Vector-induced spectral measures and instrument exogeneity

Patrick F. Burauel
DIW Berlin and Free University of Berlin
pburauel@diw.de

Abstract

To obtain valid causal effect estimates in instrumental variable (IV) studies, an instrument has to fulfill (a.o.) the exclusion restriction and the exchangeability assumption, whose justification, in practice, often lie outside the model under consideration. Existing formal approaches to analyze the validity of the exclusion restriction rely on the testable implications in IV models derived by Balke and Pearl (1997). The present paper presents a test for both the exclusion restriction and the exchangeability assumption that is based on an entirely different approach. The test relies on a comparison of spectral measures of the covariates' covariance matrix induced by OLS and IV coefficient vectors respectively. These vector-induced spectral measures should be similar to the tracial spectral measure of the variance-covariance matrix if the Independence between Cause and Mechanism postulate is fulfilled. Monte Carlo simulation studies show the high accuracy of the test and its robustness to the presence of additional (to the treatment variable) endogenous covariates, the degree of endogeneity of the treatment variable itself and varying relevance of the instrument.

1 Introduction

Typical causal queries in diverse fields ranging from epidemiology to economics are concerned with the question of what level an outcome variable Y will take if a variable T is set to certain level; e.g. what would the level of unemployment, Y, be if the minimum wage were set at level T? This type of query is an interventional query, the "holy grail of causal thinking" (Pearl and Mackenzie, 2018), and requires assumptions about an underlying causal model. Instrumental variables (IVs) can be employed to answer such interventional queries. To be useful, an IV, Z, has to fulfill three criteria.[1]

1. Z and T must be statistically dependent (relevance assumption)
2. Z only affects Y through its relation with T (exclusion restriction)
3. Z and Y do not share common causes (exchangeability assumption)

Of these assumptions only the first can easily be tested statistically by a regression of X on Z and a test whether the associated coefficient is different from 0. The remaining assumptions cannot, in general, be tested since they involve the unobserved U. Nevertheless, the assumed causal structure implies testable constraints which can be leveraged to assess instrument exogeneity (Balke and Pearl, 1997). This paper provides a test which can detect violations of the exclusion restriction and the exchagneability assumption.

In this paper, we argue that an instrumental variable which violates either the exclusion restriction or the exchangeability assumption (such an instrument will be called endogenous from now on; vice versa, an instrument which fulfills these assumptions is called exogeneous) leaves discernible statistical traces in the joint distribution of the dependent, independent, and instrumented endogenous variables. By analyzing these traces, it is possible to test whether a potential binary instrument is exogenous. We show how the method proposed by Janzing and Schölkopf (2018) ('JS' in the following) to measure the extent to which an observed statistical relationship is due to confounding or genuine causation can be leveraged to devise a test for instrument exogeneity. JS is representative of the surging interest in causal modeling in the machine learning community (Peters et al., 2017).

The paper is structured as follows. First, we provide an overview of previous literature. Second, we describe the IV model to be analyzed. Third, we detail the testing procedure. Fourth, we present results of Monte Carlo simulations and a discussion. Finally, we conclude.

2 Previous literature

The Sargan (1958)-Hansen (1982) J-test for overidentifying restrictions arguably initiated the literature concerned with specification testing in instrumental variable (IV) models. The J-test can be used to test instrument exogeneity when there are more instruments than endogenous regressors. Conditional on the assumption that at least one instrument is exogenous, the

[1] In additiona, Z has to fulfill the monotonicity assumption. We take this assumption for granted here.

test can help decide whether *all* instruments are exogenous. However, the test is not able to detect a situation in which all instruments are endogenous.

A more recent strand of the literature proposes nonparametric tests for exogeneity in mean regressions. For example Blundell and Horowitz (2007) propose a test for exogeneity in nonparametric regression analysis which does not rely on non-parametric IV estimation (which often suffers from slow convergence that, in turn, results in low power of such tests). Breunig (2015) uses series estimators to propose a test for exogeneity. Gagliardini and Scaillet (2017) employ a Tikhonov Regularized estimator of the functional parameter to minimize the distance criterion corresponding to the moment conditions. Breunig (2018) extends these results to nonparametric quantile regression with nonseparable structural disturbances. Broadly conceived, what unites these papers is their reliance on testing whether the moment conditions implied by the instrumental variable model are fulfilled. By analyzing also higher-order moments, these models can resort to overidentifying restrictions even when there is only one instrument per endogenous variable.

Though diverse methods to test the exclusion restriction in *overidentified* IV models have been proposed, those for just identified models prove more elusive. Kitagawa (2015) is the study closest to this paper as it is the first to propose a test for the exclusion restriction in just identified models with a binary treatment and a binary instrument.

Kitagawa (2015) proposes testing the joint validity of the exclusion restriction, random assignment of instrument, and the absence of defiers (instrument monotonicity) by resorting to testable implications derived by Balke and Pearl (1994), Balke and Pearl (1997) and Heckman and Vytlacil (2005). These testable implications imply constraints on the outcome distributions of defiers, compliers, never- and always-takers.[2] Huber and Mellace (2015) and Mourifié and Wan (2017) provide closely related extensions to Kitagawa (2015). Huber and Mellace (2015) and Mourifié and Wan (2017) provide closely related extensions to Kitagawa (2015). The former propose a similar test which relies on mean potential outcomes rather than their distributions. Mourifié and Wan (2017) build on Kitagawa (2015) by representing his test in terms of conditional moment inequalities which have the advantage to easily incorporate additional covariates as well as to fit into the intersection bounds framework of Chernozhukov et al. (2013).

The test proposed in this paper builds on a separately advancing literature on discovering causal signals in non-experimental data. The underlying idea, which goes back to Haavelmo (1944), is that invariance structures in observed data justify making statements about the underlying causal structure of the system under study without (quasi-)experimental or instrumental variable approaches. This has been formalized from an information-geometric perspective as the Independence between Cause and Mechanism (ICM) postulate (Janzing et al., 2012; Peters et al., 2017). Janzing and Schölkopf (2018) show that traces of violations of ICM can be discerned in the spectral measures of variance-covariance matrices in the presence of unobserved confounding. In this paper, we show how employing this reasoning can be used to analyze instrument exogeneity.

Thus, the main contribution of this paper is the proposal of a novel testing approach which does not rely on moment restrictions. Instead, our approach is based on the decomposition of the spectral measure of the covariates' variance-covariance matrix induced by the corresponding parameter vector. Since the present work is, in turn, based on the ICM postulate, it adds to the growing literature using ICM as a powerful concept that helps to shed light on causal queries or to advance causal understanding more generally (Peters et al., 2016; Besserve et al., 2017; Besserve et al., 2018).

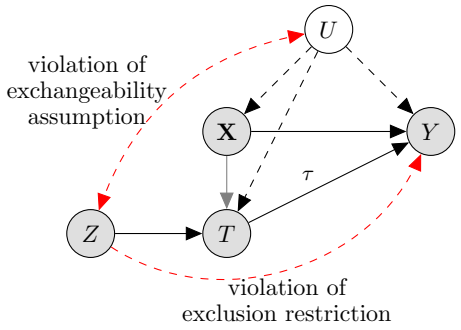

Figure 1: *Graphical representation of the IV model under study. T represents the binary variable of interest. The red arrow from Z to Y indicates how the exclusion restriction can be violated. The double-edged arrow between U and Z indicates how the exchangeability assumption can be violated. If either of the two arrows is present, the instrument is endogeneous and the treatment effect, τ, cannot be estimated consistently. The proposed test investigates whether either of the two arrows is present.*

3 Testing Endogeneity of an Instrument

Consider the following model

$$Y = \mathbf{X}\beta + \tau T + \varepsilon \tag{1}$$

where $\mathbf{X}$ represents a set of d exogenous covariates, T is a binary treatment indicator and τ the corresponding causal treatment effect. A maintained assumption

[2] In the potential outcomes notation (Holland, 1986) the compliers are defined as those individuals who do not take the treatment, $T_i = 0$, when $Z_i = 0$ and who do take the treatment, $T_i = 1$, when $Z_i = 1$. Defiers are individuals who do not take the treatment, $T_i = 0$, when $Z_i = 1$ and who do take the treatment, $T_i = 1$, when $Z_i = 0$. Always-takers have $T_i = 1$ regardless of Z_i. Never-takers have $T_i = 0$ regardless of Z_i.

throughout this paper is the constancy of τ (homogeneity of treatment effects). The test procedure will allow us to evaluate whether a potential instrument, Z, fulfills exclusion restriction and exchangeability assumption.

