j\big] + \big[0, -\eta 1\big],$$

where e_j is the j-th unit vector and 1 has the entry 1 at every component.

Proof. Work in progress. □

We hope that this characterisation helps to deduce the V-representation and combine the sensitivity sets of several categorical and continuous instrumental variables.

6.3 Combinations

Thus far, we have considered special cases of sensitivity sets in a mostly isolated way. Yet, some applications, especially those with many instruments, require more complex sensitivity sets. Above all, the case of multiple categorical and/or continuous instruments is relevant for researchers, for example Figure 9. Moreover, combining ellipsoidal and polyhedral constraints can also be of interest, see Figure 10. Both the definition of these sensitivity sets and the derivation of the limit distribution is not straightforward. Therefore, it is advisable to assess which degree of generalisation is actually helpful for practitioners.

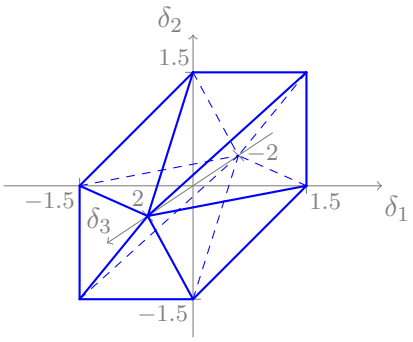

Figure 9: Sensitivity set for one 3-level categorical and one continuous instrument.

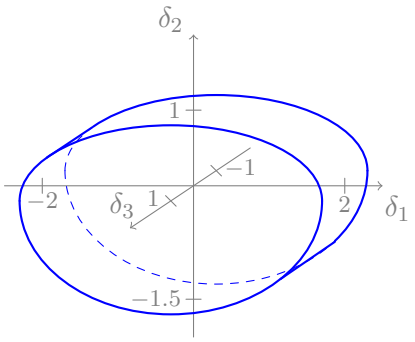

Figure 10: Sensitivity set for three continuous instruments with ellipsoidal and polyhedral constraints.

7 Future Work and Outlook

As this research project is still evolving, we would like to outline the next steps, both short- and mid-term, and give some comments for future work in this and related areas.
First, we aim to complete the missing proofs that were pointed out and finish the characterisation of a class of sensitivity sets that can accommodate for most applications. Subsequently, we implement the discussed methods, i.e. union method via optimisation or percentile bootstrap and inversion of tests with non-adaptive likelihood ratio (LR) test (one- and two-step), adaptive LR test and split-LR test, and compare their empirical performance. Initially, we use simulated data to verify the statistical properties; then, we apply them to real-world datasets focusing more on sensitivity sets.
If our approach appears to be successful, we plan to implement an R-package to make it easily applicable for practitioners. Moreover, we could consider generalising our framework, e.g.

allowing for simultaneous equations models or using splines as first-stage estimator to capture non-linear relationships better. Regarding non-parametric IV, Kernel Instrumental Variables proposed by Singh et al. (2020) appear particularly interesting as kernel methods provide the demanded generality but also maintain enough mathematical structure such that goodness-of-fit tests like the Kernelised Stein Discrepancy (KSD), cf. Liu et al. (2016), can be conducted. Nonetheless, constructing sensitivity sets in this set-up remains a difficult undertaking and we are unsure if basic ideas from our current approach carry over.

Conducting this research project, we found two spots where we would profit from additional theoretical results. Most noteably, the method proposed by Al Mohamad et al. (2020) relies on strong assumptions such as normality and parameter sets given by polyhedral cones. While we are hopeful that the findings hold true asymptotically when we consider asymptotically normal random variables and Chernoff-regular sets, the development of this theory certainly requires time and effort. In addition, we wonder whether the existing theory on constrained statistical inference has a semiparametric generalisation. If so, the techniques of our approach can potentially be generalised to semiparametric Instrumental Variables. On a more general note, the core of constrained statistical inference theory relies on the characterisation of the distribution $T(Y)|Y \in S$, where Y is a Gaussian random variable, S is a polyhedron and T is the squared norm of the projection of Y onto the polar cone of S. One of the prominent approaches to post-selection inference, cf. Lee et al. (2016), is based on the very same idea; only a different function T is used. It might be interesting to investigate if other areas of statistics also use the same idea and we can leverage cross-field synergies.

References

Al Mohamad, Diaa, Erik W Van Zwet, Eric Cator, and Jelle J Goeman (May 2020). "Adaptive critical value for constrained likelihood ratio testing". In: *Biometrika* 107.3, pp. 677–688.

Amemiya, Takeshi (1985). *Advanced Econometrics*. Harvard University Press.

Anderson, T. W. and Herman Rubin (1949). "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations". In: *The Annals of Mathematical Statistics* 20.1, pp. 46–63.

Ashley, Richard (2009). "Assessing the credibility of instrumental variables inference with imperfect instruments via sensitivity analysis". In: *Journal of Applied Econometrics* 24.2, pp. 325–337.

Ashley, Richard and Christopher Parmeter (2015). "Sensitivity analysis for inference in 2SLS/GMM estimation with possibly flawed instruments". In: *Empirical Economics* 49.4, pp. 1153–1171.

Chernoff, Herman (1954). "On the distribution of the likelihood ratio." In: *Annals of Mathematical Statistics* 25, pp. 573–578.