Before introducing the test procedure, it is instructive to briefly, and on an intuitive level, review the approach to estimate confounding strength by JS. Consider a simple linear regression setting in which a dependent variable is regressed on a set of observed independent variables. It is hard to tell to which extent an observed statistical relationship between the observed variables is due to genuine causation and to which extent it is due to confounding. JS approach this problem by relying on the Independence between Cause and Mechanism (ICM) postulate to provide a method to measure the degree of confounding. What the ICM implies on an intuitive level is that the *mechanism*, which translates cause into effect and is represented by the true parameter vector, and the *input to the mechanism* or *cause*, which is represented by $\Sigma_{\mathbf{XX}}$ should be 'independent'. JS make the concept of 'independence' operational by arguing that, if the ICM is fulfilled, the true parameter vector should lie in generic orientation with respect to the eigenspace spanned by the eigenvectors of the covariates' covariance matrix, $\Sigma_{\mathbf{XX}}$. More technically, such genericity is defined by the equivalence of two spectral measures: the spectral measure of $\Sigma_{\mathbf{XX}}$ induced by the true parameter vector (which results from weighting the eigenvalues of $\Sigma_{\mathbf{XX}}$ by that true parameter vector) should be similar to the (unweighted) tracial spectral measure of $\Sigma_{\mathbf{XX}}$.

The true parameter vector is, obviously, unknown and this precludes a direct computation of the spectral measure induced by that (true) parameter vector. However, the spectral measure induced by the estimated (and possibly biased) parameter vector can be computed from the data. The crucial result in JS is that this spectral measure can be decomposed into one part that is due to confounding and a second part that represents a genuine causal relation. It can be parameterized by a two-parametric family of probability measures. One of the parameters represents confounding strength, the relative weight of the confounding part in the decomposition. The algorithm proposed by JS finds those two parameter values that minimize the distance between the two-parametric estimate of the vector-induced spectral measure and the observed spectral measure induced by the estimated (and possibly) biased parameter vector. In the main part of the paper I take their method as given and show how it can be employed as a workhorse in testing instrument exogeneity. As a courtesy to the reader, the procedure to estimate κ is described in Appendix A.

In this section, we describe the test for evidence of instrument endogeneity in a step-by-step manner. We use the code provided by Janzing and Schölkopf[3] to estimate κ.

1. Normalize the data such that all variables have the same mean and variance as the treatment indicator T.
2. Take the original data without the treatment variable, ie. observations of Y and $\mathbf{X}_{\setminus\{T\}}$ (which includes the covariates but excludes the treatment variable T), then calculate the degree of confounding following JS, call the resulting metric $\kappa_x := \kappa(\mathbf{X}_{\setminus\{T\}}; Y)$.
3. Instrument the treatment variable with the instrument (which you want to test for exogeneity). Calculate the degree of confounding in the resulting data: Y and the concatenation of the exogenous $\mathbf{X}$ and the instrumented T and call it $\kappa_i := \kappa(\mathbf{X}_{\setminus\{T\}}, \hat{T}; Y)$, i for 'instrumented'.
4. Take the difference between the two κs:
$$\delta = \kappa_i - \kappa_x. \tag{2}$$

Recall that a higher κ indicates more confounding. The fundamental idea of instrumental variables techniques is to extract that part of the variation in the treatment variable which can be explained by the instrument. If that instrument is indeed exogeneous, the resulting estimate of the treatment variable, $\hat{T}$, should not covary with the unobserved confounder, U; in other words, $\hat{T}$ should be unconfounded. Vice versa, if the instrument is endogeneous, $\hat{T}$ will be confounded. Crucially, this is true *regardless of the source of confounding of the instrument* (be it because of a violation of the exclusion restriction or a violation of the exchangeability assumption). To get an intuition for how the test works consider the following two cases. First, suppose the instrument indeed fulfills the exclusion restriction. As a consequence $\hat{T}$ is exogenous. This implies that the degree of confounding when $\hat{T}$ is added to the list of independent variables, measured by κ_i, should be smaller than the 'base level' of confounding, κ_x. Thus, the resulting δ should be smaller than zero. Second, if the instrument does not fulfill the exclusion restriction, $\hat{T}$ will still be confounded and adding $\hat{T}$ as an additional confounded variable to the list of independent variables will increase κ relative to the base level of confounding: κ_i is expected to be larger than κ_x and the resulting δ positive. Underlying this reasoning is the assumption that adding an exogenous explanatory variable to the set of covariates will not increase the level of confounding as measured by κ. The following proof will show just that.

Theorem 1. We set out to prove that κ does not increase upon adding the instrumented treatment variable if the instrument is exogeneous. The definition of κ_x can be rephrased in terms of true and estimated parameter values as

$$\kappa_x := \frac{\left\|(\hat{\beta}) - (\beta)\right\|^2}{\|(\beta)\|^2 + \left\|(\hat{\beta}) - (\beta)\right\|^2} = \frac{A}{B + A} \tag{3}$$

[3]The code is generously made available at `http://webdav.tuebingen.mpg.de/causality/`.

where β denotes the true parameter vector and $\hat{\beta}$ the estimated parameter vector. When an the instrumented variable $\hat{T}$, whose coefficient is denoted with τ, is added to the model, we have

$$\kappa_i = \frac{\left\| \begin{pmatrix} \hat{\beta} \\ \hat{\tau} \end{pmatrix} - \begin{pmatrix} \beta \\ \tau \end{pmatrix} \right\|^2}{\left\| \begin{pmatrix} \beta \\ \tau \end{pmatrix} \right\|^2 + \left\| \begin{pmatrix} \hat{\beta} \\ \hat{\tau} \end{pmatrix} - \begin{pmatrix} \beta \\ \tau \end{pmatrix} \right\|^2}. \tag{4}$$

If the instrument is indeed exogeneous, we recover the true causal effect of T, i.e. $\hat{\tau} = \tau$; consequently,

$$\kappa_i = \frac{\left\| \begin{pmatrix} \hat{\beta} - \beta \\ 0 \end{pmatrix} \right\|^2}{\left\| \begin{pmatrix} \beta \\ \tau \end{pmatrix} \right\|^2 + \left\| \begin{pmatrix} \hat{\beta} - \beta \\ 0 \end{pmatrix} \right\|^2}. \tag{5}$$

Therefore, $\kappa_i \leq \kappa_x$ when the instrument is indeed exogeneous. □

Under certain conditions, κ_i can be smaller than κ_x if the instrument is slightly endogeneous, i.e. $\tau = \hat{\tau}$ does *not hold exactly*. Let us define

$$\Delta\kappa := \kappa_i - \kappa_x = \frac{A'}{B' + A'} - \frac{A}{B + A} \tag{6}$$

In order for κ_i to decrease relative to κ_x, the following relation must hold:

$$\Delta b > \Delta a \Leftrightarrow \Delta\kappa = \kappa' - \kappa \leq 0, \tag{7}$$

where $\Delta k := \frac{\Delta K}{K} = \frac{K' - K}{K}$ denotes the relative change in a variable K.

$$\begin{aligned} &\Delta b > \Delta a \\ \Leftrightarrow\ &\frac{\tau^2}{\|\beta\|^2} > \frac{(\hat{\tau} - \tau)^2}{\left\|\hat{\beta} - \beta\right\|^2} \\ \Leftrightarrow\ &\frac{\left\|\hat{\beta} - \beta\right\|^2}{\|\beta\|^2} > \frac{(\hat{\tau} - \tau)^2}{\tau^2} \end{aligned} \tag{8}$$

If the average relative squared bias of β is larger than the relative squared bias of τ, κ_i will decrease relative to κ_x even when the instrument is not exogeneous. In those cases when it is most critical to uncover endogeneity of the instrument, i.e. when the relative bias of τ is large, κ_i will be larger than κ_x. In practice, this implies that the proposed algorithm is prone to not detecting the endogeneity of the instrument if the degree of endogeneity is relatively small.

5. Next, to incorporate uncertainty about these metrics in the subsequent decision, bootstrap over steps 1-3 above. For each bootstrap sample $b \in \{1, \ldots, B\}$ calculate

$$\delta_b = \kappa_{i,b} - \kappa_{x,b} \tag{9}$$

6. Calculate the share of samples with $\delta_b > 0$,

$$\delta_B^{\mathbb{1}} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(\kappa_{i,b} > \kappa_{x,b}), \tag{10}$$

$\delta_B^{\mathbb{1}}$ can be interpreted as a pseudo-p-value for the hypothesis

$$H_0 : (\kappa_i > \kappa_x) \Leftrightarrow \text{the instrument is endogenous.} \tag{11}$$

If $\delta_B^{\mathbb{1}}$ is low, it is more likely that κ_i is, in fact, smaller than κ_x (which indicates exogeneity of the instrument). The lower the pseudo-p-value, the more evidence against the null. Therefore, the lower δ_B, the more evidence against H_0 : the instrument is endogenous in favour of the alternative H_a : the instrument is exogeneous. Implicit in this formulation of the test is the assumption that the instrument is assumed endogeneous until proven otherwise.

7. Finally, I propose the following decision rule:

$$\psi_\delta(\alpha) = \mathbb{1}(\delta_B^{\mathbb{1}} \leq \alpha) = \begin{cases} 1 & \Longrightarrow \text{reject } H_0 \\ 0 & \Longrightarrow \text{do not reject } H_0 \end{cases} \tag{12}$$

that depends on threshold parameter α, which controls the trade-off of committing Type I and Type II errors.

Alternatively to the decision rule in (12), one can use a paired-t-test for the null hypothesis in (11),

$$\psi_t(\alpha) = \mathbb{1}(p\text{-value of } t\text{-test} < \alpha). \tag{13}$$

Note that each test relies on $B = 200$ bootstrap samples on which the κs are estimated. Thus, the tests rely on a set of B differenced κs.

Testing exogeneity It is conceivable to test the mirror image of (10) by calculating the share of bootstrap samples with $\delta_b \leq 0$,

$$\tilde{\delta}_B^{\mathbb{1}} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}(\kappa_{i,b} \leq \kappa_{x,b}), \tag{14}$$

$\tilde{\delta}_B^{\mathbb{1}}$ can then be interpreted as a pseudo-p-value for the hypothesis

$$\tilde{H}_0 : (\kappa_i \leq \kappa_x) \Leftrightarrow \text{the instrument is exogenous} \tag{15}$$

with the following decision rule:

$$\psi_{\tilde{\delta}}(\alpha) = \mathbb{1}(\tilde{\delta}_B^{\mathbb{1}} \leq \alpha) = \begin{cases} 1 & \Longrightarrow \text{reject } \tilde{H}_0 \\ 0 & \Longrightarrow \text{do not reject } \tilde{H}_0 \end{cases} \tag{16}$$

Considering the difficulty of finding valid instruments, assuming instrument endogeneity until proven otherwise (ψ_δ) seems a more honest approach than assuming instrument exogeneity until proven otherwise ($\psi_{\tilde{\delta}}$). Therefore, we will stick to ψ_δ in the following. Simulation results, which are not reproduced here, show

that a test based on $\psi_{\tilde{\delta}}$ achieves similar AUC levels as those described below for ψ_δ.

It is possible that resorting to $\psi_{\tilde{\delta}}$ with the implied *a priori* presumption of instrument exogeneity will prove inevitable. Though this option seems less appealing due to the elusiveness of convincing instruments, work is in progress to construct an empirical distribution of the test statistic under (15) by observing that the distribution of κ_i is similar to the one of κ_x under $\tilde{H}_0$.

4 Monte Carlo Simulation

We consider the model in (1). In order to analyze the effectiveness of the instrument exogeneity test, I generate data according to the following recipe.

This simulation setting extends the one proposed by Huber and Mellace (2015) in that it considers covariates in addition to the treatment variable of primary interest. First, the simulation to study violations of the exclusion restriction are presented; followed by the simulation to study violations of the exchangeability assumption.

4.1 Simulation Regime 1: Violation of exclusion restriction

Let n-dimensional vectors of disturbances, U and ε, be drawn from

$$\begin{pmatrix} U \\ \varepsilon_T \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \omega_1 \\ \omega_1 & 1 \end{pmatrix} \right), \tag{17}$$

and the instrument, Z, be generated by

$$Z \sim \text{Bernoulli}(0.5). \tag{18}$$

The set of covariates is generated by first populating a $n \times d$ matrix $\mathbf{X}_{temp}$ with random draws from a Gaussian with mean zero and standard deviation one, $\mathcal{N}(0,1)$. Draw a random d-dimensional vector

$$\beta_{c,temp} \sim \mathcal{N}(0,1)$$

and, to keep simulations for various dimensions, d, comparable, set $\beta_c = \beta_{c,temp} \times \|\beta_{c,temp}\|^{-1}$. With these ingedients set

$$\mathbf{X} = \mathbf{X}_{temp} + U\beta_c'. \tag{19}$$

To induce dependence of the treatment on the set of covariates, first draw the d-dimensional vector $\beta_{t,temp}$ populated with draws from a $\mathcal{N}(0,1)$ and set $\beta_t = \beta_{t,temp} \times \|\beta_{t,temp}\|^{-1}$.

Further, generate treatment, T, as

$$T = \mathbb{1}(\omega_2 Z + \mathbf{X}\beta_t' + \varepsilon_T > T'). \tag{20}$$

where T' is the mean of $\mathbf{X}\beta_t' + \varepsilon_T$.

Note that the binary nature of both treatment and instrument represents a special and harder (for specification testing) case than the setting with a continuous treatment and instrument.

To simulate the outcome variable, I first generate a random vector $\beta = (\beta_1 \quad \dots \quad \beta_d)^\top$ where each $\beta_1, \dots, \beta_d$ is drawn from a Gaussian $\mathcal{N}(0,1)$. The true coefficient of the treatment variable is set to $\tau = 1$ to facilitate the interpretation of deviations of the estimated coefficients from that true value in subsequent simulations.

Finally, generate outcome, Y, as

$$Y = \mathbf{X}\beta + \tau T + \omega_3 Z + U \tag{21}$$

where ω_3 controls the degree of violation of the exclusion restriction.

4.2 Simulation Regime 2: Violation of exchangeability assumption

For the simulations to test whether the algorithm can detect endogeneity of the instrument stemming from a violation of the exchangeability assumption, we replace (18) with

$$Z = \mathbb{1}(\varepsilon_Z + \omega_3 U > 0) \tag{22}$$

where ε_Z is drawn from a standard Gaussian. Therefore, ω_3 controls the degree of violation of the exchangeability assumption. Finally, we replace (21) with

$$Y = \mathbf{X}\beta + \tau T + U. \tag{23}$$

4.3 Parameter constellations

An overview of the interpretation of the parameters is provided:

- ω_1: endogeneity of treatment, T
- ω_2: relevance of the instrument, Z
- ω_3: endogeneity of the instrument, Z

Call $\hat{\beta}_{\text{nv}}$ the coefficient vector of a naive linear regression of Y on $\{\mathbf{X}, T\}$, which is biased due to the endogeneity of T and the covariates.

Next, I use Z to instrument T. Following Adams et al. (2009), I implement the IV strategy by first estimating a linear probability model (LPM) of T on $\{\mathbf{X}, Z\}$. Second, use the predicted $\hat{T}$ in the second stage to estimate $\hat{\beta}_{IV} = (\mathbf{X}_{IV}^\top \mathbf{X}_{IV})^{-1} \mathbf{X}_{IV}^\top Y$ where $\mathbf{X}_{IV} := \{\mathbf{X}, \hat{T}\}$.

To show the empirical performance of the proposed test, we implement Monte Carlo simulations for each combination of the following parameters: number of observations: $n \in \{500, 10000\}$, number of covariates: $d \in \{3, 10\}$ (one endogenous treatment variable: T, along with $d-1$ exogenous variables: $X_1, \dots, X_{d-1}$), degree of the endogeneity of T: $\omega_1 = 0.8$, degree of the relevance of the instrument: $\omega_2 \in \{0.2, 0.6\}$ degree of the endogeneity of the instrument, Z: $\omega_3 \in \{0, 1/3, 2/3, 1\}$. Moreover, the following parameters are fixed: number of bootstrap samples $B = 100$, number of Monte Carlo draws $M = 100$.