Conley, Timothy G., Christian B. Hansen, and Peter E. Rossi (2012). "Plausibly Exogenous". In: *The Review of Economics and Statistics* 94.1, pp. 260–272.

Davidson, Russell and James G. MacKinnon (1993). *Estimation and Inference in Econometrics*. OUP Catalogue. Oxford University Press.

DiPrete, Thomas A. and Markus Gangl (2004). "Assessing Bias in the Estimation of Causal Effects: Rosenbaum Bounds on Matching Estimators and Instrumental Variables Estimation with Imperfect Instruments". In: *Sociological Methodology* 34.1, pp. 271–310.

Drton, Mathias (2009). "Likelihood ratio tests and singularities." In: *The Annals of Statistics* 37.2, pp. 979–1012.

Eaton, Morris L (1983). *Multivariate statistics: a vector space approach*. John Wiley & Sons, Inc., 605 Third Ave., New York, NY 10158, USA.

Fahrmeir, Ludwig, Thomas Kneib, Stefan Lang, and Brian Marx (2013). *Regression. Models, methods and applications*. Berlin: Springer, pp. xiv + 698.

Fukuda, Komei (2020). *Polyhedral computation*. Tech. rep. Department of Mathematics, Institute of Theoretical Computer Science ETH Zurich.

Fuller, Wayne A. (1977). "Some Properties of a Modification of the Limited Information Estimator". In: *Econometrica* 45.4, pp. 939–953.

Geyer, Charles J. (1994). "On the asymptotics of constrained M-estimation." In: *The Annals of Statistics* 22.4, pp. 1993–2010.

Grünbaum, Branko (2003). *Convex polytopes. Prepared by Volker Kaibel, Victor Klee, and Günter M. Ziegler. 2nd ed.* 2nd ed. Vol. 221. New York, NY: Springer, pp. xvi + 466.

Hernán, Miguel A. and James M. Robins (2006). "Instruments for Causal Inference: An Epidemiologist's Dream?" In: *Epidemiology* 17.4, pp. 360–372.

Holland, Paul W. (1988). "Causal inference, path analysis and recursive Structural Equations Models". In: *ETS Research Report Series* 1988.1, pp. i–50.

Kang, Hyunseung, Youjin Lee, T. Tony Cai, and Dylan S. Small (2020). "Two robust tools for inference about causal effects with invalid instruments". In: *Biometrics*.

Kolesár, Michal, Raj Chetty, John Friedman, Edward Glaeser, and Guido W. Imbens (2015). "Identification and Inference With Many Invalid Instruments". In: *Journal of Business & Economic Statistics* 33.4, pp. 474–484.

Lee, Jason D., Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor (2016). "Exact post-selection inference, with application to the Lasso." In: *The Annals of Statistics* 44.3, pp. 907–927.

Lehmann, E. L. and Joseph P. Romano (2005). *Testing statistical hypotheses. 3rd ed.* 3rd ed. New York, NY: Springer, pp. xiv + 784.

Liu, Qiang, Jason Lee, and Michael Jordan (2016). "A Kernelized Stein Discrepancy for Goodness-of-fit Tests". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, pp. 276–284.

Reiersøl, Olav (1945). "Confluence analysis by means of instrumental sets of variables". PhD thesis. Stockholm College, p. 119.

Sargan, J. D. (1958). "The Estimation of Economic Relationships using Instrumental Variables". In: *Econometrica* 26.3, pp. 393–415.

Silvapulle, Mervyn J. (1996). "A test in the presence of nuisance parameters." In: *Journal of the American Statistical Association* 91.436, pp. 1690–1693.

Silvapulle, Mervyn J. and Pranab K. Sen (2005). *Constrained statistical inference. Inequality, order, and shape restrictions*. Hoboken, NJ: John Wiley & Sons, pp. xvii + 532.

Singh, Rahul, Maneesh Sahani, and Arthur Gretton (2020). *Kernel Instrumental Variable Regression*. arXiv: 1906.00232.

Small, Dylan S. (2007). "Sensitivity Analysis for Instrumental Variables Regression With Overidentifying Restrictions". In: *Journal of the American Statistical Association* 102.479, pp. 1049–1058.

Wang, Xuran, Yang Jiang, Nancy R. Zhang, and Dylan S. Small (2018). "Sensitivity analysis and power for instrumental variable studies". In: *Biometrics* 74.4, pp. 1150–1160.

Wasserman, Larry, Aaditya Ramdas, and Sivaraman Balakrishnan (2020). "Universal inference". In: *Proceedings of the National Academy of Sciences* 117.29, pp. 16880–16890.

Wolak, Frank A. (1989). "Local and Global Testing of Linear and Nonlinear Inequality Constraints in Nonlinear Econometric Models". In: *Econometric Theory* 5.1, 1—35.

Wooldridge, Jefrey M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

Wright, Philip Green (1928). *The tariff on animal and vegetable oils*. Brookings Institution. Investigations in international commercial policies 26. New York: The Macmillan company.

Zhao, Qingyuan, Dylan S. Small, and Bhaswar B. Bhattacharya (2018). *Sensitivity analysis for inverse probability weighting estimators via the percentile bootstrap*. arXiv: 1711.11286.