In order to gain deeper understanding of the performance of the test, I also report the average difference between the κs over all boostrap draws:

$$\delta_B = \frac{1}{B} \sum_{b=1}^{B} (\kappa_{i,b} - \kappa_{x,b}). \tag{24}$$

Table 1: *This Table shows the pseudo-p-value, δ_B, and the empirical rejection probability, ψ_t (based on a $\alpha = 0.05$) for various combinations of n and d. The source of endogeneity for the instrument is a violation of the exclusion restriction. When the number of observations and dimensionality is sufficiently high (e.g. 10000 and 10 respectively), the test rejects the null of endogeneity in 79% of the cases when, indeed, the instrument is exogneous; and does not reject in any case when the instrument is, indeed, endogeneous.*

n	d	ω_3	ω_2	ω_1	pseudo-p	δ_B	ψ_t
500	3	0.00	0.2	0.8	0.64	0.06	0.32
		0.33	0.2	0.8	0.80	0.44	0.20
		0.67	0.2	0.8	0.93	0.62	0.05
		1.00	0.2	0.8	0.93	0.65	0.07
	10	0.00	0.2	0.8	0.66	0.11	0.28
		0.33	0.2	0.8	0.99	0.55	0.01
		0.67	0.2	0.8	1.00	0.74	0.00
		1.00	0.2	0.8	1.00	0.77	0.00
10000	3	0.00	0.2	0.8	0.50	-0.24	0.47
		0.33	0.2	0.8	0.80	0.51	0.19
		0.67	0.2	0.8	0.85	0.57	0.15
		1.00	0.2	0.8	0.93	0.64	0.07
	10	0.00	0.2	0.8	0.20	-0.13	0.79
		0.33	0.2	0.8	1.00	0.61	0.00
		0.67	0.2	0.8	1.00	0.73	0.00
		1.00	0.2	0.8	1.00	0.73	0.00

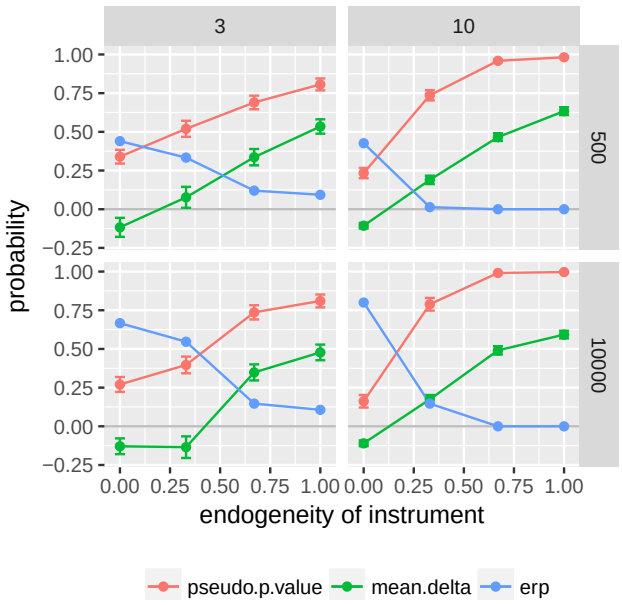

Figure 2: *This Figure shows the pseudo-p-value, δ_B, and the empirical rejection probability (based on the pseudo-p-value with threshold parameter $\alpha = 0.05$) as a function of the degree of instrument endogeneity where the source of confounding is a* ***violation of the exclusion restriction****, by number of covariates, (d, horizontal), and number of observations (n, vertical). δ_B rises sharply with the degree of confounding, as does the pseudo-p-value. Consequently, the empirical rejection probabilities go down to zero indicating that, if the degree of condounding is sufficiently high, the test does not reject the null of endogeneity.*

5 Results of Monte Carlo Study

We begin the discussion with Simulation Regime 1, i.e. simulated violations of the exclusion restriction. Figure 2 shows the evolution of the pseudo-p-value and δ_B as a function of the degree of endogeneity of the instrument. Both measures are increasing the the endogeneity, which shows that they are picking up the confoundedness signal in the data. The empirical rejection rate based on the pseudo-p-value (ψ_δ) is decreasing as a function of the confoundedness, meaning that the null hypothesis of endogeneity is not rejected as the level of endogeneity is sufficiently large. Generally, both a larger d and a larger n improve the visual performance of the test. Increasing d, given n, tends to improve performance by more than increasing n, given d. This reflects the fact that the asymptotic results in the original JS method rely on d going to infinity.

In order to evaluate the trade-off between making Type I and Type II errors we calculate the area under the ROC curve (AUC) and plot it as a function of the endogeneity of the instrument, Figure 3 (cf. Appendix B for details on the calculation). The AUC levels rise with the degree of endogeneity of the instrument and reach as high as roughly 0.95 for large numbers of observations and dimensions. Increasing the dimensionality d leads to much larger AUC levels than increasing the number of observations n. This is in line with the asymptotic results in JS relying on d going to infinity. It is noteworthy that the AUC levels tend to be larger for a lower value of the degree of relevance of Z (ω_2). A larger ω_2 implicitly goes along with a larger complier rate. Huber and Mellace (2015) underscore that "the absence of compliers maximizes the asymptotic power to find violations in IV validity" (p. 404); in that sense the superior performance of the algorithm as ω_2 decreases mirrors this result. As ω_2 increases, i.e. the share of compliers grows, Z contains less and less additional variation that can be leveraged by the exogeneity test. In the extreme, Z and T collapse to effectively one variable and the instrumented T does not contain any different information than T. In other words, the instrument cannot extract the experimental variation of T (that part of the variation that is unrelated to U) when ω_2 is too large. Nevertheless, even for large ω_2, the proposed test performs well with AUC levels ranging from 0.6 (low degree of endogeneity of

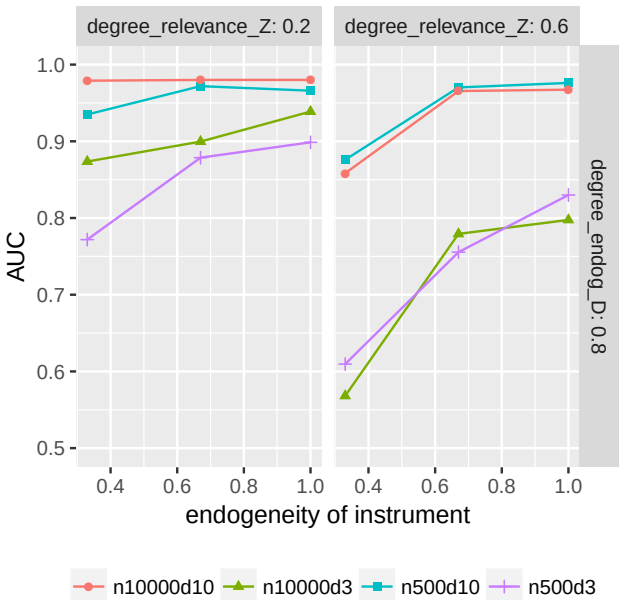

Figure 3: *This Figure shows the area under the ROC curve (AUC) as a function of the degree of instrument endogeneity where the source of confounding is a* ***violation of the exclusion restriction****, for various combinations of number of covariates, d, and number of observations, n, by instrument relevance degree (ω_2, horizontal). Underlying test statistic is the pseudo-p-value. The test achieves high AUC levels of close to the perfect score of 1 for large n and d. Under a low ω_2, which translates into lower share of compliers, the test performance increases.*

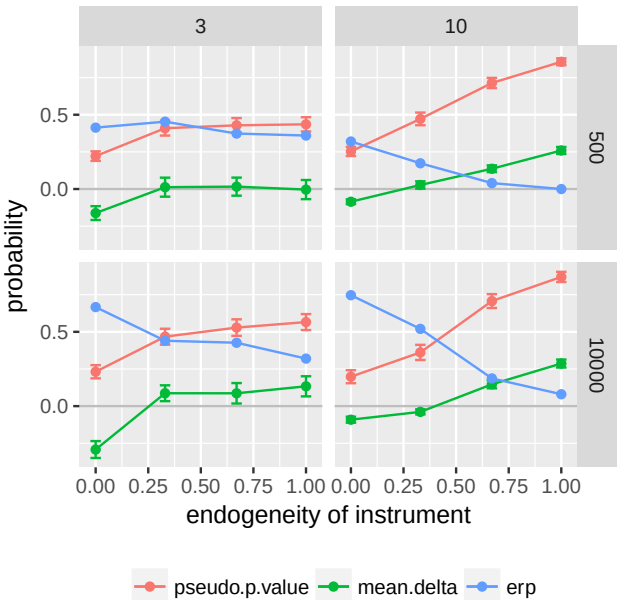

Figure 4: *This Figure shows the pseudo-p-value, δ_B, and the empirical rejection probability (based on the pseudo-p-value with threshold parameter $\alpha = 0.05$) as a function of the degree of instrument endogeneity where the source of confounding is a* ***violation of the exchangeability assumption****, by number of covariates, (d, horizontal), and number of observations (n, vertical). δ_B rises (less sharply than in the case where the exclusion restriction is violated) with the degree of confounding, as does the pseudo-p-value. Consequently, the empirical rejection probabilities go down to zero indicating that, if the degree of condounding is sufficiently high, the test does not reject the null of endogeneity.*

instrument) to 0.8 (high degree of endogeneity). Table 1 shows results of the Monte Carlo simulations in table form.

In Simulation Regime 2 we analyze whether the algorithm can also detect an invalid instrument when its invalidity stems from the fact that the exchangeability assumption is violated. Figures 4 and 5 report results for Simulation Regime 2 in the same form as previous Figures for Simulation Regime 1. The performance of the test decreases slightly. However, given sufficiently many covariates and dimensions, the AUC reaches levels around 0.95 when the degree of endogeneity of the instrument is large.

Robustness to normalization An important characteristic of the algorithm proposed by JS is that the estimated κ is not robust to rescaling of the data as this introduces a dependence between the covariance matrix of the covariates and the parameter vector. The authors acknowledge this, yet claim that the estimated κ is surprisingly robust to rescaling of the data (a claim that is vindicated in the case at hand). Nevertheless, this lack of theoretical robustness of κ to rescaling the data is the major reason why the test proposed in this paper does not rely on the size of a single κ, but rather on the comparison of two κs which are both influenced by rescaling in the same fashion, which one can therefore expect to leave their relative size unaffected. Analyzing such normalization theoretically will be subject of future work.

6 Real-world application

Following Kitagawa (2015) and Huber and Mellace (2015), we test the validity of proximity to a four-year college as an instrument of educational attainment to estimate returns to schooling, measured by log of weakly earnings. This instrument is first proposed by Card (1993). The author himself casts doubt on the validity of college proximity as an instrument as there might be factors such as family preferences or local labour market conditions which might be related to both the proximity to a college and the outcome variable. Therefore, we follow Kitagawa (2015) Huber and Mellace (2015) and present the results of the proposed test for both the unrestricted sample as well as for four subsamples: the group of white individuals,

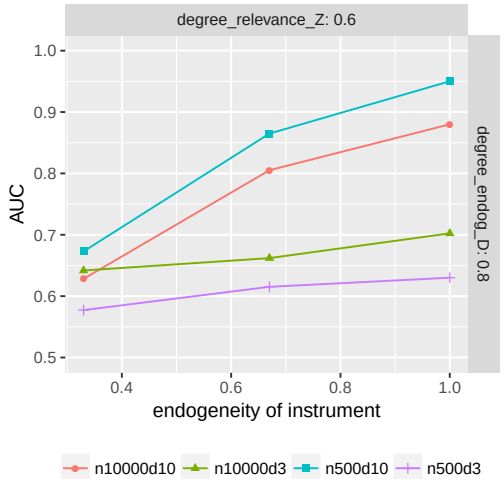

Figure 5: *This Figure shows the area under the ROC curve (AUC) as a function of the degree of instrument endogeneity where the source of confounding is a* ***violation of the exchangeability assumption****, for various combinations of number of covariates, d, and number of observations, n. Underlying test statistic is the pseudo-*p*-value. The test achieves high AUC levels of close to the perfect score of 1 for large n and d.*

not living in the South is split into four groups based on whether father's education was above or below 12 years and whether residence was in a rural or urban area.

The results in Table 2 show that the test rejects the null of endogeneity in the full sample and all sub samples, when basing the decision on the t-test p-value. However, the pseudo-p-values are in broad agreement with the results in Huber and Mellace (2015). The test only marginally rejects unconfoundedness for the full sample. It shows a much lower pseudo-p-value for the subgroups in lines 2-4. In line 5 (father's education less than twelve years and rural) the pesudo-p-value is relatively high indicating that the null cannot be rejected. The results in Huber and Mellace (2015) point in the same direction in the sense that their test indicates endogeneity of the instrument most clearly for the last of the four subgroups.

These results indicate that the test presented here suffers from a relatively high type I error (rejecting instrument endogeneity although the instrument is indeed endogeneous) if the number of observations and the degree of confoundedness is relatively small. This behaviour can be seen in the Monte Carlo simulation in Figure 2 (top left quadrant) where the empirical rejection probability only goes to zero relatively slowly.

7 Conclusion

The proposed method leverages subtle statistical traces of confounding, measured by using the methodology laid out in Janzing and Schölkopf (2018), to test whether a potential instrument violates either the exclusion restriction or the exchangeability assumption. It relies on Schölkopf and Janzing's insight that, under certain assumptions, the spectral measure of the covariance matrix of the independent variables in a linear regression can be decomposed into a causal and confounded part. As such it provides a novel way to approach the testing unconfoundedness, not only in instrumental variable models. As such, the decomposition of a spectral measure of the covariance matrix of independent variables as a path towards understanding the degree of confoundedness of a given variable of interest opens promising future research avenues.

Extensive Monte Carlo studies show that the proposed method has high accuracy. Its AUC levels reach from around 0.7 when the number of observations, covariates, and the degree of endogeneity of the instrument is low to levels close to 1 when the number of observations, covariates, and the degree of endogeneity of the instrument is high. These results prove robust to the degree of relevance of the instrument and to normalization of the data.

Another way of interpreting the results presented here is that they constitute a data-driven aid for solving the identification problem in causal studies. While acknowledging the need to keep the multiple testing problem at bay, the method proposed here could be used to search for instruments that fulfill the exclusion restriction in increasingly available high-dimensional datasets. In this sense, the current work is related to Sharma (2016) and Sharma et al. (2016) who are proposing that algorithmic search for exogeneous instruments can be guided by comparing the marginal likelihood of valid and invalid IV models in a Bayesian network setting.

Further research must address the performance of the test with real data. In addition, finding ways to make the estimation of the confounding strength κ robust to rescaling and exploring how the proposed test can be leveraged to automatically search for instruments in high-dimensional data are promising research directions.

Table 2: *This Table shows results of the empirical application, based on Card (1995). As Card himself suspects, the hypothesis of instrument exogeneity is rejected for the full sample. There is some faint indication that the exogeneity assumption is less violated in the subsample consisting of those individuals living in rural areas and whose fathers' have relatively low education.*

group	pseudo-p	t-test p-val.	δ_B
full sample	0.05	0	0.83
w.nS.feduc.12more.urban	0.02	0	0.81
w.nS.feduc.less12.urban	0.03	0	0.88
w.nS.feduc.12more.rural	0.01	0	0.60
w.nS.feduc.less12.rural	0.24	0	0.40

References

Adams, Renée, Heitor Almeida, and Daniel Ferreira (2009). "Understanding the relationship between founder–CEOs and firm performance". *Journal of Empirical Finance* 16.1, pp. 136–150.

Balke, Alexander and Judea Pearl (1994). "Counterfactual probabilities: Computational methods, bounds and applications". *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 46–54.

– (1997). "Bounds on treatment effects from studies with imperfect compliance". *Journal of the American Statistical Association* 92.439, pp. 1171–1176. ISSN: 1537274X. DOI: 10.1080/01621459.1997.10474074.

Besserve, Michel, Naji Shajarisales, and Bernhard Sch (2017). "Group invariance principles for causal generative models". arXiv: 1705.02212.

Besserve, Michel, Rémy Sun, and Bernhard Schölkopf (2018). "Counterfactuals uncover the modular structure of deep generative models". *arXiv preprint arXiv:1812.03253*.

Blundell, Richard and Joel Horowitz (2007). "A Non Parametric Test of Exogeneity". *Review of Economic Studies* 74.4, pp. 1035–1058. ISSN: 0034-6527. DOI: 10.1111/j.1467-937X.2007.00458.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-937X.2007.00458.x/full.

Breunig, Christoph (2015). "Goodness-of-fit tests based on series estimators in nonparametric instrumental regression". *Journal of Econometrics* 184.2, pp. 328–346.

– (2018). *Specification testing in nonparametric instrumental quantile regression*.

Card, David (1993). "Using geographic variation in college proximity to estimate the return to schooling".

Chernozhukov, Victor, Sokbae Lee, and Adam M Rosen (2013). "Intersection bounds: estimation and inference". *Econometrica* 81.2, pp. 667–737.

Gagliardini, Patrick and Olivier Scaillet (2017). "A specification test for nonparametric instrumental variable regression". *Annals of Economics and Statistics/Annales d'Économie et de Statistique* 128, pp. 151–202.

Haavelmo, Trygve (1944). "The probability approach in econometrics". *Econometrica: Journal of the Econometric Society*, pp. iii–115.

Hansen, Lars Peter (1982). "Large sample properties of generalized method of moments estimators". *Econometrica: Journal of the Econometric Society*, pp. 1029–1054.

Heckman, James J and Edward Vytlacil (2005). *Structural equations, treatment effects, and econometric policy evaluation*. Vol. 73. 3, pp. 669–738. ISBN: 4030008526. DOI: 10.1111/j.1468-0262.2005.00594.x.

Holland, Paul W (1986). "Statistics and causal inference". *Journal of the American statistical Association* 81.396, pp. 945–960.

Huber, Martin and Giovanni Mellace (2015). "Testing Instrument Validity for LATE Identification Based on Inequality Moment Constraints". *Review of Economics and Statistics* 97.2, pp. 638–647. ISSN: 1725-2806. DOI: 10.1162/REST. arXiv: arXiv:1011.1669v3.

Janzing, Dominik and Bernhard Schölkopf (2018). "Detecting confounding in multivariate linear models via spectral analysis". *Journal of Causal Inference* 6.1. arXiv: 1704.01430. URL: http://arxiv.org/abs/1704.01430.

Janzing, Dominik, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf (2012). "Information-geometric approach to inferring causal directions". *Artificial Intelligence* 182-183, pp. 1–31. ISSN: 00043702. DOI: 10.1016/j.artint.2012.01.002.

Kitagawa, Toru (2015). "A Test for Instrument Validity". *Econometrica* 83.5, pp. 2043–2063. ISSN: 0012-9682. DOI: 10.3982/ECTA11974. URL: https://www.econometricsociety.org/doi/10.3982/ECTA11974.

Liu, Furui and Laiwan Chan (2018). "Confounder Detection in High Dimensional Linear Models using First Moments of Spectral Measures". arXiv: 1803.06852. URL: http://arxiv.org/abs/1803.06852.

Mourifié, Ismael and Yuanyuan Wan (2017). "Testing Local Average Treatment Effect Assumptions". *Review of Economics and Statistics* 99.2, pp. 638–647. ISSN: 1725-2806. DOI: 10.1162/REST. arXiv: arXiv:1011.1669v3.

Pearl, Judea and Dana Mackenzie (2018). *The Book of Why*. New York, NY: Basic Books.

Peters, Jonas, Peter Bühlmann, and Nicolai Meinshausen (2016). "Causal inference using invariant prediction: identification and confidence intervals". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.5, pp. 947–1012. arXiv: 1501.01332. URL: http://arxiv.org/abs/1501.01332.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, Massachusetts, London, England: MIT Press. ISBN: 9780262037310.

Sargan, John D (1958). "The estimation of economic relationships using instrumental variables". *Econometrica*, pp. 393–415.

Sharma, Amit (2016). "Necessary and Probably Sufficient Test for Finding Valid Instrumental Variables", pp. 1–24.

Sharma, Amit, Jake M. Hofman, and Duncan J. Watts (2016). "Split-door criterion for causal identification: Automatic search for natural experiments". arXiv: 1611.09414. URL: http://arxiv.org/abs/1611.09414.

Appendices

A Testing for the presence of a hidden confounder

Janzing and Schölkopf (2018) propose a method to estimate the degree to which an observed statistical relationship between a multidimensional set of covariates, $\mathbf{X}$, and an outcome variable Y is due to the direct influence of $\mathbf{X}$ on Y or due to an unobserved confounder influencing both $\mathbf{X}$ and Y. They point out that the spectral measure of the covariance matrix of the (right-hand side) independent variables, $\Sigma_{\mathbf{XX}}$, induced by the parameter vector differs depending on whether there is confounding or not. More precisely, the confounded-parameter-induced spectral measure of $\Sigma_{\mathbf{XX}}$ can be decomposed into one parts: one that is due to the genuine causal influence and a second that is due to the confounding influence.

As a courtesy to the reader, I reproduce their method here; this section does not contain new results. Compared to JS, I have slightly changed the order of presentation as well as some notation to ensure consistency with the main body of this paper.

A.1 The set-up

Consider the following linear model:

$$\begin{aligned} \mathbf{X} &= \mathbf{b}u + \mathbf{E} \\ Y &= \mathbf{X}^\top \mathbf{a} + cu^\top + \varepsilon \end{aligned} \tag{25}$$

where Y is the $n \times 1$ outcome vector, $\mathbf{a}$ is the $d \times 1$ causal parameter vector of interest. $\mathbf{X}$ is a $d \times n$ matrix of covariates. The confounder u is a $1 \times n$ vector. $\mathbf{b}$ is a $d \times 1$ parameter vector. $\mathbf{E}$ is a $d \times n$ matrix of standard normal errors drawn independently from u. ε is a $n \times 1$ vector of errors. Without loss of generality, u is assumed to have unit variance.

By regressing Y on $\mathbf{X}$, we obtain the biased parameter vector

$$\hat{\mathbf{a}} := \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{X}Y} \tag{26}$$

where Σ denotes covariance matrices. Generally, we are interested in the structural parameter vector $\mathbf{a}$ which represents genuine causal influence. To illustrate, the relation between $\mathbf{a}$ and $\hat{\mathbf{a}}$ consider

$$\Sigma_{\mathbf{X}Y} = Cov(\mathbf{X}, Y) = Cov(\mathbf{b}u + \mathbf{E}, \mathbf{X}^\top \mathbf{a} + cu^\top + \varepsilon) \tag{27}$$

$$= (\Sigma_{\mathbf{EE}} + \mathbf{b}\mathbf{b}^\top)\mathbf{a} + c\mathbf{b} \tag{28}$$

$$\Sigma_{\mathbf{XX}} = Cov(\mathbf{X}, \mathbf{X}) = Cov(\mathbf{b}u + \mathbf{E}, \mathbf{b}u + \mathbf{E}) = \Sigma_{\mathbf{EE}} + \mathbf{b}\mathbf{b}^\top \tag{29}$$

and therefore

$$\hat{\mathbf{a}} = \mathbf{a} + (\Sigma_{\mathbf{EE}} + \mathbf{b}\mathbf{b}^\top)^{-1} c\mathbf{b} = \mathbf{a} + c\Sigma_{\mathbf{XX}}^{-1}\mathbf{b}. \tag{30}$$

A.2 Genericity assumptions

The underlying idea, which this method rests on, is the Independence between Cause and Mechanism (ICM) postulate Janzing et al. (2012), which states that the causal mechanism, which translates cause into effect, is independent of the cause. Intuitively, the mechanism is indifferent as to which 'level of the cause' it translates into an effect.

To understand what the ICM amounts to in the case at hand, note that the cause is represented by $\Sigma_{\mathbf{XX}}$, likewise the mechanism is represented by $\mathbf{a}$. Therefore, Janzing and Schölkopf (2018) postulate that $\mathbf{a}$ lies in 'generic orientation' relative to $\Sigma_{\mathbf{XX}}$. For instance, since $\mathbf{a}$ is chosen independently of $\mathbf{X}$, and, thus, also the covariance matrix $\Sigma_{\mathbf{XX}}$, $\mathbf{a}$ is not likely to be aligned with its first principal component.[4]

In order to make the notion of 'generic orientation' more precise, we require some definitions. First of all, assuming that all eigenvalues of a matrix are different from each other (ie. the matrix non-degenerate), we can recall that each such symmetric matrix A has a unique decomposition

$$A = \sum_{j=1}^{d} \lambda_j \phi_j \phi_j^\top \tag{31}$$

where λ_j denotes the eigenvalues and ϕ_j the corresponding normalized eigenvectors.

The renormalized trace is defined to be

$$\tau(A) := \frac{1}{d} tr(A) \tag{32}$$

for A a $d \times d$ matrix (note that the τ in this notation is unrelated to the treatment effect which it denotes in the main body of the paper).

Definition A.1. (tracial spectral measure) Let A be a real symmetric matrix with non-degenerate spectrum. The tracial spectral measure of A is defined as the uniform distribution over its eigenvalues $\lambda_1, \ldots, \lambda_d$:

$$\mu_A^\tau := \sum_{j=1}^{d} \delta_{\lambda_j}. \tag{33}$$

The tracial spectral measure is a property of a matrix. The vector-induced spectral measure complements the tracial spectral measure by accounting for its relation to an arbitrary vector.

Definition A.2. (vector-induced spectral measure) Given a symmetric matrix A with associated eigenvalues λ_j and corresponding eigenvectors ϕ_j, the spectral measure induced by an arbitrary vector $\psi \in \mathbb{R}^d$ is given by

$$\mu_{A,\psi} = \sum_{j=1}^{d} \left(\psi^\top \phi_j\right)^2 \delta_{\lambda_j}. \tag{34}$$

Intuitively, $\mu_{A,\psi}$ describes the "squared length of components of a vector projected into the eigenspace

[4] To be precise, for the structural model in (25), the argument involves a generic orientation of $\mathbf{a}$ and *the eigenspaces* of $\Sigma_{\mathbf{XX}}$.

of $[\Sigma_{\mathbf{XX}}]$" (Liu and Chan, 2018). Note that the vector-induced spectral measure of a matrix can be represented by two vectors: one which represents the support of the spectral measure, ie. a list of the eigenvalues in decreasing magnitude and a second composed of weights corresponding to the eigenvalues. For tracial spectral measures the weight vector is $w = (1/d, \dots, 1/d)$ representing the uniform distribution over the eigenvalues.

Given these definitions, the precise meaning of 'generic orientation' is formalized in the following postulate.

Postulate 1: generic orientation of vectors. Given the structural model in eq. (25) and a large d, one can define 'generic orientation' as:

1. Vector $\mathbf{a}$ has generic orientation relative to $\Sigma_{\mathbf{XX}}$ in the sense that
$$\mu_{\Sigma_{\mathbf{XX}},\mathbf{a}} \approx \mu^{\tau}_{\Sigma_{\mathbf{XX}}} ||\mathbf{a}||^2 \tag{35}$$
2. Vector $\mathbf{b}$ has generic orientation relative to $\Sigma_{\mathbf{EE}}$ in the sense that
$$\mu_{\Sigma_{\mathbf{EE}},\mathbf{b}} \approx \mu^{\tau}_{\Sigma_{\mathbf{EE}}} ||\mathbf{b}||^2. \tag{36}$$
3. Vector $\mathbf{a}$ is generic relative to $\mathbf{b}$ and $\Sigma_{\mathbf{EE}}$ in the sense that
$$\mu_{\Sigma_{\mathbf{XX}},\mathbf{a}+c\Sigma^{-1}_{\mathbf{XX}}\mathbf{b}} \approx \mu_{\Sigma_{\mathbf{XX}},\mathbf{a}} + \mu_{\Sigma_{\mathbf{XX}},c\Sigma^{-1}_{\mathbf{XX}}\mathbf{b}}. \tag{37}$$

Intuitively, (35) states that 'decomposing $\mathbf{a}$ into eigenvectors of $\Sigma_{\mathbf{XX}}$ yields weights that are close to being uniformly spread over the spectrum.' (36) captures a similar statement for $\mathbf{b}$ and $\Sigma_{\mathbf{EE}}$: the weights of $\mathbf{b}$ are uniformly distributed across the spectrum of $\Sigma_{\mathbf{EE}}$ (Janzing and Schölkopf, 2018).

Eq. (37) contains a crucial ingredient for the ability to detect confounding: the $\hat{\mathbf{a}}$-induced spectral measure (left-hand-side of (37), recall $\hat{\mathbf{a}} = \mathbf{a}+c\Sigma^{-1}_{\mathbf{XX}}\mathbf{b}$) can be decomposed into one part due to the causal vector $\mathbf{a}$ (first summand) and a second part due to the confounding (second summand).

Note that the justification of these definitions depends on an idealized generating model (as defined in Section 2.3 of Janzing and Schölkopf, 2018). For this idealized model, the authors derive asymptotic results such that the three definitions (35), (36), (37) hold with equality.

A.3 Quantifying confounding

Two indicators for confounding strength are proposed: i) a correlative, and ii) a structural indicator.

Definition A.3. (correlative strength of confounding) The correlative strength of confounding gives the degree to which the confounder contributes to the covariance between $\mathbf{X}$ and Y.

$$\gamma := \frac{\|\Sigma_{\mathbf{X}Z}\|^2}{\|\Sigma_{\mathbf{X}Y}\|^2 + \|\Sigma_{\mathbf{X}Z}\|^2} \tag{38}$$

The following indicator for confounding strength, which measures the deviation of the estimable $\hat{\mathbf{a}}$ from the genuine causal parameter $\mathbf{b}$, is proposed

Definition A.4. (structural strength of confounding)

$$\begin{aligned}\kappa :&= \frac{\left\|\Sigma^{-1}_{\mathbf{XX}}\Sigma_{\mathbf{X}u}\right\|^2}{\left\|\Sigma^{-1}_{\mathbf{XX}}\Sigma_{\mathbf{X}Y}\right\|^2 + \left\|\Sigma^{-1}_{\mathbf{XX}}\Sigma_{\mathbf{X}u}\right\|^2} \\ &= \frac{\left\|c\Sigma^{-1}_{\mathbf{XX}}\mathbf{b}\right\|^2}{\|\mathbf{a}\|^2 + \left\|c\Sigma^{-1}_{\mathbf{XX}}\mathbf{b}\right\|^2} \in [0,1]\end{aligned} \tag{39}$$

Note that from (37) and a normalizing condition

$$\mu_{A,\psi}(\mathbb{R}) = \|\psi\|^2$$

(eq. (10) in Janzing and Schölkopf (2018)), we know $\|\hat{\mathbf{a}}\|^2 \approx \|\mathbf{a}\|^2 + \left\|c\Sigma^{-1}_{\mathbf{XX}}\mathbf{b}\right\|^2$. Therefore, one can rewrite κ as

$$\kappa \approx \frac{\left\|c\Sigma^{-1}_{\mathbf{XX}}\mathbf{b}\right\|^2}{\|\hat{\mathbf{a}}\|^2}. \tag{40}$$

In words, κ is the share of the influence of u on $\mathbf{X}$ of the overall strength of the association between Y and $\mathbf{X}$. Another interpretation: κ is the deviation of $\hat{\mathbf{a}}$ from $\mathbf{a}$ relative to the sum of squared lengths of these vectors.

Note that the contribution of u to the covariance between $\mathbf{X}$ and Y is determined by the product $c\mathbf{b}$. As a consequence, rescaling c by some factor and $\mathbf{b}$ by its inverse leaves γ unaffected. Similarly, (a more sophisticated) rescaling of c and $\mathbf{b}$ leaves κ unaffected. The regimes with (i) large c and small $\mathbf{b}$ and with (ii) small c and large $\mathbf{b}$ can be thought of as two extremes on a continuum where knowing the value of u (i) hardly reduces the uncertainty about $\mathbf{X}$ or (ii) significantly reduces the uncertainty about $\mathbf{X}$. To capture these different regimes, JS propose an additional parameter that measures the explanatory power of u for $\mathbf{X}$,

$$\eta := tr(\Sigma_{\mathbf{XX}} - tr(\Sigma_{\mathbf{XX}|u})) = tr(\Sigma_{\mathbf{XX}}) - tr(\Sigma_{\mathbf{EE}}) = \|\mathbf{b}\|^2. \tag{41}$$

A.4 Estimating confounding

Loosely, the general idea of the algorithm is as follows. The vector-induced spectral measure of $\Sigma_{\mathbf{XX}}$ w.r.t. $\hat{\mathbf{a}}$ can be approximated by a normalized measure, $\nu_{\kappa,\eta}$, which decomposes into a causal part and a confounding part. The relative shares of causal and confounding parts in that decomposition is given by κ. The algorithm proceeds by finding the normalized measure most similar to (computable) $\mu_{\Sigma_{\mathbf{XX}},\hat{\mathbf{a}}}$. The parameter constellation that minimizes the distance tells us the relative confounding strength.

How do they do that? They show that $\mu_{\Sigma_{\mathbf{XX}},\hat{\mathbf{a}}}$ asymptotically depends on four parameters: $\Sigma_{\mathbf{XX}}$, $\hat{\mathbf{a}}$, κ, and η (two of which, $\Sigma_{\mathbf{XX}}$ and $\hat{\mathbf{a}}$, can be computed). Based on this insight, they formalize a two-parametric family of probability measures $\nu_{\kappa,\eta}$ such that, with high probability, it converges to $\mu_{\Sigma_{\mathbf{XX}},\hat{\mathbf{a}}}$ as the dimensionality of $\mathbf{X}$ increases (up to a normalizing factor):

$$\frac{1}{\|\hat{\mathbf{a}}\|^2}\mu_{\Sigma_{\mathbf{XX}},\hat{\mathbf{a}}} - \nu_{\kappa,\eta} \to 0 \text{ (weakly in probability)} \tag{42}$$

where

$$\nu_{\kappa,\eta} := (1-\kappa)\,\nu^{\text{causal}} + \kappa\,\nu_{\eta}^{\text{confounded}}\,. \tag{43}$$

We inspect each part in turn.

1. ν^{causal} is the hypothetical spectral measure that would be obtained in the absence of confounding. Following (35), it is defined as

$$\nu^{\text{causal}} := \mu^{\tau}_{\Sigma_{\mathbf{XX}}} \tag{44}$$

since, in the absence of confounding, the spectral measure induced by $\mathbf{a}$ should be equivalent to the tracial spectral measure of $\Sigma_{\mathbf{XX}}$ (up to a normalizing factor). This part can be estimated easily since $\mathbf{X}$ is observed.

2. To define the corresponding confounding part, JS propose an approximation to the spectral measure of $\Sigma_{\mathbf{XX}}$ induced by the vector $\Sigma^{-1}_{\mathbf{XX}}\mathbf{b}$. Recall that $\mathbf{b}$ has generic orientation relative to $\Sigma_{\mathbf{EE}}$, cf. eq. (36). However, both $\mathbf{b}$ as well as $\Sigma_{\mathbf{EE}}$ are unknown. These two unknowns correspond to two steps that are important for constructing this approximation.

   Before looking at these two steps, it is worthwhile emphasizing the two critical assumptions used: first, by virtue of the properties of the generating model we have $\Sigma_{\mathbf{XX}} = \Sigma_{\mathbf{EE}} + \mathbf{b}\mathbf{b}^{\top}$; second, $\mathbf{b}$ has generic orientation w.r.t. $\Sigma_{\mathbf{EE}}$.

   (a) The eigen decomposition of $\Sigma_{\mathbf{EE}}$ reads QM_EQ^{-1} where $M_E := \text{diag}(\lambda_1^E, \ldots, \lambda_d^E)$ with $\lambda_1^E > \cdots > \lambda_d^E$ eigenvalues of $\Sigma_{\mathbf{EE}}$. Though we do not know $\mathbf{b}$, we know that it is generic relative to $\Sigma_{\mathbf{EE}}$. Therefore, we can replace $\mathbf{b}$ with a vector that is 'particularly generic', namely $\mathbf{g} := (1, \ldots, 1)^{\top}/\sqrt{d}$, which satisfies

$$\mu_{M_E,\mathbf{g}} = \mu^{\tau}_{M_E}.$$

   Therefore, one can approximate the spectral measure of $\Sigma_{\mathbf{XX}}$ induced by the vector $\Sigma^{-1}_{\mathbf{XX}}\mathbf{b}$ by spectral measure of $M_E + \eta\mathbf{g}\mathbf{g}^{\top}$ induced by $(M_E + \eta\mathbf{g}\mathbf{g}^{\top})\sqrt{\eta}\mathbf{g}$. This construction is still not feasible as M_E, which contains the eigenvalues of $\Sigma_{\mathbf{EE}}$, is unobserved.

   (b) JS resort to a result stating that spectral measures are close in high dimensions:

$$\mu^{\tau}_{\Sigma_{\mathbf{XX}}} \approx \mu^{\tau}_{\Sigma_{\mathbf{EE}}},$$

   cf. their Lemma 4. Therefore, one can approximate M_E with $M_X = \text{diag}(\lambda_1^X, \ldots, \lambda_d^X)$ and $\lambda_1^X > \cdots > \lambda_d^X$ eigenvalues of $\Sigma_{\mathbf{XX}}$.

   Putting these two steps together, JS define a rank-one perturbation of M_X as

$$T := M_X + \eta\mathbf{g}\mathbf{g}^{\top},$$

   compute the spectral measure of T induced by vector $T^{-1}\mathbf{g}$, and define

$$\nu_{\eta}^{\text{confounded}} := \frac{1}{\|T^{-1}\mathbf{g}\|^2}\mu_{T,T^{-1}\mathbf{g}}. \tag{45}$$

   For clarity, it is useful to reiterate what has been achieved in Step 2: first, $\Sigma_{\mathbf{EE}} + \mathbf{b}\mathbf{b}^{\top}$ can be approximated by $M_{\mathbf{E}} + \eta\mathbf{g}\mathbf{g}^{\top}$. Second, $(\Sigma_{\mathbf{EE}} + \mathbf{b}\mathbf{b}^{\top})^{-1}\mathbf{b}$ can be approximated by $(M_{\mathbf{E}} + \eta\mathbf{g}\mathbf{g}^{\top})^{-1}\sqrt{\eta}\mathbf{g}\mathbf{g}^{\top}$.

A.5 Algorithmic implementation

The algorithm finds κ by taking that element in $\nu_{\kappa,\eta}$ that is closest to $\mu_{\Sigma_{\mathbf{XX}},\hat{\mathbf{a}}}$. Since eq (42) only asserts weak convergence in probability, computing l_1 or l_2 distance is inappropriate. Therefore, JS propose smoothing the spectral measures using a Gaussian kernel.

Thus the difference between vectors w and w' is given by

$$D(w, w') := \|K(w - w')\|_1 \tag{46}$$

with

$$K(\lambda_i, \lambda_j) := \exp\Big(-\frac{(\lambda_i - \lambda_j)^2}{2\sigma^2}\Big)$$

Finally, the algorithm finds the κ that minimizes $D(w, w^{\kappa,\eta})$ where w is the weight vector corresponding to the (computable) spectral measure $\mu_{\Sigma_{\mathbf{XX}},\hat{\mathbf{a}}}$ and $w^{\kappa,\eta}$ is the weight vector corresponding to the $\nu_{\kappa,\eta}$.

B ROC curves

ROC curves are an insightful way to evaluate the performance of a binary classifier (exogenous vs. endogenous instrument, in the case at hand) that plots the share of true positive (TP) decisions as a function of the share of false positive (FP) decisions. Thereby, it shows the trade-off between Type I and 1 − Type II errors of the test, ie. rejecting H_0 although it is true and rejecting H_0 when it is indeed false. The curve is traced out by varying a threshold parameter α. The false positive rate is calculated as the share of false positive decisions, ie. rejections of H_0, across M Monte Carlo draws in which H_0 is in fact true (ie. the instrument endogeneous). Similarly, the true positive rate is calculated as the share of true positive decisions across all Monte Carlo draws in which H_0 is actually false (ie. the instrument exogeneous):

$$\begin{aligned}\text{FP}(\alpha) &= \frac{1}{M}\sum_{m=1}^{M}\psi_{\delta,m}(\alpha) \text{ when } \omega_3 \neq 0 \\ \text{TP}(\alpha) &= \frac{1}{M}\sum_{m=1}^{M}\psi_{\delta,m}(\alpha) \text{ when } \omega_3 = 0.\end{aligned} \tag{47}$$

The ROC curve plots the TP rate as a function of the FP rate. The further the curve lies above the forty-five degree line, the better the test. The area under the ROC curve (AUC) is a measure for the accuracy of the test and ranges between 0.5 (useless classifier that does just as good as chance) and 1 (perfect accuracy).